Financial Services Landscape Discovery with Machine Learning:

Augmenting a Financial Markets Ontology Using Probabilistic Topic Modeling

by

Troy A. Danka

Thesis Project

Submitted in partial fulfillment of the Requirements for the degree of

MASTER OF SCIENCE IN DATA SCIENCE

August 2018

Dr. Alianna J. Maren, First Reader

Dr. Chad R. Bhatti, Second Reader

ABSTRACT

With a glut of competing priorities, the financial industry faces major challenges in extracting timely, relevant, and specifically-focused information from text. Without clear-cut business cases, making the investment in text analysis methods does not justify the return on investment. Furthermore, the business landscape continues to become increasingly complex, and at a rapid pace of change. Thus, difficulties in managing and extracting domain knowledge through heavily people-based systems leads to increased risk, the impacts of which range in billions of dollars of waste from lost business opportunity or even regulatory fines. The increased complexity provides an unprecedented opportunity to use a focused and content-rich ontology around financial markets concepts as context for meaningfully connecting ideas found in text. Ontologies encompass a formal naming of domain concepts and relations in a manner that should improve problem solving within the domain. By infusing ontologies into text analytics, the outcome is more likely to yield better capture of semantic content. The result is a declared vocabulary that provides quick and actionable insights for financial industry practitioners to stay on top of changes in their fast moving domain. This work presents a novel means for augmenting a financial markets ontology through the use of probabilistic topic modeling. In particular, this work has used the Latent Dirichlet Allocation (LDA) method for topic modeling which is able to ingest large volumes of text data and quickly extract themes across documents. LDA, which yields topic models, was applied to a large corpus of 11,693 text articles and speeches in the financial industry domain. The outcome is an informed ontology that demonstrates the evolution of concepts at deeper levels of granularity, and over distinct time periods such as changes in market policy under the Obama Administration as compared to the Trump Administration.

Keywords: ontologies, topic modeling, LDA, text analytics, machine learning, knowledge graphs

Contents

Executive Summary

One of the major challenges that the financial industry faces is extracting timely, relevant, and specifically-focused information from text. The simplest reason for this is that text data is not yet widely viewed as a critical corporate asset that can be readily deployed for generating business value. Building a business case for wider text analytics adoption first requires translating its value into the traditional measures that senior leadership uses to assess return on investment. As such, showing how its integration into existing business value streams as a means to yield better results from current methods alone is the most pragmatic place to start. Examples may include quickly and meticulously scanning competitor news to inform investor communications, or demonstrating that a more exhaustive set of risk factor scenarios were used for pricing strategies and risk management. Identifying friction points where functional capabilities have broken down in the past provides real examples of where text-based methods could add value.

Broader text adoption would be better justified if its significance resulted in measurable achievements that industry stakeholders could concretely recognize. On the revenue side, clear use cases exist for incorporating text-bases measures into trading models for better results. However, for use in more judgment-based capabilities such as effective risk management or domain-heavy business analysis projects, the infrastructure and tools to meaningfully source, organize and transform text data into measurable insights is not well known nor widely understood. Moreover, the financial industry landscape has become increasingly complex and the pace of change is increasingly faster. Thus, leveraging text data as a resource for improving stakeholder capabilities to more precisely and quickly gain insights into their specific business areas would establish tangible benefits for investing in methods for its expanded use. Lastly,

introducing the use of text as a complementary measure will also help to overcome regulatory apprehension for the use of text-based models in regulatory governed business processes.

Obtaining desired and timely text-based information is particularly challenging in the financial industry because the domain is constantly evolving at the U.S. and global levels, and the breath of market participants' individual needs creates an enormously complex web of interactions to understand. As such, the need for a focused and content rich ontology around financial markets concepts provides the context for connecting ideas discovered in text to our financial industry domain in a meaningful way. By infusing ontologies into analytics, the outcome is more likely to yield better capture of semantic content. This is a critical step for generating the value that business stakeholders likely require to justify the efforts. Without a broader landscape picture of the financial domain, text-based algorithms alone will produce topic maps based solely on the terms used by input documents, and not necessarily reflect the full richness of an ontology.

This work presents a novel means for augmenting a financial markets ontology through the use of probabilistic topic modeling allowing financial industry stakeholders to quickly glean business insights from text. In particular, this work has used the Latent Dirichlet Allocation (LDA) method for topic modeling which is able to ingest large volumes of text data and extract themes across documents. LDA, which yields topic models, is a means of identifying and measuring the existence of a topic throughout the full corpus. Figure 1 depicts the overall modeling problem of this thesis, which is applying an unsupervised machine learning problem to a large corpus of contextualized text from the financial industry, and using the output in the form of topic models to augment a financial industry domain ontology. Existing credible taxonomies and ontologies will be leveraged in some cases to better meet shared interests and ensure wider usability (e.g., Financial Industry Business Ontology for encoding of financial products).

Knowledge Modeling Problem: Grounding Topic Models with an Ontology

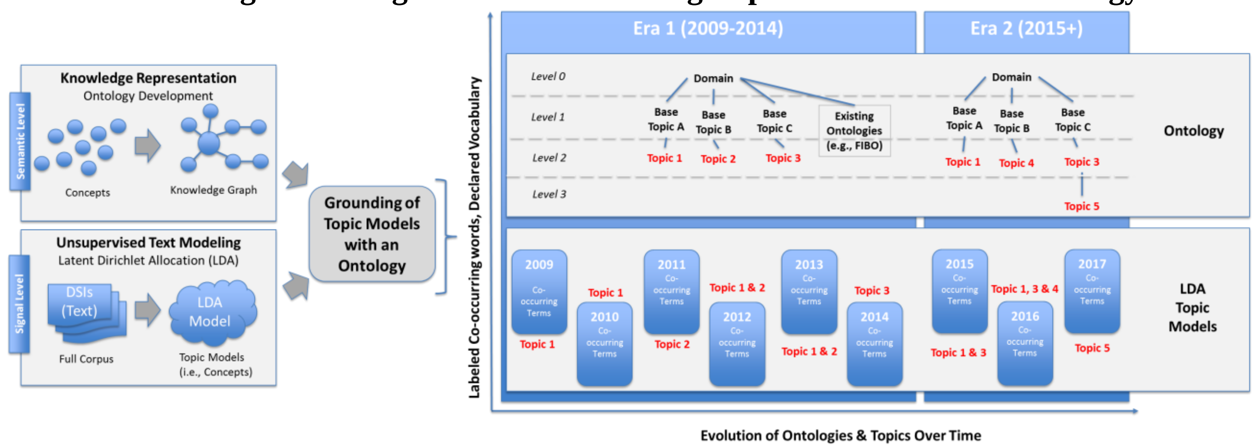


Figure 1. Framing of overall modeling problem: encoding the financial industry domain in a visual ontology, and informing its declared vocabulary with an unsupervised learning model.

Such a tool provides a method for stakeholders to tap into large volumes of external data to an organization such as market and news data for business strategy and competitive differentiation. For example, an investor relations function could immediately consume competitor earnings releases to better position for questions from their own investors. Alternatively, a risk manager could utilize text data in internal models as part of a challenge and review process that could scan a wide variety of news, industry and social media data to assess the likelihood of changes within a sector as part of a forward-looking forecast for the next several financial quarters. Perhaps the biggest immediate benefit of adopting a text-based approach into organizational processes is the increased confidence that judgment-based processes can be cross-validated. Thus, mitigating the heavy dependence on subject matter expertise with algorithmic methods that go beyond what a single person or group of people within an organizational function may know.

By infusing a domain ontology with novel topics that appear over time, we are able to transform ideas reflected in text sources into a network of concepts to better capture and visualize the latent insights included in text data. Figures 2a-c illustrates three time periods of our financial domain ontology to illustrate how the combination of a robust ontology linked to new topics of focus provides more relevance for topic connections. For example, we can see that the Volcker Rule appearing in the ‘Obama Era’ of Figure 2b has multiple connections to the broader financial markets landscape base ontology, including market oversight of trading assets, and in particular oversight of trading desks in the front-office, and their trading and execution processes. Contrarily, the same topic is a subject of repeal just four years later in the ‘Trump Era’ illustrated in Figure 2c where a more business friendly policy is in play. Ontologies provide a contextual way to encode the output of text-based algorithms into a deep set of hierarchical class relationships that is multi-dimensional across both time and space.

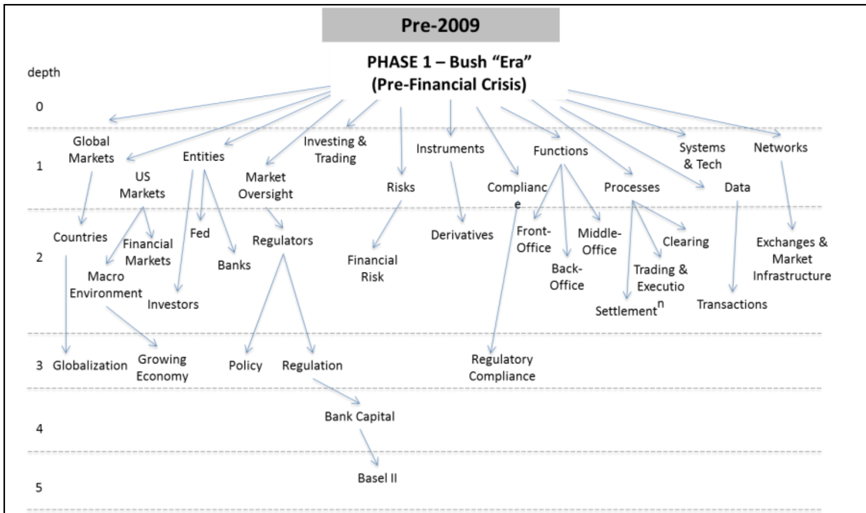


Figure 2a. Pre-Financial Crisis View of Financial Markets, Bush Era.

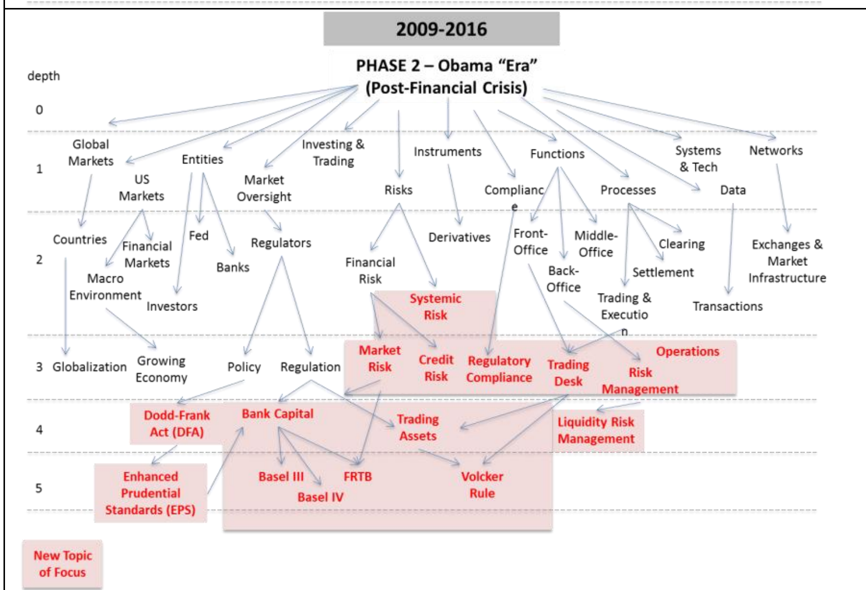


Figure 2b. Post-Financial Crisis View of Financial Markets, Obama Era. New topics of focus emerge.

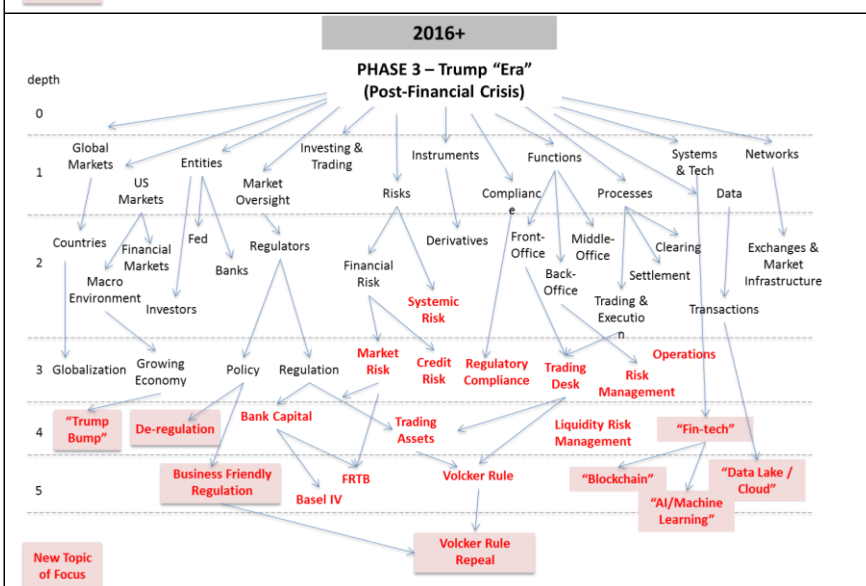


Figure 2c. Post-Financial Crisis View of Financial Markets, Trump Era. New topics of focus emerge, some changes to existing.

Note: Shaded boxes reflect new topics discovered in each era compared to prior period.

A topic modeling algorithm identifies groups of words that occur together frequently throughout a set of documents. Thus, topic-modeling algorithms treat each document as a mixture of topics; grouped words are mapped to topics, which are in turn mapped to multiple documents in the corpus. The analyst is not required to know anything about the text in order to summarize the prominent themes. The algorithm provides an analyst the capability to scan a large volume of text documents with fast and actionable insights, and links back to originating sources. Topic modeling can be applied to the entire corpus of available text documents, or in multiple passes allowing the analyst to zoom into key areas. This is equivalent to reviewing all the books in a library, or say just the ones related to a given category, author, or publisher. By adjusting the scope, the analyst is able to capture deeper levels of granularity of concepts.

The labeling of topics is an important activity, often performed manually by people, and becomes the semantic bridge for linking back to broader organizational knowledge representation frameworks in the form of ontologies. For some topics, labeling is a simple name tagging, while subject matter experts may be required for detecting and naming more complex topics. Figure 3 depicts single words found in multiple opinion articles that are associated with four topics associated with the financial markets. A critical idea to recognize is that these single words are associated with multiple topics. The simple example shows how hidden topics are discovered through the frequent co-occurrence of single-word clusters, which then aligns to the distribution of words associated with a given topic. However, it is only when we connect the labeled topics back to our domain ontology of Figure 2 that we can see the broader context of the concept.

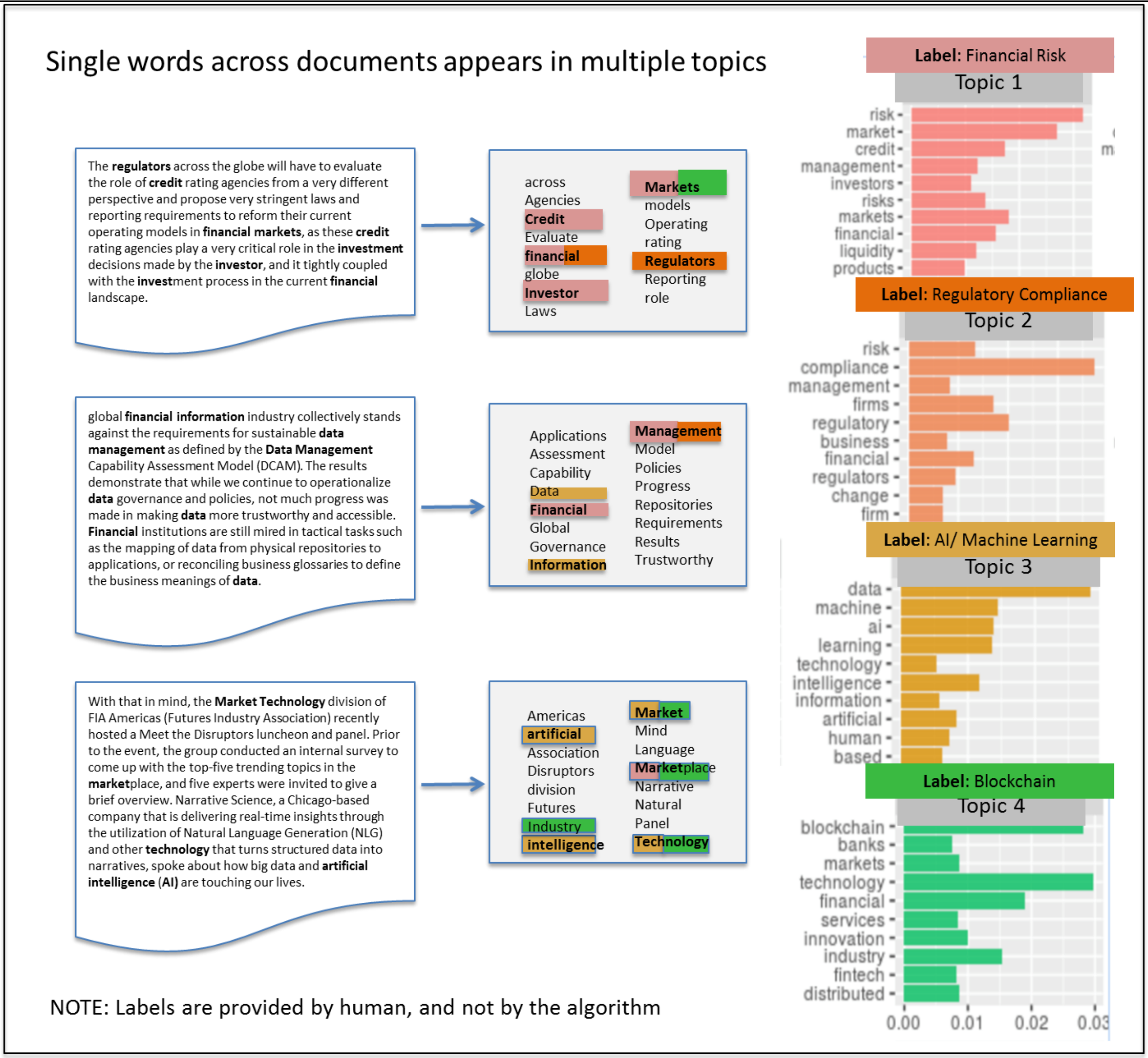


Figure 3. Topic Modeling Example with the Latent Dirichlet Allocation (LDA) Method.

The events of the financial markets have far-reaching implications on nearly every aspect of American society and global society; many of which are both political and financial, affecting legal, economic, and social concerns. The 2008 Financial Crisis proved to be especially complex for financial markets following unprecedented new regulations after the passage of the Dodd-Frank Act in the United States. In the subsequent decade following the crisis, business and political differences emerged around market oversight policy as we see in the Obama Era vs. Trump Era of Figures 2a-c.

The passage of far-reaching new regulations has impacted nearly all business and operational processes in financial institutions, triggering large volumes of content from financial news, industry groups, professional services, and the landscape of vendors aiming to interpret and synthesize their reactions in the form of digitized text available on the public web. Moreover, the regulators themselves provided commentary outside formal regulations in the form of periodic speeches in public forums. These collective topics were amplified by the heightened sense of global political urgency facing government officials to act. On one hand, we had President Obama inherit a shattered economy and oversee the passage of many new regulations. However, less than ten years later, American voters elected President Donald Trump who campaigned on a de-regulatory platform.

Information exchange is very fast today, therefore making it improbable that individuals can read all of it. Topic modeling provides a means to identify and quantify central points of discussion, and therefore providing a landscape view for understanding the key topics. The repeated occurrences of people, places, things and abstract concepts are the basis for extraction. Thus, the content is more important than the medium in topic modeling for the reason that the mixture of words in documents forms the topics.

To ensure a well-represented sample, text was obtained by scraping 656 of the Board of Governor speeches from the Federal Reserve website, 82 of which were from the Trump Era beginning in 2016. In addition, 11,040 opinion analysis pieces were scraped from a leading capital markets industry publication called TabbFORUM, 3,284 of which were from the Trump Era. Each full-text article was published between January 1, 2009 and October 31, 2017 aiming to cover a lengthy period following the financial crisis. Although each source required distinct pre-processing steps, the topic algorithms were applied to both sets of documents as a single corpus. Figure 4 below visually depicts the distribution of our data source items (DSI) to show the breakdown over time and across the three eras previously defined. The findings from our corpus are based on text only from the Obama and Trump Eras, and not the Phase 1 Bush Era.

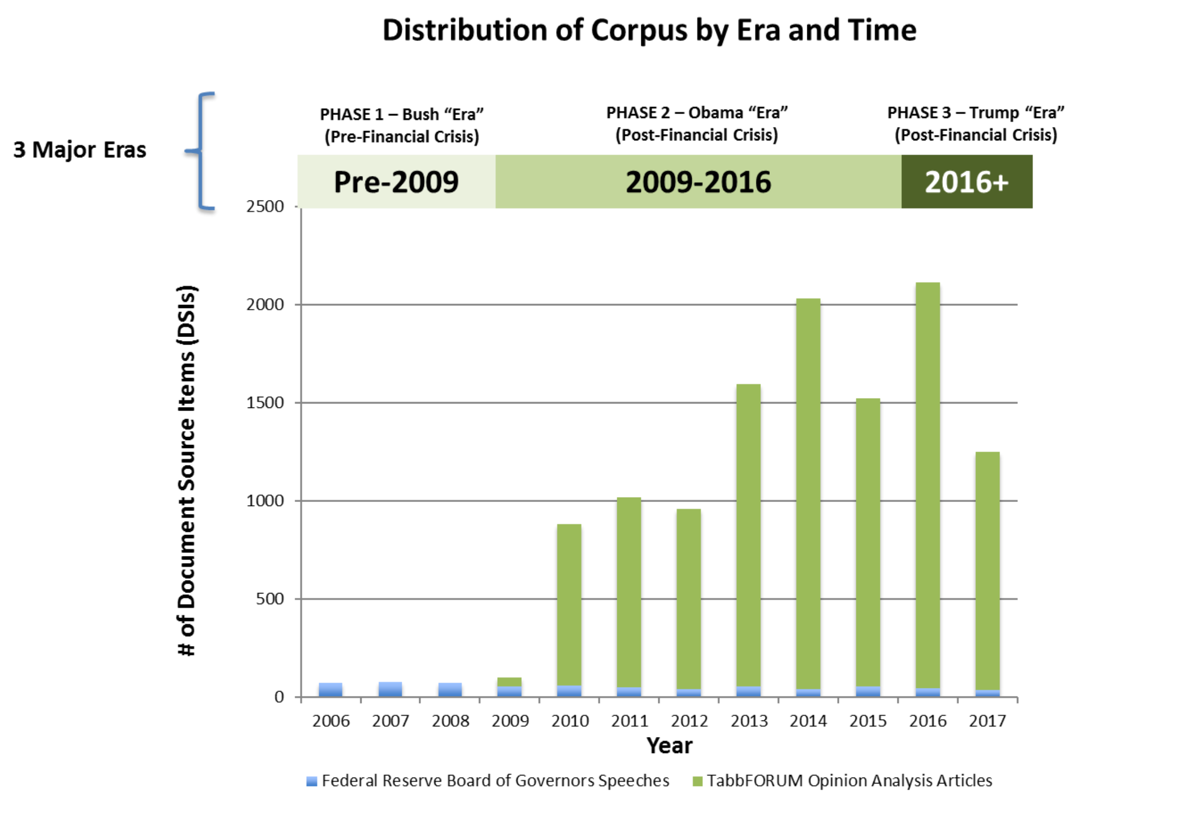


Figure 4. Distribution of corpus DSIs over time, and aligned to three eras. A breakdown by document type is represented in blue (Federal Reserve Board of Governors) and green (TabbFORUM) shading. Corpus scope is only Phase 2 and Phase 3.

Major findings from this analysis of regulatory speeches and professional industry participants revealed that market oversight and compliance, organizational functions and processes, and systems and technology played a large role in the recent financial markets conversation. As an example of a concrete benefit of what topic modeling can deliver, we observed as one of the key results of this work that market discourse originally viewed regulation largely as an impediment. However, over the following decade the abundance of centrally stored data has led to the view that investment in infrastructure and data analytics provides a new corporate asset for competitive advantage.

Insights such as the ones revealed about regulation have led to important changes in how financial organizations do business. As opposed to simply taking steps to minimally comply with such regulations, organizations are investing in new applications and enterprise architectures that will bring about new services and opportunities to better respond to customer needs. Instead of establishing data silos centered around specific functional problems, leading organizations are taking a single firm view to build logical enterprise data domains centered around data with similar characteristics. This provides an opportunity to organize data around transactions, products, master reference data, and emerging data such as text in ways that provide more transparent, quality and streamlined access, and advanced analytics opportunities.

Specifically, the enterprise architecture has evolved from a central tenant of exclusively facilitating regulatory compliance, to enabling the pursuit of better serving customers and growing legacy and new businesses through new digital channels. Figure 5a below visually depicts the key regulations captured by our LDA topic models on the left side, with the major topic themes shown below each era. The intent is to highlight how the leading themes of the ‘regulatory compliance’ era required heavy data infrastructure investment. The resulting themes of the ‘digital transformation’ era were made possible by the resulting modern ‘big data’ architecture.

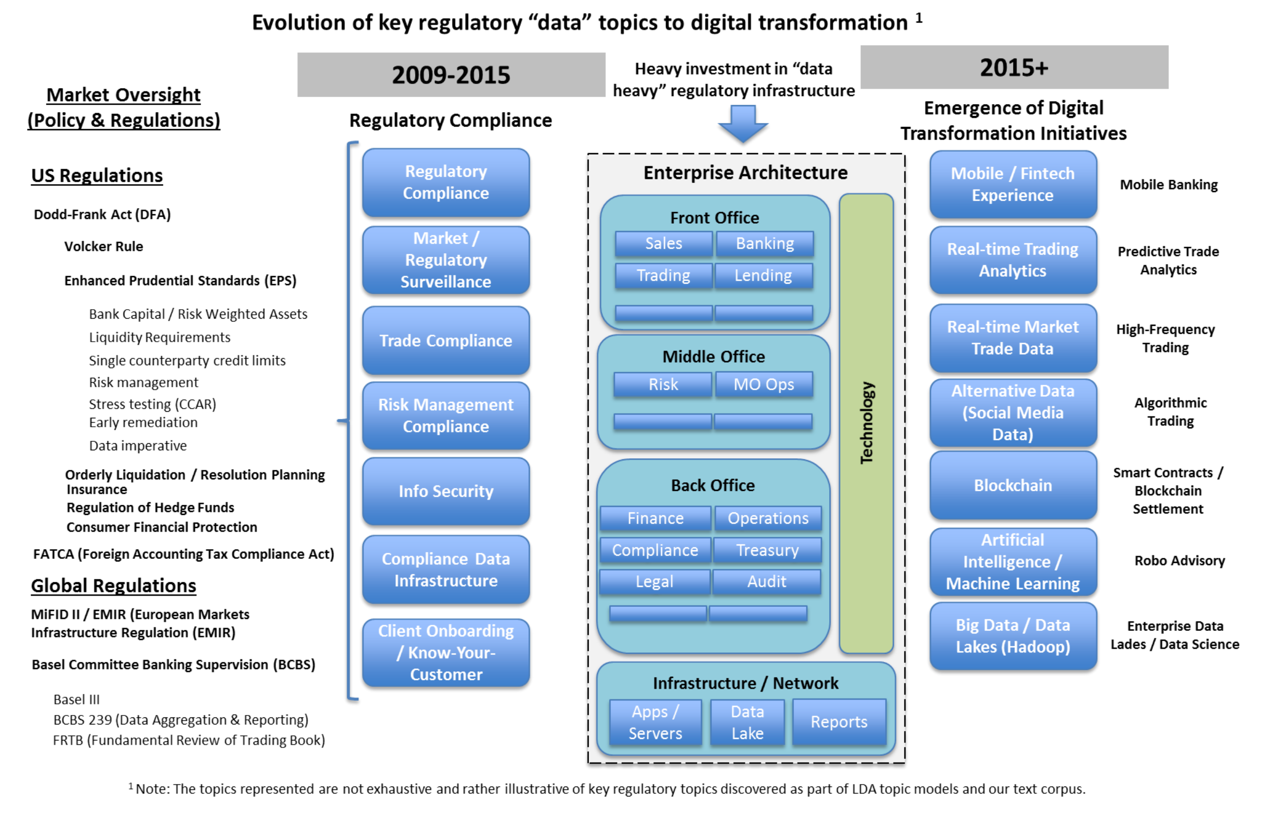


Figure 5a. Depiction of regulatory drivers surrounding Enterprise Architecture leading to digital topics.

Thus, a tangible result of how text analytics can impact financial organizations begins with integrating text-derived insights with evolving processes. By viewing regulatory compliance topics with a forward looking impact on the markets, participants are collectively anticipating ongoing change and taking proactive steps to shape what financial services could look like in a fast emerging digital world. For example, the application of new distributed ledger technologies such as blockchain, and its potential for transforming business processes through trusted and permanent records could be a significant turn for the industry. Figure 5b is a tree-map that shows the prevalence of topics within the entire corpus of 11,693 documents.

Financial Markets Landscape

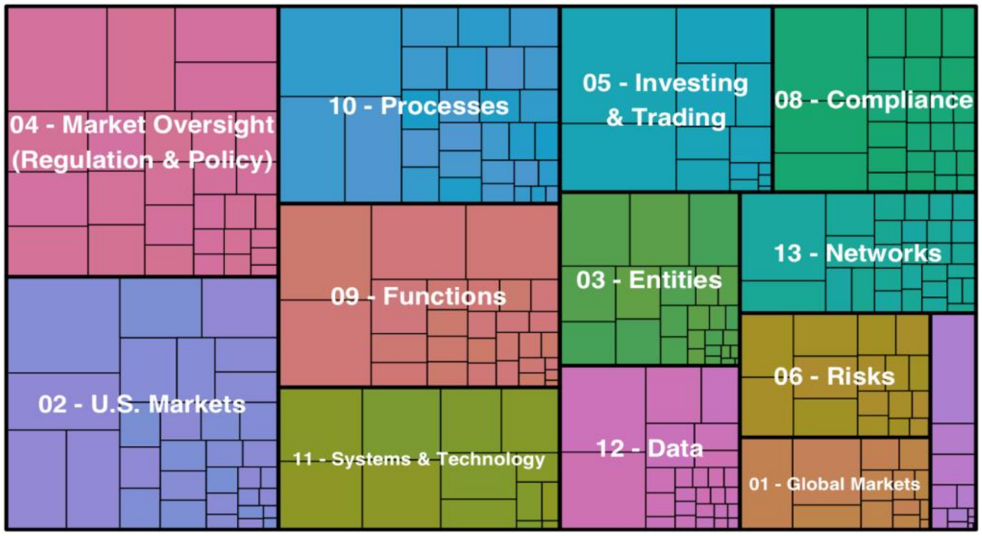
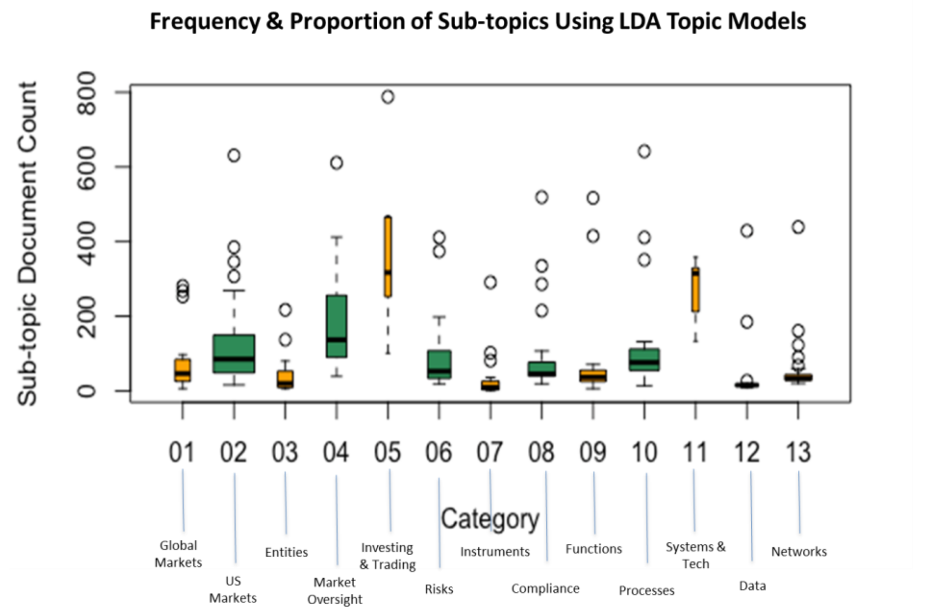


Figure 5b. Landscape of Financial Market Topics.

By associating topics with the major categories of a financial markets domain ontology, topics could be analyzed by themes and mapped to important sub-topics within each. Moreover, by incorporating the month and year in which the articles were published, topics could be analyzed over time for variation or sustained levels of importance. Figure 5c is a box-plot that shows the document frequency distribution of sub-topics within each category along with their proportion relative to other categories. The results show that categories such as 'Investing & Trading' and 'Data' had a minimal number of important topics, but were pervasively included in a large number of documents. Other categories such as 'U.S. Markets', 'Systems & Technology' and 'Compliance' had a large overall proportion of the distribution of topics (i.e., many topics within the category), many of which had significant implications for the financial industry.

To demonstrate the practical benefits of topic modeling, Figure 5d highlights at a high level some of the key insights and findings across the categories. Ultimately, businesses can leverage topic modeling to stay informed on their key focus areas and identify novel areas of interest.



Note: Box length is 1st through 3rd quartile of sub-topic counts per category; the extended whiskers illustrate min/max range. The bubbles represent a few outliners well outside the normal range. Box width is proportion of sub-topics to total sub-topics. Colors alternate without meaning for visual effect.

Figure 5c. Distribution of sub-topics by category and document count

Insights into Major Financial Markets Landscape Topics

Impact of political change on markets
- Emerging business and political differences
- Less regulation, more business friendly environment

Derivatives reform and effect on Global Markets lead to many new processes
- Derivatives regulation major focus (i.e. swaps & futures trading)
- Global markets oversight of derivatives via MiFID II in Europe

Equity markets have been very active and evolving
- US Equity markets – single largest topic of discussion
- Automated trading is a major topic
- High Frequency Trading (HFT) risks

Business operating model impacts
- Risk and Finance functions face increased cross-functional processes

Data
- Data has emerged as a pervasively important topic
- Increased investment in data is improving value of advanced analytics
- Desire for more and faster data

Lending failures, many new protections
- Borrower lending oversight in wake of financial crisis has undergone major reform
- Automated trading is a major topic

Systems and technology seen as a major growing force as competitive asset
- Infrastructure investments for regulatory needs now turning to force for competitive advantage
- Emerging technologies (e.g., digital)

Transformation of risk and compliance function
- Heavy and progressive bank capital regulations
- Banks are learning to operating in a continuous state of regulatory change

Minority forces have more influence
- Community banks push back on effects of Dodd-Frank Act
- Women in labor markets seen as a increasingly important topic at all levels

Figure 5d. Summary of financial markets insights

Introduction

The increasingly complex, fast-paced, highly competitive and politically charged business environment of the financial industry today has placed an enormous value on being able to generate concrete return on investment from data. Market participants in the global financial system traditionally rely on a plethora of data sources of all types and mediums to stay on top of a dynamic financial industry ecosystem. However, the notion of extracting business value from text data is a relatively new phenomenon as financial companies have sought to improve efficiencies with business processes, and expand opportunities to meaningfully harness timely, relevant, and specifically-focused information from text. As such, identification of more descriptive enhancements to important data elements or hard-to-qualify inputs into business processes provides tangible opportunities to begin exploring text data as a resource for introducing incremental value to existing data sources. Specifically, by using text data to fill data completeness gaps, or to improve the accuracy and frequency of key measures, focused text could benefit highly sensitive processes. For example, using insights gleaned from text to assign more nuanced data classifiers would provide greater precision of process input data. Taking these initial steps could allow for the greater confidence required before heavily investing in text data as a corporate asset, and developing new business processes based on the newly fabricated text data.

Acquiring relevant and suitable text-based information in a timely manner is particularly challenging in the financial industry because text has not been subject to the levels of governance and control that more traditional data sets have received. Stakeholders spend a great deal of their time ensuring quality of data can be tied back to authoritative and credibly declared data sources. Even if text data is used as a supplementary indicator that could positively impact business outcomes, it still must undergo reasonableness checks to avoid challenges in explaining the results. This is especially true in regulatory governed processes where regulators are slowly gaining confidence in the use of models and machine learning for business processes. Therefore, a need exists to 'ground-truth' text data against a content-rich ontology that can be consensually developed and maintained by qualified subject matter experts in a given domain to ensure that the output of text-based models is reasonable and appropriate. Moreover, the task of maintaining such a domain ontology using exclusively people-based approaches could be costly and divert resources from other important efforts such as running the business, or generating new opportunities.

As such, the need for a value-added text-based solution that combines the business context of an ontology with the raw signals provided by text analytics allows for a more tractable and cost-effective approach and more likely to produce meaningful results. This abstract notion builds on a critically important idea that people maintain rich, conceptual, and highly-structured world-views, which can be connected to lower-level signals such as spoken words. Establishing connections between these two levels is at the cornerstone of Natural Language Processing (NLP) applications and provides the structure for inference and establishing meaning. By infusing ontologies into analytics, the outcome is more likely to yield better capture of semantic content, and garner the respect of senior leaders. This is a critical stride for generating the value that business stakeholders likely require to justify the efforts. Without a broader landscape picture of the financial domain, text-based algorithms alone will produce topic maps based solely on the terms used by input documents, and not necessarily reflect the full richness of an ontology (Chen et al. 2018).

In today's fast moving and data driven world, investing in analytically informed views provides a more objective way to develop value-added business processes and overcome the operational issues discussed previously. To meet these challenges, this thesis explores a multi-pronged approach on three dimensions:

- First, building a domain ontology;
- Next, making sense of large text corpora;
- Lastly, fusing our ontology with topic models.

Therefore, establishing an ontological worldview of the financial markets ecosystem, and augmenting with probabilistic topic modeling can assist in adopting an agile and cost-effective analysis tool by creating a pragmatic solution to understand the market landscape, monitor plausible risk catalysts, and manage a broader enterprise knowledge representation framework.

Building a Domain Ontology

Before taking steps to build an ontology for representing a domain, it is helpful to understand its basic features and methods for expressing data relationships. Essentially, ontologies are designed to promote greater consistency in the description of data. Thus, terms in an ontology maintain textual definitions to ensure consistency among people, and logical definitions to aid computer access and quality control. The resulting graph is an organization of nodes based on terms and edges for linking the ontological relations between classes and sub-classes (Alp, Smith, & Spear, 2015). Application of the graph is often to annotate or describe data contained in disparate information stores in a dynamic and interconnected way.

With the complexity of the financial markets, it is easy to imagine a network of concepts and interactions that can be described with a domain ontology. However, the task of organizing the range of financial industry terms into a hierarchical classification is not straight forward and requires input from a multitude of reference sources. As such, a good approach is to start small by establishing a baseline to simplify the initial complexity of the financial domain and lead with ideas found in existing structured taxonomies to seed the development. Specifically, the process begins with a simple world-view that starts with the definition of the most general concepts in the finical markets, followed by subsequent specialization of the concepts. Figure 6a below is a network-based representation of a Level One ontology for the financial industry to illustrate how nodes and edges come together to form a knowledge graph. Therefore, a key goal of building our domain ontology will be to have a concise way to convey the things that make up the financial markets – i.e., people, entities, events, and abstract concepts – to establish context and allow for consistent application of its ideas in both people-based and machine-based interactions.

Perhaps the most pragmatic application of the ontology will be the opportunity for visual exploration of the financial markets domain in a manner that allows for greater discovery of insights, which may often be overlooked when using traditional research methods. Today with the wide availability of text-based commentary on the markets, relationships can be harnessed to visualize, model and convey important ideas among industry participants. Thus, the results of our work will be conveyed through several forms of network visualization to provide a more tractable means for observing the dynamics of large financial networks. Moreover, text analytics can be visualized in a dynamic way as previously shown in Figures 1a-c to support data exploration, highlight econometric results or provide early warning information (Heijmans et al. 2016).

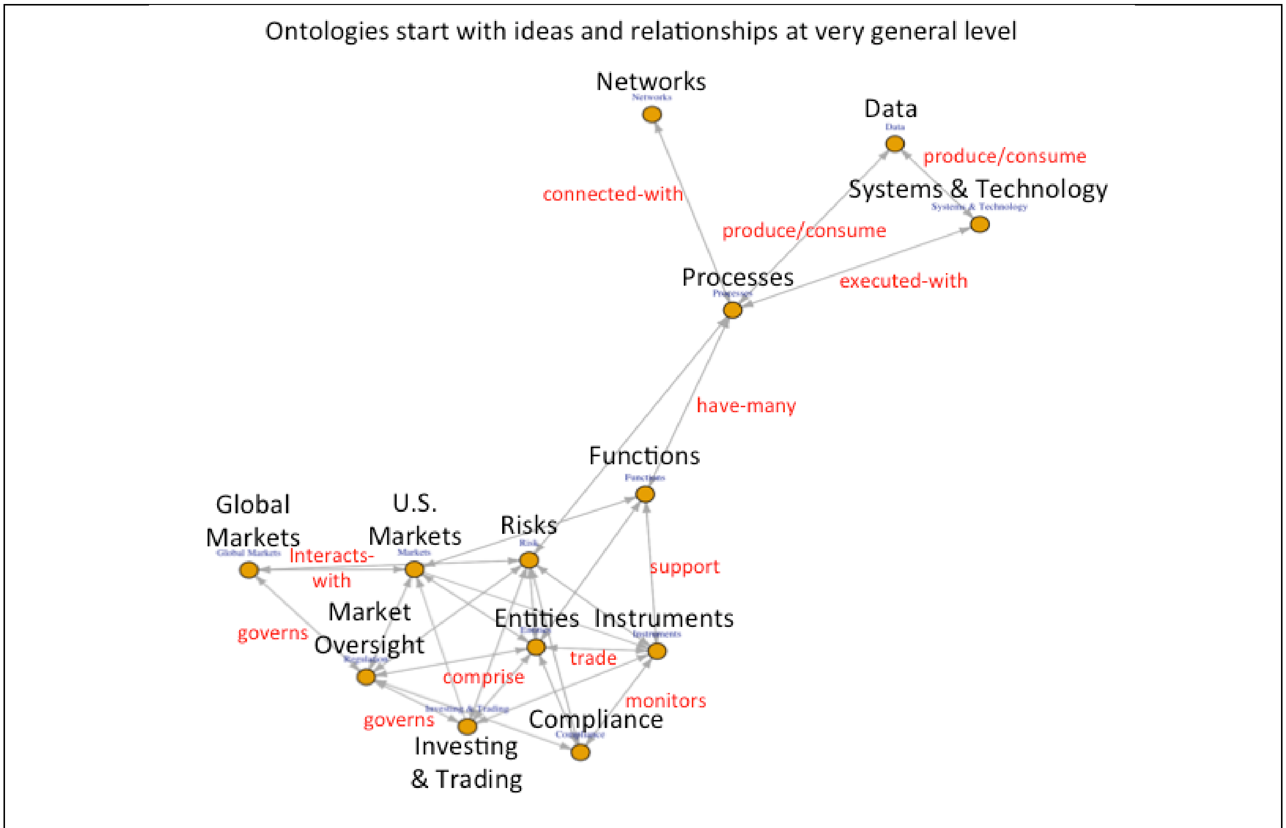


Figure 6a. Level One illustration in building our world-view of the financial markets with note relationships established.

Making Sense of Large Text Corpora

The diversity of journalistic and Internet based text commentary of the financial markets provides a boundless collection of sources for understanding the financial markets. By ingesting an array of financial news, regulatory perspectives, and professional industry-participants' opinion analysis, we can appreciate with a wider vantage the knowledge and momentum of the dialogue around the financial markets over time. However, the primary challenge with this approach is that the sheer volume of information available is nearly impossible for a single person to comprehend, nor have sufficient time for its consumption. The required synthesis of ideas into meaningful summaries takes considerable time, and can be greatly enhanced with the use of unsupervised machine learning methods such as LDA topic models.

Topic modeling can be applied to large collections of text documents, such as speeches, articles and opinion pieces, so that important themes and relationships are identified. It can be a very effective tool for addressing a critical notion that text data is highly dimensional. Traditional techniques to reduce dimensionality by clustering words into groups (i.e., ‘bag of words’) fall short of making connections with the semantic level of the text. Thus, topic modeling allows a text corpus that is divided into documents to be parsed such that the semantic similarities among the words and documents are correlated. Thus, topic modeling algorithms treat each document as a mixture of topics, providing a method to both group words into topics and to measure similarities between the mixtures of topics in two separate documents (Curme et al. 2017). Figure 6b below demonstrates the statistical power of LDA topic models to extract the multiple latent topics across a corpus of text data by illustrating topic assignment proportions for a single text document; this example is adapted from pioneering work by Blei in a survey paper (Blei, 2012).

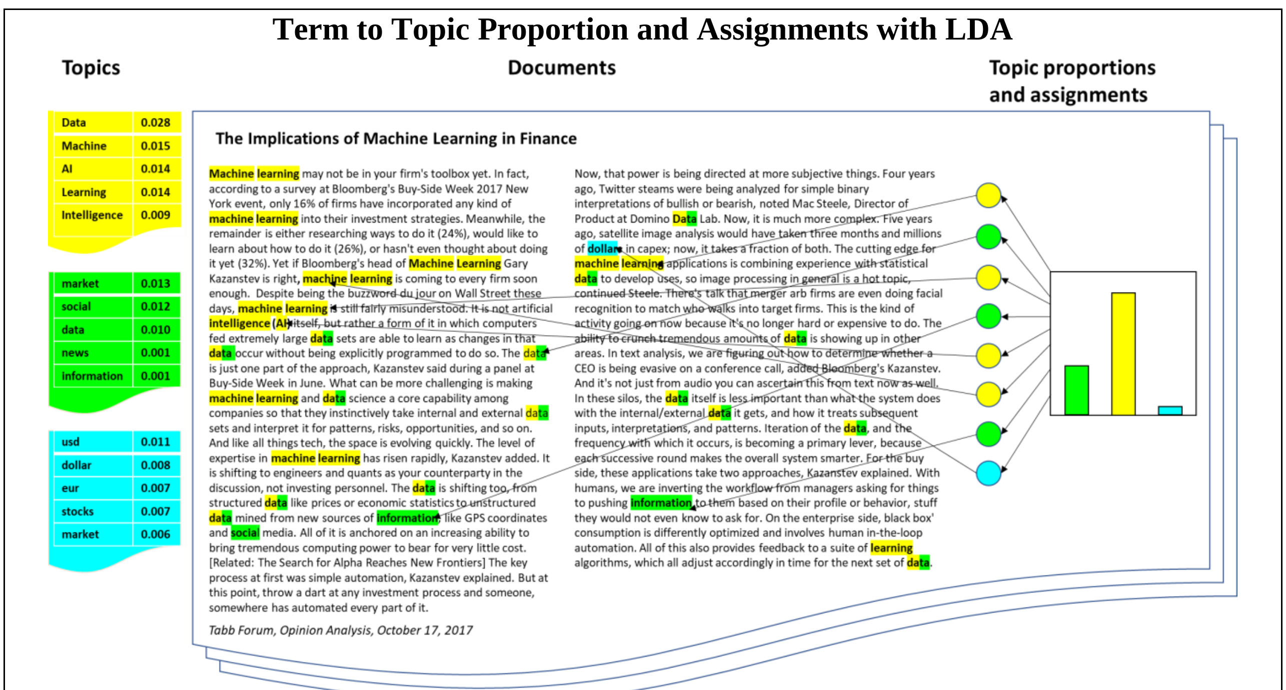


Figure 6b. Single document view of intuitions behind Latent Dirichlet Allocation (LDA) method.

Fusing our Ontology with Topic Models

Working with text data in a pragmatic and business friendly way requires effective summarization and extraction of key ideas in a simplified but meaningful form. By bringing together the higher-level semantic representation of our financial domain concepts with the lower-level signal representation from text, we are able to achieve a key goal centered around the idea of dimensionality reduction. The integration of two parallel workstreams at the signal and sematic levels provides both a top-down and bottom-up solution that can be more specifically-focused to support business processes. Thus, taking an iterative approach to represent financial industry knowledge through the propagation of lower-level signals to high-level symbolic reasoning provides an actionable tool for making sense of the complex financial services landscape.

In our financial markets domain ontology, the notion can be applied pragmatically by taking the higher level semantic view established by our ontology, and connecting it to the lower representation levels found in the signals produced by topic models. As previously discussed, we started our financial markets world-view with specification of the most general concepts. Thus, the output of our topic models provides an explicit opportunity for subsequent specialization of the concepts by linking the ontological and topic model concepts. Moreover, the supplementary measures around frequency, variability, and association of topics provide further avenues to define ontology relationships and model their dynamics over time. Figure 6c illustrates the simple idea of connecting terms found in our corpus back to a domain ontology. By taking a specific concept like algorithmic trading (i.e., High-Frequency Trading [HFT]), we are able to demonstrate a key benefit of making higher-level connections with the multitude of major market forces at play. We chose to expand on algorithmic trading because of its collective significance to our ontology, which can be quantified using the data in Table 1 based on frequency counts of related topics. We can see that the concept HFT trading appeared 98 times in our corpus. Interestingly, the concept is directly associated with 19 other topics of focus. For example, the buy-side trading concept appeared 795 times in our corpus, the single most frequent theme. Yet other more specialized topics such as HFT strategies (450 occurrences), the ‘Flash Crash’ (340 occurrences), HFT Regulation (295 occurrences), and HFT latency (210 occurrences) are a few examples of the broader context surrounding algorithmic trading. With our understanding grounded in several big ideas, we have a much more informed perspective on its significance, and are able to make other inferences such as links to alternate data and the value of social media market news (525 occurrences).

Fusing the Semantic (Ontology) and Signal (Text) Levels

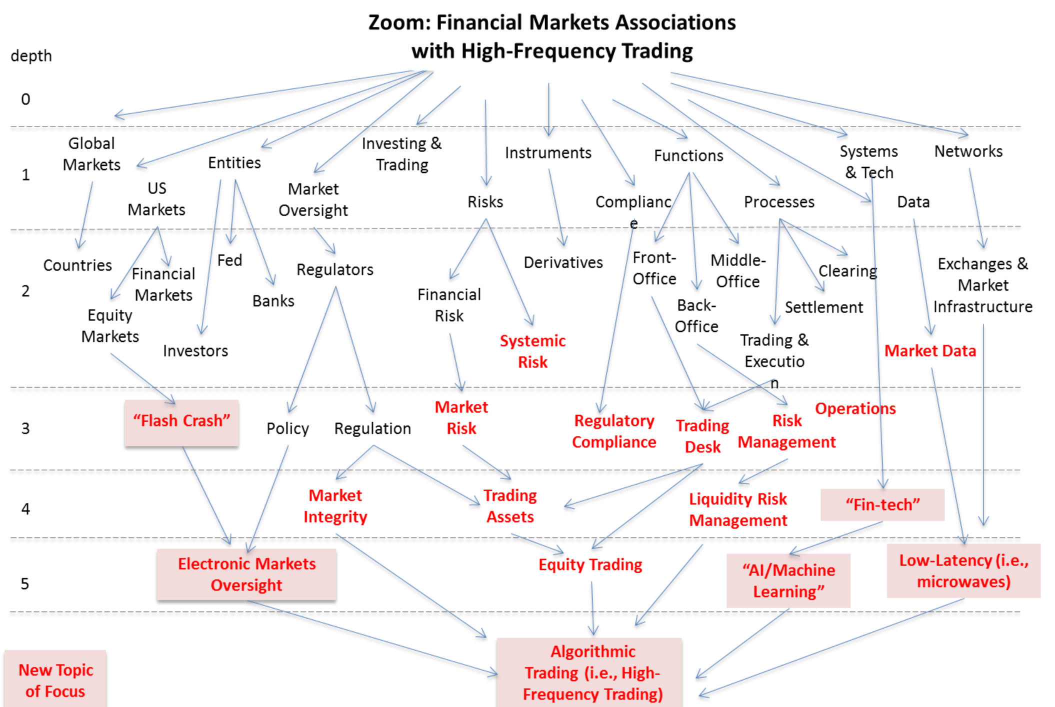


Figure 6c. Single-topic illustration of many semantic connections to financial markets. Differentiating the Levels:

- **Semantic Level:** Base ontology represented by text with black font
- **Signal Level:** Labeled concepts produced by LDA based on corpus terms with red font
- **New Signal Level Concepts:** Important new topics appearing further in time after the 2008 Financial Crisis highlighted with pink boxes as the compliance landscape evolved

#	Category	Topic	Topic Document Count
1	Investing & Trading	Buy Side Trading	795
2	U.S. Markets	Equity Market Volume	650
3	Data	Social Media Market News	525
4	Investing & Trading	High Frequency Trading – Strategies	450
5	Data	Market Data – Real-time	450
6	U.S. Markets	‘Flash Crash’	340
7	Market Oversight	Regulatory Compliance (Financial Firms)	325
8	Market Oversight	High Frequency Trading – Market Regulation	295
9	Investing & Trading	High Frequency Trading – Latency	210
10	Systems & Technology	Artificial Intelligence / Machine Learning	200
11	Market Oversight	NMS Rule Equity Regulation	110
12	U.S. Markets	Trading Markets – Equity	101
13	**Investing & Trading**	**Algorithmic Trading**	**98**
14	Functions	Market Execution	90
15	Functions	Trade Execution – Routing	75
16	Market Oversight	Flash Crash – Regulatory Surveillance	50
17	Networks	Market Execution – Venues	40
18	Networks	Market Trading – speed	38
19	Compliance	Market Surveillance – Trading Data	35
20	Data	Alternative Data – Hedge Funds	25

Table 1. Summary of topics and occurrence frequency related to equity trading.

The following parts of this thesis will describe current knowledge modeling, text analytics and exploratory data analysis techniques; and explore why collectively it can yield insightful results by establishing an ontological world view supported by structural topic modeling. Enabling a pragmatic approach for rapidly extracting actionable insights from large sets of documents, and maintaining it as a holistic ontological model leads to many possibilities for people and systems enabled market intelligence. Data collection and statistical modeling methods are delineated, and a discussion of results is provided. The conclusions drawn from the data reveal that the financial markets are a dynamic and multifaceted domain, which can be analytically informed through text-based topic modeling. Finally, given the importance of understanding the value of financial markets dynamics, there is substantial value in being able to quickly draw connections from low-level signals with a potentially big impact on the broader markets.

Literature Review

Overview

One key objective of our work is to model the complexity of the financial markets. Accordingly, there are several aspects to this undertaking, which can be logically segmented around two key ideas. The first deals with addressing the notion of domain context and modeling the semantic connections between concepts in an ontology or knowledge graph. Secondly, our work employs analytics as a critical facet to understand the nuances in financial market text depicted in news, blogs and other social media forums. Having already introduced the background on our corpus as relevant to this study, the subsequent discussion will review common methods for analyzing text which often start with either a top-down (i.e., ontology development) or bottom-up (topic models) approach. Therefore, the following discussion will cover ontology building, text analytics and information retrieval methods, exploratory data analysis techniques, and topic modeling.

Ontologies

Overview

Most people have developed mind maps to organize their thinking around central themes. The mind map is essentially a model to conceptualize ideas around that theme, and share notions relevant to the topic. The mind map diagram is a method to clarify thinking and share that interpretation with other people. Extending the mind map further, one can recompose the mind map with focus on understanding how concepts are arranged and what relationships exist between the concepts. The result is a more meaningful description of the central theme and its related topics as shown in Figure 7.

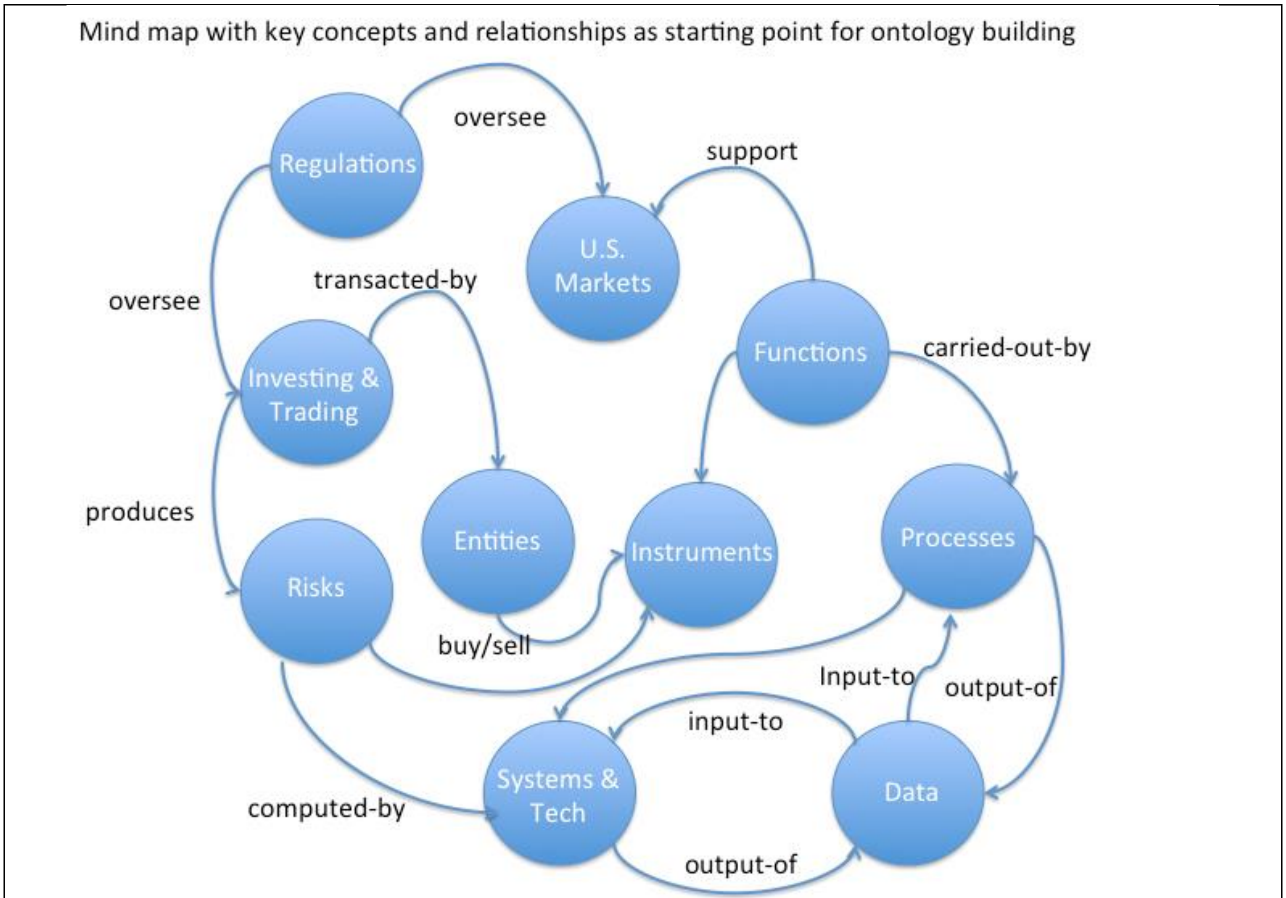


Figure 7. Illustration of an extended mind map often the precursor to building an ontology that shows the relationships between concepts.

In today's automated world, the need to have knowledge representations that are understandable by both people and machines provides tremendous potential for enabling artificial intelligence (AI). As such, taking a mind map and converting it to a form interpretable by both people and machines allows us to achieve more expressive experiences. For instance, better ways of sharing meaning across humans, across machines, and improved interactions among the two with encoded models. Essentially, this notion establishes the power of ontology modeling and setting the stage for powerful knowledge management capabilities.

In simple terms, an ontology is a little like a mind map, but brings in sufficient added structure and self-description that allows the meaning of whatever it is representing to be clearer.

An ontology can be a visual representation for human beings to understand and share, a coded representation for machine interpretation, or both depending on the tools and methods used to create and view ontologies. While there is no single formal definition of an ontology, (Chungoora, 2018) noted the most quoted definition in existing literature was provided by Eftom Gruber, who stated in 1993 that an ontology is 'an explicit specification of a conceptualization.' That definition was later extended by Studer and colleagues, who in 1997 in their paper about knowledge engineering, defined an ontology as 'a formal, explicit specification of a shared conceptualization.' This extension is a concise way of saying an ontology is a machine readable and precise representation of subject matter derived from consensus. In the context of this thesis analysis, an ontology is an agreed blueprint for representing knowledge that is designed to be interpretable by humans and machines. Therefore, an ontology can be applied (i.e., re-used) to meet various needs: capture of meaning (i.e., semantics), domain representation, system design, building controlled vocabularies, business modeling, and many more (Chungoora, 2018). Knowledge sharing is actually one of the key roles of ontologies in computer science. In addition, combined with ontological commitments, the task of knowledge sharing differentiates ontologies from data models, which are mostly intended to be used within a single application (Schreiber, 2008).

In a survey from work related to concept modeling, a widespread general agreement among the majority of cited authors is that uniform knowledge representation should, in general, be achievable by using ontologies; as an intellectual example, knowledge representation is symbolized in Figure 8a. While there is no uniform definition of a concept, it is often accepted that concepts are somehow connected with the process of human cognition; they are also abstract and universal. A similar explanation is given by (Sowa, 2000) who considers a concept to be a 'mediator that relates a symbol to its object'. In the lower two corners of the triangle in Figure 8b, there is an icon resembling a person named John and a printed symbol representing John's name. The cloud on top represents a 'neural excitation' that is induced by an object associated with the symbol representing the respective object. This excitation, a mediator between the symbol and its object, is called a concept.

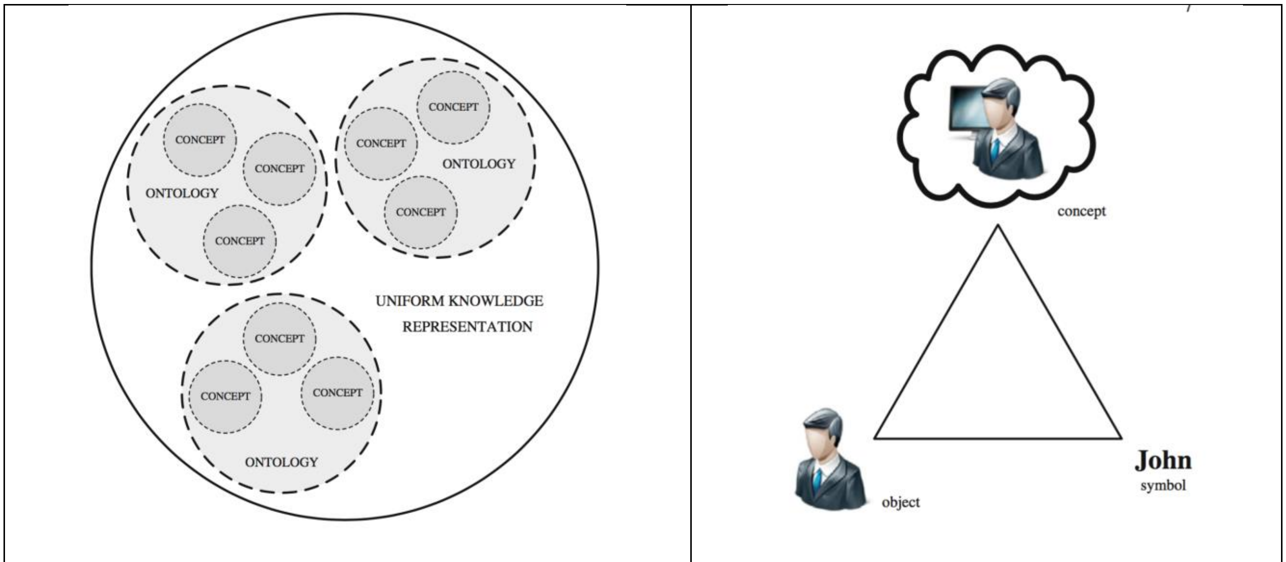


Figure 8a. Uniform knowledge representation consisting of ontologies populated by concepts. *Figure 8b.* Represents John working at his office and appears in one's mind when thinking about a person named John.

Structure

Generally, it's always good when designing an ontology to think in terms of important definitional elements and how their abstractions can be best represented. Therefore, a first order aim of ontologies is to structure how things and instances of things come together through their relationships. Moreover, providing clarity of understanding by extracting how those instances are grouped together in a logical way. Additional considerations such as any constraints that exist provide further ways to model how classes are used.

Before constructing an ontology, some issues regarding the structure and content of ontologies have to be addressed, as noted by (Jakus et al. 2013). Firstly, one needs to choose the proper means to adequately describe a domain by selecting relevant components of the domain model. Another matter to consider when building an ontology is the level of its generality and, consequently, the level of its reusability. Finally, the ontology has to be constructed, either manually, semi-manually or even completely automatically (Jakus et al. 2013).

Using a basic example from our domain, we can illustrate in Figure 9 how these elements come together to structure and depict an ontology by considering the ownership structure of a company. With these relations defined we can see a clearer description of the things we are trying to represent. These associations are the links between the things we describe in our domain. If we wanted to go even further in the definition of our relations, we could start defining rules, constraints, and how they are used. For example, we could say that a company has only one CEO at a time. This assumes that the axiom is agreed and accepted in our domain. While it is conceivable that a company can have two CEOs, it is commonly understood that one CEO is generally accepted. In our case, we want to capture the relationships between companies in a certain way. For example, companies belong to sectors, and in turn sell products and services. These entities can be represented as classes, and may also be described as concept, category, kind and type. The basic means for organizing concepts into knowledge structures are semantic relations. Semantic relations are meaningful associations between two or more concepts or entities (i.e., objects, instances or extensions of concepts) (Khoo & Na, 2006).

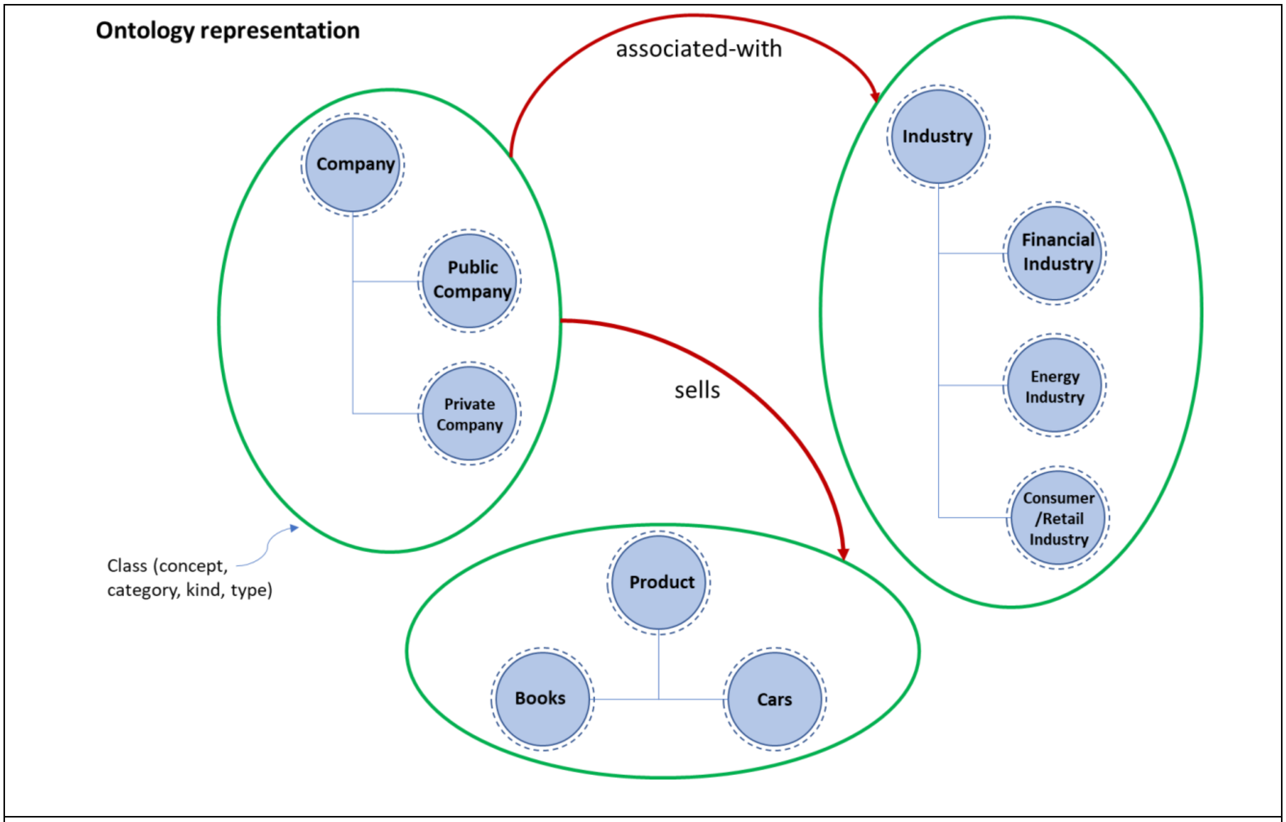


Figure 9. Illustration of how ontologies can organize class taxonomies and represent symbolic relations between classes.

Classes can be grouped together in a tree like structure called a taxonomy. In ontology modeling, a taxonomy works in a similar way and provides the backbone of an ontology. However, there is one important understanding for taxonomies; that is, there is a distinct link between a class and its sub-class, which allows us to model what's called inheritance. Inheritance allows a child class to be extended from its parent, while inheriting the core attributes and behavior of the parent class in a transitive manner. For example, a contract class can have a sub-class of financial contract, which in turn has a sub-class derivative contract. This means that if class financial contract is derived from class contract, and derivative contract is derived from class financial contract, then class derivative contract inherits the attributes of class financial contract, as well as class contract. This is shown in Figure 10.

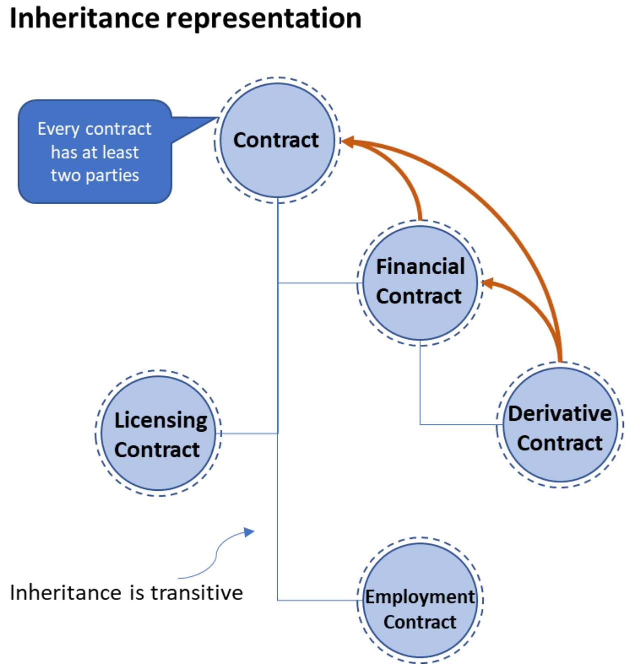


Figure 10. Shows how the transitive property of inheritance transfers properties to all child-classes

By establishing a basic structure, we are positioned to better model knowledge by understanding what each class is comprised of, how it changes over time, capturing other related measures of confidence at minimum.

Representation

There are two main ways of constructing an ontology. The first is through visual representation and the second is through coded representation. There are several ways to visually represent knowledge modeling. The most common are through Unified Modeling Language (UML) diagrams, which are basically visual representations of concepts, with some relationships between them. These diagrams are very suitable for human interpretation, but insufficient for ingestion and understanding by computers.

Visual representations are commonly referred to as 'lightweight' ontologies and have three defining properties. First, they easily allow for sharing understanding across humans.

Second, they are less formal. Finally, they provide for semantic meaning to be readily understood. All this assumes that for lightweight ontologies, the meaning of graphical constructs and entities defined can be understood unambiguously.

On the other hand, a coded representation of ontologies is necessary to make our models mean something to computers. Because of this additional constraint, they are commonly referred to as 'heavyweight' ontologies, which can be described by two facets. First, a heavyweight ontology is machine readable and formal. Second, a knowledge representation language is used. Past and current initiatives in the literature have developed a range of recommendations for various formal logic based frameworks and languages to choose from. Some examples of well-known knowledge representation languages are: the Knowledge Interchange Format developed in 1992 by Geresereth and Fikes, Common Logic (CL), which is also an international standard (ISO/IEC 24707:2007), and the Web Ontology Language (OWL), developed by the world-wide-web consortium (W3C, 2004). OWL is probably the most widely used knowledge representation language at present time

Logic and range of criteria when selecting an ontology language

Choosing the right knowledge representation language requires careful consideration of a range of criteria. The choice is likely influenced by a number of requirements, and the primary ones to consider are: the expressivity of the language, the semantics of the language, and the level of mathematical rigor provided by the language. Given the importance and nuance of each of these measures, each will be described in turn.

The expressivity of a language measures the range of constructs that it provides to correctly and explicitly describe the components of an ontology. For example, if a language only describes the constructs for classes and taxonomy, and not support the description of other kinds of relations, then certain limitations exist on what we would like to express. This prevents specifying a whole range of other kinds of value-added associations.

As another measure, semantics provide the degree of clarity in the meaning of the specification. For example, the notion of precedence which can declare that only a single event can precede some other event in time. In contrast, precedence can allow for multiple events to occur prior to an event, or even allow for events to occur in parallel. Therefore, the clarity of intended meaning of specification of a knowledge representation language is therefore another important consideration when choosing a knowledge representation language for modeling purposes.

Finally, the mathematical rigor of encoding imparted by a knowledge representation language is a key consideration. Two important characteristics for assessing mathematical rigor are the satisfiability and coherence of a representation. Specifically, the integrity and soundness of an ontology. In mathematical logic, satisfiability is when some statement when declared in the universe of a study (i.e., this thesis) has an interpretation that makes it true. On the other hand, coherence can be achieved by ensuring some consistency in the way that the entities have been encoded within the universe of study. The level of mathematical rigor of an ontology depends on the semantics of the knowledge representation language, as well as other factors such as the skills of the modeler and the features of supporting ontology tools.

In summary, heavyweight ontologies (encoded in a mathematically rigorous way), unlike lightweight ontologies (less formal and assume entities readily understood), benefit from all structures of lightweight ontologies but also possess the ability to encode formal axioms and rules. These additional structures help to better clarify the intended meaning of terms in our defined ontologies. Essentially, it allows support for formal capture of domain knowledge as well as domain logic. As a visual way to illustrate clusters for these languages and defined selection criteria, Figure 11 depicts the spread of knowledge representation languages on two dimensions (formality and expressivity).

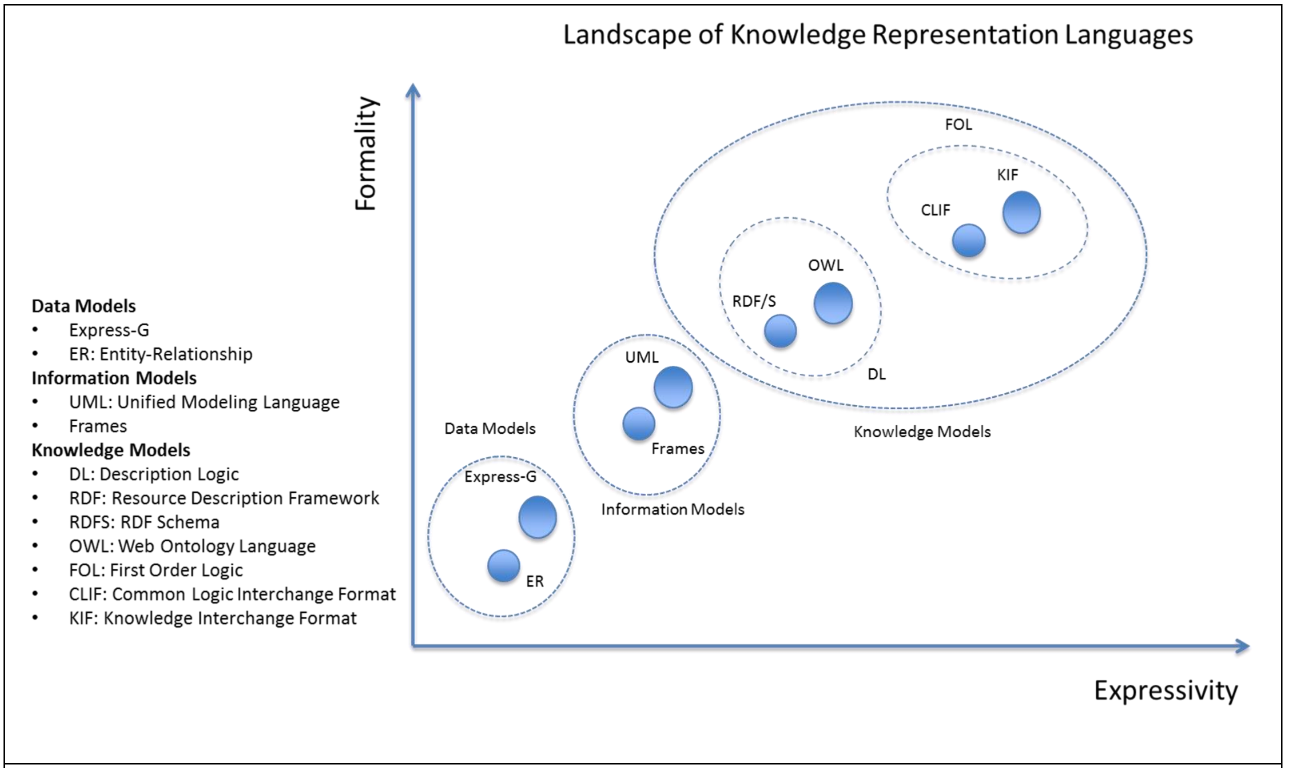


Figure 11. Shows the distribution of knowledge representation languages and how they compare based on the measures of formality and expressivity (Chungoora, 2018).

Topic Modeling

Limitations in search and information retrieval based exclusively on the terms themselves has fostered a need for alternative text search methods. These limitations arise when the concepts described in text documents are not specifically referenced in the terms themselves. Moreover, the volume and variety of text that business professionals and researchers must consider has grown exponentially without any signs of slowing in the near term. As such, topic modeling provides a crucial alternative method to text analysis by incorporating a thematic approach to examining documents in a corpus. Therefore, modern information retrieval techniques have progressed past indexing and supervised learning methods that require humans to codify documents based on a set of predefined rules. Beyond clustering documents by their similarity, additional techniques now exist to group individual words that occur together frequently, and to do this grouping without human intervention (O'Neill, 2016).

Theme based topic modeling is part of a family of algorithms known as probabilistic topic modeling. These models are statistical methods that analyze the words of the original texts to discover the themes that run through them, how those themes are connected to each other, and how they change over time. To this end, machine learning researchers have developed a suite of algorithms that aim to discover and annotate large archives of documents with thematic information (Blei, 2012). The techniques described here sit on top of Latent Dirichlet Allocation (LDA), which makes predictions from the Dirichlet distribution on word associations related to a topic. Conveniently, a text modeler or researcher is only required to provide specific inputs such as the text documents, list of stopwords, and the number of topics for the model to generate. (Blei, Ng, & Jordan, 2003) introduced the LDA topic model, a generative statistical model to detect latent (hidden) structures in discrete data.

In a nutshell, the idea behind LDA is that documents exhibit multiple topics. For example, consider the article in Figure 12. This article entitled 'The Implications of Machine Learning in Finance,' is about using machine learning in investing activities by Wall Street firms, and the importance of what a system does with the data it gets. Using a similar depiction first introduced by Blei, we have manually highlighted different words that are used in the article to aid in understanding how topic models work (Blei, 2012). Words about machine learning, such as 'data' and 'AI,' and 'machine,' are highlighted in yellow; words about social media, such as 'social', 'data' and 'news,' are highlighted in green; words about currency markets, such as 'dollar' and 'usd,' are highlighted in blue. You can see that the article blends machine learning, social media, and currency markets in different proportions. LDA is a statistical model of document collections that tries to capture this intuition. LDA uses a generative process, which means the modeling observations are drawn from a probability density function. Since LDA is probabilistic, is specifies a joint probability distribution over observation and target (label) values.

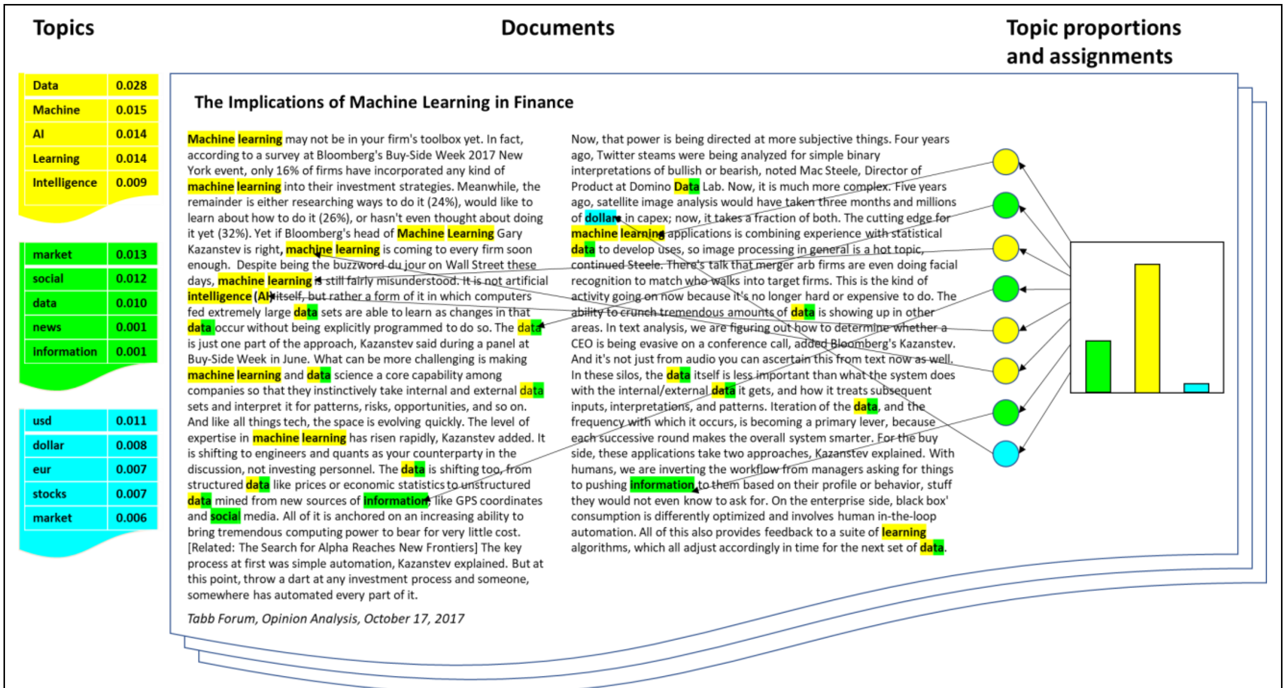


Figure 12. Illustrates the intuitions behind Latent Dirichlet Allocation.

The LDA model works by defining a topic as a distribution over a fixed vocabulary. Therefore, the topics are dependent on the corpus and change as new documents are added (or removed). The model works in a two-step process which first assigns a distribution for each of the number of topics specified as input to the model. Next, in an iterative manner, a random topic is selected from the distribution of topics. In addition, a randomly chosen word from the corresponding distribution over the corpus vocabulary. This process reflects the intuition that documents exhibit multiple topics, and each document exhibits multiple topics in different proportions (Blei, 2012).

Based on the above description of how LDA topic models work, the formula below shows that the probability that a word (W) within a document (D) belongs to topic (Z) is a function of a latent structure expressed in weights and hyperparameters (Underwood, 2012). Equation 3 is a simplified topic modeling formula:

$P(Z\|W,D) = \dfrac{\#\ of\ word\ W\ in\ topic\ Z + \ \beta w}{total\ tokens\ in\ Z + \beta} \times$ (# words in D that belong to $Z + \alpha$)
D = document
W = word
Z = topic
α , β = hyperparameters
Equation 3.

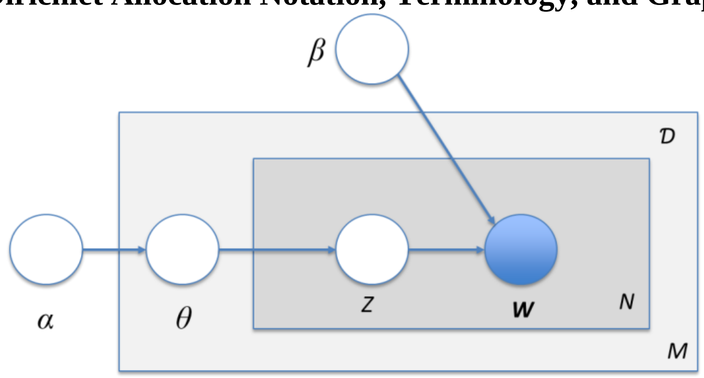
Probabilistic topic models, (Blei, 2012), dynamic topic models (Blei & Lafferty, 2006), and hierarchical topic models are variations on LDA that aim to best predict which words belong to which topics, and the degree to which each topic is present in each document. Before structural topic modeling was formalized, (Quinn et al. 2010) incorporated the variable time into topic model estimation (O'Neill, 2016). For readers interested in better understanding how LDA works, the following section is a specifically-focused discussion on the LDA algorithm.

Latent Dirichlet Allocation (LDA) Algorithm Focus

The LDA algorithm is central to this thesis for its role in generating our topic model output. Therefore, the aim here is to describe in greater detail the LDA algorithm for readers not familiar with it, but who wish to understand its underlying mechanics. A very simple way to think about LDA is that it can be very helpful in dimensionality reduction of text. Treating each word as a feature yields a robust, but very large feature set. However, LDA reduces any document to a fixed set of real-valued features – the posterior Dirichlet parameters associated with the document. While some discriminatory information is lost by reducing the document description to the LDA parameters, it is interesting to see the residual themes (Blei et al. 2003).

Conceptually, it is helpful to think about LDA at three basic levels, which is also represented in Figure 12 below. The parameters α and β are corpus-level parameters, assumed to be sampled once in the process of generating a corpus. The variables θ_d are document-level variables, sampled once for each word in each document. Finally, the variables Z_{dn} and W_{dn} are word-level variables and are sampled once for each word in each document. Importantly, LDA involves three levels which notably allows the topic node to be sampled repeatedly within the document. Under this model, documents can be associated with multiple topics (Blei et al. 2003).

Latent Dirichlet Allocation Notation, Terminology, and Graphical Model Representation



M: A *corpus* is a collection of *M* documents denoted by $\mathcal{D} = \{\mathbf{W}_1, \mathbf{W}_2, \ldots, \mathbf{W}_M\}$

N: A *document* is a sequence of *N* words $w = \{w_1, w_2, \ldots, w_M\}$

For each of the N words w_n:

1. Choose N Poisson(ζ).
2. Choose θ Dir(α)
3. For each of the *N* words w_n:
   (a) Choose a topic Z_n Multinomial (θ)
   (b) Choose a word w_n from $p(w_n|Z_n,\beta)$, a multinomial probability conditioned on the topic Z_n.
4. The parameters *α* and *β* are corpus-level parameters, assumed to be sampled once in corpus generation process.

Figure 12. Graphical Model Representation of Latent Dirichlet Allocation (LDA). The boxes are "plates" representing replicates. The outer plate represents documents, while the inner plate represents the repeated choice of topics and words within a document (Blei et al. 2003).

LDA assumes that words are generated by topics (i.e., by fixed conditional distributions) and that those topics are infinitely exchangeable within a document. The aim is to ‘marginalize’ the topic variables and endow θ with a Dirichlet distribution. (Blei et al. 2013) provide detailed analysis in their seminal paper on LDA that by applying the de Finetti theorem, LDA can capture significant intra-document statistical structure via the mixing of distributions. The aim of the end-result is to find a probabilistic model of a corpus that not only assigns high probability to members of the corpus, but also assigns high probability to other “similar” documents.

To further help follow the intuition behind the LDA statistical model, it is most easily described by its generative process. This is the imaginary random process by which the model assumes the documents arose. Therefore, a key first step is to define topics over a distribution of the fixed vocabulary of the corpus. In simple terms, a randomly chosen distribution over topics is generated. Then in a two-stage second process, each document in the corpus is evaluated to generate the word distributions associated with the topics. For each word in the document, randomly choose a topic from the distribution over topics in step one. Then, in the second stage, randomly choose a word from the corresponding distribution over the vocabulary. The statistical model reflects the intuition that documents exhibit multiple topics, with each document exhibiting the topics in different proportions (Step 1). Each word in each document is drawn from one of the topics (Step 2b), where the selected topic is chosen from the per-document distribution over topics (Step 2a) (Blei, 2012).

The computational problem of LDA can be either sampling-based or variational algorithms. Sampling-based attempt to collect samples from the posterior to approximate to approximate it with an empirical distribution (e.g., Gibbs sampling, constructing a Markov chain). Variational methods are a deterministic alternative to sampling-based algorithms. Rather than approximating the posterior with samples, variational methods posit a parameterized family of distributions over the hidden structure and then find the member of that family that is closest to the posterior. The inference problem is transformed to an optimization problem (Blei, 2012).

In closing, LDA is a statistical solution to the problem of managing sets of text documents. As an unsupervised machine learning algorithm, its use in summarizing and understanding text quickly is becoming increasingly popular. Moreover, many software packages such as R and Python provide multiple implementations of the LDA algorithm for easily integrating into broader text problems.

Methodology

Scope of Project

The financial services industry has encountered significant transformative events in recent years, largely triggered by the 2008 Financial Crisis. In its aftermath, a mandate for structurally altering regulatory reform and changing global market conditions have led to a volatile and uncertain playbook for industry participants. Some forces driving its evolutions include: more than one political regime change, multiple regulators promulgating new rules and implementation mandates, divergent public opinions (i.e., 'wall street vs. main street'), global contagion events (i.e., BREXIT), and entrance of 'fintech' competition. Collectively these events impact financial institutions on three key fronts: 1) raising existential business challenges on the path for investor return; 2) compliance with complex rules, especially for designated systematically important banks in a 'too big to fail' global banking system; 3) addressing operational challenges to manage its complexity in a cost effective and sustainable manner.

Moreover, the industry is comprised of many market participants with diverse needs and contributions, no single authority to govern, and heavy restrictions on information flows in the public domain. As a result, various text data sources such as news coverage, political and regulatory speeches, industry blogs and publications, and social networking all contribute to our collective industry understanding.

Recognizing the substantial complexity of the domain, decisions were made throughout the project to keep the scope manageable in terms of data and objectives. Aspects of the potential scope considered at the outset that were not included were proprietary news sources (i.e., *Wall Street Journal*), certain subscription-based industry publications such as *Risk.net*, and public facing websites that did not allow consumption of their data. Instead, the focus of this study is topics covered by the Federal Reserve Board of Governors' speeches and a wide range of opinion analysis pieces made available to the public by TabbFORUM, a leading financial services industry news forum. While additional sources may have provided greater diversity and removed some bias, the diversity of professional contributors to the TabbFORUM opinion blogs serve the leading financial institutions in the U.S. capital markets and serve as a good representative proxy.

Methodology for Ontology Modeling

Overview

Knowledge representation is an important part of postulating about the world. And there are many considerations such as reason-based inferences, a language for human expression, efficient communication, and so on. But as noted by (Davis et al. 1993), of the MIT AI lab, knowledge representation is, in part, a series of ontological commitments in terms of how we think about the world. Moreover, these commitments begin with the earliest choices accumulating in layers (Davis et al. 1993). As such, ontology building requires an appropriate methodology to allow for establishing an initial commitment and its subsequent iterations in the collection and representation over time.

In practical terms, developing an ontology includes some basic steps: 1) defining classes in the ontology, and arranging the classes in a taxonomic hierarchy, 2) defining slots and describing allowed values for these slots, and 3) filling in the values for slots for instances (Noy et al. 2000). The knowledge base can then be established by filling in specific slot value information and additional slot restrictions. Despite the steps highlighted above, there is no single way to model a domain. The best solution depends on the application or objective.

A review of several ontology engineering methodologies helps to define the range of considerations one should think about when selecting the best ontological approach. The list in Table 2 outlines the most popular ontology methodologies. The methodologies were reviewed for use in this study. After consideration of several factors including purpose, phases and output, complexity, and flexibility, the Ontology Development 101 methodology was selected in favor of its simplicity and incorporation of both top-down and bottom-up approach that aligns to the objectives of this study (i.e., connecting an ontology with topic model output, and linking the two).

Methodology	**Description**
Cyc	This oldest of knowledge bases and ontologies has been mapped to many separate ontologies. (http://www.cyc.com/researchcyc/)
COINS (Context Interchange System)	A long-running series of efforts from MIT's Sloan School of Management.
DOLCE: Descripting Ontology for Linguistic and Cognitive Engineering	Idea is to develop a library of ontologies rather than a monolithic ontology.
DILIGENT Methodology	Aims at creating a set of effective ontologies that the user can share, and at the same time, expand for local use at their will and individual needs.
IDEF5 (Integrated Definition for Ontology Description Capture Method)	Part of a broader set of methodologies developed by Knowledge Based Systems, Inc. (https://en.wikipedia.org/wiki/IDEF5)
METHONTOLOGY	One of the better-known ontology building methodologies; however, not many known uses. (http://semanticweb.org/wiki/METHONTOLOGY)
ONIONS (ONtologic Integration Of Naive Sources)	Set of methods especially geared to integrating multiple information sources, with a particular emphasis on domain ontologies.(http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.22.3972&rep=rep1&type=pdf)
Ontology Development 101	Developed by the authors involved in ontology editing environment such as Protégé.
OTK (On-To-Knowledge)	(On-To-Knowledge) was a methodology that came from the major EU effort at the beginning of last decade; it is a common sense approach reflected in many ways in other methodologies (http://semanticweb.org/wiki/OTK_methodology)

TOVE (Toronto Virtual Enterprise)	A first-order logic approach to representing activities, states, time, resources, and cost in an enterprise integration architecture. (https://en.wikipedia.org/wiki/TOVE_Project)
UPON (United Process for ONtologies)	A UML-based approach that is based on use cases, and is incremental and iterative.

Table 2. Summary of several ontology development methodologies.

Building an Ontology

Ontology engineering is a field that studies the methods and methodologies for building ontologies. As an example of its application, consider a large-scale representation of abstract concepts such as actions, time, physical objects and beliefs. Its ideas are central to conceptual modeling, and form the basis for establishing a knowledge representation model. One of the most popular methodologies is Ontology Development 101 (OD 101) framed as a guide for creating your first ontology (Wikipedia/Ontology_engineering).

Following the principles of OD 101, it is important to maintain the notion that an ontology is a model of reality, and the concepts in the ontology must reflect this reality. Therefore, any initial version that does not align with this objective can be evaluated, debugged by using it in applications, or discussed with experts in the field. Thus, it will likely need to be revised from its initial form and definition. The process is iterative, and will likely continue through the entire lifecycle of the ontology. As a way of describing this process further, Table 3 outlines the major steps of the OD 101 methodology, along with their inputs, activities and outputs.

Step	**Phase**	**Input**	**Phase Description**	**Output**
1	Determine domain and scope of the ontology	Nothing. It is the first step.	Definition: - what is the domain that the ontology will cover? - what ontology will be used? - what types of questions the ontology should provide answers to (competency questions are very important in this domain;	The resulting document may change during the whole process, but at any time, this document helps to limit

Step	Phase	Input	Phase Description	Output
			they allow the designer to understand when ontology contains enough information)? - who will use and maintain the ontology?	the scope of the model.
2	Consider reusing existing ontologies	Documents with domain and scope of the ontology	Looking for other ontologies that are defining the domain. There are libraries on reusable ontologies on the Web and in literature	One or more domain ontologies, or part of them with their description
3	Enumerate important terms in the ontology	Documents with the domain, the scope of the ontology, and libraries on the domain	Write a list of all terms used within the ontology, and describe the terms, their meanings, and their properties	Terms and important aspects to model in the ontology
4	Define the classes and the class hierarchy	Important terms in the ontology, domain, and scope description	There are several possible approaches in developing a class hierarchy: - top-down development process starts with the definition of the most general concepts in the domain and subsequent specialization of the concepts; - bottom-up development process goes in the opposite direction; - a combination development process is a combination of the top-down and bottom-up approaches	Classes and class hierarchy
5	Define the properties of classes-slot	The taxonomy, and the domain and scope description	Add all necessary properties and information that allow the ontology to answer the competency questions	Classes and class hierarchy
6	Define the facets of the slots	Slots and classes	There are different facets describing the value type, allowed values, the number of the values, and other features of the values the slot can take: slot cardinality, slot-value type, domain, and range	Ontology
7	Create instances	The ontology	Create individual instances of classes in the hierarchy, which means choosing a class, creating an individual instance of	Ontology and the modeled domain

Step	Phase	Input	Phase Description	Output
			that class, and filling in the slot values.	
Table 3. Summary of the Ontology Development 101 methodology key steps.				

Developing competency questions as part of the initial steps of the OD 101 methodology is crucial to help focus the scope and efforts of the ontology development. The questions are not intended to be exhaustive, but rather serve as a litmus test on whether the ontology has enough detail to answer these types of questions. For example, a competency question for our domain might be, 'what risks does a broker-dealer face when serving as an equity market-maker to a counterparty?' Another might be, 'Is a convertible security a bond or an equity?' The idea is to share the kinds of classes, concepts and relationships that the ontology will need to answer.

The reuse of existing ontologies is a pragmatic and useful consideration as part of the next step. Not only can it save substantial time and resources, but it helps inform your thinking and establish shared acceptance of the ontology. For example, in documenting countries, one might explore the CIA World Fact Book as an official and credible source. The consideration of existing ontologies can also help with the third phase of the OD 101 methodology, which is to enumerate important terms. It is helpful to list all terms to be explained to a user or that we want to make statements about. At first it is more important to get a comprehensive list of terms than to address overlap in concepts or relations among them.

In our domain, one such example is the Financial Industry Business Ontology (FIBO), which provides a taxonomy for the financial contract. The business problem that FIBO is trying to solve is ensuring a standardized way to share and represent financial products consistently across financial firms. Therefore, through the Enterprise Data Management (EDM) Council, and industry consortium has set out to define a robust, comprehensive and descriptive ontology for encoding financial products and instruments. The following diagram shows FIBO domains that have been modeled with participating industry consensus, and are encoded in several formal ontology languages (i.e., OWL) for consumption and integration into financial processes. Figure 13 illustrates the FIBO domains available as of this thesis (EDM Council, 2018). The FIBO ontology becomes specifically relevant to our financial industry ontology by allowing for integration into the top-level Instrument category (i.e., category 7) of our major level-one topics. This approach is very beneficial on two fronts. First, it saves substantial time in developing a product and instrument taxonomy from scratch. Secondly, FIBO has already undergone industry review and has achieved validation through shared consensus, a critical notion in knowledge sharing.

FIBO (Financial Industry Business Ontology) Domains

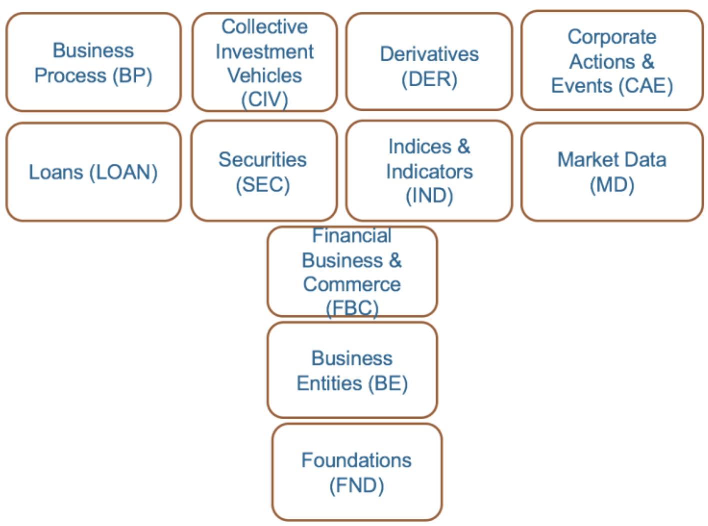


Figure 13. Representation of FIBO Domains available in a formalized ontology language.

The next two steps, developing the class hierarchy and defining properties of concepts (i.e., slots), are closely linked. It can be helpful to work through each of them in tandem, rather than as sequential exercises. It is common to define a few concepts, and immediately continue by describing properties of these concepts and so on. These two steps are the most important in the ontology design process. Finally, with classes and slots defined, the facets and their values can be specified before ending with the generation of instances.

Methodology for Financial Markets Text Data Collection

Overview

A substantial component of the modeling problem with this thesis involved extracting text from public sources, preparing it for modeling (i.e., parsing, cleaning, formatting, etc.), applying LDA topic models, and summarizing the results for subsequent analysis. Given the scope and complexity of this study, a structured approach was taken apply generally accepted text methodology practices, with specific phases and steps adapted to produce the most meaningful topic vocabulary for integration into our financial domain ontology. Figure 14 is a graphical representation of the methodology that was developed for this study, with the steps described in turn in the following sections. In general, there were four top-level phases: data sourcing, data preparation, data modeling and data visualization.

In the context of text analytics and topic modeling, the graphical representation is intended to visually depict the key activities undertaken within each phase. Some steps were not executed due to either time constrains or determined to be of limited value-add (i.e., non-LDA algorithms given our focus is on the LDA topic modeling), and are highlighted with red borders. Overall, the approach was to extract and store text in comma-separated value (csv) files, java-script object notation (JSON), and transform, model and visualize with Python and R scripts. Appendix B illustrates an example JSON file of the processed output (described later) and Appendix C illustrates an example Python web-scraper used for text extraction (also described later). During the modeling phase, the use of Amazon Web Services was used to perform processing in the cloud given the need for additional processing power for the LDA modeling.

Visual Representation of Text Analytics Methodology Followed for LDA Topic Modeling

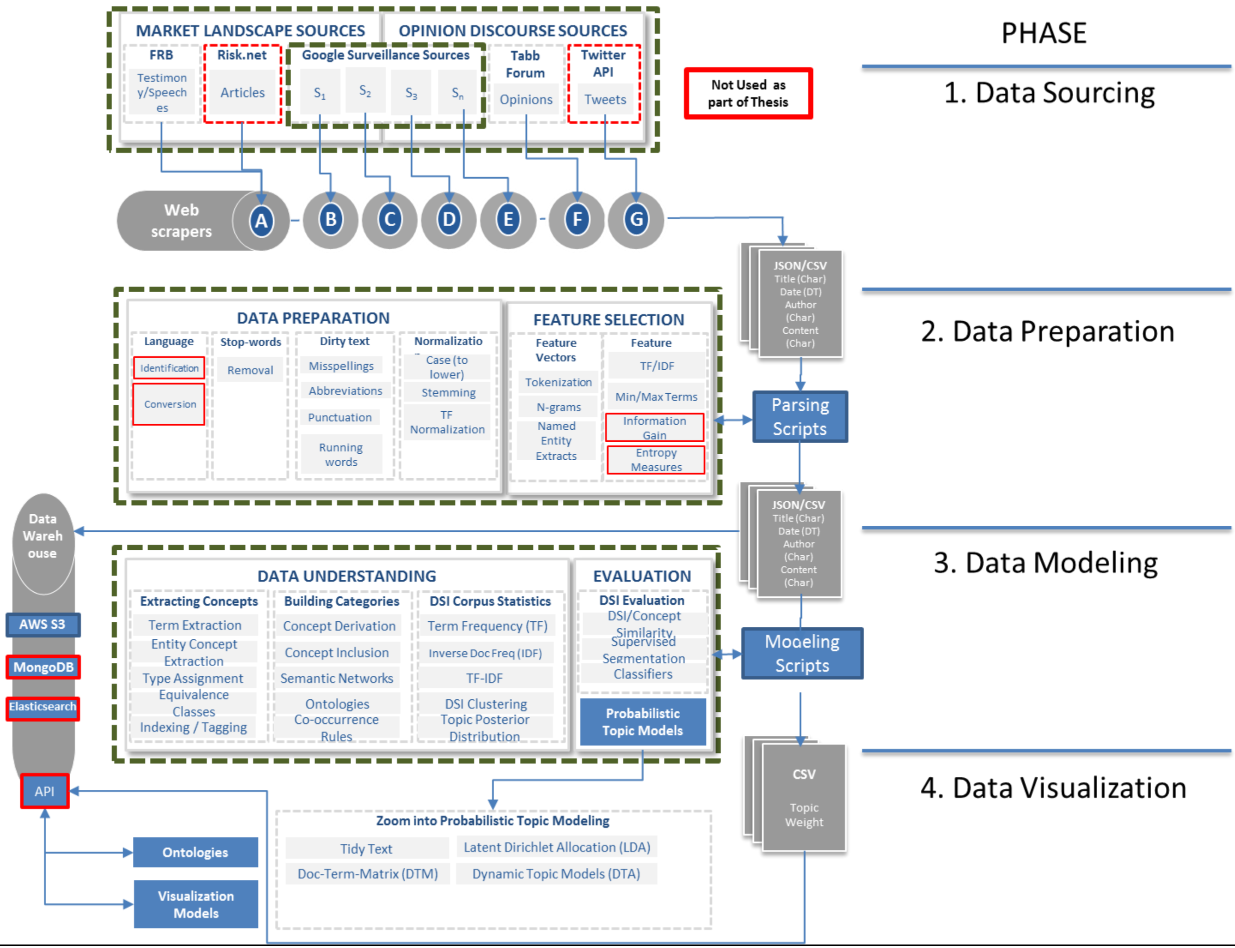


Figure 14. Illustration of the overall methodology that was followed adapted from CRISP-DM with functional steps on how data is prepared, understood, and modeled, and how the resulting models are evaluated.

Identifying Sources

The amount of data available for inclusion in this study was massive, so initially a very wide net was cast. A key initial goal was to get a robust and diverse set of sources representative of all the major mediums, therefore, data sources from three categories were considered: 1) news and publications; 2) direct communications from market participants; and 3) informal communication from market participants. These channels and potential sources that were explored are illustrated in Figure 15.

The news and publications channel turned out to be the most difficult sources to secure, and were ultimately not included in the study. As a first option, direct news publications were considered such as *The Wall Street Journal*, *Bloomberg*, *Reuters*, and others. However, due to legal constraints for using their content and website terms of use that prevented scraping their data, the sources were avoided. In an alternate path to secure news and publications, a wide scan of potential news sources was explored by utilizing Google's free service. News alerts were created as RSS feeds for a set of potential market topics (e.g., financial regulation, equities, fixed-income, financial derivatives, fintech, and automation/analytics), and scraped and consolidated for several weeks using a Python script. This process was repeated several times during the first few months of the project. The result was 109 potential news sources that could be scraped for text and included in our data corpus. However, a consistent pattern emerged similar to the news sites in that the terms-of-use did not allow the sites to be scraped or content downloaded in bulk. As such, alternative sources were pursued outside the news publication category.

As a second category, formal direct communication from market participants as potential text sources was explored. To this end, the Federal Reserve Board of Governors speeches delivered by the Board of Governors over the last 10 years was selected as a representative source. The rationale for the selection is based on the diversity of topics covered from approximately 70 speeches delivered a year, and given its representation as a credible and regulatory perspective, and public availability of the content on the FederalReserve.gov website. Moreover, the Federal Reserve oversees the entire U.S. financial system and ultimate regulator for most financial institutions in the U.S. markets.

For the final category of informal direct communication from market participants, a few potential sources were considered. The primary source selected was from data published publicly on the website of TabbFORUM, a credible financial services industry share forum. The site provides the capability for market participants to share opinion and analysis, and who professionally represent a variety of leading consultancies, market data providers and financial institutions. As such, the views expressed are personal, but aim to cover pressing topics, issues and challenges facing the U.S. capital markets. Moreover, the sight explicitly permitted the use of web scrapers to extract content from there site as long as contributors' names were acknowledged and that source was the TabbFORUM when making any direct references.

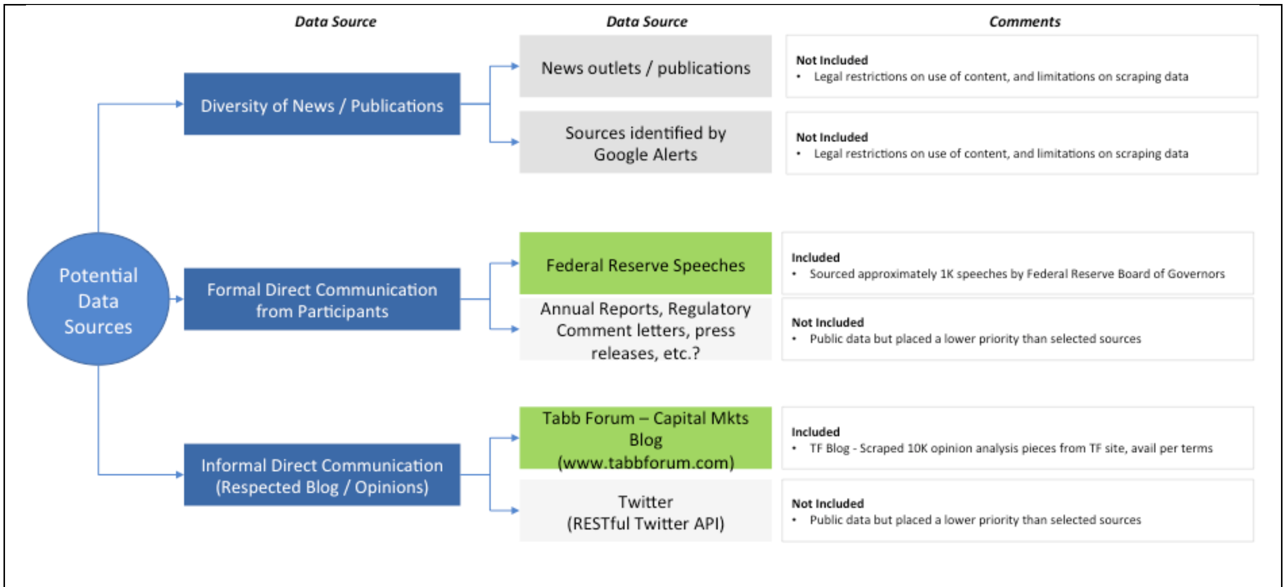


Figure 15. Illustrative view of the potential sources explored, with final scope highlighted in green.

Requests for Participation

Many websites strictly prohibit both the automated collection of their content and the redistribution of it on the Web. To avoid infringement of the terms and conditions, each of the websites' policies were closely reviewed. Some websites explicitly allow use of content for personal or academic purposes, while others only allowed with prior permission from the site. While some attempts were made to request prior authorization and explain the academic nature of the study, no responses were obtained. *Risk.net* was willing to grant permission to use the site's content, but only with a subscription. While these terms were fair, it violated the Northwestern University policy to only use publicly available data sources. Thus, only the TabbFORUM and Federal Reserve Board of Governors' content was used to avoid any legal infringement issues.

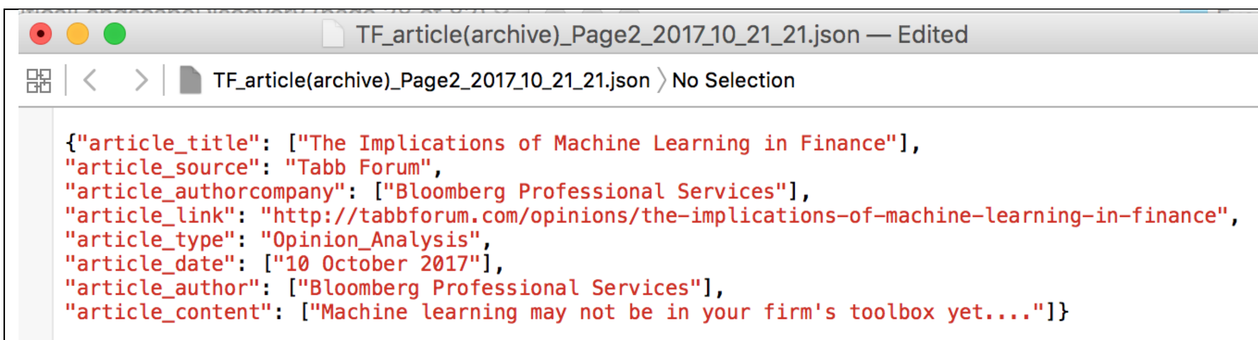
Data Collection via Web Crawling and Scraping

While the Federal Reserve Board speeches and TabbFORUM content is available online, it is not organized for download in bulk. Thus, automated data collection had to be performed with a Web scraper. While many tools exist to aid in the consumption and scraping of web data, Python was selected because it has several packages that are actively maintained, including Scrapy. Scrapy is a robust tool with an excellent framework for programmatically crawling links and obtaining related text sources by following links on a page. Appendix C provides an illustrative example of a Scrapy Web scraper in Python used for extracting the Federal Reserve Board of Governor speeches.

Using Python to develop both custom scripts, as well as utilizing the Scrapy package, a two-step approach was followed for each source. The first step was to crawl the index page by iterating through any necessary pagination, and extracting the links of desired content. For the Federal Reserve Board of Governor speeches, this resulted in a file containing the links and available metadata for every speech between 2009 and 2017. While Federal Reserve speeches were available all the way back to 1996, prior to 2007 a different system was used, and each speech had unique HTML tags introducing many scrapping errors. However, the focus of this study is on the period following the 2008 Financial Crisis. For the TabbFORUM opinion analysis pieces, a similar index scraping step was performed with a separate script, and also generating files with links and metadata to the specific data sources.

Upon generation of the list of links to be scraped for each site, a second set of Python scripts were developed for each source to extract the content of the data source as well as available metadata that could be used for covariate purposes as part of our analysis. Since Python is an in-memory interpretive language, persisting the extracted data sources to a permanent file is necessary. A JSON format was selected given its simplicity to persist as text files, and flexibility as an unstructured text format. Figure 16 shows an illustrative example of the JSON format for maintaining the content and metadata associated with each scraped data source. A detailed description and example of a single pre-processed DSI is presented in Appendix B to illustrate the parameters and format of inputs available for model inputs.

For each scraped document, the following data were collected: title, author, timestamp, link, type, source, author company (if available) and full text of each article as content. Website structure and schema vary widely, so each data source needed its own custom scraper. However, the output of the scraper generally standardizes the format as shown in Figure 16.




```
{"article_title": ["The Implications of Machine Learning in Finance"],
"article_source": "Tabb Forum",
"article_authorcompany": ["Bloomberg Professional Services"],
"article_link": "http://tabbforum.com/opinions/the-implications-of-machine-learning-in-finance",
"article_type": "Opinion_Analysis",
"article_date": ["10 October 2017"],
"article_author": ["Bloomberg Professional Services"],
"article_content": ["Machine learning may not be in your firm's toolbox yet...."]}
```


Figure 16. Illustrates a sample JSON file format generated by the Python text scraper for each data source to be subsequently pre-processes and input to the topic model.

Defining Covariates

Structural topic modeling requires as input various parameters referred to as covariates. These metadata tags allow for the description of each DSI in a structured format so that subsequent analysis can consider while segmenting, grouping and describing the data. In a nutshell, a covariate can be an independent variable (i.e., of direct interest) or it can be an unwanted confounding variable. Adding a covariate to a model can increase the accuracy of your results. As such, an effort was made to consider the covariates that would add the most value to this study considering the range of available inputs. Covariates in essence allow us to collect data on characteristics before running an experiment. For example, the covariate data can be used to see how treatment affects different groups or populations. Or, the data can be used to control for the influence of any covariate.

Therefore, a critical step is to define and tag covariates at the beginning of the data collection effort before any subsequent transformation of the data. Figure 17 shows the design of covariates intended to be captured as part of this analysis. Some of these are straight forward such as the data source title, author, and timestamp. However, a few additional business rule covariates were designed in order to allow for some basic classification of the data. Specifically, three business rules covariates were included based on the presence of keywords themes (i.e., ‘data’, ‘regulation’, and ‘artificial intelligence/machine learning’).

As a means to incorporate covariates into the data, several processing tasks were completed in the open-source software framework R. At his point the covariates were extracted from the source in raw (i.e., without change) format and added to the resulting JSON file. First, the timestamps for each source had to be converted to a single format using the lubridate package. Following an approach used by (O’Neill, 2016), all dates were formatted using the

JODA timestamp styles (Joda.org, 2016). For example, the data ‘2017-07-01 08:50:25’ would be matched by the lubridate package as ‘yyyy-MM-dd HH:mm:ss’. Then the year and month were extracted from each publication date, and dates were represented by a variable called ‘monthterm,’ an integer corresponding to the sequential position of the month from January 2009 to November 2017. This was used for chronological analysis. The monthterm variable is used as a covariate in the STM model so that topics can be evaluated as they become more or less prominent over time.

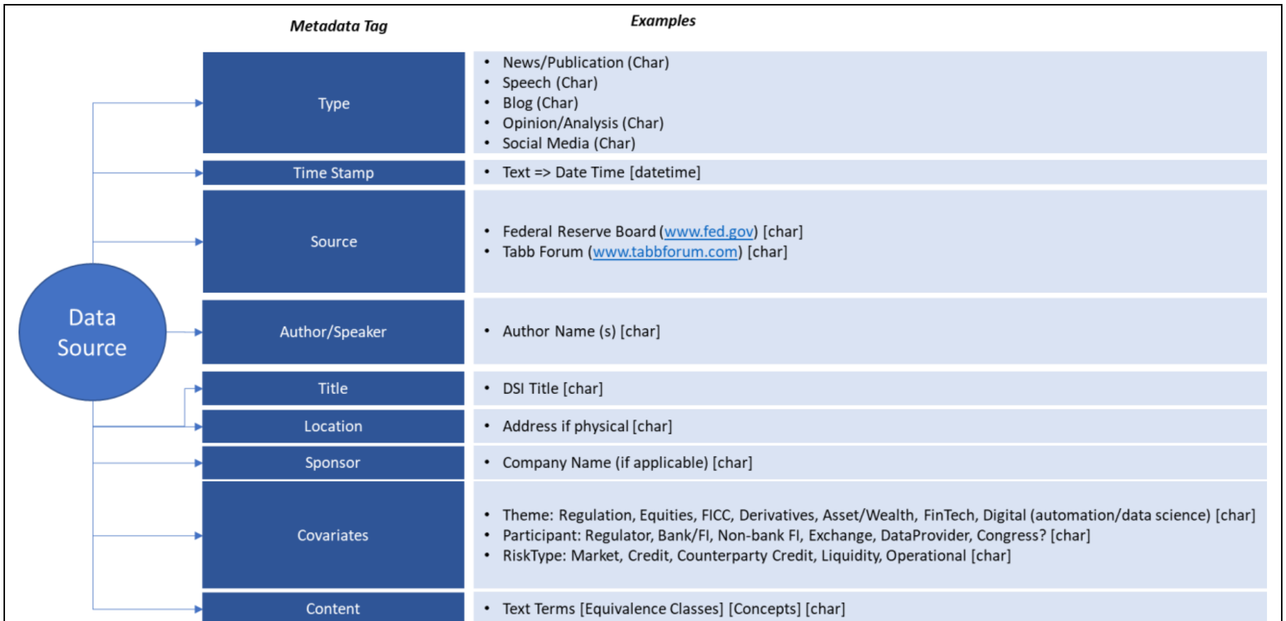


Figure 17. Shows the specification covariates to be collected for each data source.

Data Cleaning

Text data extracted from the web often includes many words, characters and symbols that are unnecessary for meaningful text analysis, and thus require pre-processing steps to remove the text not needed. Examples include HTML tags, hyperlinks, media and advertising banners, unrelated side-bar text blocks, and many others. Given each data source provides specific challenges, a customized approach is potentially needed for each source. However, the data sources for this study were ‘clean’ in nature given the editorial oriented nature of the authors;

there was a general confidence in the data to maintain a consistent structure for all similar data source items on each site, and thus, pre-processing was limited to a common set of steps almost always required to prepare text data for modeling.

The pre-processing steps were performed in an object-oriented Python script that carried out three basic tasks. First, distinct elements such as title and content from each data source item were aggregated into single new feature called 'dsi_aggregated_content' to avoid missing any meaningful key-words split across the two. The original features of title and content were also maintained for potential analysis at a later time on only one feature (e.g., text from each title).

Once the aggregated feature was defined and populated, a second transformative step was performed to replace pre-defined named strings with a consistent single token. For example, 'U.S.' was recoded as 'unitedstates,' or 'Capital Markets' to 'capitalmarkets.' The analysis of ngram-combinations can potentially require a fair amount of time and computing resources. Therefore, at this stage only a minimal set of pre-defined named strings were replaced to avoid any negative impacts on later modeling. Finally, a few custom functions were called on each data source item to remove single character stop items (e.g., removing carriage-returns/line-feeds, non-alphanumeric characters, and single character words with a space), remove the standard set of stop words with typically no value-added in text analysis (e.g., 'and', 'the', 'but', etc.), and lastly replacing multiple blank characters with one blank character. Appendix D provides a list of the stop word items removed from each DSI in pre-processing steps.

Many other text cleaning steps could be taken here depending on the scope of data sources such as language translation, misspellings, abbreviations, and colloquialisms. However, these additional steps were not necessary for data used in this study. The transformed output was saved to a new JSON file including a new feature with all transformed text called 'dsi_aggregated_contentPP.' The net result was a clean and standardized JSON file across both data sources for subsequent consumption by the model, while preserving the original text.

Preparing Data for Modeling

One of the most common structures that text mining packages work with is the document-term-matrix (or DTM). This is a matrix where each row represents one document (e.g., book, article, etc.), each column represents one term, and each value contains a number of appearances of that term in the document. Since most pairings of documents and terms do not occur (they have zero values), DTMs are usually implemented as sparse matrices (Silge & Robinson, 2017). There are several ways to represent text data in R for consumption by popular topic modeling packages such as tidytext (tidy), quanteda (tm), and topicmodels (LDA). For our analysis, we will generally require two basic formats: 1) a data frame format which allows for performing various selections, filters, groupings and aggregations of the corpus; 2) a document-term-matrix (DTM) which is directly input to the topic model. Moreover, it is common to switch back and forth between the two since modeling is a dynamic process.

For our analysis, we need a framework and set of libraries that allow us to work with our text data in both data frame and DTM formats. Figure 18 is adapted from Text Mining with R, and illustrates how switching between various R packages can occur depending on the particular phase of analysis (Silge & Robinson, 2017).

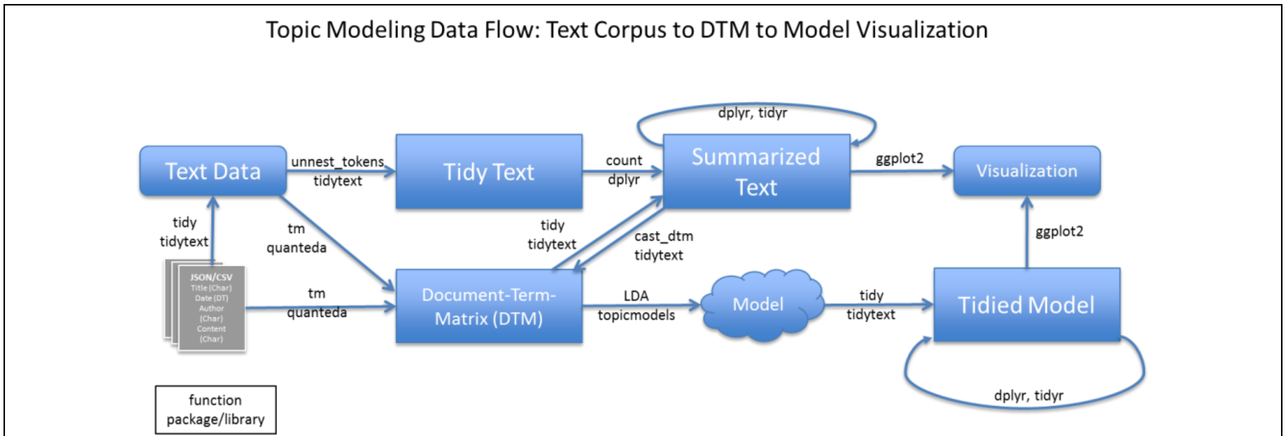


Figure 18. Shows a flowchart for a text analysis that incorporates topic modeling. We use a variety of names packages for both summarizing and modeling text data. Conceptualization redacted from Text Mining with R.

Another package used in our analysis is the STM vignette which describes the ‘textProcessor’ and ‘tm’ functions that assist in the preparation of text corpa for modeling (Roberts, Stewart, & Tingley, 2016). The textProcessor function, similar to the tidy function in tidytext, tokenizes each document based on a set of parameters. The STM function is used for structural topic modeling and allows for consuming multiple parameters by the structural topic model. Perhaps the most important decision at this stage is whether or not to use stemming (O’Neill, 2016). Setting the stem parameter to ‘TRUE’ means that only the stem of each word is indexed. For example, ‘company’ and ‘companies’ are represented by a single term ‘compan’ rather than two distinct words. While choosing to stem a vocabulary decreases the number of terms and helps with computational efficiency, it is harder to read. Given the potential complexity of our context, the stem parameter was set to ‘False’ so that results we more easily understood.

One last consideration is a choice to limit the number of tokens in the vocabulary based on some threshold that can be helpful when vocabularies contain large numbers of uninteresting words. This approach can achieve a similar purpose as the inverse document frequency metric.

Choosing K Topics

One of the key inputs to the LDA model is selecting the *K* number of topics. We followed a general approach for selecting *K* by selecting an initial value for *K*, and then through trial and error over many iterations of the model with different values for *K*. For this thesis, an initial value of one-hundred *K* topics was selected, with variations on subsequent iterations. As some point selecting high values for *K* does not produce more meaningful results, and an overlap in topic concepts emerges. Therefore, our approach was to perform a second pass on subsets of the corpus related to certain topics with different values for *K*, rather than continually increase *K*.

Interpretations and Labeling

At first glance, topic models provide a sense of mystique as they mask the sophisticated algorithms used as part of their computations and produce a simple set of words most likely to appear with each topic as output. But a further simplification step is required which is to assign a single label or concept to each topic. The association of a single label provides a short word or phrase to capture the essence of each topic and use as part of further analysis and communication. A second important benefit is that labeled topics can be more easily associated with the classes and concepts in our ontology, while connecting back to our knowledge representation models. Applying short labels to topics is a subjective process, and at times requires the input of a subject matter expert. One example from these data would be the following group of words: 'markets', 'u.s.', 'regulatory', 'cftc', 'financial', 'American', 'trading', 'economic', 'regulation.' Depending on the analyst, he or she might choose to label this 'U.S. trading markets' or 'Regulated Markets.' A subject matter expert might even go one step further and choose 'U.S. Derivatives Markets' since the 'cftc' is associated with the

Commodities Futures Trading Commission, who oversees the futures and swaps market in the U.S. It is important to control for bias and ensure personal opinions are avoided.

Another common challenge when labeling topics is the result of words that do not have any obvious association with other words returned by the topic model. One approach is to categorize topics by the difficulty level in labeling. Figure 19 provides examples of easy, medium, and hard to label topics. By taking a first pass at the easy and medium topics, one can then focus on the remaining hard to label topics. Generally, there are three options: 1) consult domain experts to ensure no relevant connections are missed, 2) revisit the pre-processing step to ensure data is clean on no unwanted terms are included, 3) assign to a general 'other' category simply to avoid making an interpretation.

Visualizing the results of topic models is very powerful, but often crowded when more than 15-20 topics are used for K. As such, it may be necessary to summarize through aggregations and groupings before preparing charts and graphs. Another approach to simplify the interpretation of topic models is to organize topics into groups or clusters. For example, Appendix E shows that there are many topics that have to do with one of thirteen major categories such as Compliance or Data specifically. The process of labeling topics is iterative and only becomes clearer after repeated cycles. The most important outcome of topic aggregation and labeling is that the reader understands and is not misled. As such, incorporating a feedback control where a shared consensus can be obtained is incredibly helpful. As a reminder, a shared consensus is part of the definition of an ontology which will become important as we aim to link our topic model results to our ontologies.

Easy **Label:** Data Science

Topic 4: Topic_4

Words sorted by highest frequency/probability:

[1] “data”, “machine”, “firms”, “AI”, “learning”, “regulators”, “intelligence”, “artificial”, “human”, “based”, “information”, “technology”

Medium **Label:** Regulated Change to Risk Management

Topic 2: Topic_2

Words sorted by highest frequency/probability:

[1] “compliance”, “regulatory”, “firms”, “risk”, “financial”, “regulators”, “management”, “business”, “change”, “firm”

Hard **Label:** Business Friendly Regulation

Topic 36: Topic_36

Words sorted by highest frequency/probability:

[1] “financial”, “regulation”, “trump”, “industry”, “global”, “rules”, “business”, “voice”, “UK”

Figure 19. Examples of Easy, Medium, and Hard to Label Topics.

Results and Analysis

An Ontological Model of the Financial Markets

Macro view of the financial markets

Financial markets in simple terms provide a physical or virtual forum to buy and sell financial securities (e.g., stocks) and derivatives such as futures and options at low transaction costs (Wikipedia/Financial_market). General market participants of all sorts exchange products and services and thus require financial services as a means to facilitate these financial contracts. For example, a manufacturer obtains a loan for funding inventory production, which is funded by the bank making the loan on the secondary bond market by a variety of retail investors. The data surrounding the transaction in the financial markets provides technical information for a wide variety of analysis. In today's modern banking system, financial markets provide vast numbers of signals about the performance of companies, banks, assets, and economies. Risk managers and regulators can use these signals to better understand economic dependencies, correlations and phase transitions (Soramaki, 2016). As such, the following analysis provides a framework for conceptualizing the financial markets landscape. By using a popular ontology development methodology (i.e., Ontology Development 101) to map multiple-dimensions of financial markets participants and their connections, the breadth of financial markets concepts can be tagged and linked through further synthesis. By establishing an ontological world-view, we take an important step towards encoding the semantic level of the dynamics of the financial markets. While the notion of semantic or symbolic representation is not new, it will provide an important means to represent what we know about things, people and abstract concepts that are foundational for navigating the financial industry domain (Shi & Weninger, 2018).

While the financial markets may be unfathomably complex, at the highest level we can quickly build a view of the major components. Figure 20a provides a first-level ontology of the financial markets based on the ideas of (Williams, 2011) in *An Introduction to Trading in the Financial Markets*. Even from this very simple level-one view, we can begin to see important connections and groupings. For example, entities are central actors and directly connect with seven of the thirteen concepts. Yet we clearly see that operational links to technology and data occur by way of the entities' functional processes. Given the many ways to model and build out an ontology, an important first step is to select the domain and scope of the ontology. Therefore, we raise several key questions that will provide a basis for knowing if our ontology is directionally complete. In general, we are interested in the most important topics that have been debated in the years subsequent to the 2008 Financial Crisis and how expansive they are with respect to the entire financial markets domain. What risk factors remain critical, and have new risk factors emerged? What are the biggest operational challenges, and means to address them? How will emerging and disruptive technologies such as machine learning play a role in the financial markets? Collectively, does the banking system still pose a risk to the broader main street economy, or has the negative post-crisis sentiment subsided? While there are many potential questions, these are a few questions at minimum that will help shape efforts to build our basic world-view. The subsequent journey to identify new concepts and maintain current connects will never be complete. Therefore, defining an expressive conceptual model can establish a baseline for ongoing consensual validation and usability.

Level One Financial Industry Ontology

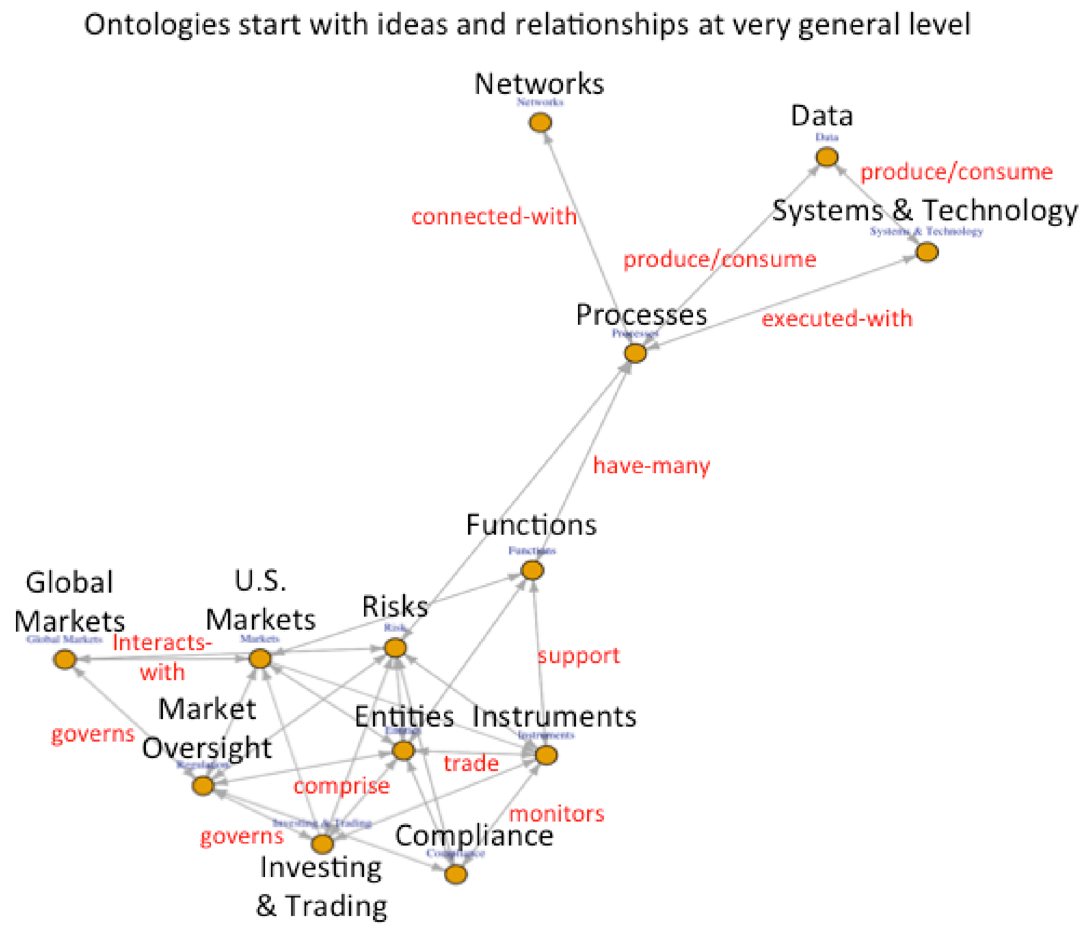


Figure 20a. Level One effort in building our world-view of the financial markets

In advance of building out the class hierarchy for linking important concepts in the domain, it is helpful to enumerate important concepts and terms in the ontology. This next-level ontology development will provide important links for subsequent modeling and ensure that a world-view includes the minimum dimensions for representation. In Figure 20b, we take our initial Level One ontology and build out further the concepts into their key Level Two sub-classes. At this point, we are not trying to establish any measures to distinguish the relative importance of some topics over others, or the strength of their relationships. Rather, the goal is to simply capture the representative terms at the highest levels that could be built out if helpful for the objectives of the ontology. Moreover, at this point we are developing a lightweight ontology that is not intended to be machine-readable. However, using the power of network visualization we can represent the classes of our domain ontology and sub-classes in the initial hierarchy with a simple network.

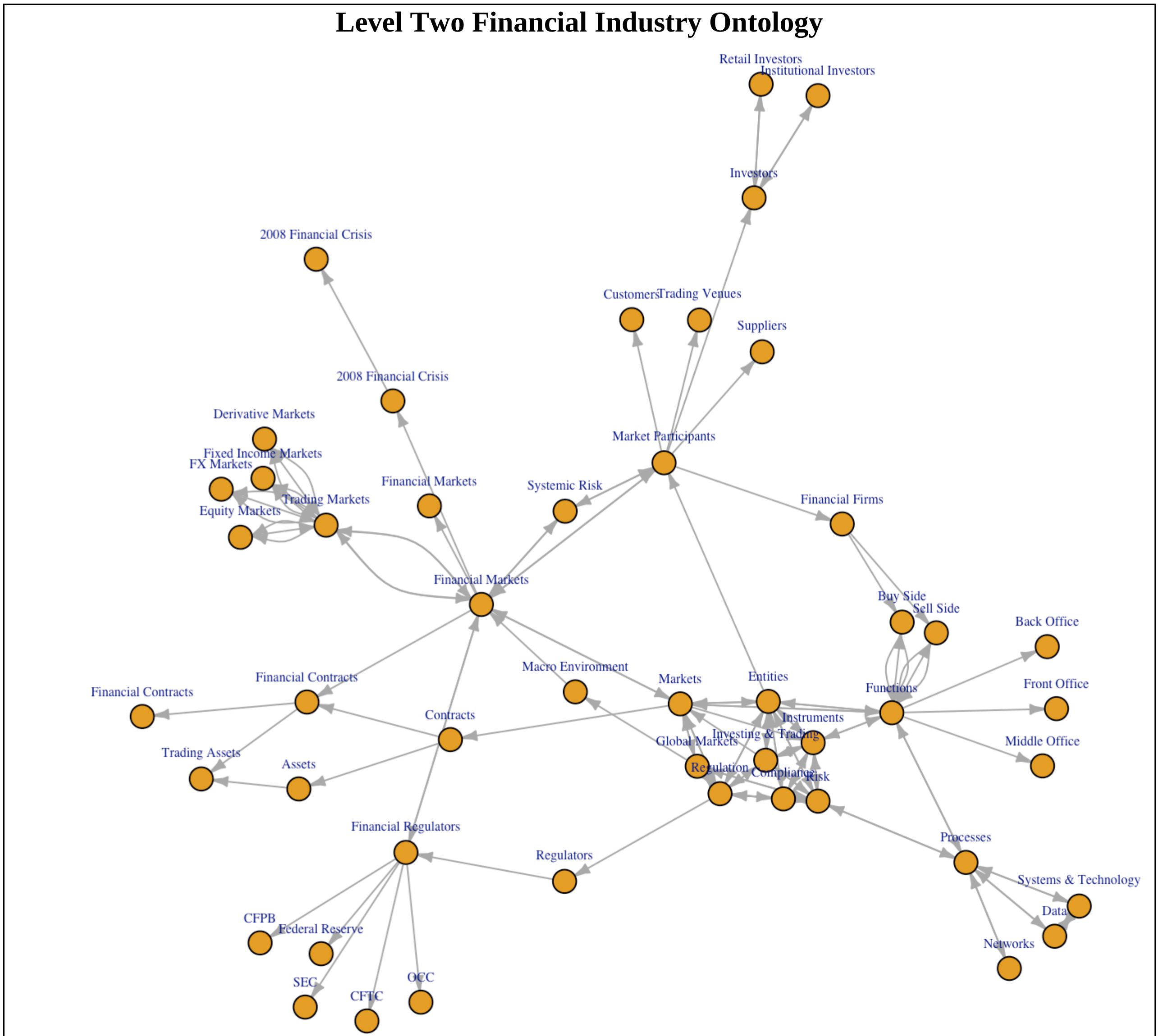


Figure 20b. Level Two Key Term Enumeration (Step 3 of Ontology Engineering 101)

Financial Markets Landscape Represented by its Topics

Overview

The analysis of 11,693 documents comprised of Federal Reserve Board of Governors regulatory speeches and opinion pieces from the capital markets industry group TabbFORUM revealed 302 topics from the aggregation of first and second pass topic model results. In the context of our domain, several patterns emerged that align to the top-level categories of our domain ontology described previously with the topic model output. As such, the identified topics were mapped to one of thirteen major categories defined in our top-level ontology. Figure 21 provides an overall view of the proportion of documents containing topics within the thirteen major categories. To be discussed later, each of the major categories of topics is shown in Table 5, listing example concepts identified by our LDA output and the number of sub-topics. The thirteen major categories serve as the top-level alignment to our financial markets domain ontology, and provide conceptual anchor points for connecting the knowledge level with the 'statistical' or term-level results of the topic models that will be reviewed in the following sections.

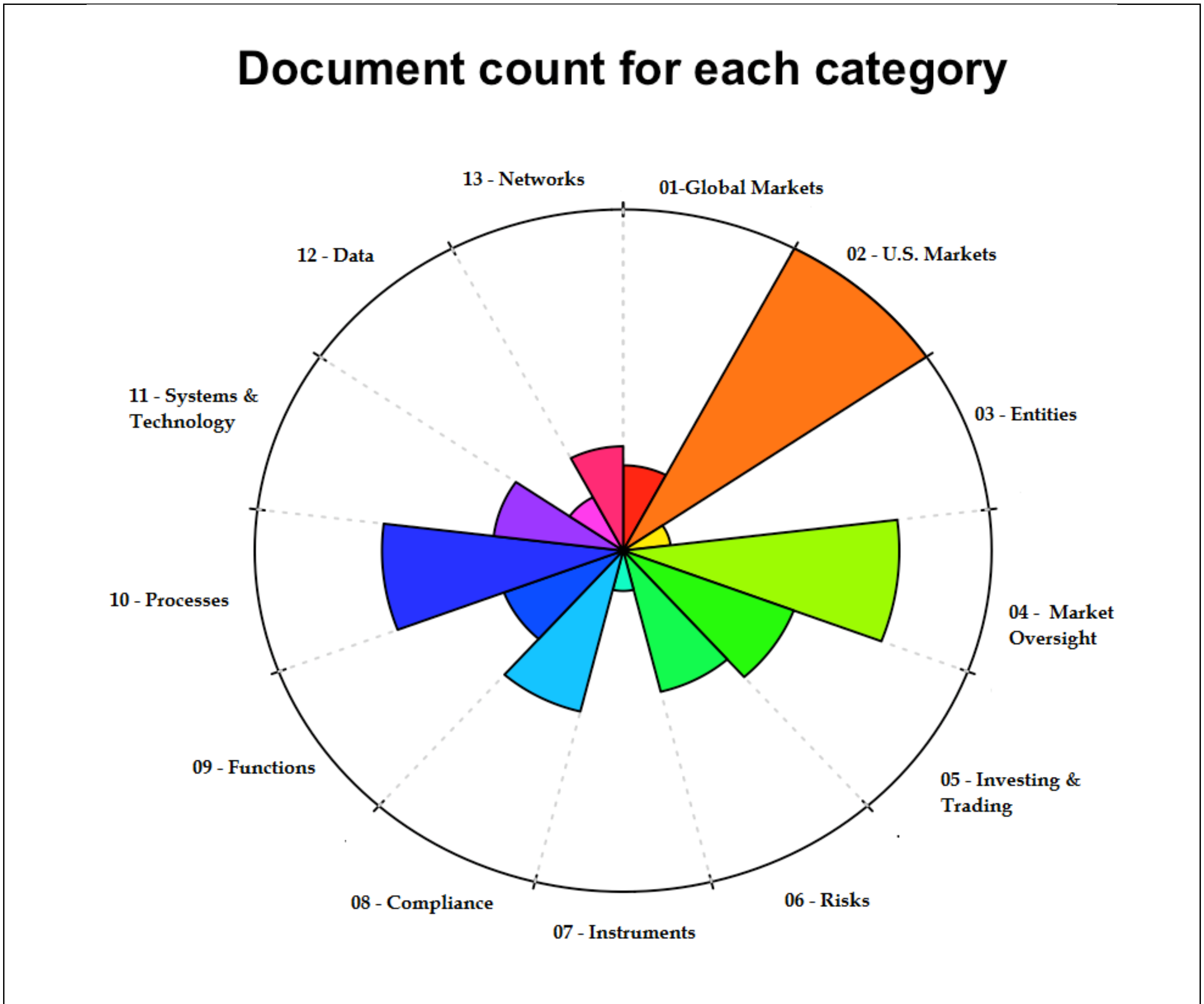


Figure 21. Proportional view of documents with LDA sub-topics aligned to each category.

By examining the proportions of these topic categories throughout the corpus, it is observed that a variety of themes contributed to financial markets conversation without a single dominant driver. A summary of the major themes and associated insights is shown in Table 4. Furthermore, related topic themes can be clustered around groups of major financial markets ideas such as market oversight, investing and trading, risk and compliance, and needed change to their related processes. This is evident through inspection of the major topics gleaned from our first-pass and second-pass topic model of the full corpus, which is illustrated in Table 5.

Summary of Corpus Insights

Theme	Insights
Impact of political change on markets	1. Business and political differences have emerged around market oversight policy
	2. Voters elected Donald trump on de-regulatory platform, driving regulatory uncertainty in the markets and business planning
	3. Volcker Rule under Obama Era to prohibit banks from speculative trading symbolically repealed in Trump Era
	4. Trump policy effect on economy, i.e., 'trump bump'
Operating model impacts	1. Risk and finance functions boundaries less clear, more granular attribution of P&L, risks, measures requires cross-functional processes and data sharing
Systems and technology seen as a major growing force	1. Regulation called for heavy technology infrastructure investment for compliance, but has sparked a new wave of continued investment in emerging technologies (i.e., digital) to make use of big data and advanced analytics
	2. Going beyond minimal regulatory compliance, firms are investing in enterprise technology architectures for adding business value
	3. Firms are taking forward steps to consider what financial services could look like in a fast emerging digital world (i.e., blockchain, artificial intelligence, etc.)
	4. Systems and technology progressively had a large overall proportion of dialogue
	5. Use of machine learning is a rapidly growing investment area
	6. Digital technology such as mobile banking are seen as a competitive advantage
	7. New risks emerging with increased evidence of cyber security threats, malware, and state-sponsored hacking
Equity markets have been very active, and are evolving	1. US Equity markets – single largest topic of discussion; high trading volumes, algorithmic trading investments, effects on liquidity, 'Flash Crash' events, etc.
	2. High Frequency Trading (HFT) is a major topic with many systemic implications, and also subject to increased surveillance
	3. Use of automated trading increasing for sell-side (e.g., HFT) and buy-side (e.g., HFT, Robo Advisors)
	4. Market trading and execution oversight has substantially increased (e.g., MiFID II / European Markets Directive), improving transparency for investors
Derivatives reform and effect on Global Markets	1. Derivatives regulation major focus (i.e. swaps & futures trading), including throughout global market infrastructure environments
	2. Swap Execution Facilities (SEFs) provide transparency into derivative trading
	3. Global markets oversight of derivatives via MiFID II in Europe

	4. Increased country risk seen through Greek Eurozone debt crisis, Brexit (i.e., referendum on globalization), and Donald trump election opposes globalization
	5. Global trade seen as a growing area of debate, with major effects on markets
Data	1. Data has emerged as a pervasively important topic
	2. Increased investment in data is improving value of advanced analytics
	3. Desire for more and faster data i.e., market data, real-time access, etc.
	4. Data management is an increasingly important enterprise function
Transformation of risk and compliance function	1. Heavy and progressive bank capital regulations (i.e., Basel III, Basel IV, Fundamental Review of Trading Book (FRTB), etc.), more risk sensitive and with increased capital requirements for large banks
	2. Banks are learning to operating in a continuous state of change around bank capital rules, and impacts on regulatory reporting
	3. Market liquidity viewed as a critical risk to understand
	4. Systemic risk is managed more holistically through the Dodd-Frank Act and the many regulatory regimes that followed its passage
Minority forces	1. Community banks push back on effects of Dodd-Frank Act, initially caught in regulations before repeal for banks under $50 billion in total assets
	2. Women in labor markets seen as a increasingly important topic at all levels from Boards, executives, and broader labor market participation
	3. Small business owners impacted by credit crisis and increased cost of doing business in a highly regulated environment
	4. Increased presence of social investing, driving need for greater transparency into investment fund activities
Lending failures	1. Borrower lending oversight in wake of financial crisis has undergone major reform, with major discussion around borrower credit, mortgages, and foreclosure

Table 4. Summary of Key Insights from Corpus Topic Modeling Results.

Major Ontology Categories with Illustrative LDA Topic Model Output

#	Topic Category	Example Topic(s) From LDA Output	Number of Sub-topics
1	Global Markets	• Economic Growth / Globalization (Topic 45)	13
2	U.S. Markets	• Economic Growth / Inflation (Topic 6), • Social Media Market News (Topic 47), • Business Environment Change (Topic 80)	19
3	Entities	• Securities Trading Firms (Topic 2), • Community Banks (Topic 5), • Public Investors (Topic 48)	21
4	Market Oversight (Regulation & Policy)	• Wall Street Reform – i.e., Too Big to Fail (Topic 8), • Volcker Rule (Topic 27), • Mortgage Credit Protection (Topic 34),	41

		• CFTC Rules for Swap Dealers / Futures Commission Merchants (Topic 62), • Bank Capital / Basel II (Topic 78)	
5	Investing & Trading	• Flash Crash (Topic 54), • Market Makers / Desk Structure (Topic 63), • Robo Advisors (Topic 95)	23
6	Risks	• Financial Risk (Topic 1) • Liquidity (Topic 2) • Consumer Credit (Topic 29)	16
7	Instruments	• Bank Loans – i.e., Commercial (Topic 86), • Bitcoin (Topic 74) • NASDAQ Stocks (Topic 7)	8
8	Compliance	• Regulatory Compliance – i.e., for financial firms (Topic 2), • Information Security Compliance (Topic 14), • Market Surveillance – Trading Markets (Topic 81)	24
9	Functions	• Investment Management Research (Topic 3), • Payment Systems (Topic 37)	29
10	Processes	• Derivatives Clearing / Collateral Management (Topic 61), • Risk Management Operating Model (Topic 67), • Bank Capital / Stress Testing - i.e., CCAR (Topic 73), • Trade Settlement with Smart Contracts (Topic 91)	27
11	Systems & Technology	• AI & Machine Learning (Topic 4), • Data Security / Cyber security (Topic 13), • Blockchain (Topic 40), • Cloud Infrastructure / Data Services (Topic 64), • Fintech & Mobile Banking (Topic 89)	35
12	Data	• Market Data – Real-time (Topic 51) • LEI (Legal Entity Identifiers) Data (Topic 21)	23
13	Networks	• Market Trading - Exchanges / Venues (Topic 8) • Dark Pools / Trading Venues (Topic 11)	23
	Total	**All Topics**	**302**

Table 5. Topic Categories, Examples and Sub-Topic Counts from LDA results

Interpreting LDA Topic Models

As a way to concretely understand the benefits of topic modeling, it is helpful to focus on a single topic example. As previously discussed, LDA topic models identify the distribution of words for a given topic and their associated probabilities. The LDA function returns an object containing the full details of the model fit, such as how words are associated with topics and how words are associated with documents. Thus, Figure 22a below illustrates the per-topic-per-word probabilities called 'beta' as the x-axis, along with the top ten identified terms for the given topic. In this specific example, we see that the word 'bank' has the highest probability of association as a topic indicator or 0.025. However, when grouped with other terms such as 'data', 'fintech' and 'mobile', we can extract the notion that banks are trying to target customers through mobile devices, a common technology-driven approach used by 'fintech' early stage companies. Finally, humans often manually assign labels for LDA topics.

View of Single LDA Topic Output in R – 'Banks / Data/ Fintech'

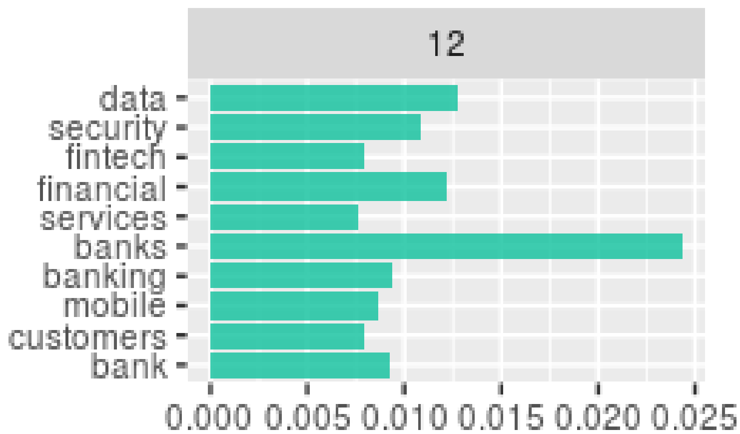


Figure 22a. Visualization of a single topic LDA results in R. Note: y-axis shows terms associated with topic in corpus, and x-axis represents beta or the probability of term association with topic.

Expanding on the previous single topic example, Figures 22b-e illustrate a subset of 36 key topics associated with our text corpus broken out over four diagrams to aid in the visual presentation. A diverse range of topics is shown. Moreover, an occurrence of overlapping terms across topics is evident providing an indication of the power of LDA topic models to find the hidden or latent concepts in text (e.g., risk, data, market, etc.). This subset of topics provides some insights into our corpus and an opportunity for inspection of several key concepts debated over the last decade. For the full list of 302 topics identified by our LDA topic models, see

Appendix E. For example, Topic 1 recognizes terms that appear together and associate with discussion of various financial risks such as credit risk, market risk and liquidly. Whereas, Topic 2 captures a related but distinct concept of risk management compliance in financial businesses. By inspecting the illustrated sub-set of first pass topics, we already can quickly glean key ideas like hedge fund and asset manager research (Topic 3) or a conversation around data and machine learning (Topic 4). Many core processes were a topic of discussion for banking and brokerages operations such as derivatives clearing (Topic 61), trade settlement contracts and systems (Topic 91), or electronic payment systems (Topic 31). Yet market discourse also included conversation of disruptive technologies such as bitcoin (Topic 74), digital and mobile banking technology (Topic 89), or blockchain and distributed ledger technology in banking (Topic 40).

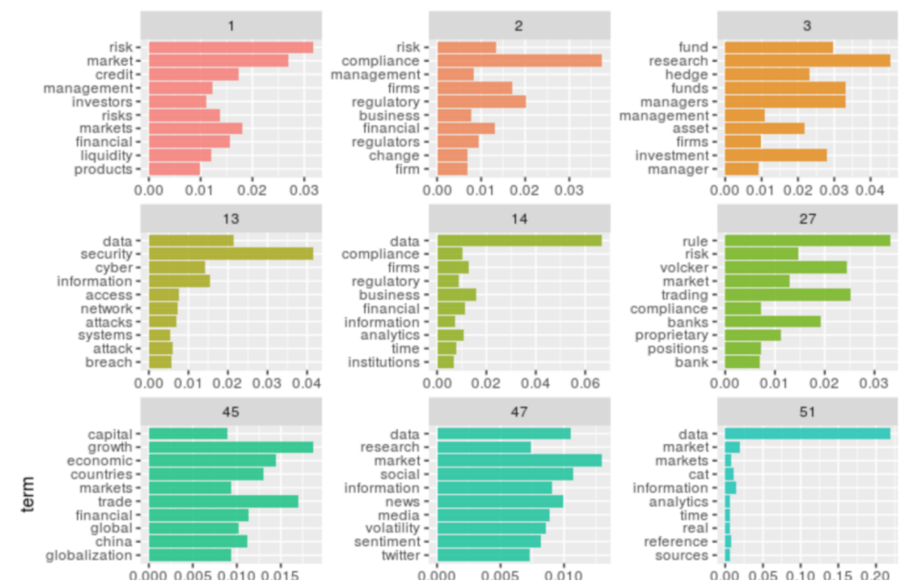


Figure 22b. Sub-set of Topics with manual assigned labels:

- 1: Financial Risk
- 2: Regulatory Compliance (for financial firms)
- 3: Investment Management Research
- 13: Data Security / Cyber security
- 14: Information Security Compliance
- 27: Volcker Rule
- 45: Economic Growth - Globalization
- 47: Social media market news
- 51: Market Data - Real Time

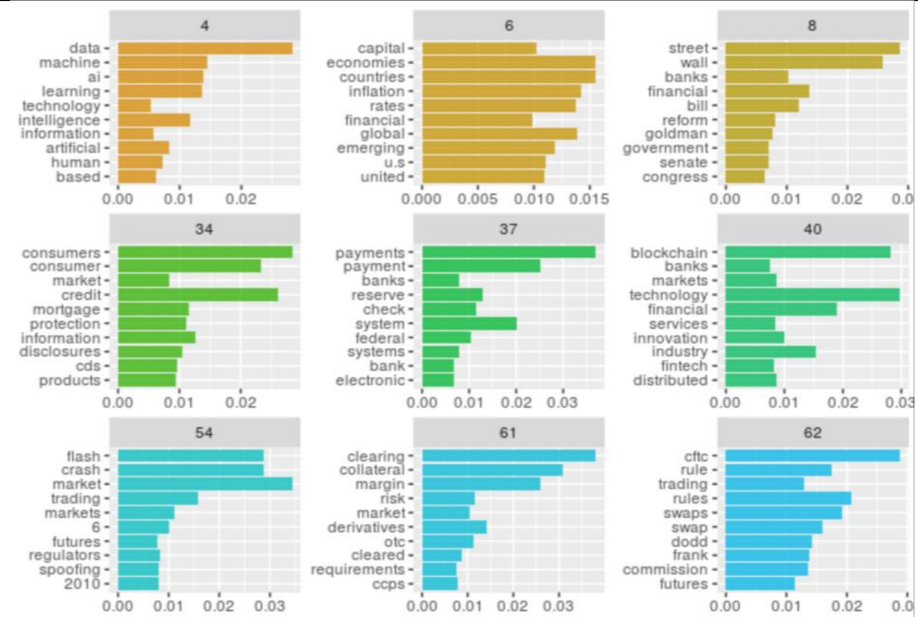


Figure 22c. Sub-set of Topics with manual assigned labels:

- 4: Machine Learning
- 6: Economic Growth - Inflation
- 8: Wall Street Reform (i.e., too big to fail)
- 34: Mortgage Credit Protection
- 37: Payment Systems
- 40: Blockchain
- 54: Flash Crash
- 61: Derivatives Clearing - Collateral Management
- 62: CFTC Rules for Swap Dealers / Futures Commission Merchant

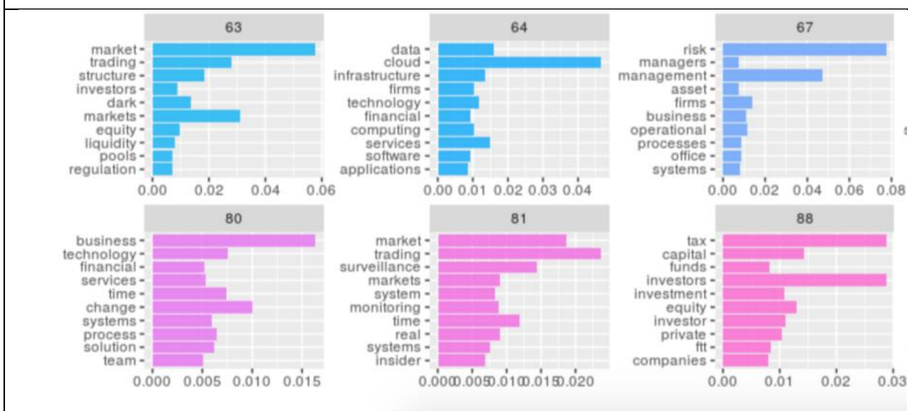


Figure 22d. Sub-set of Topics with manual assigned labels:

- 63: Market Makers – Desk Structure
- 64: Cloud Infrastructure / Data Services
- 67: Risk Management Operating Model
- 80: Business Environment Change
- 81: Market Surveillance - Trading Markets
- 88: Investor Taxation in Private Funds

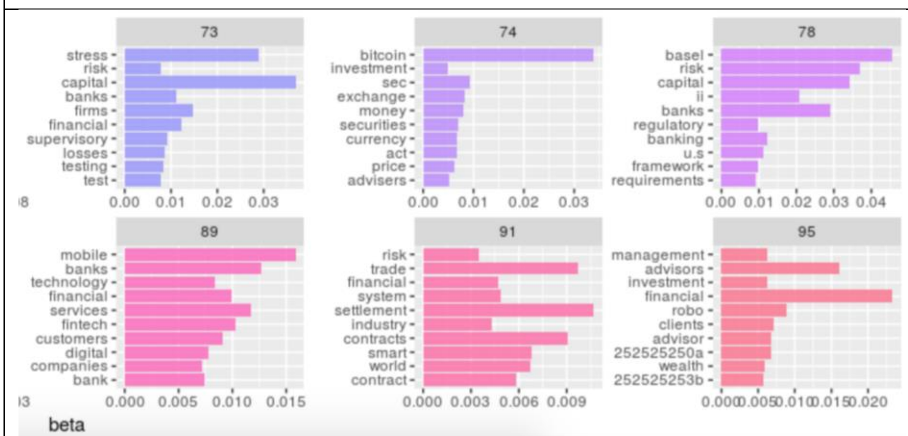


Figure 22e. Sub-set of Topics with manual assigned labels:

- 73: Stress Tests/CCAR
- 74: Bitcoin
- 78: Bank Capital - Basel II
- 89: Fintech & mobile banking
- 91: Trade Settlement with Smart Contracts
- 95: Robo Advisors

Figures 22b-e. Subset of topics from LDA results on full corpus, but with human labeling.

By inspecting the results of our topic models through the lens of our high-level domain ontology, we are able to make sense of the disparate set of concepts uncovered by the topic models. A thorough analysis that looks at not only the existence of the topics themselves, but also how they vary by time, frequency and other covariates provides additional perspective for the most important topics. In the following sections the 13 major categories tell the story of the financial markets landscape over the decade following the 2008 Financial Crisis from 2009 to 2017. It will be shown that the focus of conversation throughout the U.S. financial markets remained on market oversight through transformation of the risk and compliance functions. Towards the latter half, a growing focus on data and technology to manage risk and compliance, but it also emerged as an enabler of disruptive growth. These ideas set the stage for a highly competitive and technology-driven financial services marketplace in the years to come.

Global Markets

Global markets allow corporations to transact business across country borders by selling goods and services across country jurisdictions. As such, there are both economic and political forces at play, which can introduce volatility and directional changes on short notice. While corporations are independent entities, often established in a single country, they are subject to the laws, taxation and consumer forces at play within individual countries. Globalization, or the expansion of global trade, has been a driving force since the late 20th century. Global leaders have championed their countries' products and services, and capital flows have been directed at countries with high growth.

Perhaps the most notable trigger for a change in sentiment in global markets was appreciated after the 2008 Global Financial Crisis, which triggered country exposure risks on liquidity and debt levels. There are several historical examples of banking crises where global markets have faced the need for government intervention. For example, the hedge fund Long-Term Capital Management collapsed in the late 1990s, leading to a recapitalization agreement among 16 financial firms and supervised by the Federal Reserve. Moreover, it lost $4.6 billion in less than four months following the 1997 Asian financial crisis and 1998 Russian financial crisis, requiring financial intervention by the Federal Reserve, with the fund liquidating and dissolving in early 2000 (Wikipedia/Long-Term_Capital_Management). However, the period following the 2008 Global Financial Crisis raised questions about country debt risks where some countries in the European Union held large debt obligations to foreign creditors and risked the possibility of default. Perhaps Greece's potential default was the start of a broader trend to limit exposure to individual countries and also their contagion effect. We can see in Figure 23a that the 'Eurozone - Greek Debt Crisis' was detected by our topic models as a fairly sizeable discussion.

0

Figure 23a. Distribution of Global Markets topics by number of document counts.

Reviewing the topics results in Figure 23a illustrates that access to non-U.S. markets was a big discussion. The two most notable were the Asian and Canadian markets, but we can also see other related themes such as Malaysian market growth, the Chinese stock market, and global trade. While no single country dominated, there is evidence that counter-forces were at play. Not surprisingly 'Brexit' was a major topic, where UK voters supported a referendum to pull out of the European Union on June 23, 2016. Reflected in U.S. market topics discussed later, the election of Donald Trump represented a similar sentiment of U.S. voters who were angry at global forces. Additional trends to better regulate the global markets are showcased in the MiFID II regulation, the European Markets Directive for overseeing the European Markets, and which has introduced substantial structural changes to trading markets. Some examples include topics related to client categorization and order handling, pre- and post-trade transparency, and best execution (Wikipedia/MiFID_II). Also emerging are data privacy topics under Global Data Protection Rule (GDPR) and the Global Trade war.

Taking a time-series view of the discussion, Figure 23b shows that the period of 2011 to 2014 held the highest level of discussion of the Global Markets, which was also aligned to the time of the European Debt Crisis. While not in the scope of our data, recent steps taken by U.S. President Donald Trump to introduce tariffs is reigniting the global markets category, which may lead to an all-out trade war. As an example of a concrete benefit for staying aware of trends around the global trade debate, topic modeling would provide a fast way to consume relevant global publications and monitor for the growing debate of global trade and openness of global markets, and the potential implications of tariffs.

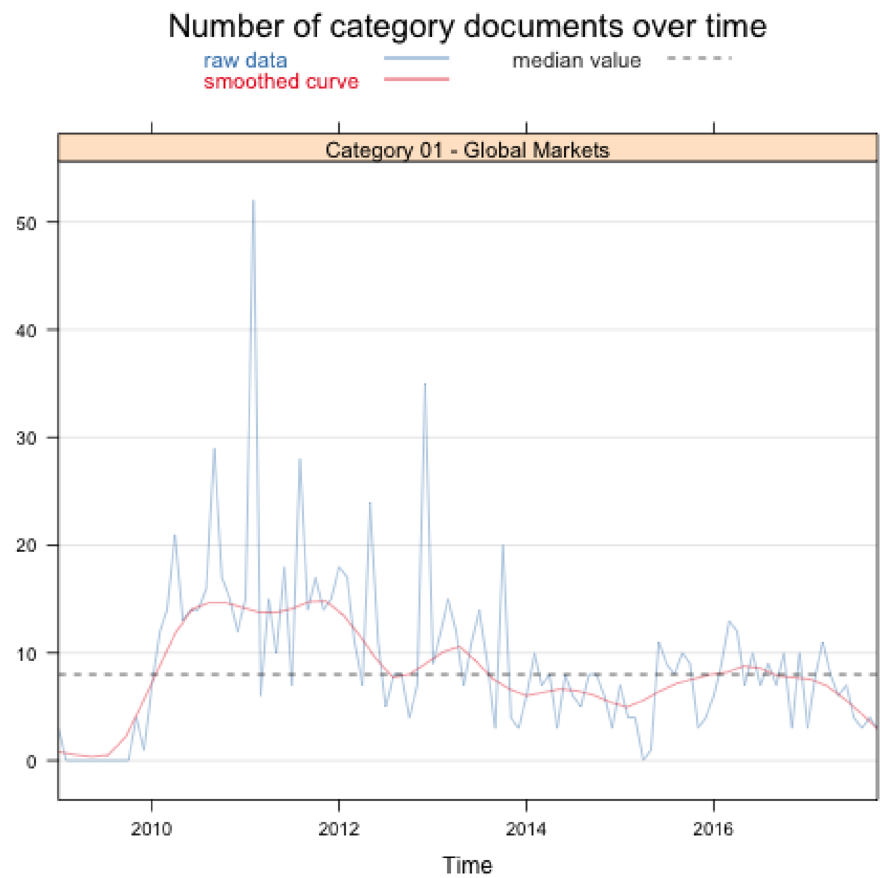


Figure 23b. Time series of Global Markets category topic document count over time.

U.S. Markets

The dominant category in our full corpus of data is the U.S. markets with the greatest proportion of documents containing U.S. markets topics. This is not surprising since our two data sources have a bias towards U.S. markets. Economic growth drivers are a clear dominant segment of the U.S. market topics which is shown in Figure 24. Moreover, the most frequent topic was related to U.S. equity market volumes, which is a clear indication of a liquid market and sign of a healthy economy where both buyers and sellers are actively engaged. Financial news discussion of U.S. markets very commonly has a focus on evaluation of economic growth, and positive or negative signals, which could impact its momentum.

The 2008 Financial Crisis triggered a need for market transparency and control through more centralized trading execution, clearing and settlement. The volume of many bi-lateral agreements between the major financial institutions through the over-the-counter (OTC) market reflected a large proportion of derivatives contracts such as swaps. The post-crisis regulations called for increased centralized clearing for standardized derivative contracts, or institutions would face higher costs in the form of margin, collateral and capital requirements. For example, the third most common U.S. markets topic is related to the swaps and futures trading markets illustrated in Figure 24. Cross-referencing this topic with related topics in categories such as market oversight, compliance and processes reflects the recent heavy focus on derivatives markets.

U.S. markets topics centered around three key themes. First, related to different market's asset classes such as equity markets, fixed-income markets and mortgage securitization markets. Furthermore, the equity markets had a heavy focus on electronic trading also known as high-frequency-trading as illustrated by several topics in Figure 24. One very specific example is the 'Flash Crash' topic, which was triggered when high-frequency-trading algorithms reacted to some erroneous trades and falsely moved the market down at substantial levels. The markets quickly rebounded when the algorithm anomaly was recognized, but the backlash was high since mainstream investors lost money for unfair reasons.

The second major theme of this category was related to the liquidity of the markets. Liquidity is a key risk that is largely associated with the cause of the 2008 financial crisis when inter-bank lending markets seized up. As such, there was heavy conversation related to its identification, mitigation, and to ensure adequate buffers through regulation. In particular, liquidity of the fixed income market, in particular, was the second highest topic seen in Figure

24, and closely linked to the collapse of Lehman Brothers when their credit spreads spiked as investors anticipated a default on fixed income contractual obligations. Trading markets in general were a major theme also recognized here, and in other categories such as market oversight with the Volker Rule (i.e., prohibition of proprietary trading) and many related compliance requirements required by the Dodd-Frank Act and MiFID II legislation.

Finally, political sentiment around U.S. markets also gained substantial public debate as political winds shifted in the U.S. as evidenced with the election of President Donald Trump. Politics often measures the economic track record of elected officials via economic growth measures such as GDP (Gross Domestic Product) and the unemployment rate, which is evidenced in our topic results. In addition, other qualitative topics appeared such as the participation of women in labor markets and the level of skilled labor pools available. The smallest topic by document frequency, but also reflective of the changing political sentiment, is the ‘Trump Bump,’ a colloquial reference to economic gains from Trump’s business friendly campaign rhetoric.

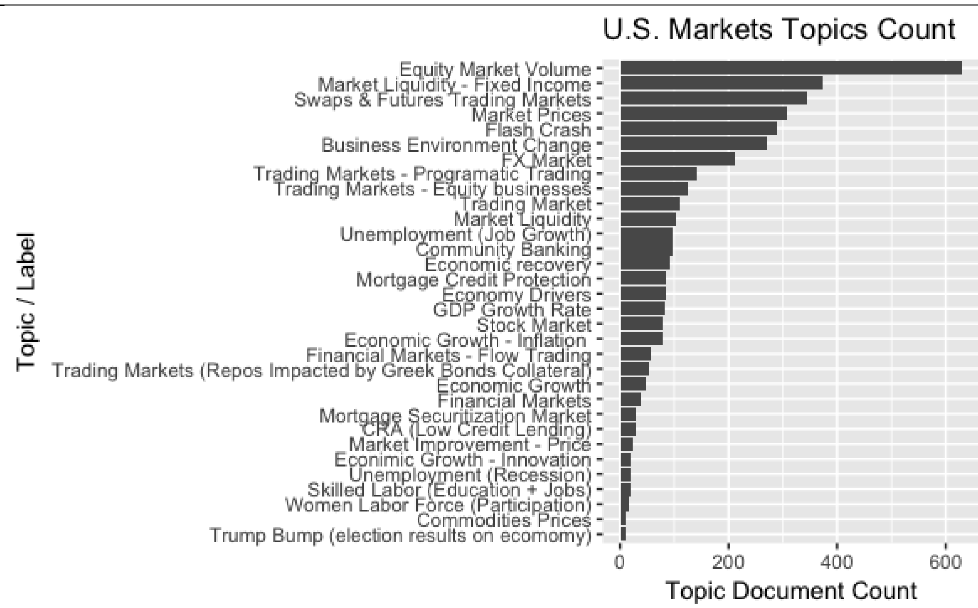


Figure 24. Distribution of U.S. Markets topics by number of document counts.

Entities

By definition an entity is a thing of distinct and independent existence. Yet in the context of financial markets, entities represent the market participants in the form of legal entities, financial firms, investors, consumers, regulators, vendors, and clearing houses, to name a few. Topic model results from our corpus identified just a few of the many entities that participate in the financial markets today. For example, Swaps & Trading Execution Facilities (SEFs) was the most common topic in this category, as shown in Figure 25. This is not surprising since as discussed in the prior section, derivatives markets were directly linked with the financial crisis, and SEFs were required to be formed as part of the Dodd-Frank Act legislation to facilitate the execution of bi-lateral swaps contracts not executed on a central clearing house.

The results show three major categories of entities identified: 1) government or quasi-regulatory entities; 2) financial institutions, and 3) individual people who participate in the markets. Government entities played a dominant role in our results with inclusion in the most number of documents. For example, in addition to SEFs, the European Union (EU) was a major topic of discussion. Several threads contribute to the large EU discussion where related topics identified in other categories such as European Debt Crisis, MiFID II regulation, or even most recently Brexit all contribute to broad topic of the EU.

As a second category, many types of financial institutions were identified, including retail brokerages, investment banking institutions, securities trading firms, credit rating agencies and community banks. There was a broad range of topics covered related to this set of entities, with some insight into the far-reaching nature of financial services to touch all aspects of society. For example, community banks were part of a fair amount of discussion in approximately fifty documents in the corpus, which is credible since there was substantial debate among community banks that the regulations imposed by the Dodd-Frank Act (DFA) were extremely onerous for community banks. Moreover, the issue reached a new direction when the Trump Administration repealed part of the DFA for small banks with less than $50billion in total assets (i.e., Community Banks). Here is a clear example of where topic modeling provides a fast insight that can be further drilled into to get more granular on the subtler underlying issues.

Lastly, the individual investor was arguably a large victim of the financial crisis as investment portfolios halved and mortgages exceeded the value of the homes they covered in large quantities. As such, focus on the credit worthiness of individual borrowers and its transfer to the broader economy through the mortgage securitization market was a major topic of discussion. We can see that ‘low credit lending’ and ‘mortgage foreclosures’ were observable topics in Figure 25, and also linked to the mortgage securitization market in the prior section. The results here show that many roles of the individual were affected including small business owners, families (income levels), and public investors.

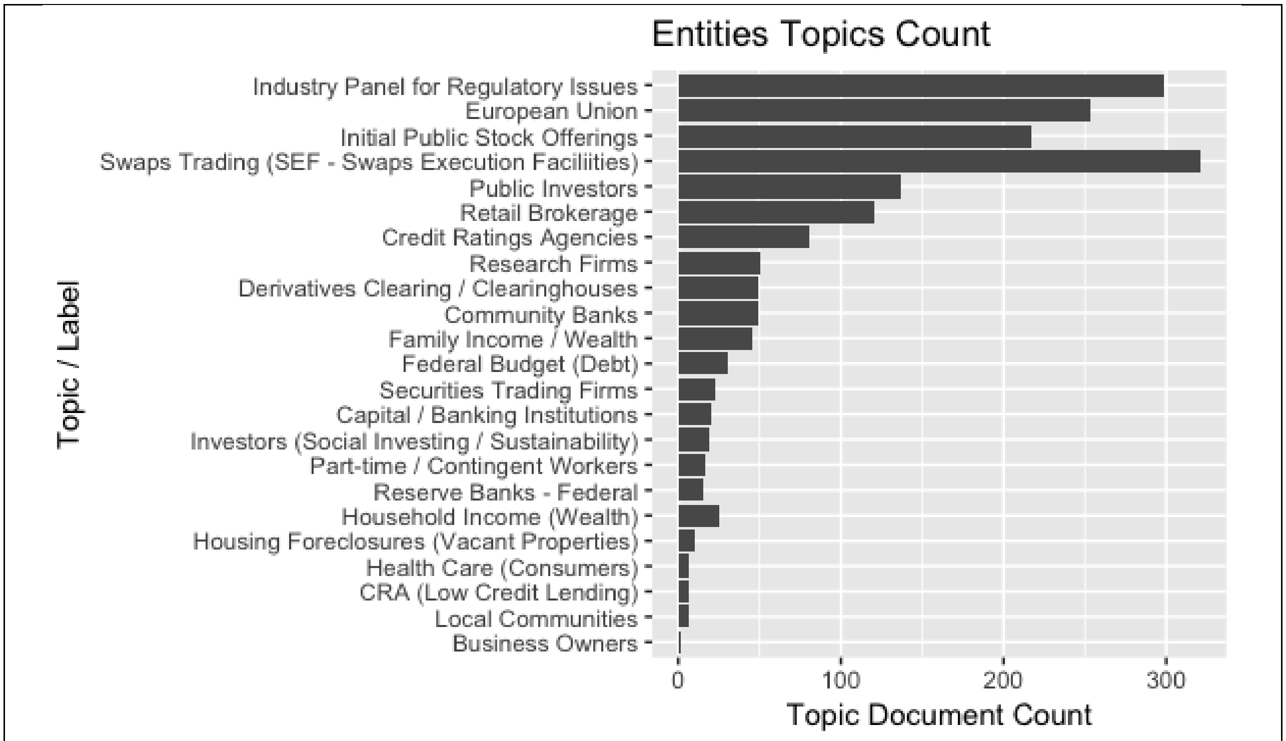


Figure 25. Distribution of Entity topics by number of document counts.

Market Oversight (Regulation & Policy)

Market Oversight is largely the most dominant category of our results both in terms of the number of sub-topics as well as the proportion of documents covering those topics. This outcome is not surprising since our corpus was approximately ten percent regulatory speeches, and the remaining text focused on the capital markets during a period following a substantial banking crisis. In fact, we are able to see in figure 26a that the number of documents containing Market Oversight topics on a monthly basis sustained a median level of forty for nearly an entire decade following the 2008 Financial Crisis. It is undeniable that the banking crisis left a permanent impact on the financial markets. Interestingly though, by way of the many regulations passed, the financial institutions themselves remained at the heart of the rules and policies passed to prevent a future banking crisis. This is observed in Figure 26b where ‘Regulatory Compliance (for financial firms)’ was the third highest topic.

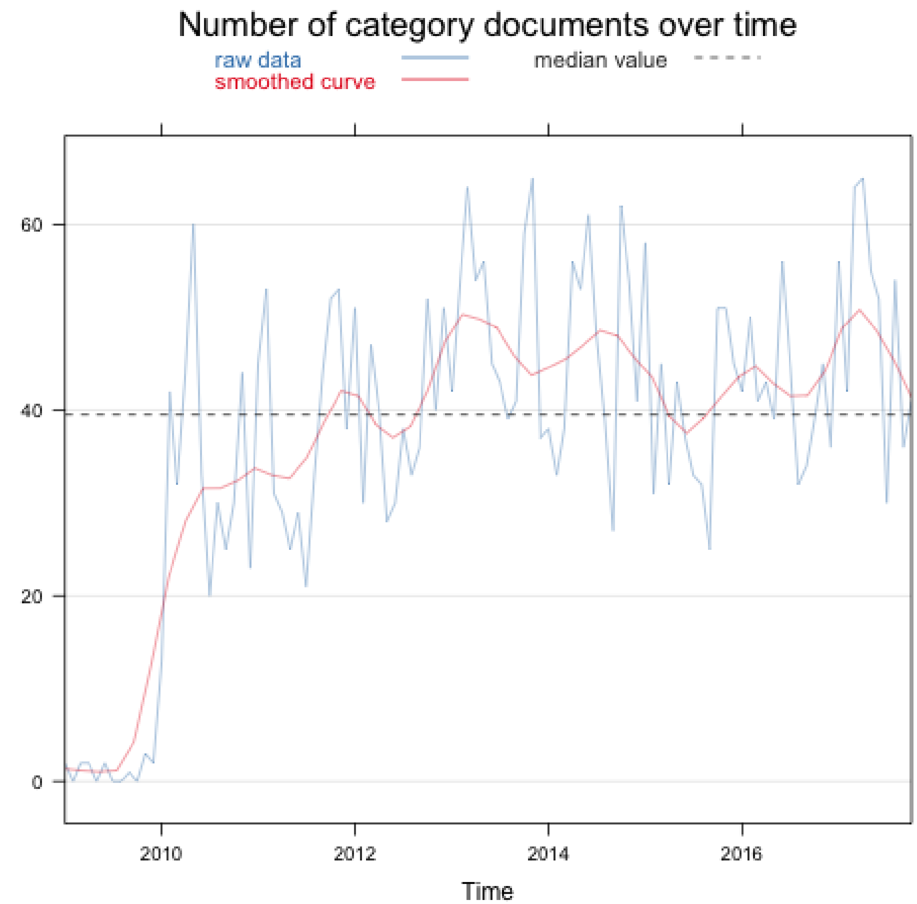


Figure 26a. Time series of Market Oversight category topics by document frequency.

Banking reform centered around three broad themes to ensure that the largest banks designated systematically important or 'too big to fail' would never again pose a risk to the banking system. The first tranche of regulations aimed to prevent banks from taking risks similar to what lead to the financial crisis itself. For example, new onerous bank capital regulations became much more sensitive to real risks based on more granular data requirements (i.e., loan or instrument level) and empirically derived from historical data. The implications were that banks became less incentivized to hold non-investment grade rated (i.e., credit ratings below BBB- by Standard & Poor's) and uncollateralized securitizations on their balance sheets that were at the center of the fire sale that engulfed the markets in the crisis. Nearly a decade after the banking crisis, several iterations of bank capital rules known as the Basel Accord are still incomplete. Basel III replaced Basel II in 2014, and figuratively Basel IV replaced Basel III in 2017. Now banks are addressing another capital rule referred to as the Fundamental Review of the Trading book expected to be even more onerous by replacing VaR with Expected Shortfall measures aimed at better mitigating tail risk in the trading book.

Aside from bank capital rules, many other regulations placed restrictions on trading activities or banned it altogether. For example, the Volcker Rule bans financial firms with financial assets of a certain size from taking speculative positions referred to as proprietary trading. It does provide exemptions to perform basic market functions such as ‘market making’, and trading for ‘risk management’ purposes. Another example previously discussed and one of the most prevalent topics was ‘CFTC Swap Dealer / Futures Commission Merchant’ regulation. These rules imposed many new requirements financial firms registered with the CFTC to deal swaps and futures that ensured transparency in execution, clearing and reporting of such instruments. These are broad and deep topics, but all aimed at the central idea of preventing excessive risk taking and ensuring full transparency for regulators to oversee.

A second tenant of the financial reform dealt with forecasting how resilient financial institutions could be in the event that financial markets became distressed. The Dodd-Frank Act called for bank ‘stress tests’ referred to as Comprehensive Capital Allocation & Review (CCAR). In essence, the banks were provided several stressed scenarios by Federal Regulators and were required to project their income statements, balance sheets, and capital projections over nine quarters based on the provided scenarios. Any failure to maintain minimum threshold levels could result in several consequences such as disallowing dividend payments. These stress tests were both very onerous for banks in terms of the required historical data for modeling, and also for the multiple layers of challenge processes required to justify the accuracy and reasonableness of results.

Lastly, a third category of regulation aimed to provide a framework for resolving banks that enter receivership or bankruptcy in the unlikely event of a bank failure. Following the Lehman Brothers bankruptcy, the Federal Deposit Insurance Corporation (FDIC) spent over a decade unwinding its estate and ensuring creditors received their entitled portion of assets. However, the fire sale of assets during the financial crisis led to the triggering of default on many derivative and other financial contracts and posed a much broader risk to the entire banking system. As such, large financial institutions are now required to have 'Resolution & Recovery Plans' that provide a framework and detailed playbook for how to unwind the bank in the event of failure, and funds to maintain minimal operational levels.

These three categories of issues were mandated by the twelve broad sections of the Dodd-Frank Act, and provided a legislative mandate for bank regulators to establish rules and enforcement oversight. Interestingly, despite the nearly decade of increased regulation and macro-prudential policy aimed at ensuring banks do not pose systemic risk to the broader economy, a turning sentiment has developed. The topic of 'Business Friendly Regulation' has emerged as seen in the top half of common topics in Figure 26b. While banks privately complained about the cost of regulation on their business models, the election of President Donald Trump introduced a new administration which campaigned on reducing bank regulation. At the time of writing this paper, some aspects of the Dodd-Frank Act have already been repealed such as parts of the Volker Rule. Banks must now operate in a highly regulated environment subject to ongoing changes and new regulations, while also maintaining growth strategies for investors.

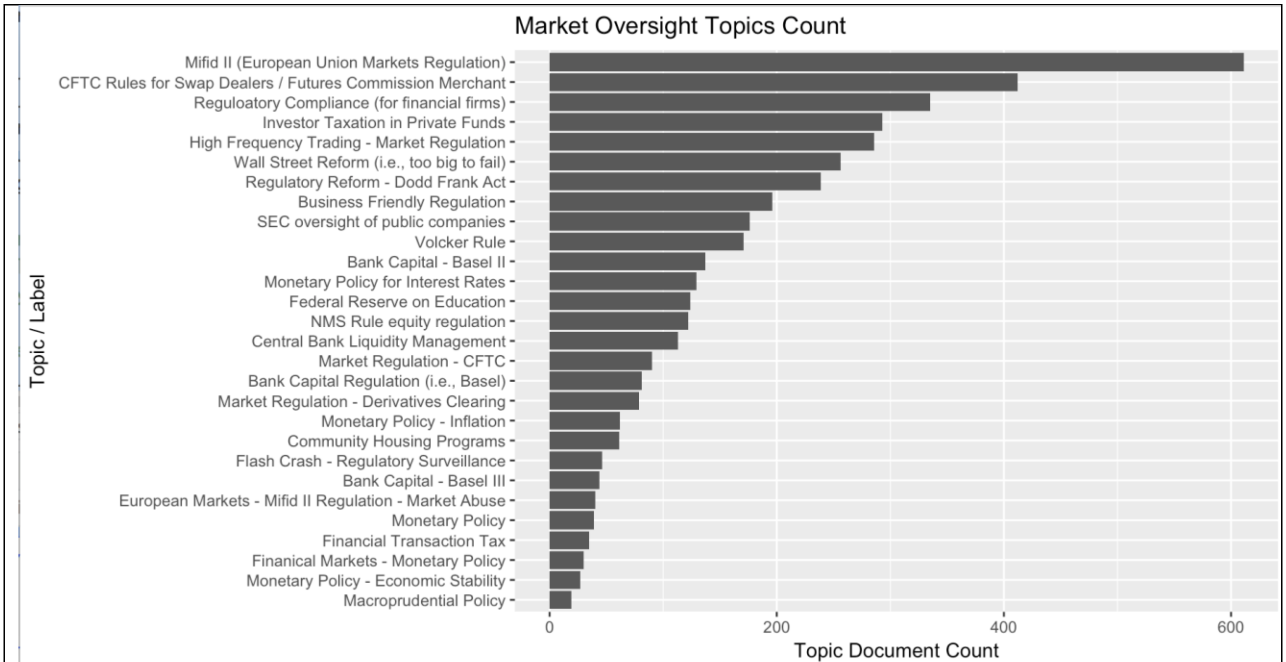


Figure 26b. Distribution of Market Oversight topics by number of document counts.

Investing & Trading

A core function of the financial markets is providing investors a forum for investing and trading securities, financial instruments, currencies, commodities and other financial products. Therefore, it is no surprise that the most common topic in this category is ‘Buy Side Trading’, which refers to financial firms that provide retail or individual investors brokerage services for trading. The retail investor was able to make substantial investment returns in the years leading up to the financial crisis, and was viewed as a major victim following the crisis when stock markets nearly halved. As such, many of the political and regulatory outcomes discussed in prior and subsequent sections address the plight of the retail investor.

While the retail investor played big role in the pre- and post-crisis conversation, another major theme identified in topics illustrated in Figure 27 relates to electronic or high-frequency trading. Equity markets have been able to trade at exceedingly fast levels down to nanoseconds using sophisticated algorithms making decisions on various types of data. Four topics in this category's topics relate to high-frequency trading. Automated trading is also expanding into asset wealth management businesses as observed in the 'Robo Advisors' category. The notion is that asset management strategies can been automated and maintained at much lower costs.

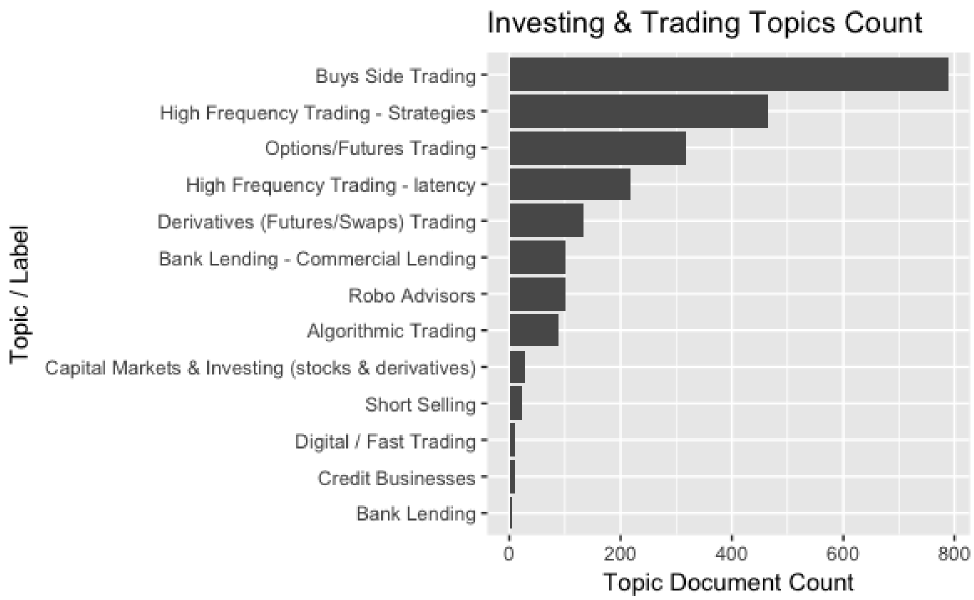


Figure 27. Distribution of Investing & Trading topics by number of document counts.

Risks

Many types of risks face the financial markets today and include both external risks and internal risks. Examples of external risks include financial market risk, political and regulatory risk, macro-economic risk, and others. Internal risks include operational risks, strategic risks, and reputational risks. Within each of these categories there are several sub-categories. Risk can be a very broad topic, while at the same time very specific when acutely dealing with a certain kind of risk (Hardy, 2013). We see in Figure 28a that the category of risk and its sub-topics steadily increased over the decade following the financial crisis, before a sudden decline.

Two specific kinds of risks appear to dominate the risk category discussion in our data which revolves around market risk and liquidity risk. Market risk generally is defined as the amount of loss that can occur when changes in macro-economic variables occur or some specific market event triggers a loss. As such, risk managers of financial firms aim to quantify market risk to limit downside of financial investments. Figure 28b clearly shows that market risk topics were heavily discussed with ten or more topics in the risk category associated with market risk.

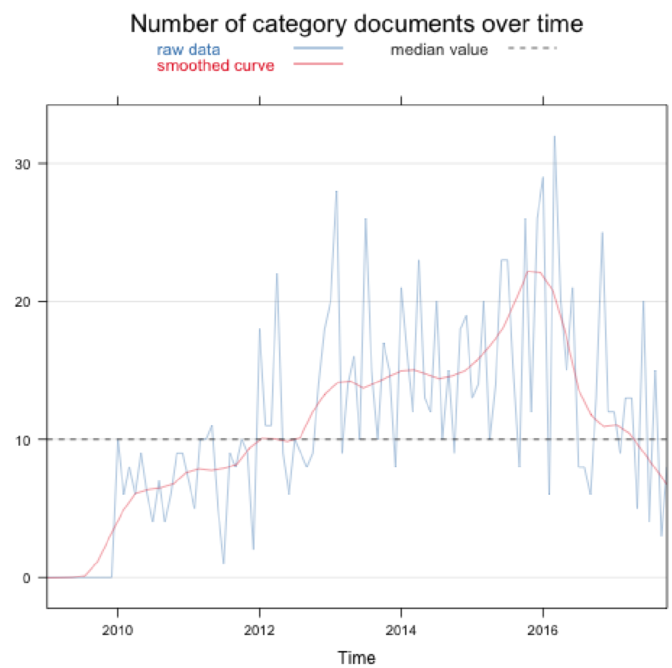


Figure 28a. Distribution of Risks category topics by document frequency.

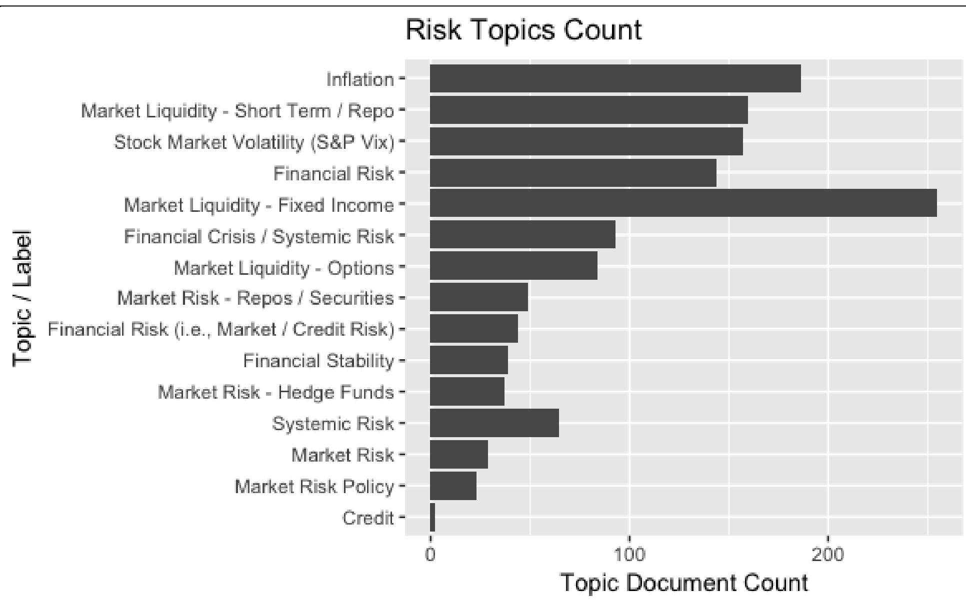


Figure 28b. Distribution of Risk category topics by number of document counts.

Instruments

Financial securities, instruments, and currencies are a fundamental concept to the financial markets and used as a basis for storing and exchanging economic value. As we see in Figure 29a, currencies are the most common topic in this category by far. Clearly currencies play a major role in financial markets both within individual countries and across borders allowing for global trading and capital flows. Outside of currencies we can see that financial markets cover many other instruments including contracts on oil, bonds, stocks and various kinds of loans. While instruments play an important role in financial markets, interestingly their inclusion in discussion is less common. Often conversation circles around the instruments discussing the nature, parties, business and related processes. We can see in Figure 29b that the instrument category is generally low outside of the start of 2016 where a substantial spike in this category occurred.

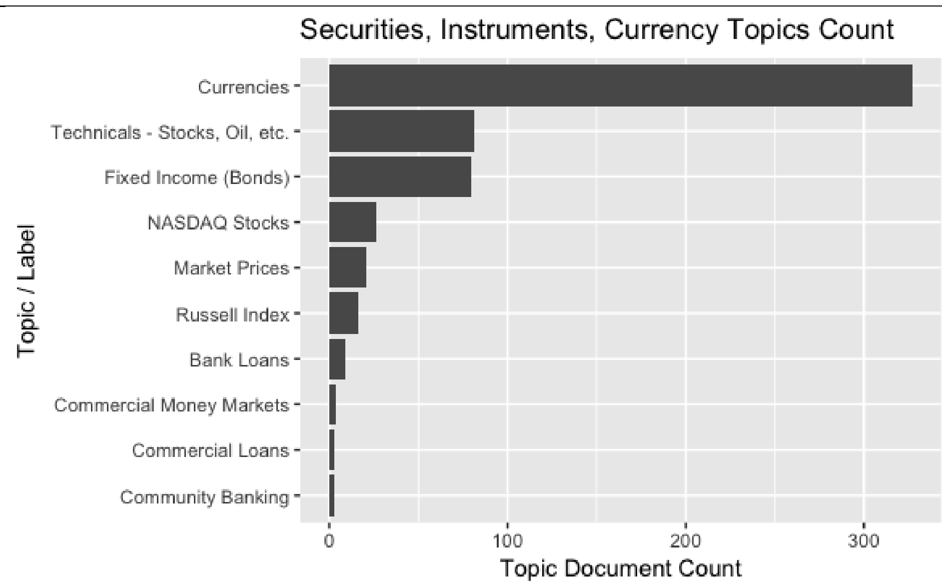


Figure 29a. Distribution of Securities, Instruments, & Currencies topics by number of document counts.

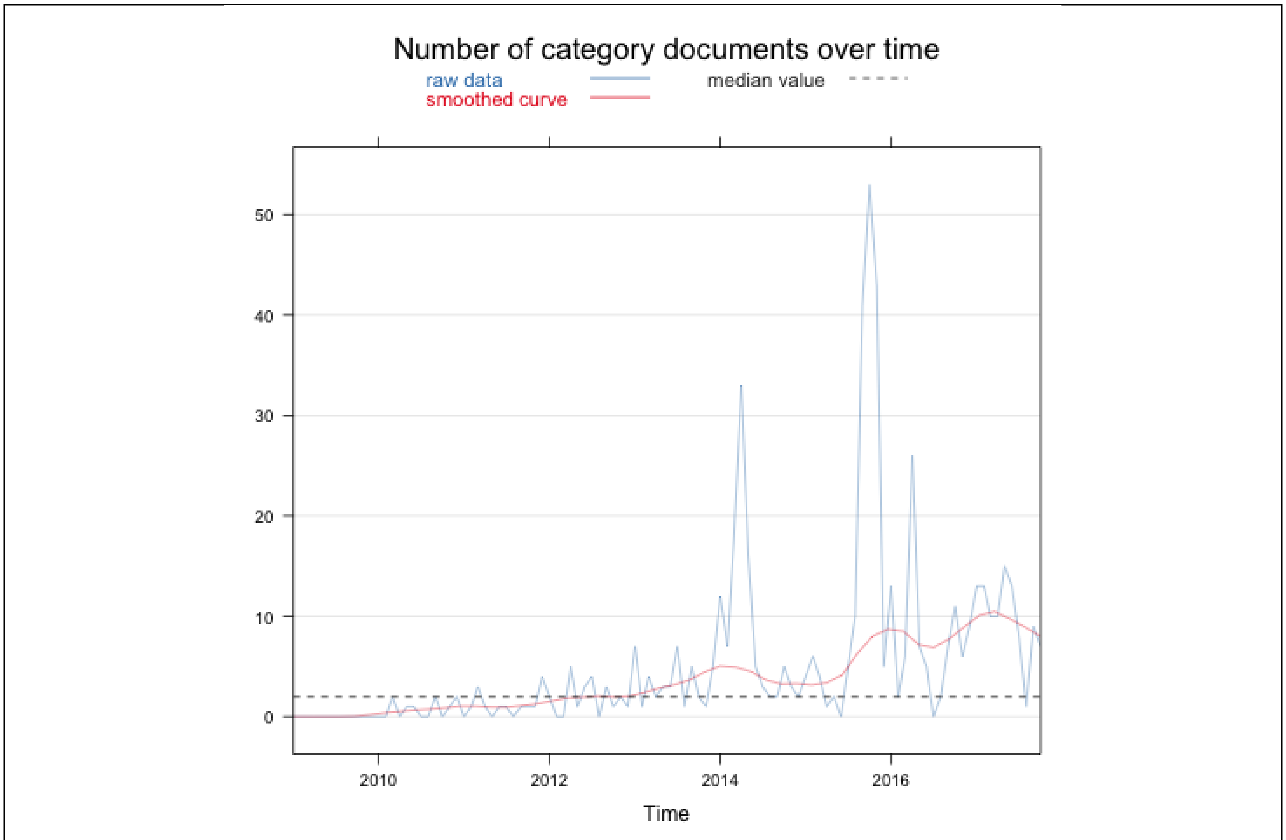


Figure 29b. Time series of Instruments category topics by document frequency.

Compliance

Increased levels of compliance activity and discussion are a natural consequence of the wave of new regulations that followed the financial crisis. The multitude of topics that are observed in the compliance category as depicted in Figure 30a shows just how far reaching regulations impacted financial firms. We can generally think of the compliance topics on three dimensions. First, monitoring of the external market from the perspective of the financial firm is an important activity. For example, the second most referenced topic is related to market surveillance of the trading markets. This notion applies to financial risk management and the macro risk factors such as economic growth, unemployment and interest rates as well as the company specific valuation movements as observed in the equity markets. A supporting idea that was also introduced in our results is the importance of data. Access to timely and accurate market data is a key input into market surveillance models and an important focus of financial firms.

A second group of compliance oriented topics centered on understanding the customer or counterparty of financial firms. A clear example is the concept of ‘Know Your Customer’, which is a key element of ‘Anti-money Laundering’ rules that aim to ensure customer profiles and approved financial services align with actual customer behavior. Steps carried out to meet these requirements are part of the more general client onboarding process, which was also heavily discussed. Customer data input and output of each of these processes is also viewed as an important component of the mix. Collectively financial firms are expected to perform a greater level of due diligence, behavior monitoring and reporting of customer activity, leading to a higher degree of confidence in market oversight.

Lastly, compliance largely focused on ensuring that firm behavior is aligned to regulatory requirements. Compliance topics related to firm behavior spanned from the front office to the back covering topics like broker trade execution, asset manager compliance, and even tax compliance (i.e., Foreign Accounting Tax Compliance Act – FATCA). In addition to overseeing the internal actors within a firm, external risks also required firm oversight. The most prevalent example of this was information security compliance. With heightened awareness on cyber security attacks and recent examples global ‘Ransomware’ firms have major programs related to cybersecurity (Wikipedia/ransomware).

The time-series view of compliance related topics in Figure 30b shows a steady increase in coverage over a ten-year period following the financial crisis. This trend continued until 2017, the last year in our corpus, when a compliance discussion decreased for the first time, but still at high levels historically. The state is that compliance activity has not decreased, but is maturing into an established 'business-as-usual' mode where successive transformation is stabilizing. As new risks emerge with the adoption of new market structures and technology, it's possible a new wave of compliance efforts will emerge.

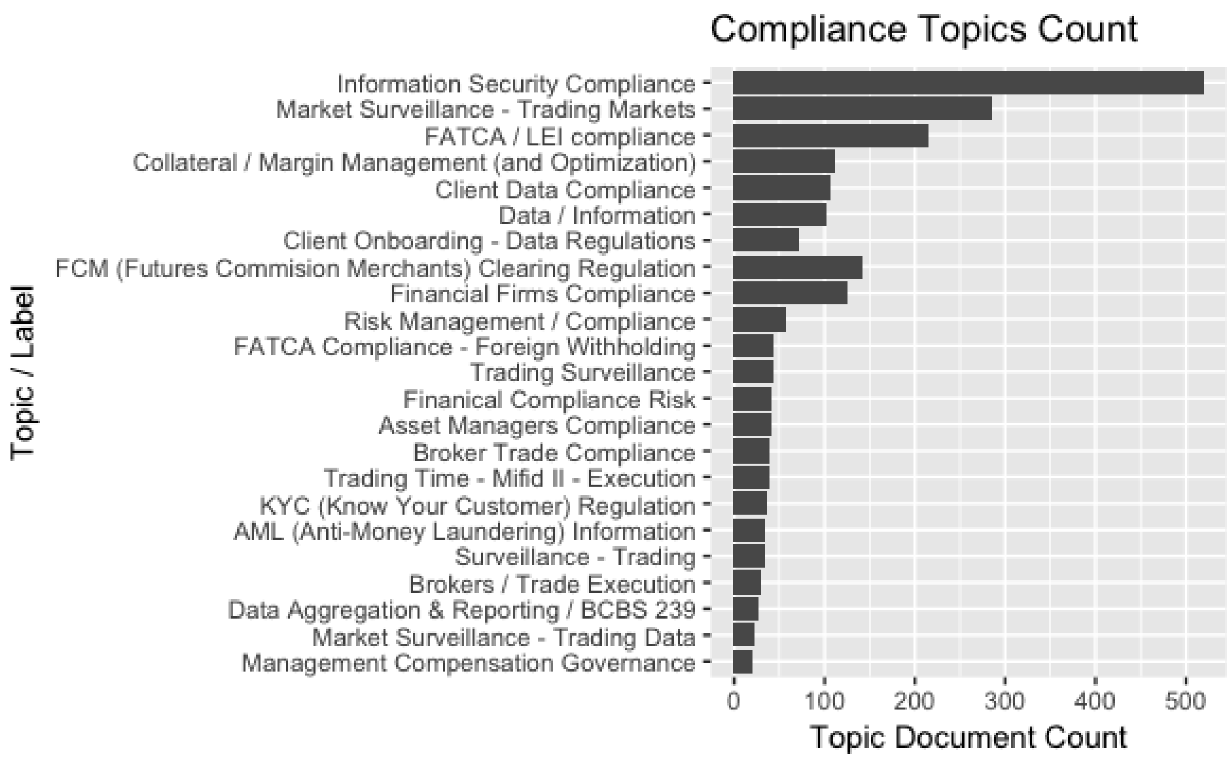


Figure 30a. Distribution of Compliance topics by number of document counts.

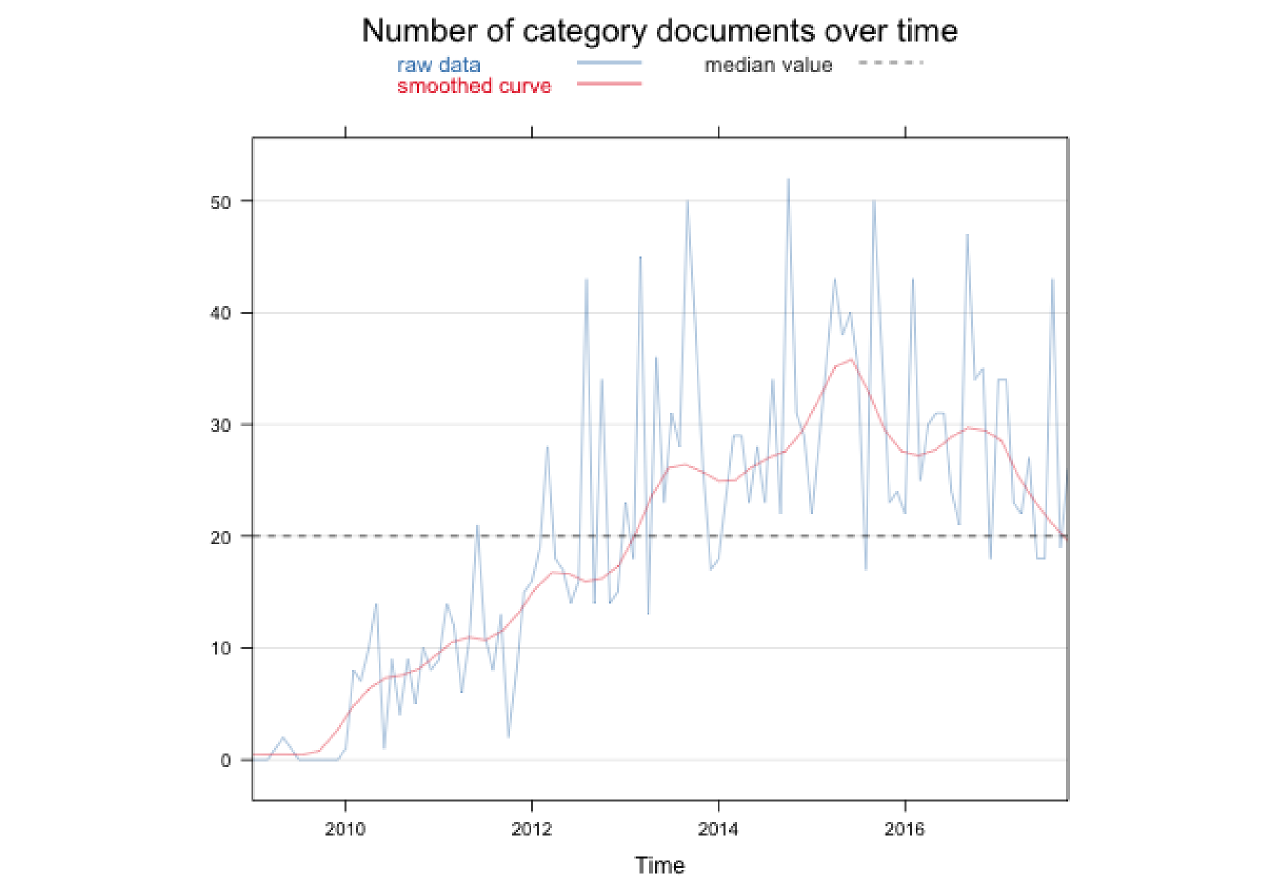


Figure 30b. Time series of Compliance topics by number of document counts.

Functions

The way financial firms organize themselves to carry out various capabilities to meet firm and market needs is the core idea of a function. Arguably the segregated nature and often silo-oriented behavior of functions was a contributor to the financial crisis in the form of limited information sharing across functional boundaries. As such, post crisis regulation took many strides to ensure integration and aggregation of functional measures was collectively reviewed as part of new processes. Functions in financial organization are typically grouped into front-office, middle-office, and back-office functions; we can see evidence in the functional oriented topics in our corpus shown in Figure 31a.

Certain front-office functional topics related to the buy side (i.e., Asset Management) and sell side (i.e., investment banking and trading) businesses appeared in our results. Key examples centered around trade execution on both the buy-side and sell-side, asset management research, portfolio management, and retail brokerage (i.e., FX). All these functions support ensuring clients receive fair investment advice, quotes and trade execution care. Also within the trading function was a sub-topic related to desk structure, with links to the Volcker Rule and the notion that trading desks at investment banks should only engage in permissible trading (i.e., market making) on behalf of customers, not speculative trading outside desk mandates.

Middle-office functions largely provide the oversight and validation of front-office operational processes to ensure that all parties agree on the terms of a trade or agreement. This is evident in the operations topics observed related to clearing, settlement and payment activities. Each of these activities plays a crucial role for the integrity of the markets by providing trade confirmation prior to the movement of securities, instruments and funds between parties. As such, issues arose in the financial crisis when Lehman’s default triggered a breach of terms that gave firms legal grounds to freeze clearing, settlement and payments. Another important middle-office function centers around risk management, which was also observed in our results. Risk management functions that relate to financial risks such as market, credit and liquidity risk are a complementation activity to the clearing and settlement functions. These functions ensure that changes in risk factors around contractual terms are monitored appropriately, and the introduction or risk measures in terms of both quantity and scope was widely discussed.

Finally, back-office functions address many of the reporting activities used to manage the businesses and full enterprise. Finance, accounting, and treasury functions are primary back-office functions. However, compliance, legal and audit next line of defense support to other functions. Many processes important to each of these back-office functions are discussed in the next section. However, back-office functions by name were not readily apparent in our results. However, one notable observation from our topic models was that middle-office and back-office functions played a key role in business decisions in the post crisis world with increasing levels as shown in Figure 31b. From 2009 to the end of our scope in 2017 functional topics increased at a sustained rate throughout this period.

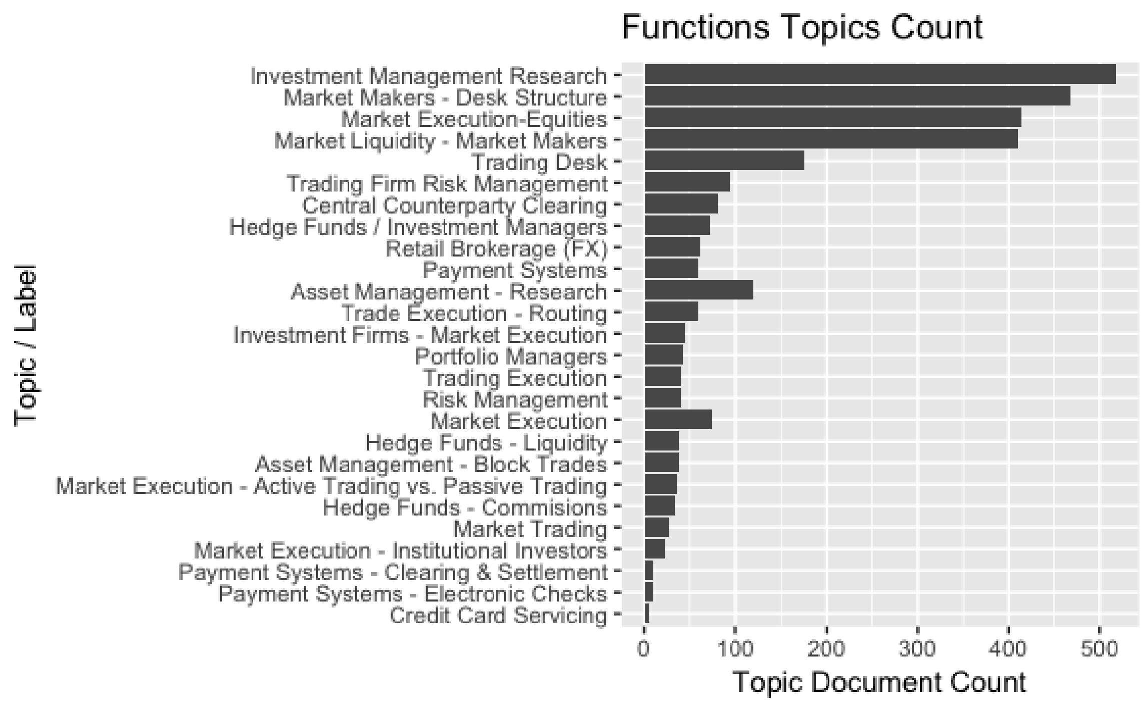


Figure 31a. Distribution of Function topics by number of document counts.

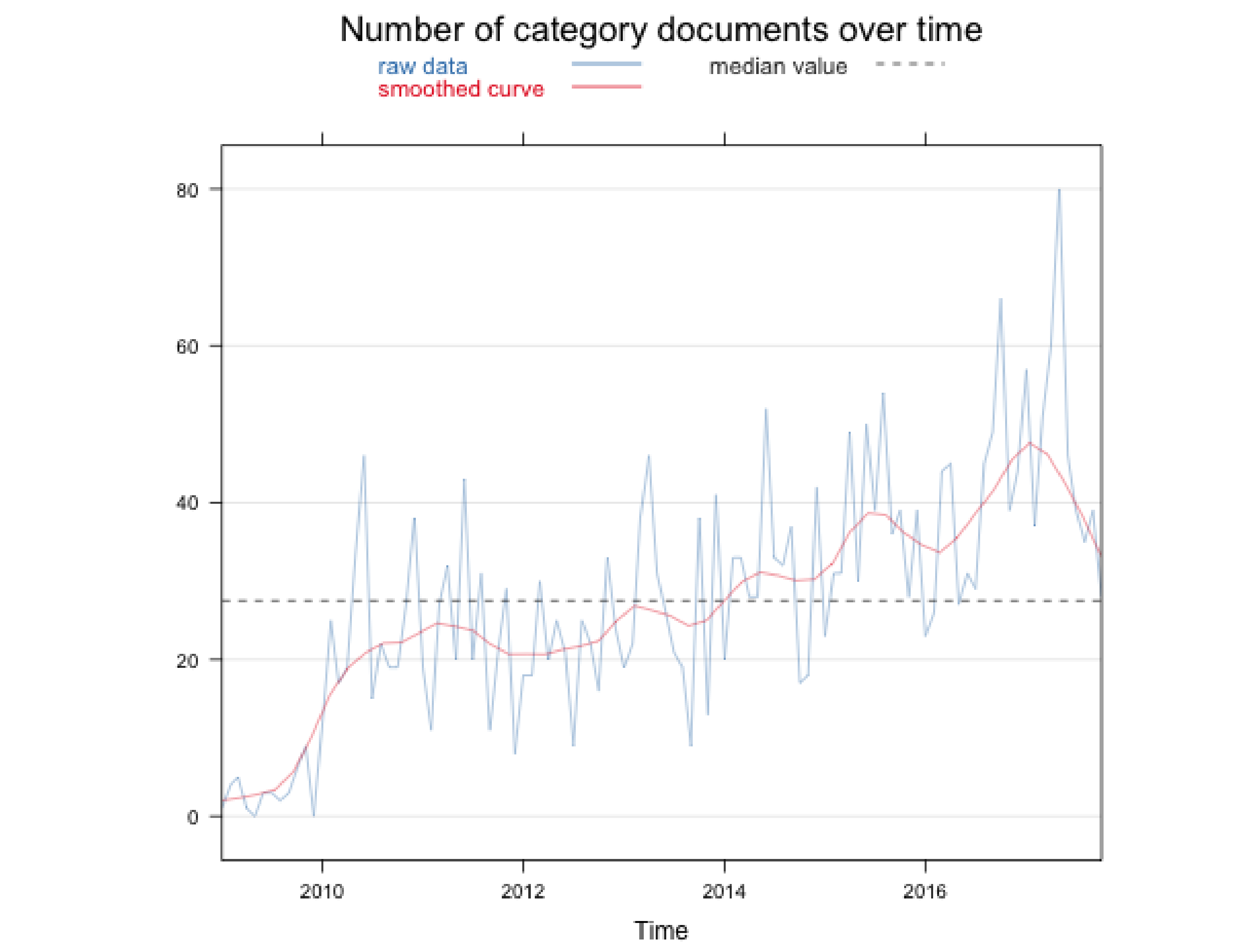


Figure 31b. Time series of Functions topics by number of document counts.

Processes

Discussion of middle-office and back-office functional processes centered around risk management, finance, and operations were largely visible in our results. Moreover, we can see in Figure 32a that transformation of the risk management operating model was the most talked about topic in the processes category. The breakdown of the risk management operating model was central to the process failings that ultimately lead to the financial crisis both individually as well as cross-functional information flow breakdowns.

We observe three general categories of process discussion as shown in the distribution of topics presented in Figure 32a. The first theme centers around client exposure management.

Derivative contracts in the form of swaps agreements introduced a lot of complexity during the 2008 Financial Crisis because the Lehman default triggered many to call in default clauses in the contracts, and therefore claim ownership of the collateral. As such, in the following years great focus was put on the management of derivatives with increasingly more deals traded on central exchanges. This allows the collection of margin or collateral, and ensuring it stays funded relative to changes in market value of the related contract. As such, many of the topics in this category relate to the management and optimization of derivatives collateral, and other clearing and settlement activities.

The second area of process topics relates to how financial firms manage their risk and resiliency. Changes in market conditions or with the financial condition of firm counterparties poses a big risk to financial firms, that was evidenced in the financial crisis. Therefore, a variety of risk, finance and operations processes related to transparently and more accurately capturing financial risks was a key topic. Perhaps the biggest was in the area of bank capital rules under the Basel Accord. Firms faced new and evolving regulations for bank capital such as Basel III and subsequent iterations, as well as being subject to very demanding bank stress tests under the Comprehensive Capital Allocation Review (CCAR) including Fed based scenarios. These new rules took risk management to develop more risk sensitive processes that used more granular data sets. Additionally, the focus on increased levels of controls was also a major focus.

A third category of process topics centers around data and process management. Data management is becoming an increasingly important topic. As firms increasingly rely on data architectures to supply a multitude of timely and accurate information sources, data governance, data standards, and processes around data management. Data is a central topic and discussed in more detail below in the Data category topic.

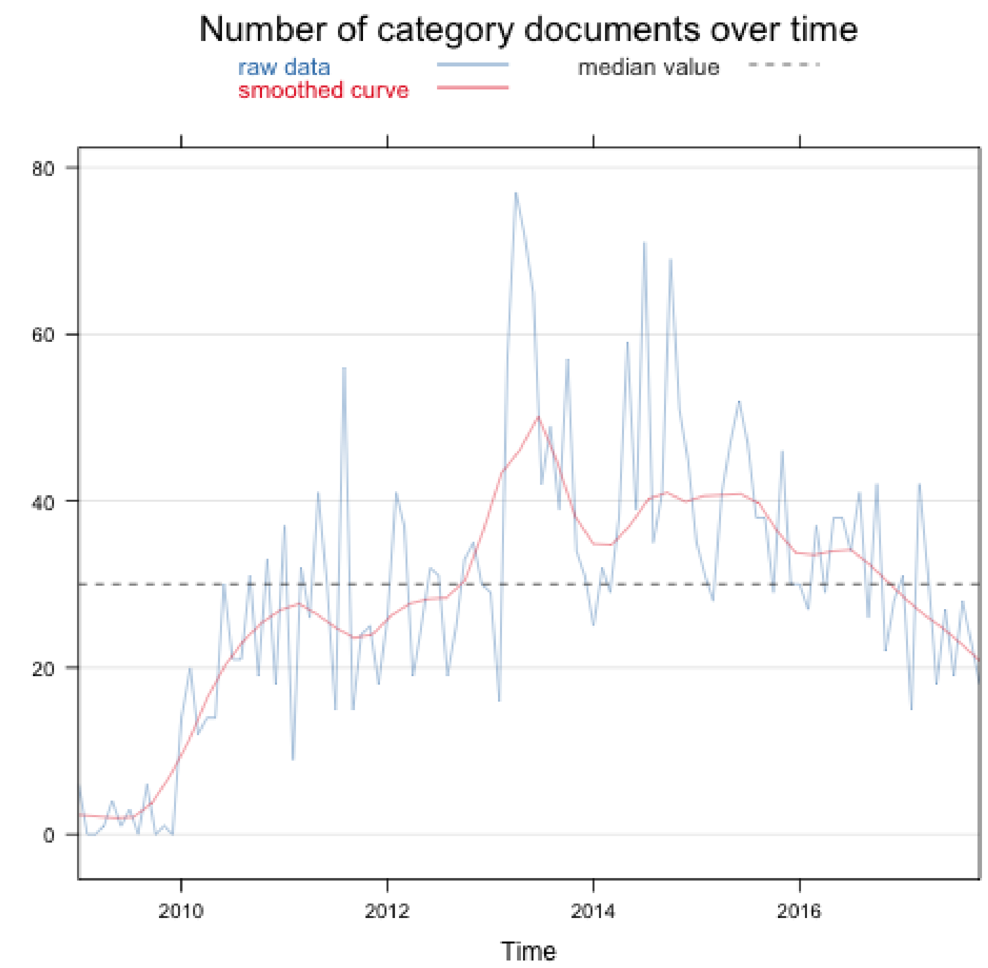


Figure 32a. Time series of Process topics by number of document counts.

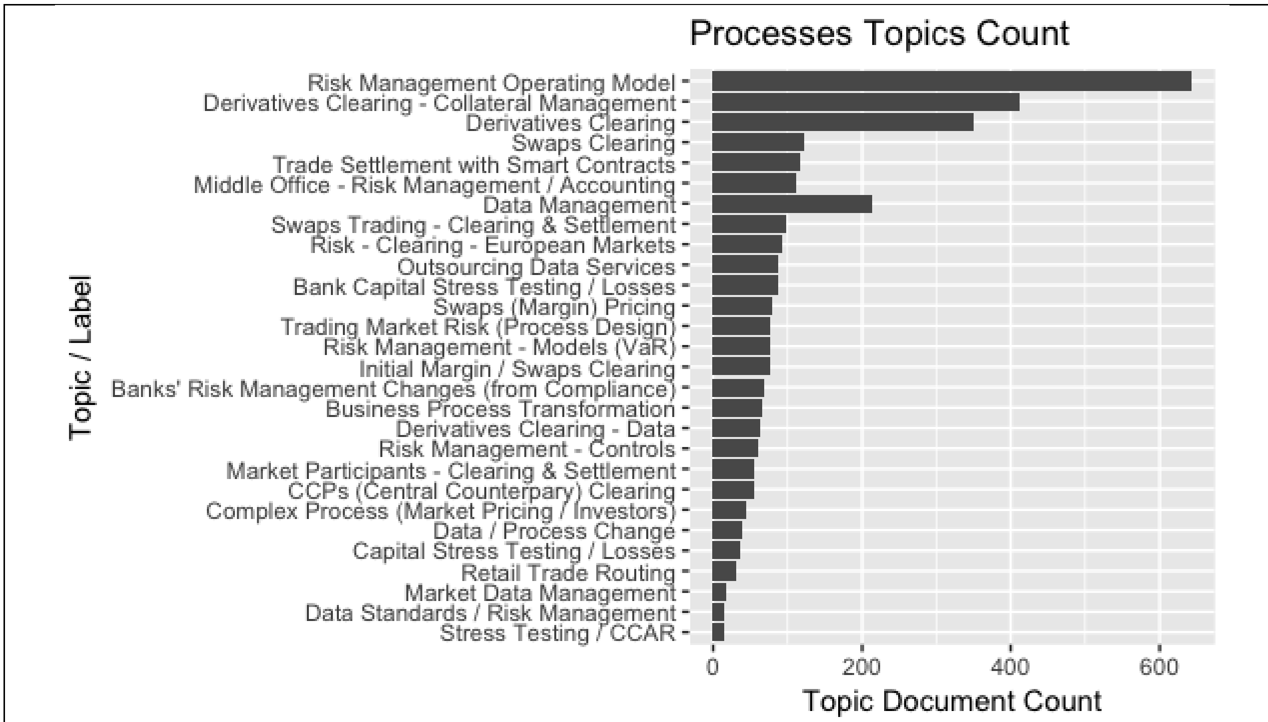


Figure 32b. Distribution of Process topics by number of document counts.

Systems & Technology

Underpinning nearly every aspect of modern financial markets are sophisticated systems and technology, which enables managing the speed and volume of processing necessary to transact the millions of daily trades and positions of financial markets. While large scale technology infrastructure has supported the markets for decades, an interesting result from our topic models shows that many new technologies not historically part of financial services infrastructure are emerging as shown in Figure 33a. Specifically, blockchain was the most frequent topic covered by our corpus. Blockchain is a recent distributed ledger technology that enables permanent, immutable and trusted records of transaction events, and viewed as a disruptive technology for its potential to fundamentally alter financial services by cutting out intermediary parties. Other frequent topics also relate to more innovative technology discussion such as data storage on a cloud infrastructure and mobile fintech banking. These are technologies historically associated with Silicon Valley startups.

While more traditional banking technology was identified such as trade reporting and collateral management systems, there is a clear focus on emerging digital technologies centered on the use of machine learning. Another notable observation is that we can see from Figure 33b that the technology discussion has steadily increased in the last ten years. This is a clear indication that the industry views technology as a strategic imperative both in terms of competitive advantage and cost reduction through process automation. While most automation has focused on the low value operational processes, there is clear intent to continue up the value chain by incorporating machine learning into the process.

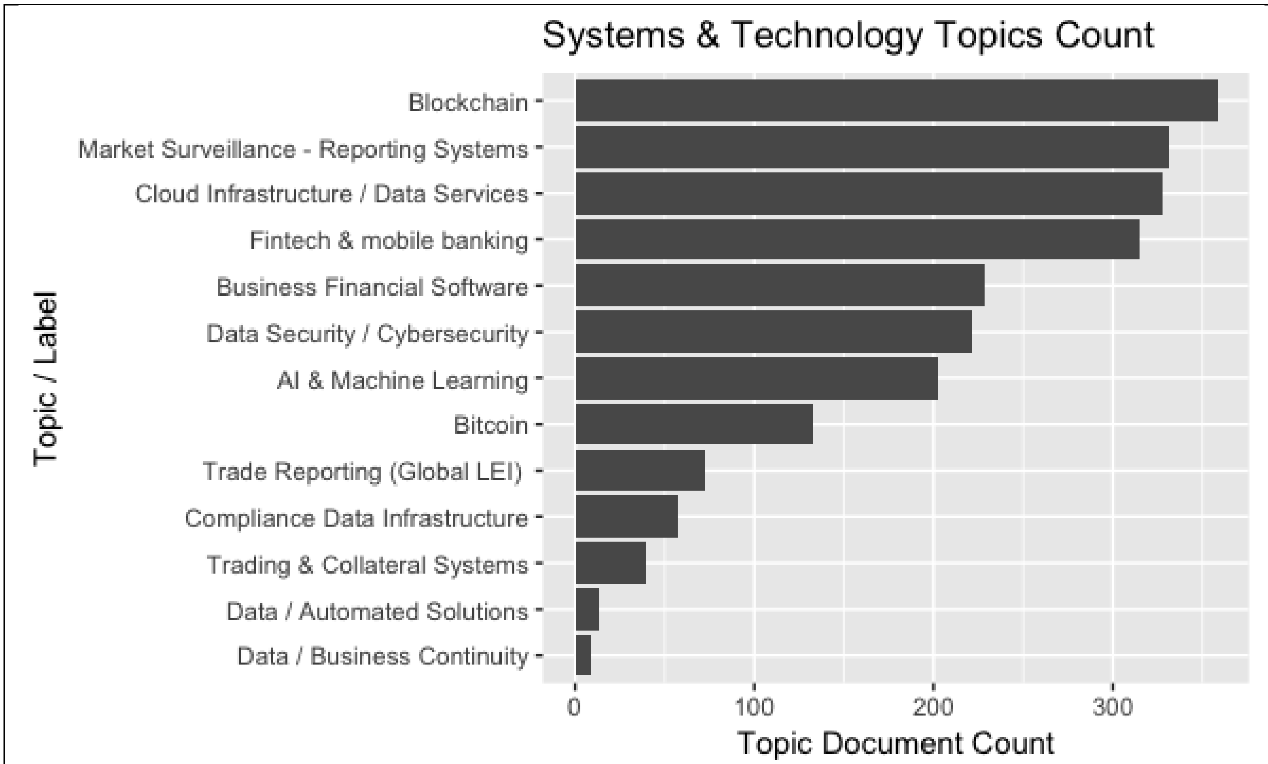


Figure 33a. Distribution of Systems & Technology topics by number of document counts.

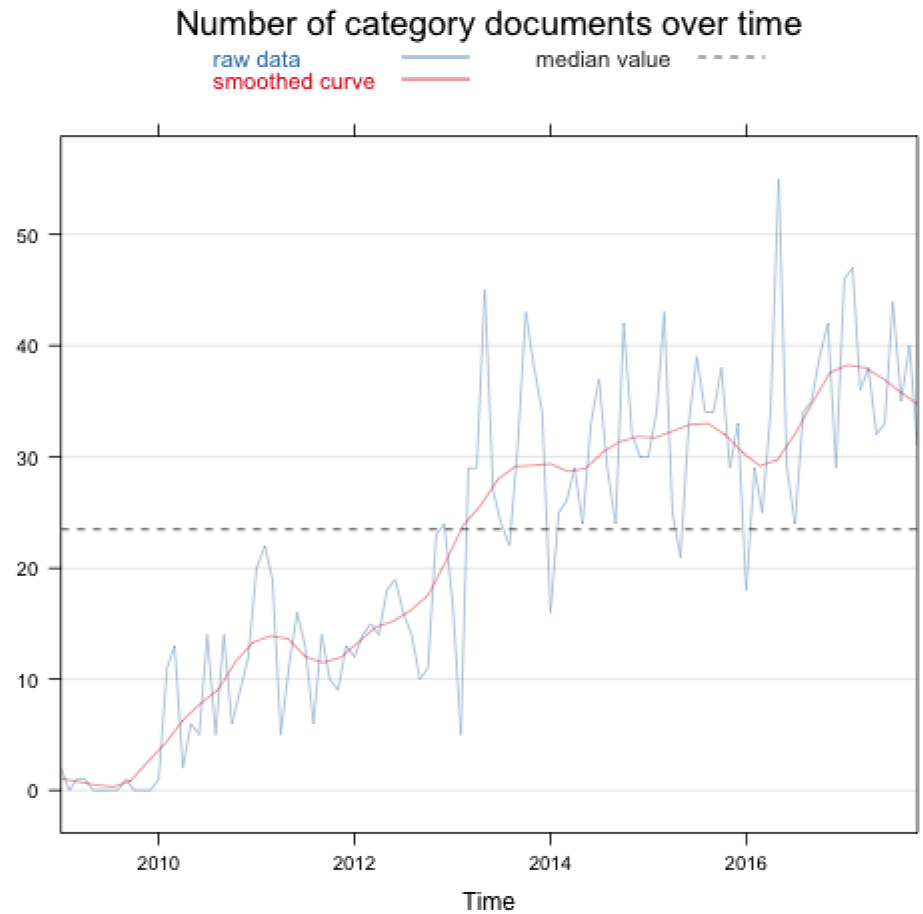


Figure 33b. Time Series of Systems & Technology topics by number of document counts.

Data

As a major theme in our corpus, data has substantially increased as a discussion topic in the most recent five years of our study period. We can see in Figure 34a that from 2014 onward, data has received tremendous attention. A clear leading theme centers on market data both in terms of real-time access and also in alternative forms such as social media market news. As shown in Figure 34b, these two topics were substantially more common than all the other topics. There appears to be a good correlation with the fact that algorithmic trading was a heavily discussed topic as previously covered, and that market data is a key input to algorithmic trading.

Regulatory reporting was a substantial topic emanating from multiple regulations, therefore it is not surprising that data quality for regulatory reporting emerged as a key topic. Data management as a theme represents the data inputs required for business processes as well as the storage of such data, and both are observed in our results. Another theme of the data discussion is customer data, with the most notable relating to the Legal Entity Identifier (LEI), aimed to standardized counterparty identification across firms and internal systems. Client data in general also emerged as a key data topic.

The results show that various kinds of data are important to the financial markets including transactional data, market data, reference data, as well as descriptive characteristics of the data such as real-time data, big data, unstructured data which all can be observed in our results. Lastly, given the importance of data for the financial markets, two additional topics that were identified relate to data security and privacy. These topics address how customer and firm data is protected, and also business continuity which requires the that data storage is appropriately backed-up for retrieval purposes.

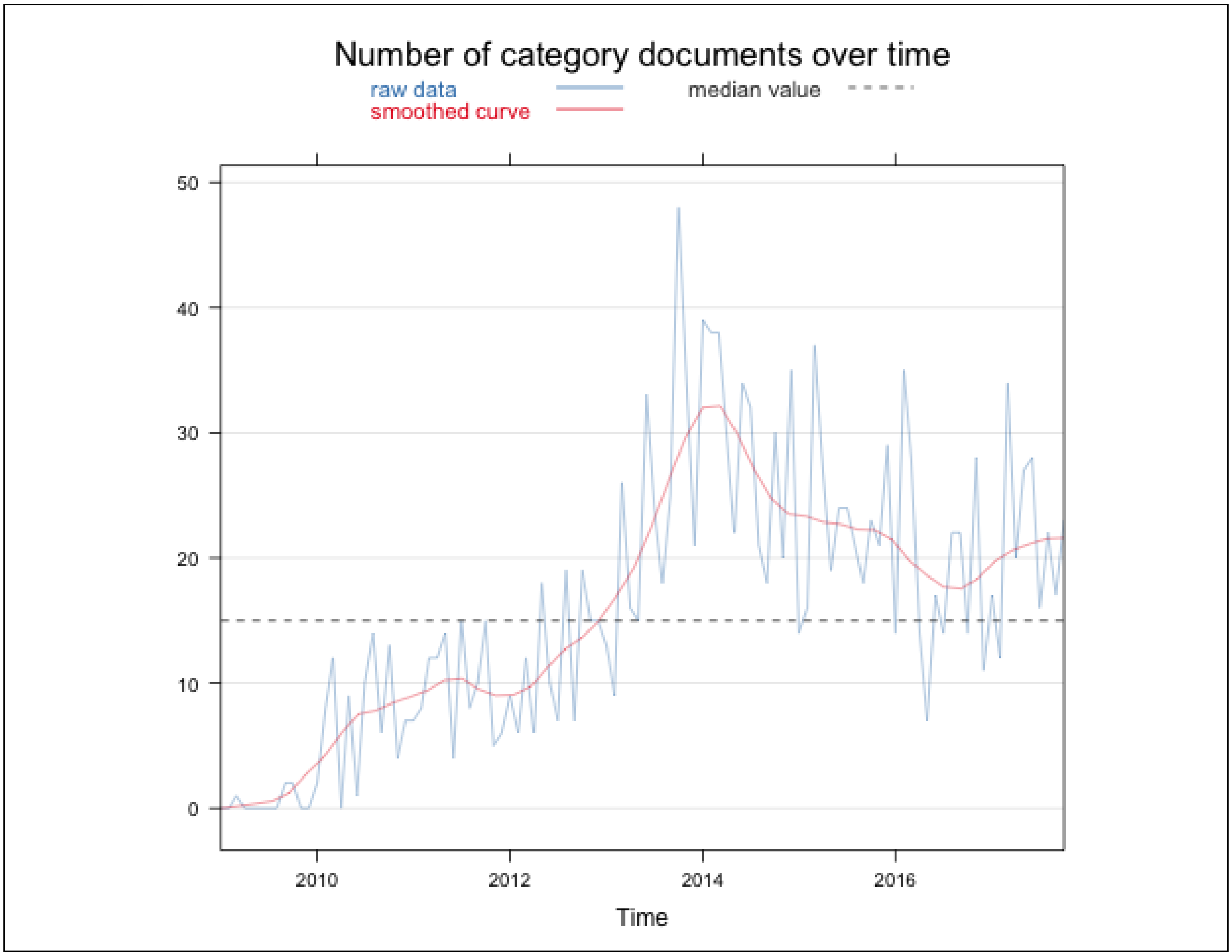


Figure 34a. Time series of Data topics by number of document counts.

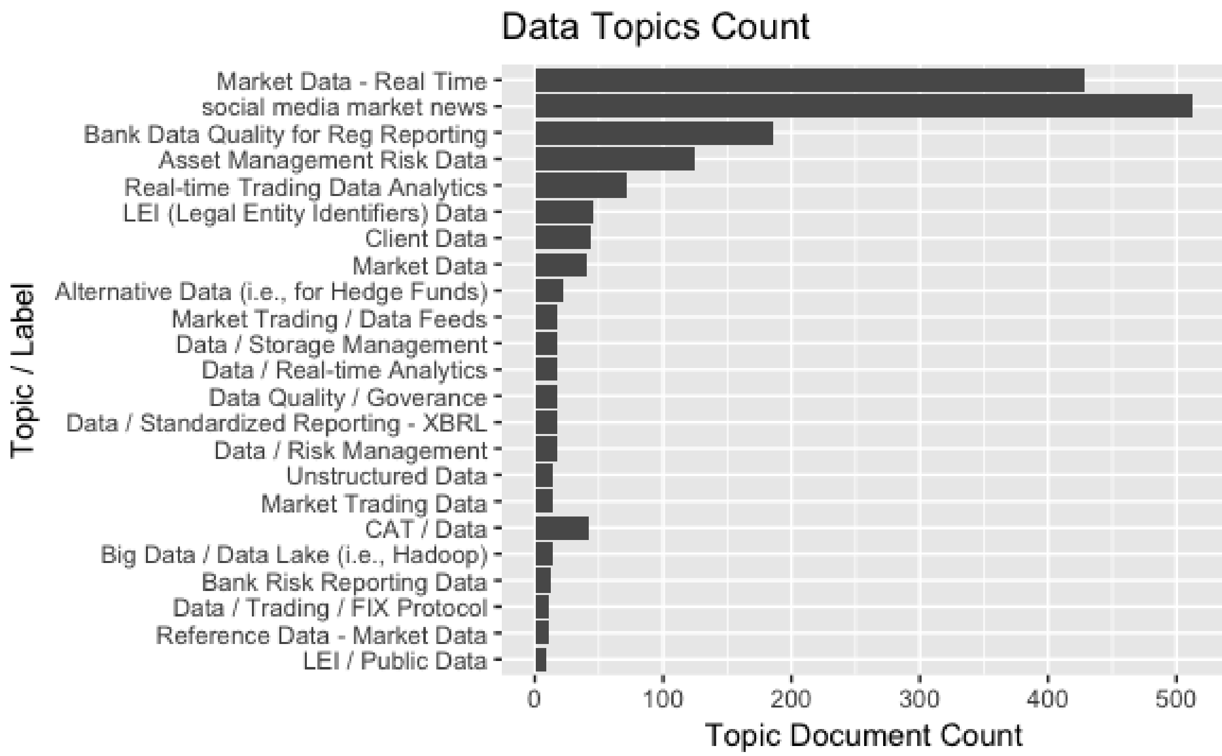


Figure 34b. Distribution of Data topics by number of document counts.

Networks

A final top-level category relates to the interaction of market participants to carry out market functions across networks. With multiple financial markets for the same products and services trying to connect buyers and sellers, networks can become complex. For example, securities can be purchased on multiple exchanges, bi-lateral agreements in the form of derivative instruments, and as shown in in Figure 35 alternative venues such as dark pools. The topic of dark pools drew substantial attention in our corpus relative to other network topics. These trading venues allow buyers and sellers to trade anonymously where computer algorithms match trades to mask who is changing positions. These constructs become increasingly important for firms buying or selling large positions and want to hide the trade to avoid price manipulations. As observed, networks are important for financial derivatives markets and often cover issues related to exchange requirements, transaction frequency requirements, and ability to receive real-time market-data feeds for trade volumes and prices.

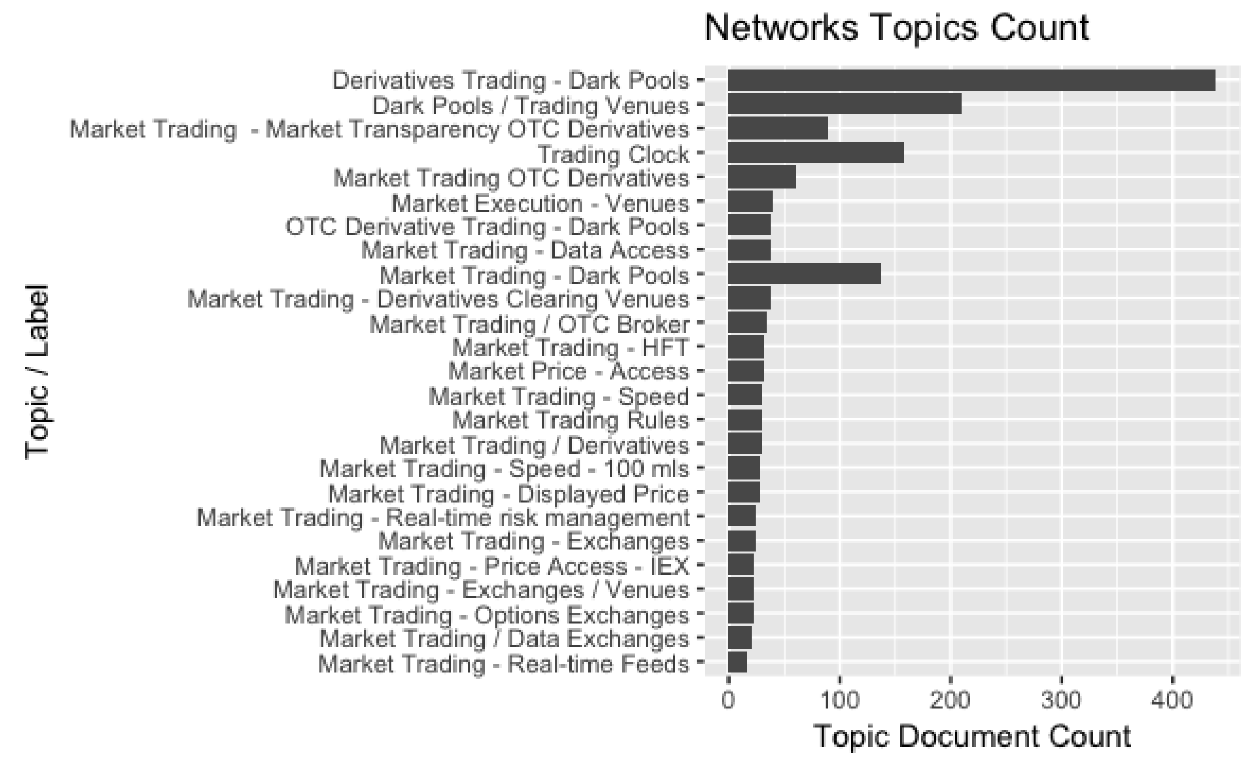


Figure 35. Distribution of Network topics by number of document counts.

A Financial Markets Knowledge-graph: Linking Ontologies and Topic Models

Overview

As previously argued, a major challenge that the financial industry faces is extracting timely, relevant, and specifically-focused information from text. Furthermore, a central challenge is to educate senior stakeholders on the power of using text as a corporate data asset to gain insights into business strategy and operations, and uncover emerging trends. Thus, the notion of fusing a financial ontology into text analytics topics models is a pragmatic approach that arguably yields more meaningful results by taking the higher-level semantic view established by our ontology, and connecting it to the lower representation levels found in the signals produced by topic models.

As previously shown, Figure 6c (repeated below), and supported by Table 1, illustrates the simple idea of connecting terms found in our corpus back to a domain ontology. By taking a specific concept like algorithmic trading, we are able to demonstrate a key benefit of making higher-level connections with the multitude of major market forces at play. This approach provides an opportunity for contextual comparisons over time and use of additional covariates. Knowing that algorithmic trading is an important topic is helpful. However, a much more powerful result would be establishing a capability to capture and monitor changes in related and predictive measures over time. For example, by making the semantic association between algorithmic trading and the notion of electronic oversight, we can subsequently measure changes in regulatory ‘tone’ (i.e., sentiment) around the topics across various mediums such as regulatory comment letters, speeches and public news. With the abundance of text surrounding the financial markets, there is an unlimited set of sources and topics that can be analytically modeled to gain fast, actionable, and specific insights to support management of business operations.

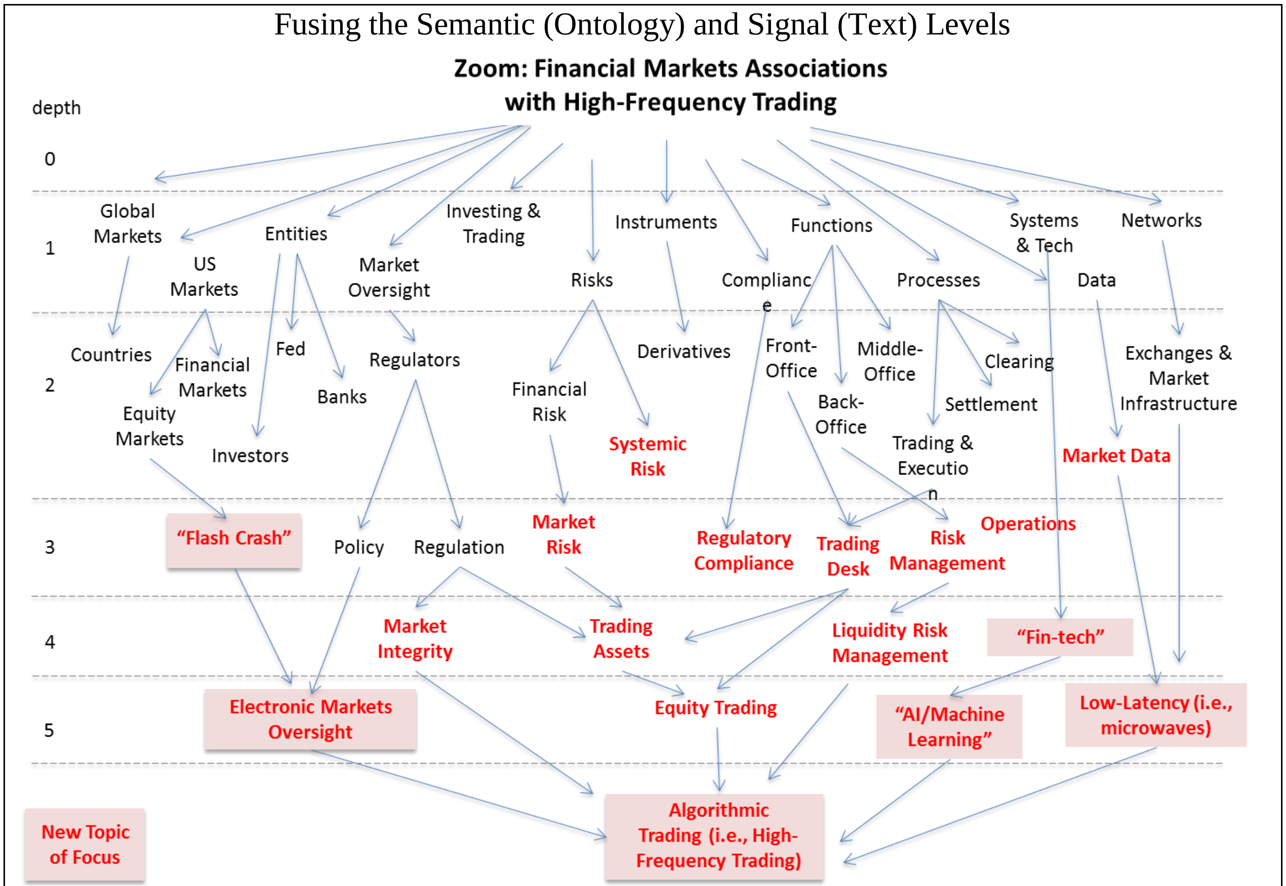


Figure 6c. Single-topic illustration of many semantic connections to financial markets.

Topics Surrounding Equity Trading

#	Category	Topic	Topic Document Count
1	Investing & Trading	Buy Side Trading	795
2	U.S. Markets	Equity Market Volume	650
3	Data	Social Media Market News	525
4	Investing & Trading	High Frequency Trading – Strategies	450
5	Data	Market Data – Real-time	450
6	U.S. Markets	‘Flash Crash’	340
7	Market Oversight	Regulatory Compliance (Financial Firms)	325
8	Market Oversight	High Frequency Trading – Market Regulation	295
9	Investing & Trading	High Frequency Trading – Latency	210
10	Systems & Technology	Artificial Intelligence / Machine Learning	200
11	Market Oversight	NMS Rule Equity Regulation	110
12	U.S. Markets	Trading Markets – Equity	101
13	**Investing & Trading**	**Algorithmic Trading**	**98**
14	Functions	Market Execution	90

15	Functions	Trade Execution – Routing	75
16	Market Oversight	Flash Crash – Regulatory Surveillance	50
17	Networks	Market Execution – Venues	40
18	Networks	Market Trading – speed	38
19	Compliance	Market Surveillance – Trading Data	35
20	Data	Alternative Data – Hedge Funds	25

Table 1. Summary of topics and occurrence frequency related to equity trading

Drilling into LDA example with concrete insights: digital transformation discovery

As an example of a concrete benefit of what topic modeling can deliver, one of the key results from this work is that regulation started as a means of market reform, but is also quickly enabling its digital transformation with calls to collect, organize and deepen analytics of cross-functional data now centrally stored as a result of many regulations. However, unearthing specific cases of how business processes, regulatory infrastructure investments, and related technology interdependences evolve from the ground-up is complex and not naturally transparent without a structured approach. Figure 36 below is intended to show the complex web of dependencies that exist among compliance topics and related 'nearest neighbor' concepts. We see that concepts logically connect with other concepts sometimes in a circular nature. For example, we see that 'Data' (Path 1 on diagram) feeds many cross-functional regulatory driven business processes shown by tracing flows from left to right: 'Asset Manager Research', 'Client Onboarding – Data Regulations', 'Information Security Compliance', and 'Data Aggregation and Reporting – BCBS 239', and others. However, another path through the network below (Path 2 on diagram) shows that 'Compliance' leads to 'Post-Trade Analytics', and in turn to 'Collateral Optimization' and 'Collateral / Margin Management.'

Without a domain ontology, making connections between compliance topics is difficult

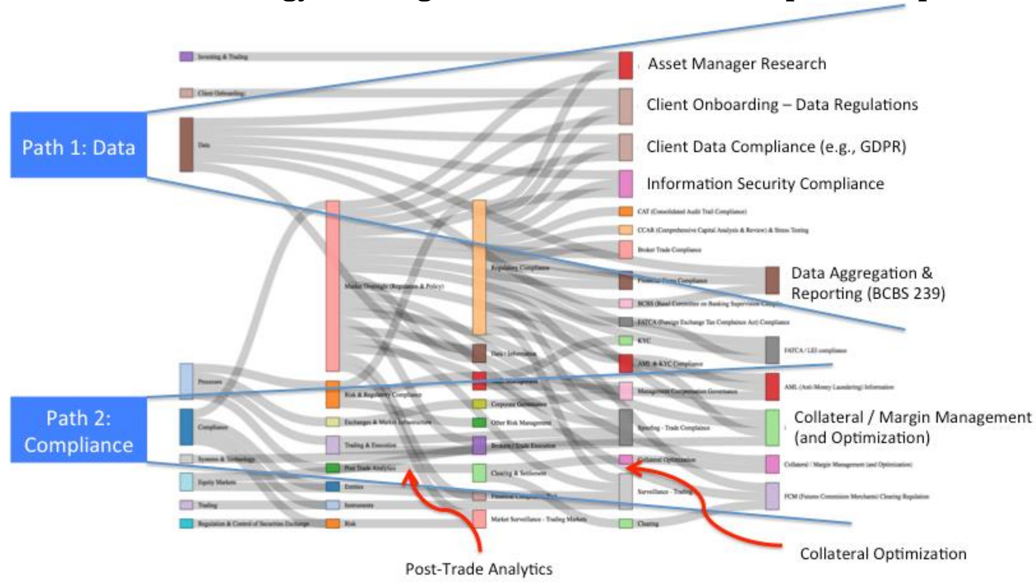


Figure 36. Illustration of compliance topics and subset of nearest-neighbor connections

While links between general clusters of topics can be observed, it's hard to extract the context and origin of the topic. Therefore, a much better way to depict meaningful connections among new topics in a given domain focus area is through the use of ontologies. Two 'eras' have been identified between the early post-crisis years (i.e., 2009 to 2015) and the later stage post-crisis years (i.e., 2015+). The first era focused on implementing new regulations in the wake of the financial crisis, many of which required capturing, storing and reporting data at new levels of granularity and across many new measures. These early stage infrastructure investments improved collection and storage capabilities through the adoption of new big data technologies such as the enterprise data lake. With all this cross-functional data in one place, financial firms were well positioned to think of data as a corporate asset and pursue new digital channels for connecting with customers and using advanced analytics for improved services. Figure 37 below visually depicts the key regulations captured by our LDA topic models on the left side, with the major topic themes shown below each era. The intent is to highlight how the leading themes of the 'regulatory compliance' era required heavy data infrastructure investment.

The resulting themes of the 'digital transformation' era were made possible by the resulting modern 'big data' architecture.

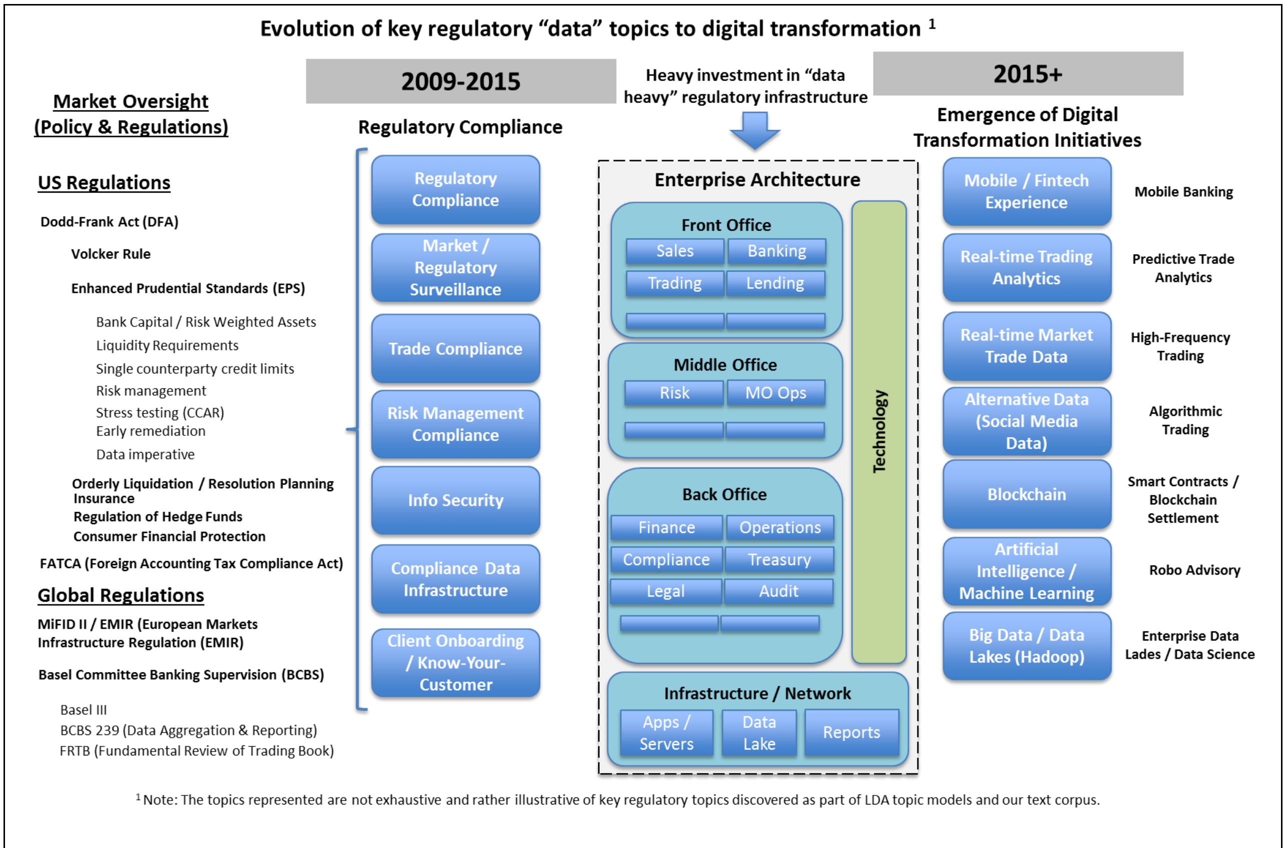


Figure 37. Depiction of regulatory drivers surrounding Enterprise Architecture leading to digital topics.

A distinct set of topics was discovered in each era, however, they can be related to each other by establishing common links back to a base ontology. Throughout our work, we established a base ontology for the financial markets with subsequent specializations. As such, understanding the evolution of topics over time is easier when depicting their existence as of a point in time relative to a base ontology. To this end, the new topics from each era depicted in Figures 38a-b are not linear extensions of each other, but rather two different sets of topics, both deriving from or extending an original base ontology. However, the extensions were different for the two different timeframes, and provide insight into evolutions of sub-topic structure.

Taking a closer look at the first era, which was dominated by new regulations and the stand-up of new compliance programs, we are able to quickly see ‘data’ emerge as a central tenant. Figure 38a shows how an ontology can depict the links between regulatory compliance topics and need for data heavy infrastructure. The financial industry has always produced tremendous quantities of data. However, the wave of post-crisis regulation established new data requirements on three levels. First, a broader scope of business activity to be covered such as oversight of trading desk activity against permitted mandates under the Volcker Rule. This forced a review of all trading business activity to ensure trade flows were included for completeness.

Second, a new set of measures and changes to existing measures on many levels required new data sourcing, additional transformation of data, and the computation or modeling of required measures. For example, the Basel Accord under Basel III imposed new measures for the calculation of counterparty exposures associated with securities financing transactions and derivative instruments. These new measures were required to account for collateral in new ways and consider netting restrictions across other trades with the same counterparty. Many examples of new computational complexity exist.

Lastly, the granularity of data requirements increased in some instances. This required sourcing, for example, loan level or trade level information in the forecasting of stress scenarios aimed to predict banking losses. The net result was larger, deeper, and more frequent data sets to consume and store for integration into calculations and reporting. For example, derivatives trade reporting was a major topic of discussion.

With the evolution of the markets, an emerging era centered on digital transformation appeared around the time of 2015. Figure 38b shows a convergence of several data-driven topics towards digital transformation initiatives with an ontology. We can clearly establish a context for how data connects to customers, products, markets and new technologies under the new banner of 'FinTech'.

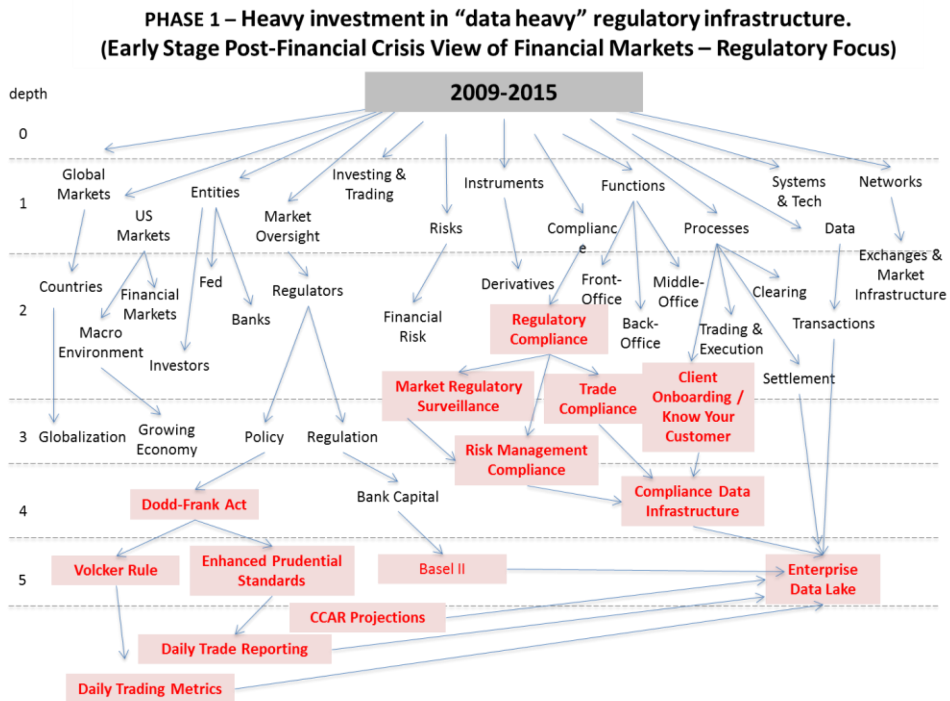


Figure 38a. Illustrative view of 'Regulatory Focused Tech & Data Financial Marketplace.

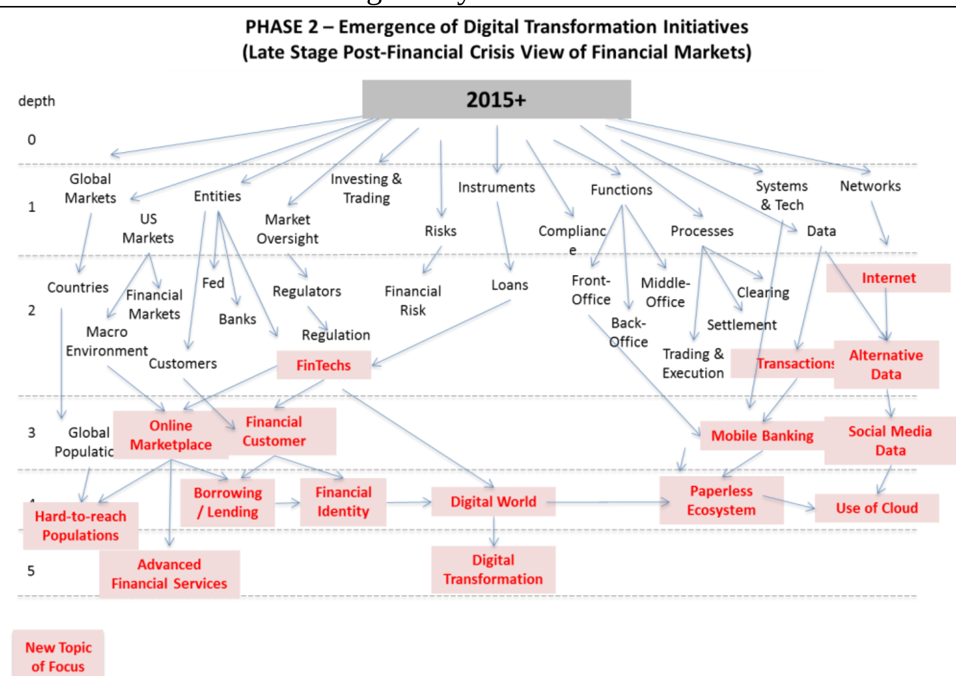


Figure 38b. Illustrative view of 'Emerging Digital Transformation' Financial Marketplace.

In addition to word-to-topic-probabilities, LDA also provides a mechanism for identifying the statistical presence of a topic within a specific document. This measure is referred to as 'gamma' and indicates the probability of a topic's existence in the document. In the prior section, many new topics associated with an emerging era of 'financial services digital transformation' were highlighted. However, with the use of the gamma measure we are able to identify specific documents of relevance. For example, the following redacted article was published by Ron Suber representing Prosper on December 9, 2016 under the heading 'The Past, Present and Future of FinTech' on TabbFORUM:

> As an investor in many companies leading **digital transformation**, I specifically seek out opportunities that help people gain access to credit in this new **digital world**. My investments are in **paperless ecosystems** with a vision of a world that includes **everyone financially**. The **Internet** has become an integral part of life for billions of people around the world, and as such, they should be able to access **financial solutions** while at the same time gain a **financial identity**. **FinTech** is the answer. … I believe there are four phases of **FinTech**. The first phase began with establishing **FinTech** as a recognizable concept with what I call a period of EAU: Education, Awareness, Understanding. … As the industry was getting off the ground, both industry insiders and the public reacted hesitantly to the idea of **borrowing and lending** on the **Internet**, which for most sounded a Ponzi scheme. … Phase Two of **FinTech** took place early last year, when **large finance and technology companies**, as well as **banks**, started taking a sincere interest in **marketplace lending**, moving **FinTech** and **online lending** from a novelty to an interesting new niche. Relationships formed between **large financial institutions and banks** that wanted to better understand the benefits that new **online marketplaces** could offer; **alternative data** and **lending companies** were no longer seen as competition, but rather as potential partners for **financial companies** and institutions. Phases 3 and 4: FinTech Today and the Optimistic Future Phase 3 is now. **Digital transformation** means different things to different industries, but at its core it is the same, a new way of conducting business, transactions, even life. **Digital financial technology, or FinTech**, is doing three things now: enabling new products, providing new services and meeting new segments in hard-to-reach populations, where they are and at a low cost. … The use of **cloud computing** and **data science** have made **complex functionalities** accessible across diverse locations and markets … With three billion new **smartphone subscribers** expected in emerging markets by 2020, it is clear that the potential to deliver **advanced financial services** is vast and is spurring a **digital revolution**. Morgan Stanley reported in June 2015 that marketplace lending companies could command $150 billion to $490

billion globally by 2020. In Phase 4, we will see rapid mergers, acquisitions, and full adoption into the main stream.

The specifically-focused opinion piece speculates the emergence of many disruptive forces for the financial industry that strategically envisions the use of new technologies, data and analytics to better serve financial customers. By providing the statistical links of topics to documents, we are able to zoom into specific text to gain better insight based on the purpose of our research. Thus, LDA topic modeling provides a broad tool to scan large amounts of text, focus on subsets of text for increased granularity of topics, and finally enabling deep-dives on specific text documents in the corpus for further inspection.

In summary, navigating the evolving web of complex financial services and products across market participants and their business operations requires new ways of semantically modeling connections across the chain. Topic modeling as a tool alongside ontologies provides a powerful mechanism to traverse these evolutions quickly and with the ability to scale. The use of semantic methods alongside machine learning models is still at an early stage. However, ultimately organizations in the financial industry are well placed to take advantage of such a capability.

Conclusion

Review

Specifically-focused text analytics combined with a content rich ontology provides a contextual lens to review the landscape of terms discovered throughout a corpus. More specifically, topic modeling applications with unsupervised methods such as LDA can provide tremendously helpful insights to businesses and individuals that need to extract insights from large collections of text. Through the analysis of a large volume of text data related to the decade following the 2008 Financial Crisis from 2009 to 2017, major themes were identified. Someone without deep domain knowledge would be able to understand what the important policies, regulations and financial industry implications were, as well as key market players, main functions and impacted processes, and major focal points of discussion. Though these data reveal that policy forces were evolving over time, one is still able to make sense of the major dynamics of the financial industry and how it certainly will have a data and technology driven future.

This thesis introduced domain representation through ontologies and topic modeling as a solution to the problems of providing semantic context to analytics, making sense of large corpora, and combining the two for an informed contextual network of topics. The literature review presented a range of text analytics and modeling methods, and described why structural topic modeling is best suited to achieve the desired results. Methods were described in detail so that the paper can serve as a blueprint for future analyses. Finally, the results were revealed and discussed in depth so that meaning can be drawn from the large array of 302 topics. The following concluding remarks provide some additional guidance for readers and analysts alike.

Assumptions

Several important assumptions were made throughout the course of this analysis. It is critical to keep these assumptions in mind while reviewing the methods and results. Because the financial industry domain is vast, only a sample of regulatory content in the form of Federal Board of Governor speeches and opinion articles were analyzed. This means that topics discussed in other news sources have the potential to be completely different from what is captured and depicted above. Secondly, the 2008 Financial Crisis had profound impacts on the global markets, however, our work was biased towards the U.S. market impacts given the scope of sources selected. To some extent, this limits the global financial industry landscape. An analysis that considered key market news sources such as *Bloomberg, The Wall Street Journal,* and other credible media channels would have provided a broader perspective. Additionally, more focused publications around functional issues such as *Risk.net* would also have added depth of discussion to certain key regulatory topics and financial firm implications. However, many of the sources were not public and maintain legal restrictions.

Another key assumption is that our domain ontology started with general topics and was built out to more specific levels only to the extent that links were needed to bridge concepts discovered as part of the analysis. A much broader set of financial industry terms could be included in a more robust ontology. For example, the Financial Industry Business Ontology (FIBO) is one example of an industry driven effort to model the range of financial contracts. Efforts could be conducted to identify and leverage available ontologies, as well as further build out the domain ontology with help of industry experts.

Furthermore, topic modeling itself is a method of quantifying how common groups of words appear throughout a corpus. This means that uncommon words, though they might be crucial for a more sophisticated understanding, are not included in the analysis. Topic modeling aims to capture the broadest themes exhibited in the data. The number of topics chosen has an immense impact on the granularity of the results. By choosing 302 industry topics, it is likely that conclusions related to the commonalities of smaller topics are omitted. This is to say that topic modeling may not be the right choice for an analyst pursuing fine-grained detail. Like a map, topic modeling does not aim to capture the small details within a data set. Instead it aims to capture the bigger picture, or the landscape within which the data exist.

Future Research

There are several ways in which this analysis can be complemented in the future. First, the set of parameters around topic modeling is a very manual and trial-by-error approach. More robust tools could be developed for selecting the number of topics to be identified by LDA methods, or other inputs to optimize the quality of the results. This would require less time spent on orchestration of the text to obtain meaningful results, and more time inspecting and analyzing results.

Another improvement would be the ability to find labels for the output of LDA topic models with less human intervention. Either through external services that could link words to topics through common associations, or ways to crowd-source the process of labeling to multiple people for both the assigning of initial labels as well as cross-validation. As noted by (O'Neill, 2016) Amazon Web Services offers a service that allows users to assign human intelligence tasks like classification and labeling to a set of participants.

Finally, a major area of improvement would be to find better ways to get to the ontological or semantic level with greater feedback from the statistical level. This has generally not been possible over the last decade or more, despite many attempts to try. Analysis of text at the statistical or signal level is only able to identify co-occurrence of text with other text, not establish any semantic meaning. Some work by (Chen et al. 2018) addresses an analogous problem through vision analytics. Their discovery is that by modeling deep signal level data, after it has been parsed, smoothed, feature-extracted, and other processing, with the incorporation of ontologies provides a feedback mechanism to aid in the interpretation. There is substantial more work to do here, but it's an initial step towards a feedback mechanism and more meaningful results at the semantic level.

Final Remarks

Though the major results of this work show that the financial industry discourse is complicated by political views and the sentiment of impacted retail investors in the wake of the 2008 Financial Crisis, it is important to remember that the risks that lead to the financial crisis were very real. Thus, despite the political overtone, the need for major reform and continued improvement of risk management is generally viewed as a positive outcome of the crisis. While no future crisis can ever be fully prevented, there is a much greater confidence that if the scale of the crisis re-occurred in a post Dodd-Frank Act world, the effects would not be as severe. Specifically, banks are much more capitalized to absorb losses and positioned for better management of exposures with individual counterparties. Moreover, robust stress testing under the Comprehensive Capital Allocation Review (CCAR) program provides much greater confidence that banks could withstand a large shock to the markets. Finally, a new resolution regime would prevent a fire sale of assets by providing more legal tools and an improved framework to the FDIC as receiver to manage an orderly liquidation.

Many mainstream constituents were convinced that large financial firms were the sole force behind the crisis. However, the build-up of risk, yield hungry investors at all levels, and the transfer of risk through the system in a non-transparent manner all contributed to the financial crisis. Hopefully, with enhanced and more regulation, increased data capture and reporting, and advanced analytics future crisis scenarios will be detected earlier and steps can be taken by regulators and investors to unwind positions and rebalance portfolios before sudden asset drops. Overall, it should be noted that the amount of time spent complying with regulations could have been devoted to better serving businesses and investors in ways that improve overall financial opportunity. Therefore, three major findings from this analysis are:

- Regulation was a major focus across all functions and businesses of major financial firms, however, these same firms are increasingly well placed to better understand the needs, risks and opportunities of their customers and counterparties
- Data and technology are increasingly being used as tools for competitive advantage and continue to drive a digital transformation that will likely last decades. With the central storage of enterprise data, increases in processing capability, and wider adoption of advanced analytics such as machine learning, financial firms are entering a new era of disruption and change
- Financial organizations have undergone so much change in the last decade that operating models regularly anticipate change and future changes as part of the status quo. Therefore, financial organizations are much nimbler and are well positioned to better navigate future challenges with increased transparency, information flow across functions, and better governance

The analysis of speeches and opinion articles generated during the course of the decade following the 2008 Financial Crisis with probabilistic topic modeling represents a novel application of text analytics methods. It assists readers with a clarified understanding of the major issues and illustrates a powerful set of tools that can be applied to other disciplines. The text analytics community has gained a powerful tool that provides a fast and actionable method for extracting value from text. Therefore, a much wider set of text documents can be mined for hidden value and a major corporate asset that can provide an entire wave of new opportunities.

References

Arp, R., Smith, B., Spear, A.D. (2015). *Building Ontologies with Basic Formal Ontology*. Cambridge, Massachusetts: MIT Press.

Bergman, M., (2018, Feb 21). *Desiderata for Knowledge*. Retrieved July 12, 2018, from http://www.mkbergman.com/category/ontology-best-practices/

Blei, D. M. (2012, April). Probablistic topic models. *Communications of the ACM*, 55(4), 77-84. doi:10.1145/2133806.2133826

Blei, D. M., & Lafferty, J. D. (2006). Dynamic topic models. *ICML '06 Proceedings of the 23rd international conference on Machine learning*, (pp. 113-120). doi:10.1145/1143844.1143859

Blei, D. M., Ng, A. Y., & Jordan, M. L. (2003, March 1). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3, 993-1022. Retrieved from http://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf

Chang, J., Boyd-graber, J. L., Wang, C., Gerrish, S., & Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. *Advances in Neural Information Processing Systems*, 22, pp. 288-296.

Chen, X.L., Li, L.-J., Li, F.-F., and Gupta, A. (2018, Mar. 29). Iterative visual reasoning beyond convolutions, *arXiv:1803.11189*. online access, accessed April 1, 2018 by TAD

Chungoora, T. (2018). Ontologies for Business Analysis. Retrieved from udemy.com: https://www.udemy.com/user/tishchungoora

Curme, C., Zhuo, Y. D., Moat, H. S., & Preis, T. (2017, March). Quantifying the diversity of news around stock market moves. *Journal of Network Theory in Finance*, 3(1), 1-20. ISSN 2055-7809.

Davis, R., Shrobe, H., & Szolovits, P. (1993). What is a Knowledge Representation? *AI Magazine*, 14(1):17-33, 1993. http://groups.csail.mit.edu/medg/ftp/psz/aimag-final.ps

Ding, C., & Xiaofeng, H. (2004). K-means clustering via principal component analysis. *ICML '04 Proceedings of the twenty-first international conference on Machine learning* (p. 29). New York: ACM. doi:10.1145/1015330.1015408

EDM Council. (2018). FIBO Domains. Retrieved on Aug 4, 2018 from edmcouncil.org: https://wiki.edmcouncil.org.

Hardy, M. R., (2013). Classifying Risk. 2013 Applied Report. Department of Statistics and Actuarial Science University of Waterloo.

Heijmans, R., Heuver, R., Levallois, C., Leilveld, I. V. (2016, June). Dynamic visualization of large financial networks. *Journal of Network Theory in Finance*, 2(2), 57-79. ISSN 2055-7809.

Holland, S. M. (2008, May). Principal Components Analysis (PCA). Retrieved from strata.uga.edu: http://strata.uga.edu/software/pdf/pcaTutorial.pdf

IBM. (2010, May). *Mastering new challenges in text analytics, Making unstructured data ready for predictive analytics*. IBM Corporation. IMW14301USEN-01.

Jakus, G., Milutinovic, V., Omerovic, S., Tomazic, S. *(2013). Concepts, Ontologies, and Knowledge Representation.* New York, NY: Springer.

Joda.org. (2016, 11 03). *Formatting.* Retrieved from http://www.joda.org/joda-time/key_format.html

Khoo, C., Na, J-C. (2006). Semantic relations in information science. *Annu Rev Inform SciTechnol* 40(1):157–228

Lewis, D. D., & Ringuette, M. (1994). A comparison of two learning algorithms for text categorization. *Third Annual Symposium on Document Analysis and Information Retrieval*, 33, pp. 81-93. Las Vegas.

MacQueen, J. B., (1967). Some Methods for classification and Analysis of Multivariate Observations, *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, University of California Press, 1:281-297

Manning, C. D., Raghavan P., & Schütze H. (2009, April 7). *Tf-idf weighting*. Retrieved October 16, 2016, from http://nlp.stanford.edu/IR-book/html/htmledition/tf-idf-weighting-1.html

Noy, N. F., McGuinness, D. L. (2000). Ontology Development 101: A Guide to Creating Your First Ontology. Retrieved from https://protege.stanford.edu/publications/ontology_development/ontology101.pdf

Olteanu, D. (2002, May 28). XPath 1.0 Tutorial. Retrieved from http://www.cs.ox.ac.uk/: http://www.cs.ox.ac.uk/dan.olteanu/tutorials/xpath1.pdf

O'Neill, K.S., (2016, December). Political Landscape Discovery with Machine Learning: Using Structural Topic Modeling to Analyze Election Discourse. M.S. Thesis. Retrieved from Kelsey O'Neill.

Roberts, M. E., Stewart, B. M., & Tingley, D. (2016, 01 14). *Package 'stm'*. Retrieved from cran.r-project.org: https://cran.r-project.org/web/packages/stm/stm.pdf

Roberts, M. E., Stewart, B. M., & Tingley, D. (2016b, October 6). *stm: R Package for structural topic models*. Retrieved 10 09, 2016, from https://cran.r-project.org/: https://cran.r-project.org/web/packages/stm/vignettes/stmVignette.pdf

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., Rand, D. G. (2014, March 6). Structural topic models for open-ended survey responses.

American Journal of Political Science, 58(4). doi:10.1111/ajps.12103

Saadah, M. N., Atmagi, R. W., Rahayu, D. S., & Arifin, A. Z. (2013, February). Information Retrieval of Text Document with Weighting TF-IDF and LCS. *Journal of Computer Science and Information*, 2(1), 67-101.

Salton, G., & Buckley, C. (2016, October 6). *Onix Text Retrieval Toolkit*. Retrieved October 6, 2016, from http://www.lextek.com/manuals/onix/stopwords2.html

Schreiber G. (2008) Knowledge engineering. In: van Harmelen F, Lifschitz V, Porter B (eds) *Handbook of knowledge representation*. Elsevier, Amsterdam, pp 929–946

Shi, B.X. and Weninger, T. (2018, Feb. 23). Visualizing the Flow of Discourse with a Concept Ontology, *arXiv:1802.08614v1 [cs.CL]*. *arXiv:1803.04488v1 [cs.CL]* pdf, accessed Apr. 25, 2018 by T.A.D.

Sievert, C., & Shirley, K. (2015, October 24). *LDAvis: interactive visualization of topic models*. Retrieved from cran.r-project.org: https://cran.r-project.org/web/packages/LDAvis/index.html

Silge, J., Robinson, D. (2018). *Text Mining with R*. Sebastopol, CA. O'Reilly Media. Inc.

Soramaki, K., Cook, S., Laubsch, A. (2016, March). A network-based method for visual identification of systemic risks. *Journal of Network Theory in Finance*, 2(1), 67-101. ISSN 2055-7809.

Sowa, J.F. (2000). Ontology, metadata, and semiotics. In: Conceptual structures: logical, linguistic, and computational issues. *Lecture notes in computer science*, vol 1867. Springer, Berlin, pp 55–81

Strickland, J. (2015, May 27). K-Means clustering using R. Retrieved from linkedin.com:

https://www.linkedin.com/pulse/k-means-clustering-using-r-jeffrey-strickland-ph-d-cmsp

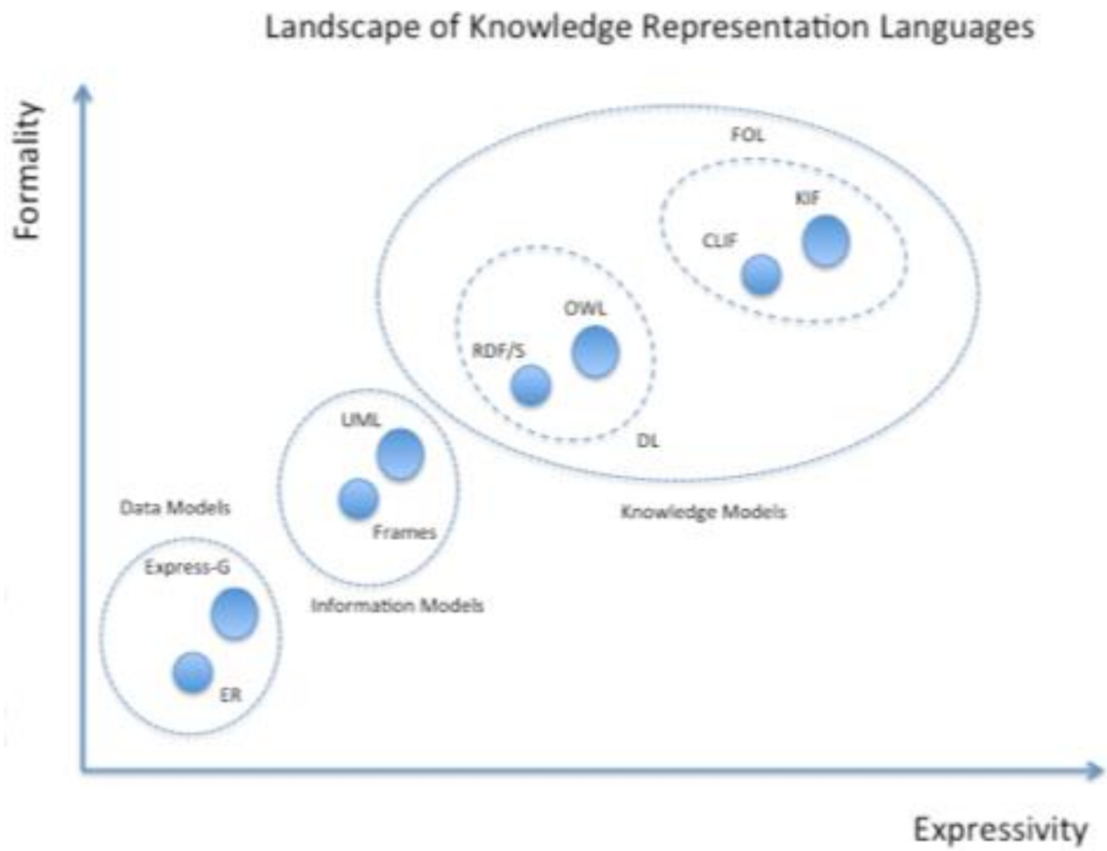
Underwood, T. (2012, April 7). Topic modeling made just simple enough. Retrieved from https://tedunderwood.com/2012/04/07/topic-modeling-made-just-simple-enough/

Vvlad. (2012, March 28). MATLAB: Simple TF-IDF implementation. Retrieved from dovgalecs.com: http://dovgalecs.com/blog/matlab-simple-tf-idf/

Wang, W., Barnaghi, P. M., & Bargiela, A. (2010, July). Probabilistic topic models for learning terminological ontologies. *IEEE Transactions on Knowledge and Data Engineering*, 22, 1028-1040. doi:10.1109/tkde.2009.122

Zikopoulous, P., & Eaton, C. (2011). *Understanding big data: Analytics for enterprise class Hadoop and Streaming Data*. McGraw-Hill Osborne Media.

Appendix A: Knowledge Representation Languages



The following tables provides a descriptive summary of each language shown above:

#	Description of Knowledge Representation Language
1	**Express-G** (Source: https://en.wikipedia.org/wiki/EXPRESS_(data_modeling_language)#EXPRESS-G) **EXPRESS** is a standard data modeling language for product data. EXPRESS is formalized in the ISO Standard for the Exchange of Product model STEP (ISO 10303), and standardized as ISO 10303-11. Data models formally define data objects and relationships among data objects for a domain of interest. Some typical applications of data models include supporting the development of databases and enabling the exchange of data for a particular area of interest. Data models are specified in a data modeling language. EXPRESS is a data modeling language defined in ISO 10303-11, the EXPRESS Language Reference Manual. An EXPRESS data model can be defined in two ways, textually and graphically. For formal verification and as input for tools such as SDAI the textual representation within an ASCII file is the most important one. The graphical representation on the other hand is often more suitable for human use such as explanation and tutorials. The graphical representation, called EXPRESS-G, is not able to represent all details that can be formulated in the textual form.

	EXPRESS is similar to programming languages such as Pascal. Within a SCHEMA various datatypes can be defined together with structural constraints and algorithmic rules. A main feature of EXPRESS is the possibility to formally validate a population of datatypes - this is to check for all the structural and algorithmic rules. **EXPRESS-G** EXPRESS-G is a standard graphical notation for information models. It is a useful companion to the EXPRESS language for displaying entity and type definitions, relationships and cardinality. This graphical notation supports a subset of the EXPRESS language. One of the advantages of using EXPRESS-G over EXPRESS is that the structure of a data model can be presented in a more understandable manner. A disadvantage of EXPRESS-G is that complex constraints cannot be formally specified. The data model presented in figure could be used to specify the requirements of a database for an audio compact disc (CD) collection.
2	**ER: Entity-Relationship** (Source: https://en.wikipedia.org/wiki/Entity%E2%80%93relationship_model) An **entity–relationship model** (**ER model** for short) describes interrelated things of interest in a specific domain of knowledge. A basic ER model is composed of entity types (which classify the things of interest) and specifies relationships that can exist between instances of those entity types. In software engineering, an ER model is commonly formed to represent things that a business needs to remember in order to perform business processes. Consequently, the ER model becomes an abstract data model, that defines a data or information structure which can be implemented in a database, typically a relational database. Entity–relationship modeling was developed for database design by Peter Chen and published in a 1976 paper. However, variants of the idea existed previously. Some ER models show super and subtype entities connected by generalization-specialization relationships, and an ER model can be used also in the specification of domain-specific ontologies.
3	**UML: Unified Modeling Language** (Source: https://en.wikipedia.org/wiki/Unified_Modeling_Language) The **Unified Modeling Language** (**UML**) is a general-purpose, developmental, modeling language in the field of software engineering, that is intended to provide a standard way to visualize the design of a system. The creation of UML was originally motivated by the desire to standardize the disparate notational systems and approaches to software design. It was developed by Grady Booch, Ivar Jacobson and James Rumbaugh at Rational Software in 1994–1995, with further development led by them through 1996. In 1997 UML was adopted as a standard by the Object Management Group (OMG), and has been managed by this organization ever since. In 2005 UML was also published by the International Organization for Standardization (ISO) as an approved ISO standard. Since then the standard has been periodically revised to cover the latest revision of UML.

4	**Frames** (Source: https://en.wikipedia.org/wiki/Frame_language) A **frame language** is a technology used for knowledge representation in artificial intelligence. Frames are stored as ontologies of sets and subsets of the frame concepts. They are similar to class hierarchies in object-oriented languages although their fundamental design goals are different. Frames are focused on explicit and intuitive representation of knowledge whereas objects focus on encapsulation and information hiding. Frames originated in AI research and objects primarily in software engineering. However, in practice the techniques and capabilities of frame and object-oriented languages overlap significantly.
5	**DL: Description Logic** (Source: https://en.wikipedia.org/wiki/Description_logic) **Description logics** (**DL**) are a family of formal knowledge representation languages. Many DLs are more expressive than propositional logic but less expressive than first-order logic. In contrast to the latter, the core reasoning problems for DLs are (usually) decidable, and efficient decision procedures have been designed and implemented for these problems. There are general, spatial, temporal, spatiotemporal, and fuzzy descriptions logics, and each description logic features a different balance between DL expressivity and reasoning complexity by supporting different sets of mathematical constructors. DLs are used in artificial intelligence to describe and reason about the relevant concepts of an application domain (known as *terminological knowledge*). It is of particular importance in providing a logical formalism for ontologies and the Semantic Web: the Web Ontology Language (OWL) and its profile is based on DLs. The most notable application of DLs and OWL is in biomedical informatics where DL assists in the codification of biomedical knowledge.
6	**RDF** (Source: https://en.wikipedia.org/wiki/Resource_Description_Framework) The **Resource Description Framework** (**RDF**) is a family of World Wide Web Consortium (W3C) specifications originally designed as a metadata data model. It has come to be used as a general method for conceptual description or modeling of information that is implemented in web resources, using a variety of syntax notations and data serialization formats. It is also used in knowledge management applications. RDF was adopted as a W3C recommendation in 1999. The RDF 1.0 specification was published in 2004, the RDF 1.1 specification in 2014.
7	**RDFS** (Source: https://en.wikipedia.org/wiki/RDF_Schema) **RDF Schema** (**Resource Description Framework Schema**, variously abbreviated as **RDFS**, **RDF(S)**, **RDF-S**, or **RDF/S**) is a set of classes with certain properties using the RDF extensible knowledge representation data model, providing basic elements for the description of ontologies, otherwise called RDF vocabularies, intended to structure

	RDF resources. These resources can be saved in a triplestore to reach them with the query language SPARQL. The first version was published by the World-Wide Web Consortium (W3C) in April 1998, and the final W3C recommendation was released in February 2004. Many RDFS components are included in the more expressive Web Ontology Language (OWL).
8	**OWL: Web Ontology Language** (Source: https://en.wikipedia.org/wiki/Web_Ontology_Language) The **Web Ontology Language** (**OWL**) is a family of knowledge representation languages for authoring ontologies. Ontologies are a formal way to describe taxonomies and classification networks, essentially defining the structure of knowledge for various domains: the nouns representing classes of objects and the verbs representing relations between the objects. Ontologies resemble class hierarchies in object-oriented programming but there are several critical differences. Class hierarchies are meant to represent structures used in source code that evolve fairly slowly (typically monthly revisions) whereas ontologies are meant to represent information on the Internet and are expected to be evolving almost constantly. Similarly, ontologies are typically far more flexible as they are meant to represent information on the Internet coming from all sorts of heterogeneous data sources. Class hierarchies on the other hand are meant to be fairly static and rely on far less diverse and more structured sources of data such as corporate databases. The OWL languages are characterized by formal semantics. They are built upon the World Wide Web Consortium's (W3C) XML standard for objects called the Resource Description Framework (RDF). OWL and RDF have attracted significant academic, medical and commercial interest. In October 2007, a new W3C working group was started to extend OWL with several new features as proposed in the OWL 1.1 member submission. W3C announced the new version of OWL on 27 October 2009. This new version, called OWL 2, soon found its way into semantic editors such as Protégé and semantic reasoners such as Pellet, RacerPro, FaCT++ and HermiT. The OWL family contains many species, serializations, syntaxes and specifications with similar names. OWL and OWL2 are used to refer to the 2004 and 2009 specifications, respectively. Full species names will be used, including specification version (for example, OWL2 EL). When referring more generally, *OWL Family* will be used.
9	**FOL: First Order Logic** (Source: https://en.wikipedia.org/wiki/First-order_logic) **First-order logic**—also known as **first-order predicate calculus** and **predicate logic**—is a collection of formal systems used in mathematics, philosophy, linguistics, and computer science. First-order logic uses quantified variables over non-logical objects and allows the use of sentences that contain variables, so that rather than propositions such as *Socrates is a man* one can have expressions in the form "there exists X such that X is Socrates and X is a

	man" and *there exists* is a quantifier while *X* is a variable. This distinguishes it from propositional logic, which does not use quantifiers or relations. A theory about a topic is usually a first-order logic together with a specified domain of discourse over which the quantified variables range, finitely many functions from that domain to itself, finitely many predicates defined on that domain, and a set of axioms believed to hold for those things. Sometimes "theory" is understood in a more formal sense, which is just a set of sentences in first-order logic. The adjective "first-order" distinguishes first-order logic from higher-order logic in which there are predicates having predicates or functions as arguments, or in which one or both of predicate quantifiers or function quantifiers are permitted. In first-order theories, predicates are often associated with sets. In interpreted higher-order theories, predicates may be interpreted as sets of sets. There are many deductive systems for first-order logic which are both sound (all provable statements are true in all models) and complete (all statements which are true in all models are provable). Although the logical consequence relation is only semidecidable, much progress has been made in automated theorem proving in first-order logic. First-order logic also satisfies several metalogical theorems that make it amenable to analysis in proof theory, such as the Löwenheim–Skolem theoremand the compactness theorem. First-order logic is the standard for the formalization of mathematics into axioms and is studied in the foundations of mathematics. Peano arithmetic and Zermelo–Fraenkel set theory are axiomatizations of number theory and set theory, respectively, into first-order logic. No first-order theory, however, has the strength to uniquely describe a structure with an infinite domain, such as the natural numbers or the real line. Axioms systems that do fully describe these two structures (that is, categoricalaxiom systems) can be obtained in stronger logics such as second-order logic. The foundations of first-order logic were developed independently by Gottlob Frege and Charles Sanders Peirce. For a history of first-order logic and how it came to dominate formal logic, see José Ferreirós (2001).
10	**CLIF: Common Logic Interchange Format** (Source: https://en.wikipedia.org/wiki/Common_Logic) **Common Logic** (**CL**) is a framework for a family of logic languages, based on first-order logic, intended to facilitate the exchange and transmission of knowledge in computer-based systems. The CL definition permits and encourages the development of a variety of different syntactic forms, called *dialects*. A dialect may use any desired syntax, but it must be possible to demonstrate precisely how the concrete syntax of a dialect conforms to the abstract CL semantics, which are based on a model theoretic interpretation. Each dialect may be then treated as a formal language. Once syntactic conformance is established, a dialect gets the CL semantics for free, as they are specified relative to the abstract syntax only, and hence are inherited by any conformant dialect. In addition, all CL dialects are equivalent (i.e., can be mechanically translated to each other), although some may be more

	expressive than others. In general, a less expressive subset of CL may be translated to a more expressive version of CL, but the reverse translation is only defined on a subset of the larger language. **The ISO Standard** Common Logic is published by ISO as "ISO/IEC 24707:2007 - Information technology — Common Logic (CL): a framework for a family of logic-based languages". It is available for purchase from ISO's catalog, and is freely available from ISO's index of publicly available standards. The CL Standard includes specifications for three dialects, the **Common Logic Interchange Format** (**CLIF**) (Annex A), the **Conceptual Graph Interchange Format**(**CGIF**) (Annex B), and an XML-based notation for Common Logic (**XCL**) (Annex C). The semantics of these dialects are defined in the Standard by their translation to the abstract syntax and semantics of Common Logic. Many other logic-based languages could also be defined as subsets of CL by means of similar translations; among them are the RDF and OWL languages, which have been defined by the W3C. The ISO standard's development began in June 2003 under Working Group 2 (Metadata) of Sub-Committee 32 (Data Interchange) under ISO/IEC JTC1, and was completed in October 2007. A technical corrigendum, correcting some errors in the original standard, is being prepared at the time being.
11	**KIF: Knowledge Interchange Format** (Source: https://en.wikipedia.org/wiki/Knowledge_Interchange_Format) **Knowledge Interchange Format** (**KIF**) is a computer language designed to enable systems to share and re-use information from knowledge-based systems. KIF is similar to frame languages such as KL-One and LOOM but unlike such language its primary role is not intended as a framework for the expression or use of knowledge but rather for the interchange of knowledge between systems. The designers of KIF likened it to PostScript. PostScript was not designed primarily as a language to store and manipulate documents but rather as an interchange format for systems and devices to share documents. In the same way KIF is meant to facilitate sharing of knowledge across different systems that use different languages, formalisms, platforms, etc. KIF has a declarative semantics. It is meant to describe facts about the world rather than processes or procedures. Knowledge can be described as objects, functions, relations, and rules. It is a formal language, i.e., it can express arbitrary statements in first order logic and can support reasoners that can prove the consistency of a set of KIF statements. KIF also supports non-monotonic reasoning. KIF was created by Michael Genesereth, Richard Fikes and others participating in the DARPA knowledge Sharing Effort. Although the original KIF group intended to submit to a formal standards body, that did not occur. A later version called Common Logic has since been developed for submission to ISO and has been approved and published. A variant called SUO-KIF is the language in

	which the Suggested Upper Merged Ontology is written.

Appendix B: Example DSI Pre-Processed JSON File

Description

The JSON file below is a single example of a pre-processed DSI item from our corpus.

The following parameters were captured in a similar format for each corpus DSI:

- Raw Parameters (direct from source):
  - dsi_title: article title
  - dsi_location: url address of source
  - dsi_date: publication date (listed on site)
  - dsi_author: article author
  - dsi_content_sponsor: article professional affiliation / or venue (for speeches)
  - dsi_source: source of content (i.e., TabbFORUM)
  - dsi_filename: name of scraped JSON file from raw source
  - dsi_content: article body
- Transformed Parameters (from pre-processing script):
  - dsi_type: DSI type such as speech, opinion analysis article, news. Etc.
  - dsi_cov_reg: business rule to identify “regulatory” content based on key terms
  - dsi_cov_data: business rule to identify “data” content based on key terms
  - dsi_cov_ai: business rule to identify “artificial intelligence” content from terms
  - dsi_aggregated_content: concatenation of title and article body
  - dsi_aggregated_contentPP: pre-processed aggregated content such as removal of stop words, punctuation, spaces, etc. Represents text converted to DTM for modeling

Sample JSON File

{**"dsi_title":** "The Implications of Machine Learning in Finance",
"dsi_location": "www.tabbforum.com",
"dsi_date": "10 October 2017”,
"dsi_author": "Bloomberg Professional Services",
"dsi_content_sponsor": [],
"dsi_type": "TF_OpinionAnalysis",
"dsi_cov_data": true,
"dsi_cov_reg": false,
"dsi_cov_ai": true,
"dsi_source": "TabbForum",
"dsi_filename": "TF_article(archive)_Page2_2017_10_21_21.json",
"dsi_content": " Machine learning may not be in your firm\u2019s toolbox yet. In fact, according to a survey at Bloomberg\u2019s Buy-Side Week 2017 New York event, only 16% of firms have incorporated any kind of machine learning into their investment strategies. Meanwhile, the remainder is either researching ways to do it (24%), would like to learn about how to do it (26%), or hasn\u2019t even thought about doing it yet (32%). Yet if Bloomberg\u2019s head of Machine Learning Gary Kazanstev is right, machine learning is

coming to every firm soon enough. Despite being the buzzword\u00a0du jour\u00a0on Wall Street these days, machine learning is still fairly misunderstood. It is not artificial intelligence (AI) itself, but rather a form of it in which computers fed extremely large data sets are able to learn as changes in that data occur without being explicitly programmed to do so. The data is just one part of the approach, Kazanstev said during a panel at Buy-Side Week in June. What can be more challenging is making machine learning and data science a core capability among companies so that they instinctively take internal and external data sets and interpret it for patterns, risks, opportunities, and so on. And like all things tech, the space is evolving quickly. \u201cThe level of expertise in machine learning has risen rapidly,\u201d Kazanstev added. \u201cIt is shifting to engineers and quants as your counterparty in the discussion, not investing personnel.\u201d The data is shifting too, from structured data like prices or economic statistics to unstructured data mined from new sources of information, like GPS coordinates and social media. All of it is anchored on an increasing ability to bring tremendous computing power to bear for very little cost. [Related: \u201cThe Search for Alpha Reaches New Frontiers\u201d] \u201cThe key process at first was simple automation,\u201d Kazanstev explained. \u201cBut at this point, throw a dart at any investment process and someone, somewhere has automated every part of it.\u201d Now, that power is being directed at more subjective things. \u201cFour years ago, Twitter steams were being analyzed for simple binary interpretations of bullish or bearish,\u201d noted Mac Steele, Director of Product at Domino Data Lab. \u201cNow, it is much more complex. Five years ago, satellite image analysis would have taken three months and millions of dollars in capex; now, it takes a fraction of both.\u201d The cutting edge for machine learning applications is combining experience with statistical data to develop uses, so image processing in general is a hot topic, continued Steele. \u201cThere\u2019s talk that merger arb firms are even doing facial recognition to match who walks into target firms. This is the kind of activity going on now because it\u2019s no longer hard or expensive to do.\u201d The ability to crunch tremendous amounts of data is showing up in other areas. \u201cIn text analysis, we are figuring out how to determine whether a CEO is being evasive on a conference call,\u201d added Bloomberg\u2019s Kazanstev. \u201cAnd it\u2019s not just from audio \u2013 you can ascertain this from text now as well.\u201d In these silos, the data itself is less important than what the system does with the internal/external data it gets, and how it treats subsequent inputs, interpretations, and patterns. Iteration of the data, and the frequency with which it occurs, is becoming a primary lever, because each successive round makes the overall system smarter. For the buy side, these applications take two approaches, Kazanstev explained. \u201cWith humans, we are inverting the workflow from managers asking for things to pushing information to them based on their profile or behavior, stuff they would not even know to ask for. On the enterprise side, \u2018black box\u2019 consumption is differently optimized and involves human in-the-loop automation. All of this also provides feedback to a suite of learning algorithms, which all adjust accordingly in time for the next set of data.\u201d",

"dsi_aggregated_content": " Machine learning may not be in your firm\u2019s toolbox yet. In fact, according to a survey at Bloomberg\u2019s Buy-Side Week 2017 New York event, only 16% of firms have incorporated any kind of machine learning into their investment strategies. Meanwhile, the remainder is either researching ways to do it (24%), would like to learn about how to do it (26%), or hasn\u2019t even thought about doing it yet (32%). Yet if Bloomberg\u2019s head of Machine Learning Gary Kazanstev is right, machine learning is coming to every firm soon enough. Despite being the buzzword\u00a0du jour\u00a0on Wall

Street these days, machine learning is still fairly misunderstood. It is not artificial intelligence (AI) itself, but rather a form of it in which computers fed extremely large data sets are able to learn as changes in that data occur without being explicitly programmed to do so. The data is just one part of the approach, Kazanstev said during a panel at Buy-Side Week in June. What can be more challenging is making machine learning and data science a core capability among companies so that they instinctively take internal and external data sets and interpret it for patterns, risks, opportunities, and so on. And like all things tech, the space is evolving quickly. \u201cThe level of expertise in machine learning has risen rapidly,\u201d Kazanstev added. \u201cIt is shifting to engineers and quants as your counterparty in the discussion, not investing personnel.\u201d The data is shifting too, from structured data like prices or economic statistics to unstructured data mined from new sources of information, like GPS coordinates and social media. All of it is anchored on an increasing ability to bring tremendous computing power to bear for very little cost. [Related: \u201cThe Search for Alpha Reaches New Frontiers\u201d] \u201cThe key process at first was simple automation,\u201d Kazanstev explained. \u201cBut at this point, throw a dart at any investment process and someone, somewhere has automated every part of it.\u201d Now, that power is being directed at more subjective things. \u201cFour years ago, Twitter steams were being analyzed for simple binary interpretations of bullish or bearish,\u201d noted Mac Steele, Director of Product at Domino Data Lab. \u201cNow, it is much more complex. Five years ago, satellite image analysis would have taken three months and millions of dollars in capex; now, it takes a fraction of both.\u201d The cutting edge for machine learning applications is combining experience with statistical data to develop uses, so image processing in general is a hot topic, continued Steele. \u201cThere\u2019s talk that merger arb firms are even doing facial recognition to match who walks into target firms. This is the kind of activity going on now because it\u2019s no longer hard or expensive to do.\u201d The ability to crunch tremendous amounts of data is showing up in other areas. \u201cIn text analysis, we are figuring out how to determine whether a CEO is being evasive on a conference call,\u201d added Bloomberg\u2019s Kazanstev. \u201cAnd it\u2019s not just from audio \u2013 you can ascertain this from text now as well.\u201d In these silos, the data itself is less important than what the system does with the internal/external data it gets, and how it treats subsequent inputs, interpretations, and patterns. Iteration of the data, and the frequency with which it occurs, is becoming a primary lever, because each successive round makes the overall system smarter. For the buy side, these applications take two approaches, Kazanstev explained. \u201cWith humans, we are inverting the workflow from managers asking for things to pushing information to them based on their profile or behavior, stuff they would not even know to ask for. On the enterprise side, \u2018black box\u2019 consumption is differently optimized and involves human in-the-loop automation. All of this also provides feedback to a suite of learning algorithms, which all adjust accordingly in time for the next set of data.\u201d",

"dsi_aggregated_contentPP": " machine learning may be firm toolbox yet fact according survey bloomberg buy side week new york event only firms have incorporated any kind machine learning into investment strategies meanwhile remainder either researching ways do would like learn do hasn even thought doing yet yet bloomberg head machine learning gary kazanstev right machine learning coming every firm soon enough despite being buzzword du jour wallstreet these days machine learning still fairly misunderstood artificialintelligence ai itself but rather form which computers fed extremely large data sets able learn changes data occur without being explicitly programmed do so data just one part approach kazanstev said during panel buy side

week june can be more challenging making machine learning data science core capability among companies so instinctively take internal external data sets interpret patterns risks opportunities so like all things tech space evolving quickly level expertise machine learning has risen rapidly kazanstev added shifting engineers quants counterparty discussion investing personnel data shifting too structured data like prices economic statistics unstructured data mined new sources information like gps coordinates social media all anchored increasing ability bring tremendous computing power bear very little cost related search alpha reaches new frontiers key process first was simple automation kazanstev explained but point throw dart any investment process someone somewhere has automated every part now power being directed more subjective things four years ago twitter steams being analyzed simple binary interpretations bullish bearish noted mac steele director product domino data lab now much more complex five years ago satellite image analysis would have taken three months millions dollars capex now takes fraction both cutting edge machine learning applications combining experience statistical data develop uses so image processing general hot topic continued steele there talk merger arb firms even doing facial recognition match walks into target firms kind activity going now because no longer hard expensive do ability crunch tremendous amounts data showing up other areas text analysis figuring out determine whether ceo being evasive conference call added bloomberg kazanstev just audio can ascertain text now well these silos data itself less important than system does internal external data gets treats subsequent inputs interpretations patterns iteration data frequency which occurs becoming primary lever because each successive round makes overall system smarter buy side these applications take two approaches kazanstev explained humans inverting workflow managers asking things pushing information them based profile behavior stuff would even know ask enterprise side black box consumption differently optimized involves human loop automation all also provides feedback suite learning algorithms which all adjust accordingly time next set data "
}

Appendix C: Example Web Scraper

```
# prepare for Python version 3x features and functions
from __future__ import division, print_function

# each spider class gives code for crawing and scraping
import scrapy  # object-oriented framework for crawling and scraping
from scrapy_application.items import MyItem  # item class
from scrapy.spiders import CrawlSpider, Rule
from scrapy.linkextractors import LinkExtractor
from scrapy.selector import Selector

# spider subclass inherits from BaseSpider
# this spider is designed to crawl just one website
class MySpider(CrawlSpider):
    name = "FRB_Speeches_2017"  # unique identifier for the spider
    allowed_domains = ['federalreserve.gov']  # limits the crawl to this domain list

    import pandas as pd
    import numpy as np

    # Read list of 2017 FRB speech links for scraping
    df = pd.read_csv('C:\\FRBLinks_2017_Speeches.csv') # csv file with links to each speach page url
    FRB_Speeches_2017 = df["ArticleLink"]
    FRB_Speeches_2017_List = list(FRB_Speeches_2017)
    start_urls = FRB_Speeches_2017_List

    # define the parsing method for the spider
    def parse(self, response):
        sel = Selector(response)
        divs = sel.xpath('//div[@class="col-xs-12 col-sm-8 col-md-8"]')  # identify all <div> nodes

        results = []  # initialize list
        this_item = MyItem()  # use this item class
        this_item['speech_content'] = divs.xpath('.//p//text()').extract()
        divs2 = sel.xpath('//div[@class="heading col-xs-12 col-sm-8 col-md-8"]')
        this_item['speech_date'] = divs2.xpath('.//p[@class="article__time"]//text()').extract()
        this_item['speech_title'] = divs2.xpath('.//h3[@class="title"]//text()').extract()
        this_item['speech_speaker'] = divs2.xpath('.//p[@class="speaker"]//text()').extract()
        this_item['speech_location'] = divs2.xpath('.//p[@class="location"]//text()').extract()

        results.append(this_item)  # add to the results list
        return results
```

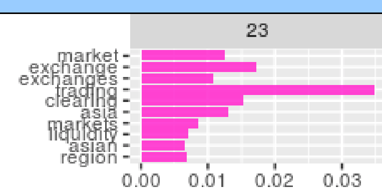
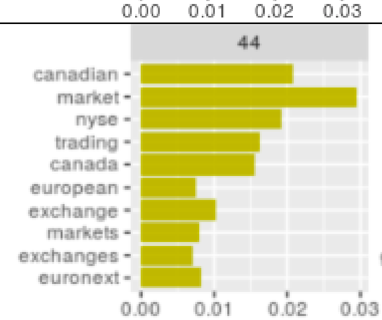
Appendix D: Stop Words List

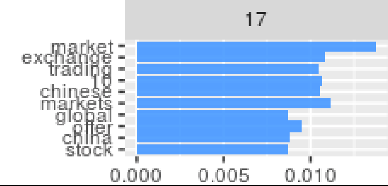
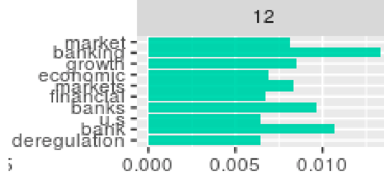
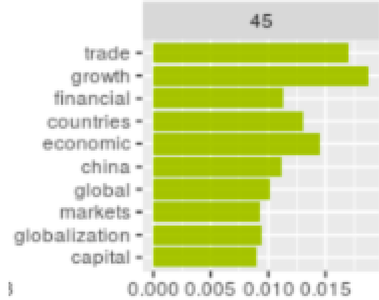
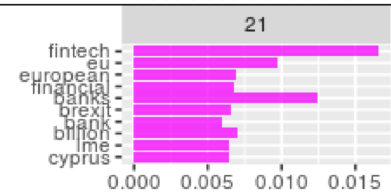
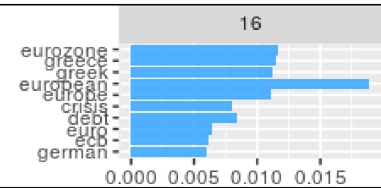
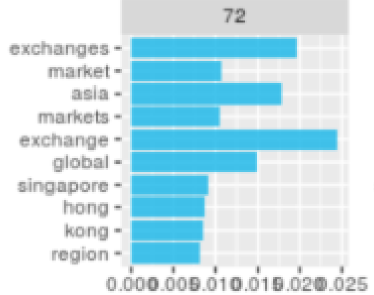
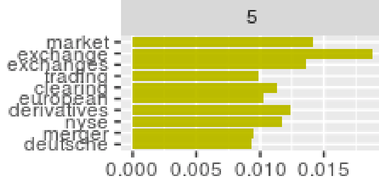
"a", "a's", "able", "about", "above", "according", "accordingly", "across", "actually", "after", "afterwards", "again", "against", "ain't", "all", "allow", "allows", "almost", "alone", "along", "already", "also", "although", "always", "am", "among", "amongst", "an", "and", "another", "any", "anybody", "anyhow", "anyone", "anything", "anyway", "anyways", "anywhere", "apart", "appear", "appreciate", "appropriate", "are", "aren't", "around", "as", "aside", "ask", "asking", "associated", "at", "available", "away", "awfully", "b", "be", "became", "because", "become", "becomes", "becoming", "been", "before", "beforehand", "behind", "being", "believe", "below", "beside", "besides", "best", "better", "between", "beyond", "both", "brief", "but", "by", "c", "c'mon", "c's", "came", "can", "can't", "cannot", "cant", "cause", "causes", "certain", "certainly", "changes", "clearly", "co", "com", "come", "comes", "concerning", "consequently", "consider", "considering", "contain", "containing", "contains", "corresponding", "could", "couldn't", "course", "currently", "d", "definitely", "described", "despite", "did", "didn't", "different", "do", "does", "doesn't", "doing", "don't", "done", "down", "downwards", "during", "e", "each", "edu", "eg", "eight", "either", "else", "elsewhere", "enough", "entirely", "especially", "et", "etc", "even", "ever", "every", "everybody", "everyone", "everything", "everywhere", "ex", "exactly", "example", "except", "f", "far", "few", "fifth", "first", "five", "followed", "following", "follows", "for", "former", "formerly", "forth", "four", "from", "further", "furthermore", "g", "get", "gets", "getting", "given", "gives", "go", "goes", "going", "gone", "got", "gotten", "greetings", "h", "had", "hadn't", "happens", "hardly", "has", "hasn't", "have", "haven't", "having", "he", "he's", "hello", "help", "hence", "her", "here", "here's", "hereafter", "hereby", "herein", "hereupon", "hers", "herself", "hi", "him", "himself", "his", "hither", "hopefully", "how", "howbeit", "however", "i", "i'd", "i'll", "i'm", "i've", "ie", "if", "ignored", "immediate", "in", "inasmuch", "inc", "indeed", "indicate", "indicated", "indicates", "inner", "insofar", "instead", "into", "inward", "is", "isn't", "it", "it'd", "it'll", "it's", "its", "itself", "j", "just", "k", "keep", "keeps", "kept", "know", "knows", "known", "l", "last", "lately", "later", "latter", "latterly", "least", "less", "lest", "let", "let's", "like", "liked", "likely", "little", "look", "looking", "looks", "ltd", "m", "mainly", "many", "may", "maybe", "me", "mean", "meanwhile", "merely", "might", "more", "moreover", "most", "mostly", "much", "must", "my", "myself", "n", "name", "namely", "nd", "near", "nearly", "necessary", "need", "needs", "neither", "never", "nevertheless", "new", "next", "nine", "no", "nobody", "non", "none", "noone", "nor", "normally", "not", "nothing", "novel", "now", "nowhere", "o", "obviously", "of", "off", "often", "oh", "ok", "okay", "old", "on", "once", "one", "ones", "only", "onto", "or", "other", "others", "otherwise", "ought", "our", "ours", "ourselves", "out", "outside", "over", "overall", "own", "p", "particular", "particularly", "per", "perhaps", "placed", "please", "plus", "possible", "presumably", "probably", "provides", "q", "que", "quite", "qv", "r", "rather", "rd", "re", "really", "reasonably", "regarding", "regardless", "regards", "relatively", "respectively", "right", "s", "said", "same", "saw", "say", "saying", "says", "second", "secondly", "see", "seeing", "seem", "seemed", "seeming", "seems", "seen", "self", "selves", "sensible", "sent", "serious", "seriously", "seven", "several", "shall", "she", "should", "shouldn't", "since", "six", "so", "some", "somebody", "somehow", "someone", "something", "sometime", "sometimes", "somewhat", "somewhere", "soon", "sorry", "specified", "specify", "specifying", "still", "sub", "such", "sup", "sure", "t", "t's", "take", "taken", "tell", "tends", "th", "than", "thank", "thanks", "thanx", "that", "that's", "thats", "the", "their", "theirs", "them", "themselves", "then", "thence", "there", "there's", "thereafter", "thereby", "therefore", "therein", "theres", "thereupon", "these", "they", "they'd", "they'll", "they're", "they've", "think", "third", "this", "thorough", "thoroughly", "those", "though", "three", "through", "throughout", "thru", "thus", "to", "together", "too", "took", "toward", "towards", "tried", "tries", "truly", "try", "trying", "twice", "two", "u", "un", "under", "unfortunately", "unless", "unlikely", "until", "unto", "up", "upon", "us", "use", "used", "useful", "uses", "using", "usually", "uucp", "v", "value", "various", "very", "via", "viz", "vs", "w", "want", "wants", "was", "wasn't", "way", "we", "we'd", "we'll", "we're", "we've", "welcome", "well", "went", "were", "weren't", "what", "what's", "whatever", "when", "whence", "whenever", "where", "where's", "whereafter", "whereas", "whereby", "wherein", "whereupon", "wherever", "whether", "which", "while", "whither", "who", "who's", "whoever", "whole", "whom", "whose", "why", "will", "willing", "wish", "with", "within", "without", "won't", "wonder", "would", "would", "wouldn't", "x", "y", "yes", "yet", "you", "you'd", "you'll", "you're", "you've", "your", "yours", "yourself", "yourselves", "z", "zero", "i", "me", "my", "myself", "we", "our", "ours", "ourselves", "you", "your", "yours", "yourself", "yourselves", "he", "him", "his", "himself", "she", "her", "hers", "herself", "it", "its", "itself", "they", "them", "their", "theirs", "themselves", "what", "which", "who", "whom", "this", "that", "these", "those", "am", "is", "are", "was", "were", "be", "been", "being", "have", "has", "had", "having", "do", "does", "did", "doing", "would", "should", "could", "ought", "i'm", "you're", "he's", "she's", "it's", "we're", "they're", "i've", "you've", "we've", "they've", "i'd", "you'd", "he'd", "she'd", "we'd", "they'd", "i'll", "you'll", "he'll", "she'll", "we'll", "they'll", "isn't", "aren't", "wasn't", "weren't", "hasn't", "haven't", "hadn't", "doesn't", "don't", "didn't", "won't", "wouldn't", "shan't", "shouldn't", "can't", "cannot", "couldn't", "mustn't", "let's", "that's", "who's", "what's", "here's", "there's", "when's", "where's", "why's", "how's", "a", "an", "the", "and", "but", "if", "or", "because", "as", "until", "while", "of", "at", "by", "for", "with", "about", "against", "between", "into", "through", "during", "before", "after", "above", "below", "to", "from", "up", "down", "in", "out", "on", "off", "over", "under", "again", "further", "then", "once", "here", "there", "when", "where", "why", "how", "all", "any", "both", "each", "few", "more", "most", "other", "some", "such", "no", "nor", "not", "only", "own", "same", "so", "than", "too", "very", "a", "about", "above", "across", "after", "again", "against", "all", "almost", "alone", "along", "already", "also", "although", "always", "among", "an", "and", "another", "any", "anybody", "anyone", "anything", "anywhere", "are", "area", "areas", "around", "as", "ask", "asked", "asking", "asks", "at", "away", "back", "backed", "backing", "backs", "be", "became", "because", "become", "becomes", "been", "before", "began", "behind", "being", "beings", "best", "better", "between", "big", "both", "but", "by", "came", "can", "cannot", "case", "cases", "certain", "certainly", "clear", "clearly", "come", "could", "did", "differ", "different", "differently", "do", "does", "done", "down", "down", "downed", "downing", "downs", "during", "each", "early", "either", "end", "ended", "ending", "ends", "enough", "even", "evenly", "ever", "every", "everybody", "everyone", "everything", "everywhere", "face", "faces", "fact", "facts", "far", "felt", "few", "find", "finds", "first", "for", "four", "from", "full", "fully", "further", "furthered", "furthering", "furthers", "gave", "general", "generally", "get", "gets", "give", "given", "gives", "go", "going", "good", "goods", "got", "great", "greater", "greatest", "group", "grouped", "grouping", "groups", "had", "has", "have", "having", "he", "her", "here", "herself", "high",

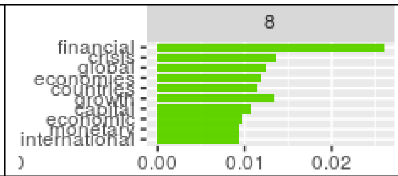
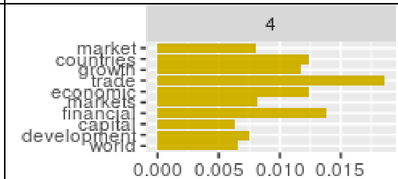
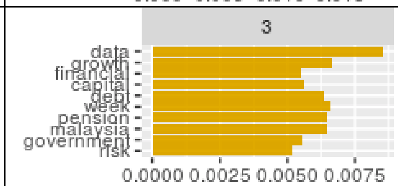
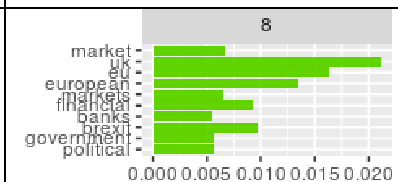
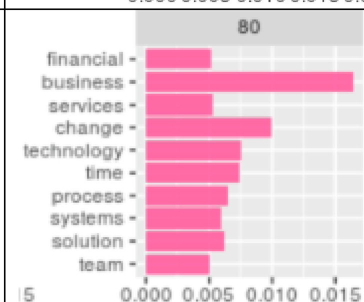
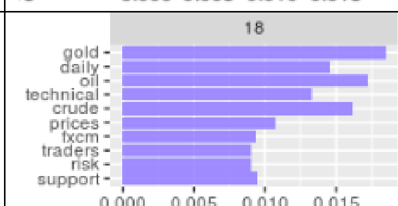
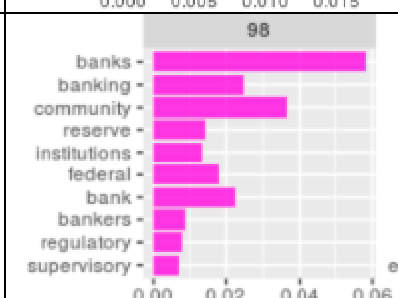
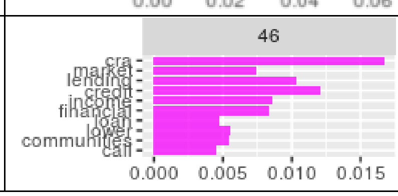
"high", "high", "higher", "highest", "him", "himself", "his", "how", "however", "i", "if", "important", "in", "interest", "interested", "interesting", "interests", "into", "is", "it", "its", "itself", "just", "keep", "keeps", "kind", "knew", "know", "known", "knows", "large", "largely", "last", "later", "latest", "least", "less", "let", "lets", "like", "likely", "long", "longer", "longest", "made", "make", "making", "man", "many", "may", "me", "member", "members", "men", "might", "more", "most", "mostly", "mr", "mrs", "much", "must", "my", "myself", "necessary", "need", "needed", "needing", "needs", "never", "new", "new", "newer", "newest", "next", "no", "nobody", "non", "noone", "not", "nothing", "now", "nowhere", "number", "numbers", "of", "off", "often", "old", "older", "oldest", "on", "once", "one", "only", "open", "opened", "opening", "opens", "or", "order", "ordered", "ordering", "orders", "other", "others", "our", "out", "over", "part", "parted", "parting", "parts", "per", "perhaps", "place", "places", "point", "pointed", "pointing", "points", "possible", "present", "presented", "presenting", "presents", "problem", "problems", "put", "puts", "quite", "rather", "really", "right", "right", "room", "rooms", "said", "same", "saw", "say", "says", "second", "seconds", "see", "seem", "seemed", "seeming", "seems", "sees", "several", "shall", "she", "should", "show", "showed", "showing", "shows", "side", "sides", "since", "small", "smaller", "smallest", "some", "somebody", "someone", "something", "somewhere", "state", "states", "still", "still", "such", "sure", "take", "taken", "than", "that", "the", "their", "them", "then", "there", "therefore", "these", "they", "thing", "things", "think", "thinks", "this", "those", "though", "thought", "thoughts", "three", "through", "thus", "to", "today", "together", "too", "took", "toward", "turn", "turned", "turning", "turns", "two", "under", "until", "up", "upon", "us", "use", "used", "uses", "very", "want", "wanted", "wanting", "wants", "was", "way", "ways", "we", "well", "wells", "went", "were", "what", "when", "where", "whether", "which", "while", "who", "whole", "whose", "why", "will", "with", "within", "without", "work", "worked", "working", "works", "would", "year", "years", "yet", "you", "young", "younger", "youngest", "your", "yours"

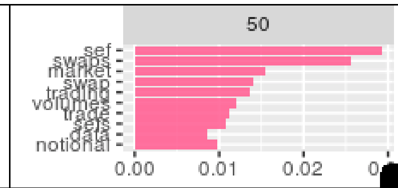
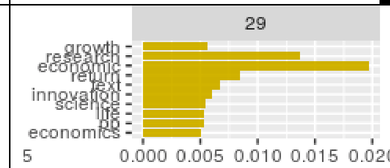
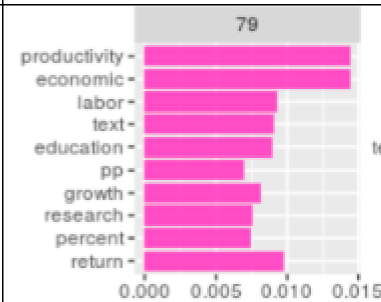
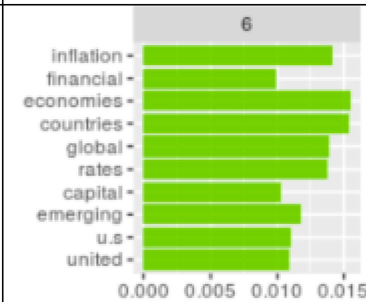
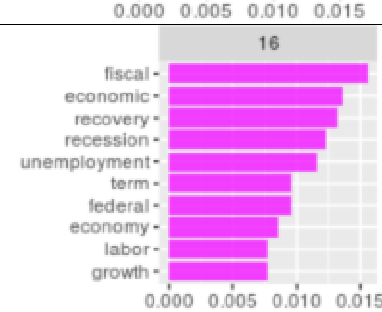
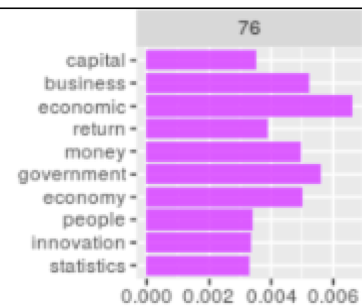
Appendix E: Financial Markets LDA Topics

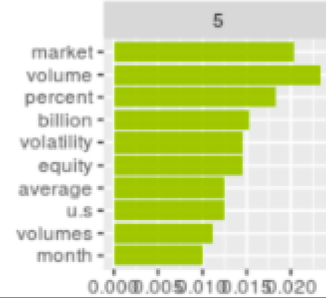
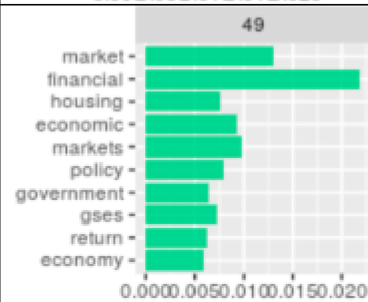
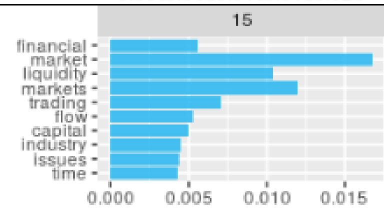
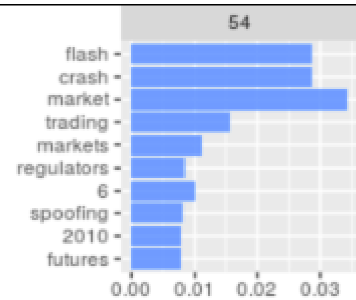
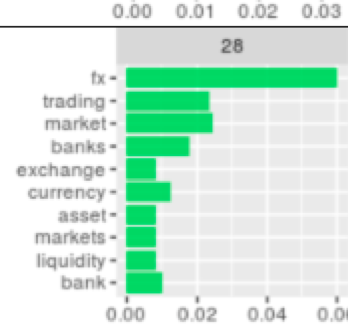
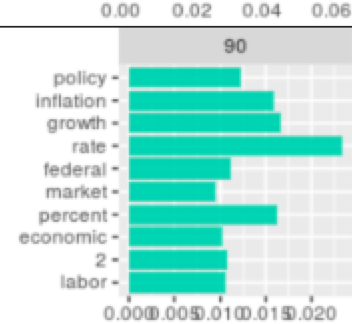
The following table represents detailed results from the LDA topic modeling analysis of our corpus. Topic models were generated from multiple passes, and multiple iterations within some passes. The first pass ('Pass 1') modeled the full corpus of all DSI items with a *K* value of 100. Topics generated by the first pass were then assigned to one of thirteen categories from our level-one ontology. Then, a second pass ('Pass 2') was modeled for each subset of the full corpus filtered to include only the DSIs related to the topics in each category. The intent of Pass 2 was to generate a more granular set of topics within each category. For simplicity, several iterations of Pass 2 were run varying the value for *K* topics. For categories with fewer DSIs, a value of 25 was appropriate for *K*. However, for larger categories, a value for 50 or 75 was appropriate for *K*. The process for selecting *K* in Pass 2 was subjective based on visual inspection of LDA output to produce meaningful results. The final set of 302 topics that were selected from the LDA modelling results are illustrated in the table below, showing the top terms as well as the *beta* value for each term. The color is for visual purposes only, and does not signify any meaning. Finally, the human-defined label and associated ontology nodes are listed to illustrate how our topic models fuse with our financial industry domain ontology.

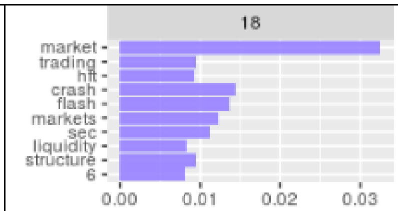
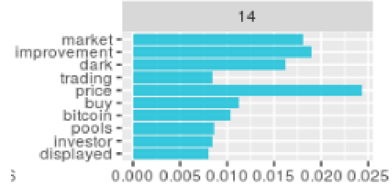
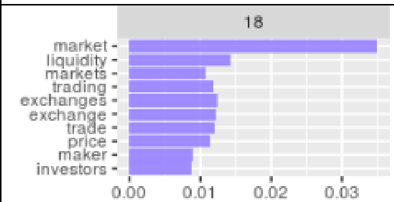
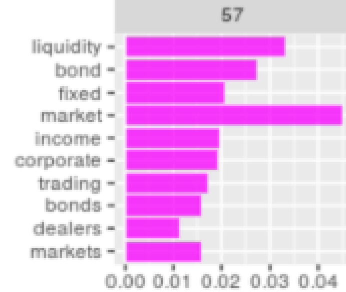
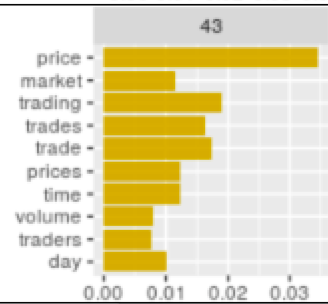
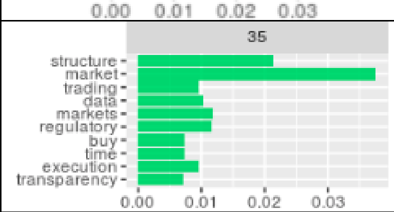
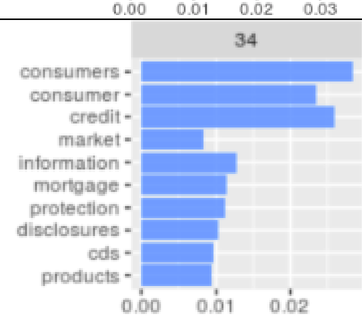
#	Final Category	Model Orchestration	Ontology Node Connection	Label	Topic Model Output
1	01 – Global Markets	Topic: 23 Pass: 2 Iteration: 1	(1) Exchanges & Market Infrastructure; (2) Asian Markets; (3) Trading Markets	Asian Trading Markets / Exchanges	
2	01 – Global Markets	Topic: 44 Pass: 1 Iteration: 1	(1) Canadian Markets; (2) Financial Markets;	Canadian Markets	

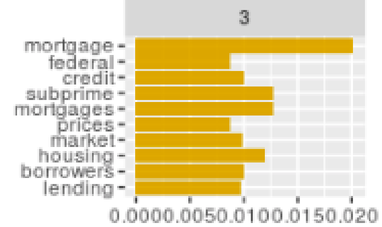
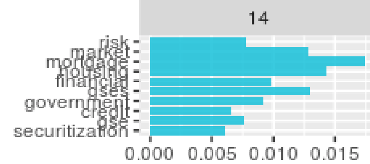
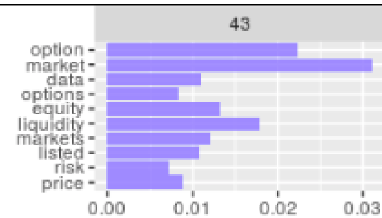
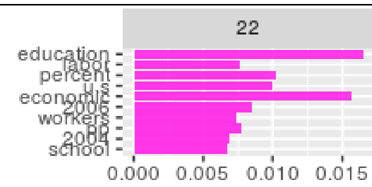
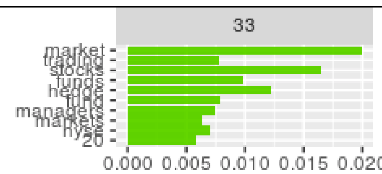
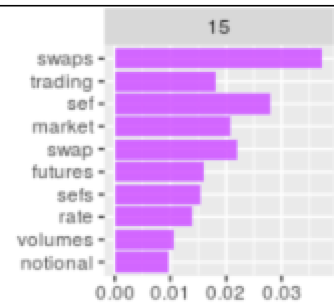
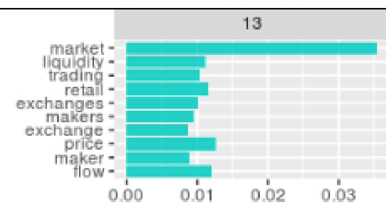
3	01 – Global Markets	Topic: 17 Pass: 2 Iteration: 1	(1) Chinese Markets; (2) Equity Markets;	Chinese Stock Market	
4	01 – Global Markets	Topic: 12 Pass: 2 Iteration: 1	(1) Financial Markets; (2) Market Oversight (Regulation & Policy);	Economic Growth - Financial Market Deregulation	
5	01 – Global Markets	Topic: 45 Pass: 1 Iteration: 1	(1) Financial Markets; (2) Global Markets;	Economic Growth - Globalization	
6	01 – Global Markets	Topic: 21 Pass: 2 Iteration: 1	(1) European Markets; (2) Systems & Technology – Fintech;	European Fintech	
7	01 – Global Markets	Topic: 16 Pass: 2 Iteration: 1	(2) Greek Markets; (3) Financial Markets;	Eurozone - Greek Debt Crisis	
8	01 – Global Markets	Topic: 72 Pass: 1 Iteration: 1	(1) Exchanges & Market Infrastructure; (2) Asian Markets; (3) Trading Markets;	Global Exchanges - Asia Region	
9	01 – Global Markets	Topic: 5 Pass: 2 Iteration: 1	(1) Exchanges & Market Infrastructure; (2) US Markets; (3) EU Markets; (4) Trading Markets;	Global Exchanges - European & US	

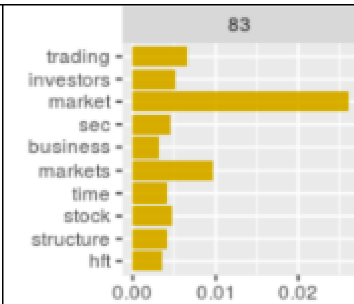
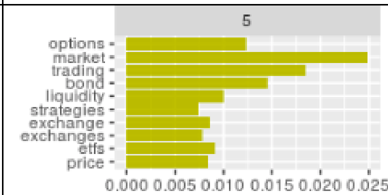
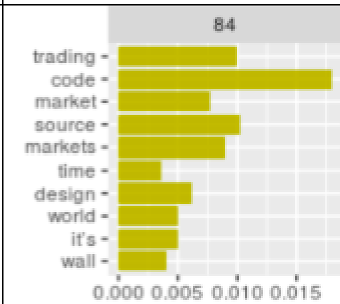
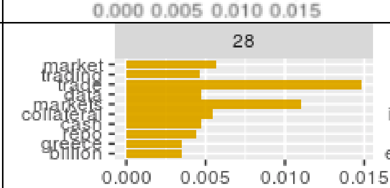
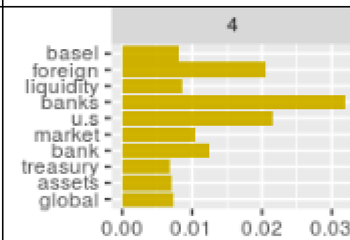
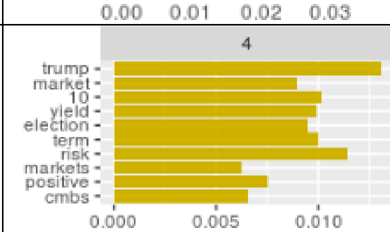
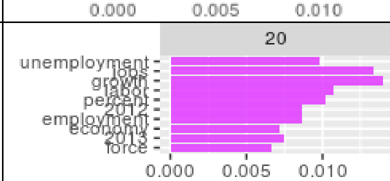
10	01 – Global Markets	Topic: 8 Pass: 2 Iteration: 2	(1) Financial Markets; (2) Global Markets;	Global Financial Crisis	
11	01 – Global Markets	Topic: 4 Pass: 2 Iteration: 1	(1) Financial Markets; (2) Global Markets;	Global Trade	
12	01 – Global Markets	Topic: 3 Pass: 2 Iteration: 1	(1) Financial Markets; (2) Malaysian Markets;	Malaysian Market Growth	
13	01 – Global Markets	Topic: 8 Pass: 2 Iteration: 1	(1) UK Markets; (2) Market Oversight (Regulation & Policy);	UK Markets - Brexit	
14	02 – U.S. Markets	Topic: 80 Pass: 1 Iteration: 1	(1) Financial Markets;	Business Environment Change	
15	02 – U.S. Markets	Topic: 18 Pass: 2 Iteration: 1	(1) Financial Markets; (2) Commodities Markets;	Commodities Prices	
16	02 – U.S. Markets	Topic: 98 Pass: 1 Iteration: 1	(1) Community Banking;	Community Banking	
17	02 – U.S. Markets	Topic: 46 Pass: 2 Iteration: 2	(1) Banking & Lending; (2) Borrower Level;	CRA (Low Credit Lending)	

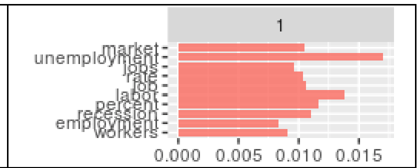
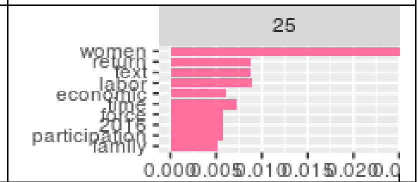
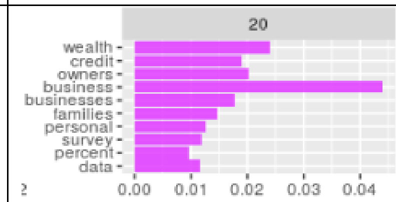
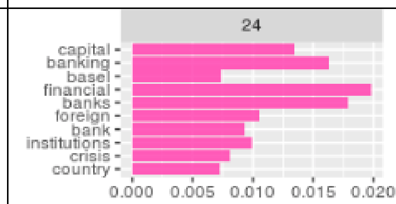
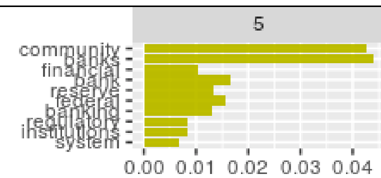
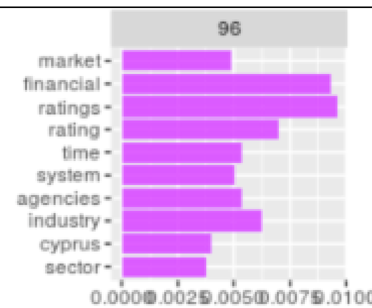
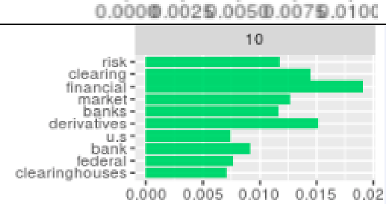
18	02 – U.S. Markets	Topic: 50 Pass: 2 Iteration: 2	(1) Derivatives Market;	Credit Derivatives Market	
19	02 – U.S. Markets	Topic: 29 Pass: 2 Iteration: 2	(1) Financial Markets;	Economic Growth - Innovation	
20	02 – U.S. Markets	Topic: 79 Pass: 1 Iteration: 1	(1) Financial Markets;	Economic Growth	
21	02 – U.S. Markets	Topic: 6 Pass: 1 Iteration: 1	(1) Financial Markets;	Economic Growth - Inflation	
22	02 – U.S. Markets	Topic: 16 Pass: 1 Iteration: 1	(1) Financial Markets;	Economic Recovery	
23	02 – U.S. Markets	Topic: 76 Pass: 1 Iteration: 1	(1) Financial Markets;	Economy Drivers	

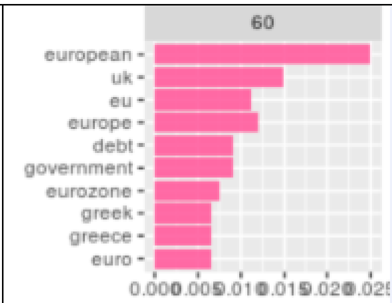
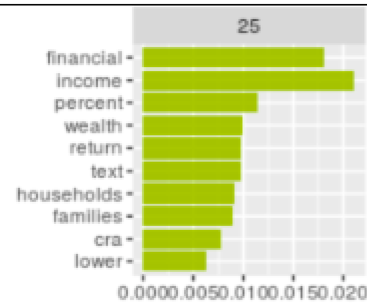
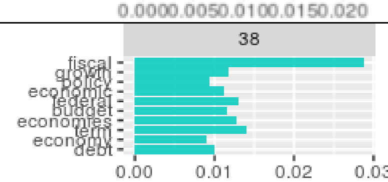
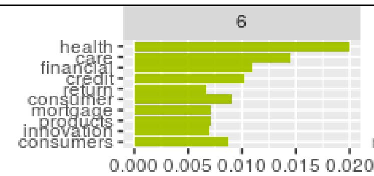
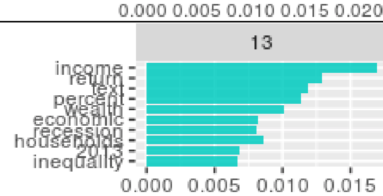
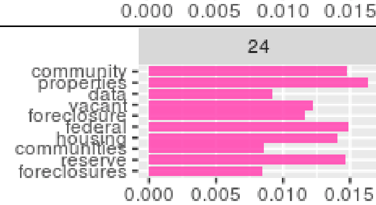
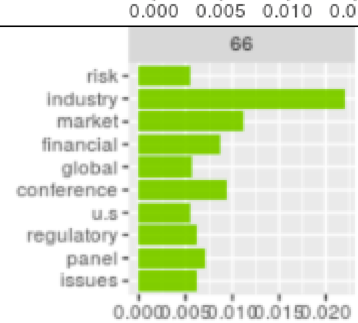
24	02 – U.S. Markets	Topic: 5 Pass: 1 Iteration: 1	(1) Financial Markets; (2) Equity Markets;	Equity Market Volume	
25	02 – U.S. Markets	Topic: 49 Pass: 1 Iteration: 1	(1) Financial Markets;	Financial Markets	
26	02 – U.S. Markets	Topic: 15 Pass: 2 Iteration: 1	(1) Trading Markets;	Financial Markets - Flow Trading	
27	02 – U.S. Markets	Topic: 54 Pass: 1 Iteration: 1	(1) Equity Markets;	Flash Crash	
28	02 – U.S. Markets	Topic: 28 Pass: 1 Iteration: 1	(1) FX Markets;	FX Market	
29	02 – U.S. Markets	Topic: 90 Pass: 1 Iteration: 1	(1) Financial Markets;	GDP Growth Rate	

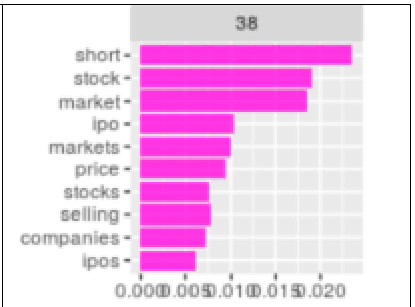
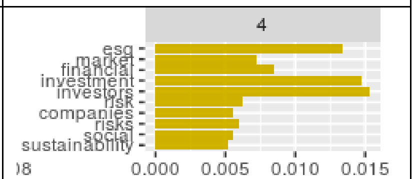
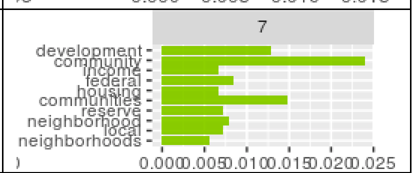
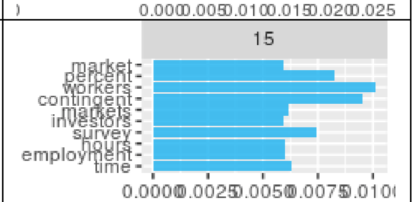
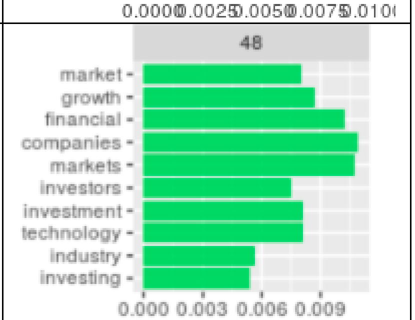
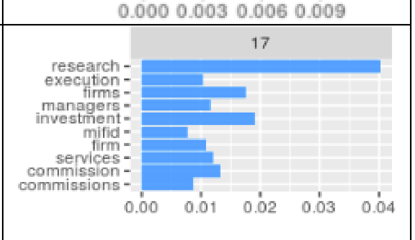
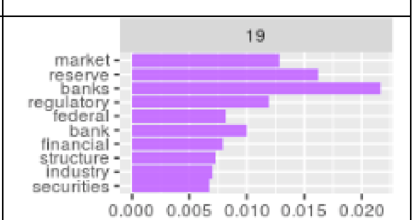
10	02 – U.S. Markets	Topic: 18 Pass: 2 Iteration: 1	(1) Equity Markets; (2) High Frequency Trading	High Frequency Trading - Markets	
11	02 – U.S. Markets	Topic: 14 Pass: 2 Iteration: 1	(1) Financial Markets;	Market Improvement - Price	
12	02 – U.S. Markets	Topic: 18 Pass: 2 Iteration: 1	(1) Financial Markets; (2) Liquidity Risk;	Market Liquidity	
13	02 – U.S. Markets	Topic: 57 Pass: 1 Iteration: 1	(1) Financial Markets; (2) Liquidity Risk;	Market Liquidity - Fixed Income	
14	02 – U.S. Markets	Topic: 43 Pass: 1 Iteration: 1	(1) Financial Markets;	Market Prices	
15	02 – U.S. Markets	Topic: 35 Pass: 2 Iteration: 1	(1) Financial Markets;	Market Structure	
16	02 – U.S. Markets	Topic: 34 Pass: 1 Iteration: 1	(1) Banking & Lending; (2) Borrower Level;	Mortgage Credit Protection	

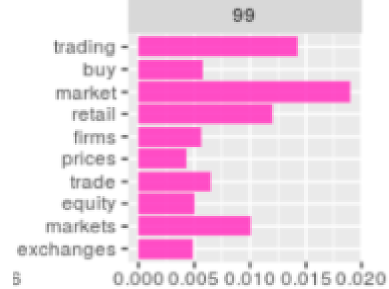
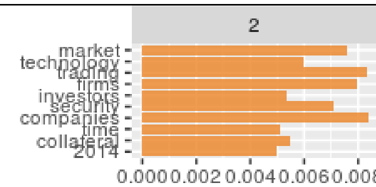
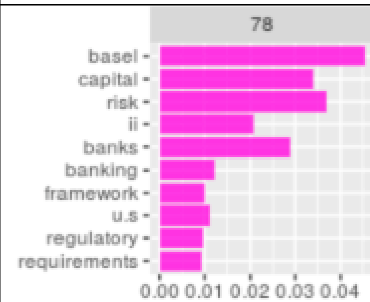
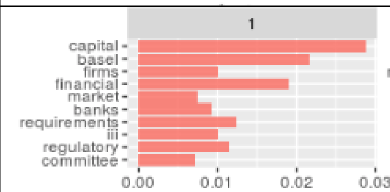
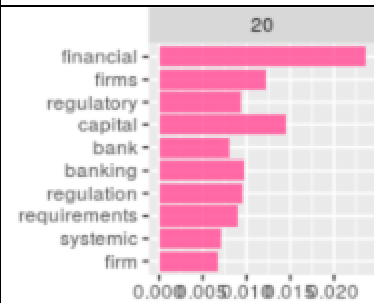
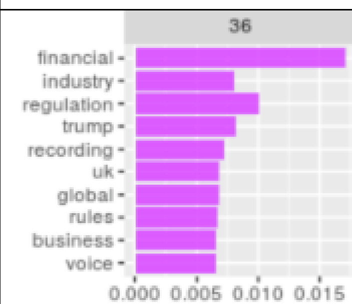
17	02 – U.S. Markets	Topic: 3 Pass: 2 Iteration: 2	(1) Banking & Lending; (2) Borrower Level;	Mortgage Market (Lending)	
18	02 – U.S. Markets	Topic: 14 Pass: 2 Iteration: 2	(1) Banking & Lending; (2) Derivatives Market;	Mortgage Securitization Market	
19	02 – U.S. Markets	Topic: 43 Pass: 2 Iteration: 1	(1) Derivatives Market;	Options Market	
20	02 – U.S. Markets	Topic: 22 Pass: 2 Iteration: 2	(1) Financial Markets;	Skilled Labor (Education + Jobs)	
21	02 – U.S. Markets	Topic: 33 Pass: 2 Iteration: 2	(1) Equity Markets;	Stock Market	
22	02 – U.S. Markets	Topic: 15 Pass: 1 Iteration: 1	(1) Derivatives Market;	Swaps & Futures Trading Markets	
23	02 – U.S. Markets	Topic: 13 Pass: 2 Iteration: 1	(1) Trading Markets;	Trading Market	

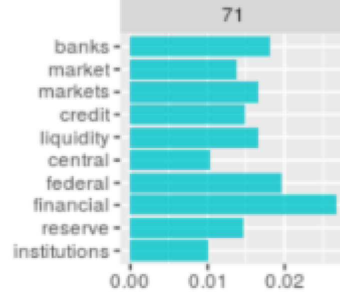
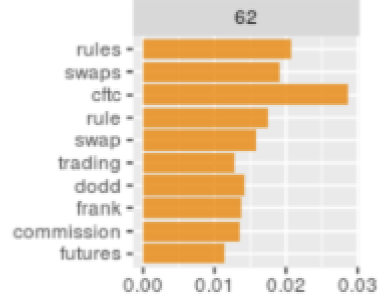
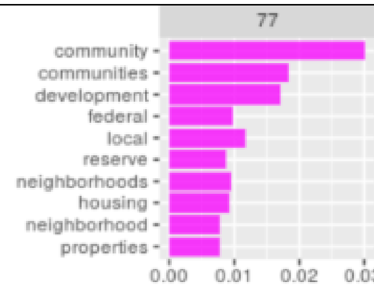
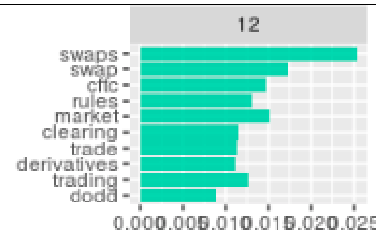
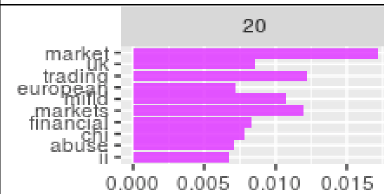
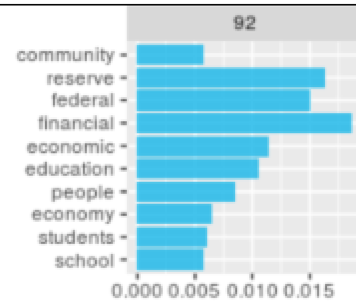
24	02 – U.S. Markets	Topic: 83 Pass: 1 Iteration: 1	(1) Equity Markets;	Trading Markets - Equity businesses	
25	02 – U.S. Markets	Topic: 5 Pass: 2 Iteration: 1	(1) Derivatives Market;	Trading Markets - Options	
26	02 – U.S. Markets	Topic: 84 Pass: 1 Iteration: 1	(1) Equity Markets; (2) High Frequency Trading;	Trading Markets - Programmatic Trading	
27	02 – U.S. Markets	Topic: 28 Pass: 2 Iteration: 2	(1) Trading Markets; (2) Fixed Income Markets; (3) Greek Markets;	Trading Markets (Repos Impacted by Greek Bonds Collateral)	
28	02 – U.S. Markets	Topic: 4 Pass: 2 Iteration: 2	(1) Fixed Income Markets; (2) Liquidity Risk;	Treasuries Market (Liquidity)	
29	02 – U.S. Markets	Topic: 4 Pass: 2 Iteration: 1	(1) Financial Markets; (2) Market Oversight (Regulation & Policy);	Trump Bump (election results on economy)	
30	02 – U.S. Markets	Topic: 20 Pass: 2 Iteration: 2	(1) Financial Markets;	Unemployment (Job Growth)	

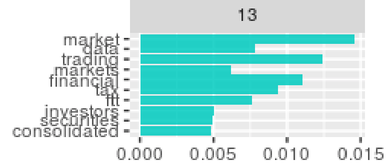
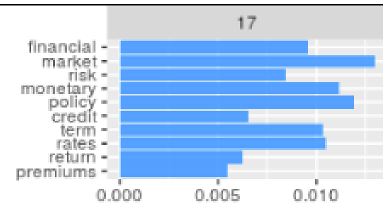
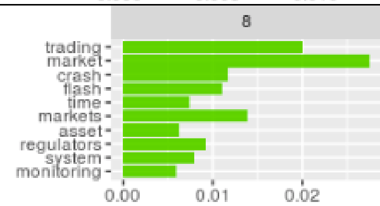
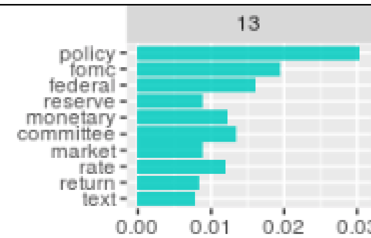
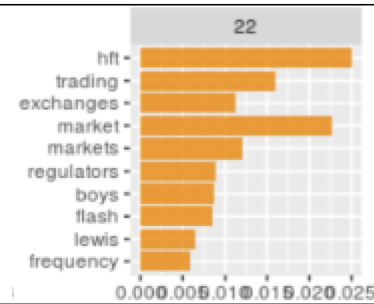
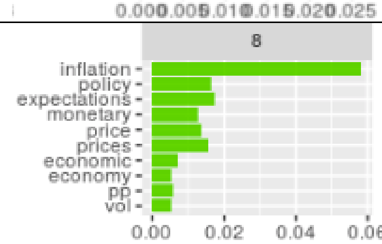
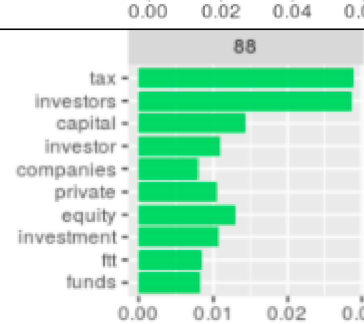
31	02 – U.S. Markets	Topic: 1 Pass: 2 Iteration: 2	(1) Financial Markets;	Unemployment (Recession)	
32	02 – U.S. Markets	Topic: 25 Pass: 2 Iteration: 2	(1) Financial Markets;	Women Labor Force (Participation)	
33	03 - Entities	Topic: 20 Pass: 2 Iteration: 1	(1) Small Medium Enterprises (SME);	Business Owners	
34	03 - Entities	Topic: 24 Pass: 2 Iteration: 1	(1) BCBS (Basel Committee on Banking Supervision Compliance);	Capital / Banking Institutions	
35	03 - Entities	Topic: 5 Pass: 2 Iteration: 2	(1) Community Banking; (2) Entities;	Community Banks	
36	03 - Entities	Topic: 96 Pass: 1 Iteration: 1	(1) Entities;	Credit Ratings Agencies	
37	03 - Entities	Topic: 10 Pass: 2 Iteration: 1	(1) Entities; (2) Clearing & Settlement; (3) Derivatives;	Derivatives Clearing / Clearinghouses	

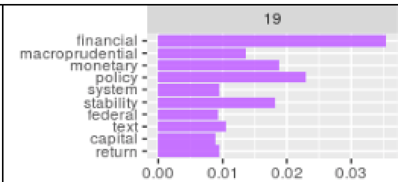
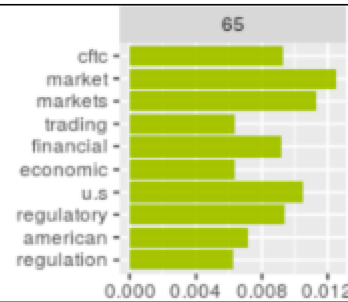
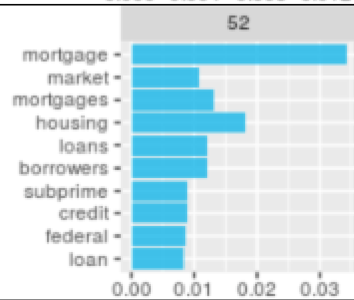
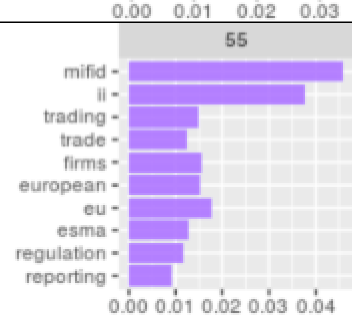
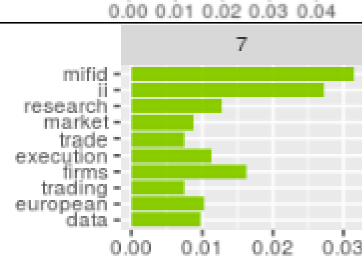
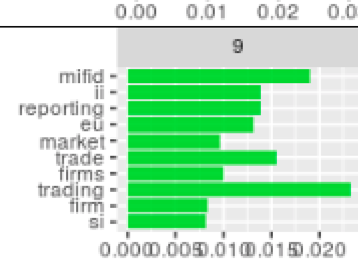
38	03 - Entities	Topic: 60 Pass: 1 Iteration: 1	(1) European Markets; (2) Market Oversight (Regulation & Policy);	European Union	
39	03 - Entities	Topic: 25 Pass: 1 Iteration: 1	(1) Households;	Family Income / Wealth	
40	03 - Entities	Topic: 38 Pass: 2 Iteration: 2	(1) Federal Reserve; (2) Fiscal Policy;	Federal Budget (Debt)	
41	03 - Entities	Topic: 6 Pass: 2 Iteration: 1	(1) Health Care Market;	Health Care (Consumers)	
42	03 - Entities	Topic: 13 Pass: 2 Iteration: 2	(1) Households;	Household Income (Wealth)	
43	03 - Entities	Topic: 24 Pass: 2 Iteration: 1	(1) Households; (2) Banking & Lending; (3) Borrower Level;	Housing Foreclosures (Vacant Properties)	
44	03 - Entities	Topic: 66 Pass: 1 Iteration: 1	(1) Market Oversight (Regulation & Policy);	Industry Panel for Regulatory Issues	

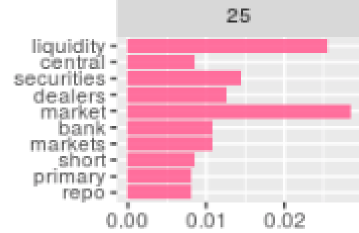
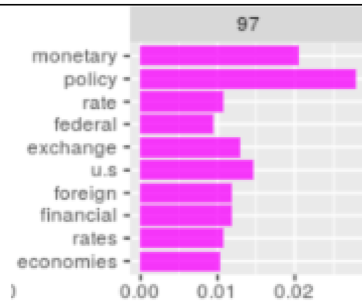
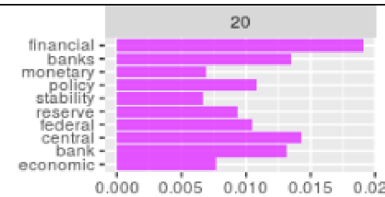
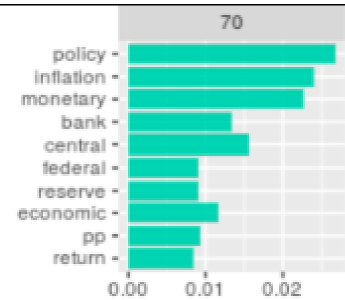
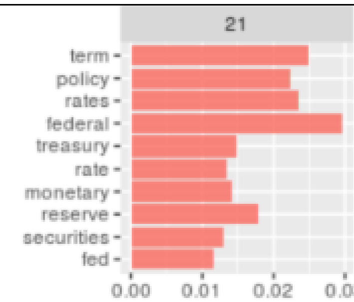
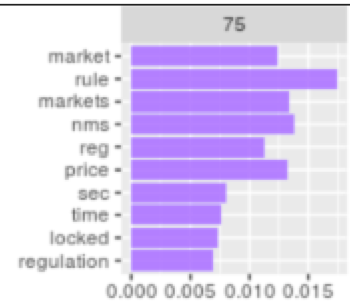
<table>
<tr><td>45</td><td>03 - Entities</td><td>Topic: 38
Pass: 1
Iteration: 1</td><td>(1) Equity Markets;
(2) Public Companies;</td><td>Initial Public Stock Offerings</td><td></td></tr>
<tr><td>46</td><td>03 - Entities</td><td>Topic: 4
Pass: 2
Iteration: 1</td><td>(1) Trading & Investing;</td><td>Investors (Social Investing / Sustainability)</td><td></td></tr>
<tr><td>47</td><td>03 - Entities</td><td>Topic: 7
Pass: 2
Iteration: 1</td><td>(1) U.S. Markets;</td><td>Local Communities</td><td></td></tr>
<tr><td>48</td><td>03 - Entities</td><td>Topic: 15
Pass: 2
Iteration: 1</td><td>(1) Entities;
(2) U.S. Markets - Labor Force;</td><td>Part-time / Contingent Workers</td><td></td></tr>
<tr><td>49</td><td>03 - Entities</td><td>Topic: 48
Pass: 1
Iteration: 1</td><td>(1) Investors;</td><td>Public Investors</td><td></td></tr>
<tr><td>50</td><td>03 - Entities</td><td>Topic: 17
Pass: 2
Iteration: 1</td><td>(1) Financial Firms;
(2) Research & Analytics;</td><td>Research Firms</td><td></td></tr>
<tr><td>51</td><td>03 - Entities</td><td>Topic: 19
Pass: 2
Iteration: 1</td><td>(1) Government Agencies;
(2) Market Oversight (Regulation & Policy);</td><td>Reserve Banks - Federal</td><td></td></tr>
</table>

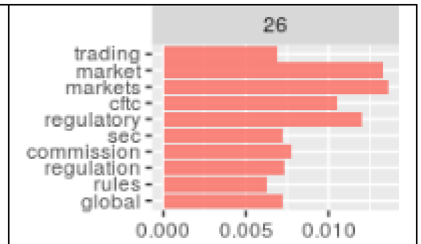
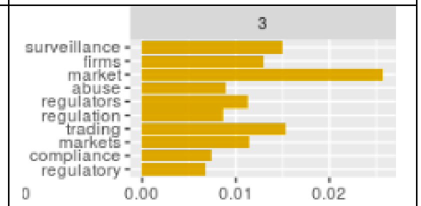
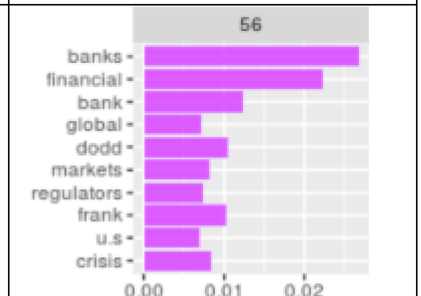
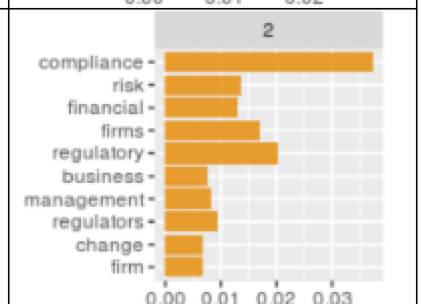
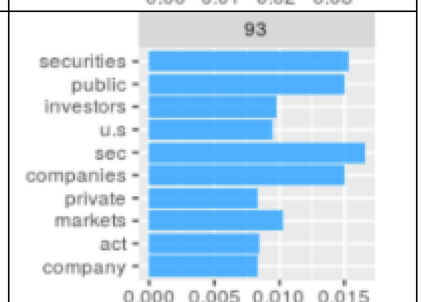
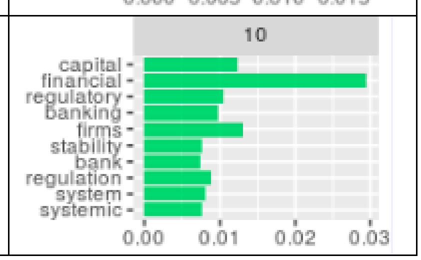
52	03 - Entities	Topic: 99 Pass: 1 Iteration: 1	(1) Sell Side;	Retail Brokerage	
53	03 - Entities	Topic: 2 Pass: 2 Iteration: 1	(1) Sell Side;	Securities Trading Firms	
54	04 – Market Oversight (Policy & Regulation)	Topic: 78 Pass: 1 Iteration: 1	(1) BCBS (Basel Committee on Banking Supervision Compliance);	Bank Capital - Basel II	
55	04 – Market Oversight (Policy & Regulation)	Topic: 1 Pass: 2 Iteration: 1	(1) BCBS (Basel Committee on Banking Supervision Compliance);	Bank Capital - Basel III	
56	04 – Market Oversight (Policy & Regulation)	Topic: 20 Pass: 1 Iteration: 1	(1) BCBS (Basel Committee on Banking Supervision Compliance);	Bank Capital Regulation (i.e., Basel)	
57	04 – Market Oversight (Policy & Regulation)	Topic: 36 Pass: 1 Iteration: 1	(1) Regulatory Compliance; (2) Market Oversight (Regulation & Policy);	Business Friendly Regulation	

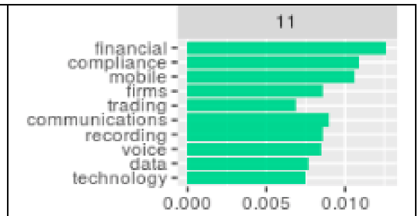
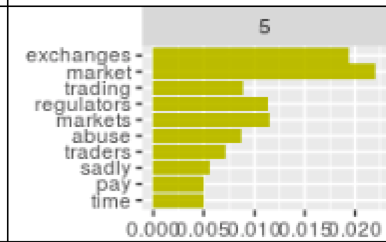
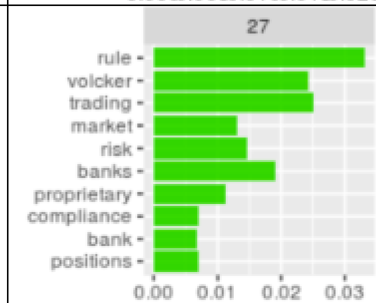
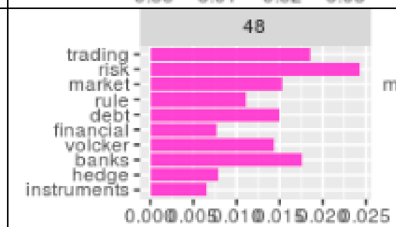
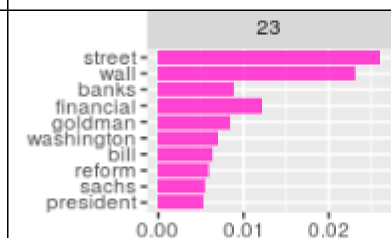
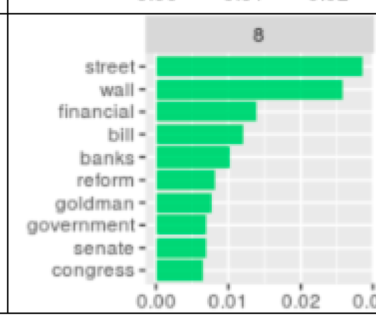
58	04 – Market Oversight (Policy & Regulation)	Topic: 71 Pass: 1 Iteration: 1	(1) Liquidity Risk; (2) Market Oversight (Regulation & Policy);	Central Bank Liquidity Management	
59	04 – Market Oversight (Policy & Regulation)	Topic: 62 Pass: 1 Iteration: 1	(1) Regulatory Compliance; (2) CFTC;	CFTC Rules for Swap Dealers / Futures Commission Merchant	
60	04 – Market Oversight (Policy & Regulation)	Topic: 77 Pass: 1 Iteration: 1	(1) Market Oversight (Regulation & Policy); (2) Households;	Community Housing Programs	
61	04 – Market Oversight (Policy & Regulation)	Topic: 12 Pass: 2 Iteration: 2	(1) Derivatives Market; (2) Compliance;	Derivatives (Futures/Swaps) Trading Rules	
62	04 – Market Oversight (Policy & Regulation)	Topic: 20 Pass: 2 Iteration: 1	(1) European Markets; (2) Market Oversight (Regulation & Policy);	European Markets - MiFID II Regulation - Market Abuse	
63	04 – Market Oversight (Policy & Regulation)	Topic: 92 Pass: 1 Iteration: 1	(1) Market Oversight (Regulation & Policy);	Federal Reserve on Education	

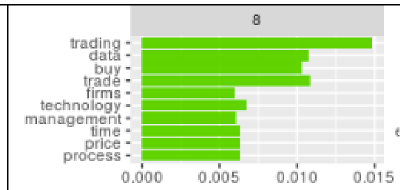
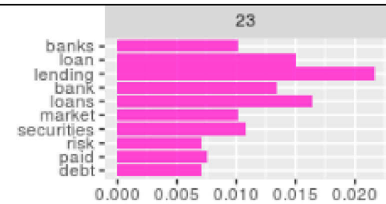
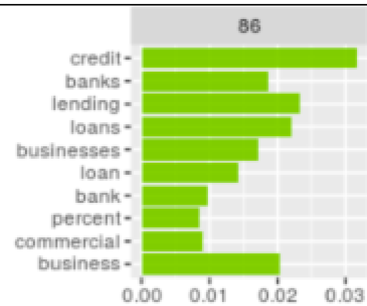
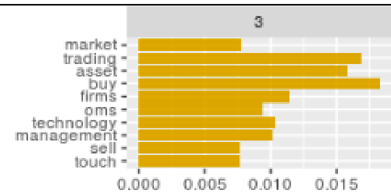
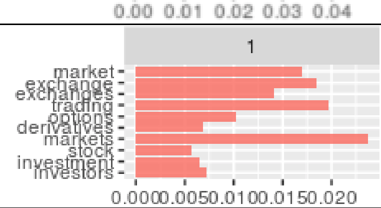
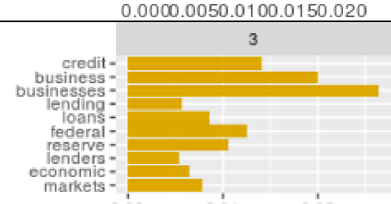
64	04 – Market Oversight (Policy & Regulation)	Topic: 13 Pass: 2 Iteration: 1	(1) Market Oversight (Regulation & Policy); (2) Tax Function;	Financial Transaction Tax	
65	04 – Market Oversight (Policy & Regulation)	Topic: 17 Pass: 2 Iteration: 1	(1) Market Oversight (Regulation & Policy);	Financial Markets - Monetary Policy	
66	04 – Market Oversight (Policy & Regulation)	Topic: 8 Pass: 2 Iteration: 1	(1) Market Oversight (Regulation & Policy); (2) Equity Markets;	Flash Crash - Regulatory Surveillance	
67	04 – Market Oversight (Policy & Regulation)	Topic: 13 Pass: 2 Iteration: 2	(1) Market Oversight (Regulation & Policy);	FOMC Policy - Monetary Policy	
68	04 – Market Oversight (Policy & Regulation)	Topic: 22 Pass: 1 Iteration: 1	(1) Market Oversight (Regulation & Policy); (2) Equity Markets;	High Frequency Trading - Market Regulation	
69	04 – Market Oversight (Policy & Regulation)	Topic: 8 Pass: 2 Iteration: 2	(1) Market Oversight (Regulation & Policy); (2) Global Markets;	Inflation Expectations	
70	04 – Market Oversight (Policy & Regulation)	Topic: 88 Pass: 1 Iteration: 1	(1) Market Oversight (Regulation & Policy); (2) Asset Management; (3) Tax Function;	Investor Taxation in Private Funds	

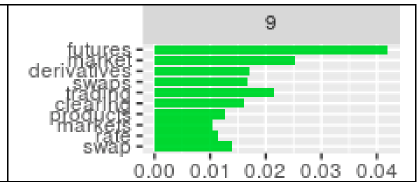
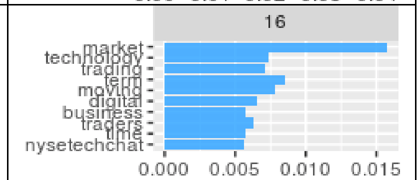
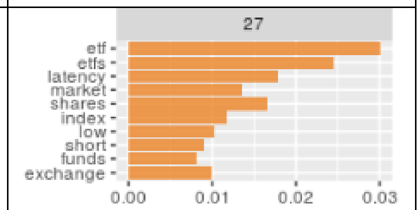
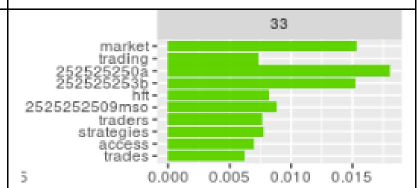
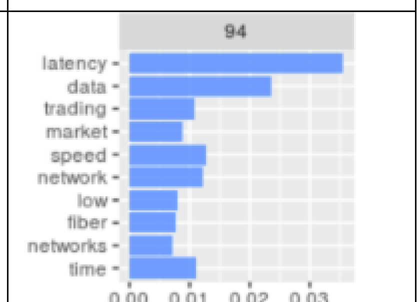
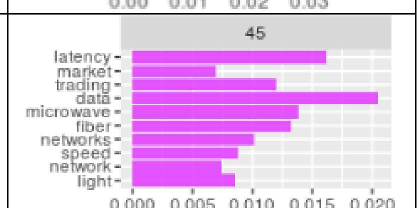
71	04 – Market Oversight (Policy & Regulation)	Topic: 19 Pass: 2 Iteration: 1	(1) Market Oversight (Regulation & Policy);	Macro prudential Policy	
72	04 – Market Oversight (Policy & Regulation)	Topic: 65 Pass: 1 Iteration: 1	(1) Market Oversight (Regulation & Policy); (2) CFTC;	Market Regulation - CFTC	
73	04 – Market Oversight (Policy & Regulation)	Topic: 52 Pass: 1 Iteration: 1	(1) Market Oversight (Regulation & Policy); (2) Derivatives; (3) Clearing;	Market Regulation - Derivatives Clearing	
74	04 – Market Oversight (Policy & Regulation)	Topic: 55 Pass: 1 Iteration: 1	(1) Market Oversight (Regulation & Policy); (2) EU Markets;	MiFID II (European Union Markets Regulation)	
75	04 – Market Oversight (Policy & Regulation)	Topic: 7 Pass: 2 Iteration: 2	(1) Market Oversight (Regulation & Policy); (2) Trade Execution;	MiFID II (Trade Execution)	
76	04 – Market Oversight (Policy & Regulation)	Topic: 9 Pass: 2 Iteration: 2	(1) Market Oversight (Regulation & Policy); (2) Trade Reporting;	MiFID II (Trade Reporting)	

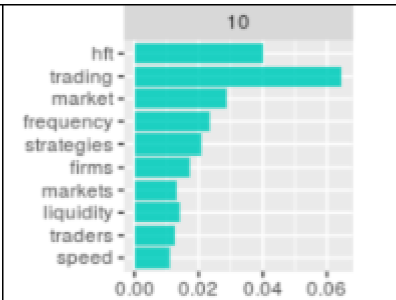
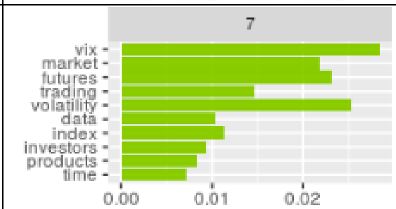
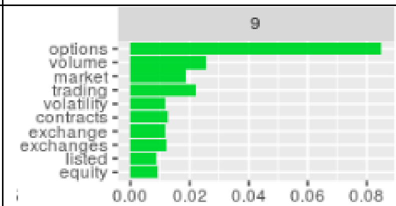
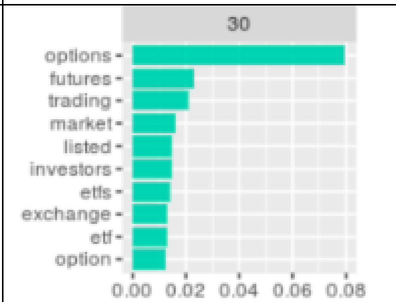
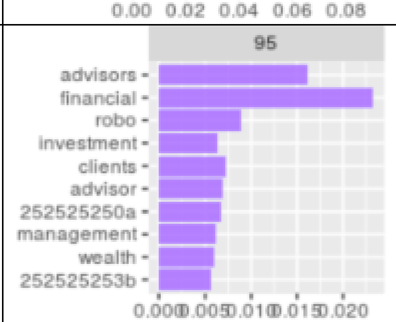
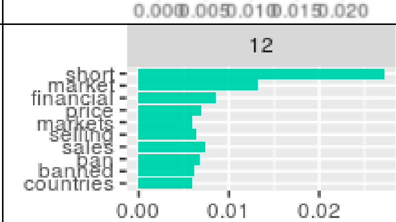
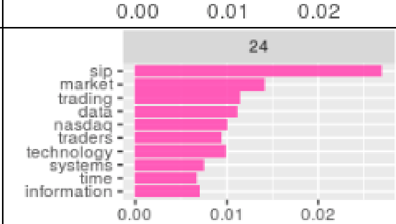
77	04 – Market Oversight (Policy & Regulation)	Topic: 25 Pass: 2 Iteration: 2	(1) Market Oversight (Regulation & Policy); (2) Trading Markets;	MiFID II (Trading)	
78	04 – Market Oversight (Policy & Regulation)	Topic: 97 Pass: 1 Iteration: 1	(1) Market Oversight (Regulation & Policy);	Monetary Policy	
79	04 – Market Oversight (Policy & Regulation)	Topic: 20 Pass: 2 Iteration: 1	(1) Market Oversight (Regulation & Policy);	Monetary Policy - Economic Stability	
80	04 – Market Oversight (Policy & Regulation)	Topic: 70 Pass: 1 Iteration: 1	(1) Market Oversight (Regulation & Policy);	Monetary Policy - Inflation	
81	04 – Market Oversight (Policy & Regulation)	Topic: 21 Pass: 1 Iteration: 1	(1) Market Oversight (Regulation & Policy);	Monetary Policy for Interest Rates	
82	04 – Market Oversight (Policy & Regulation)	Topic: 75 Pass: 1 Iteration: 1	(1) Market Oversight (Regulation & Policy); (2) Equity Markets; (3) Exchanges;	NMS Rule equity regulation	

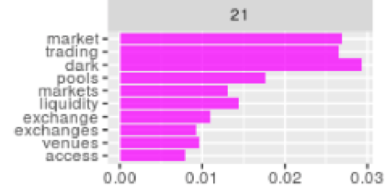
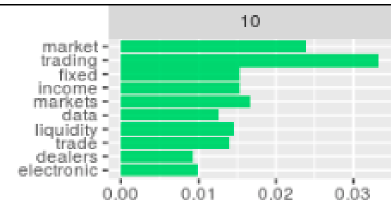
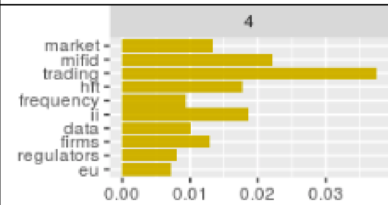
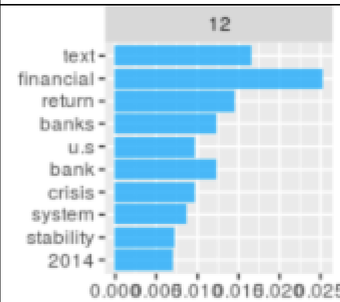
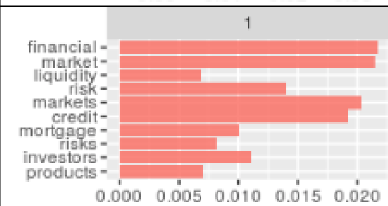
83	04 – Market Oversight (Policy & Regulation)	Topic: 26 Pass: 2 Iteration: 2	(1) Market Oversight (Regulation & Policy); (2) Global Markets;	Regulation of Global Markets	
84	04 – Market Oversight (Policy & Regulation)	Topic: 3 Pass: 2 Iteration: 1	(1) Market Oversight (Regulation & Policy); (2) Surveillances;	Regulatory Market Surveillance	
85	04 – Market Oversight (Policy & Regulation)	Topic: 56 Pass: 1 Iteration: 1	(1) Market Oversight (Regulation & Policy);	Regulatory Reform - Dodd Frank Act	
86	04 – Market Oversight (Policy & Regulation)	Topic: 2 Pass: 1 Iteration: 1	(1) Market Oversight (Regulation & Policy); (2) Financial Firms; (3) Compliance;	Regulatory Compliance (for financial firms)	
87	04 – Market Oversight (Policy & Regulation)	Topic: 93 Pass: 1 Iteration: 1	(1) Market Oversight (Regulation & Policy); (2) SEC; (3) Public Companies;	SEC oversight of public companies	
88	04 – Market Oversight (Policy & Regulation)	Topic: 10 Pass: 2 Iteration: 2	(1) Market Oversight (Regulation & Policy); (2) Non-Bank Financial Companies;	Shadow Banking	

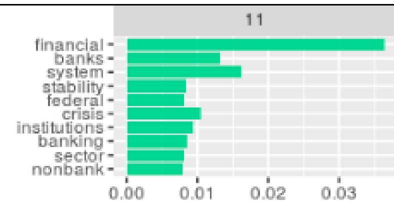
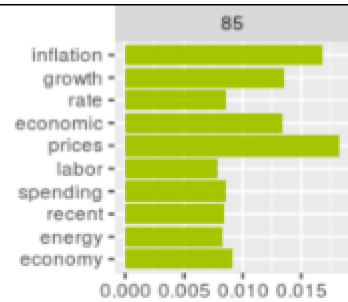
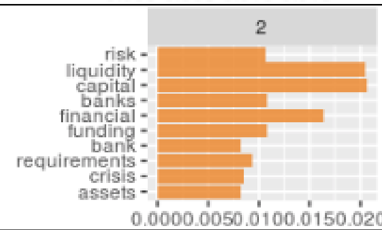
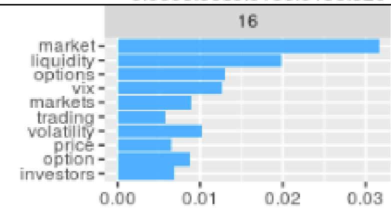
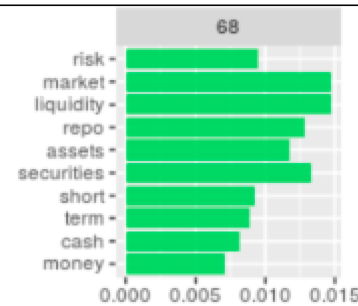
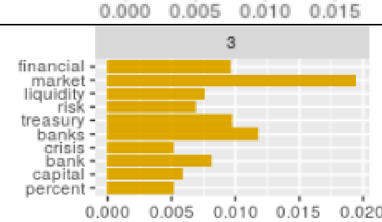
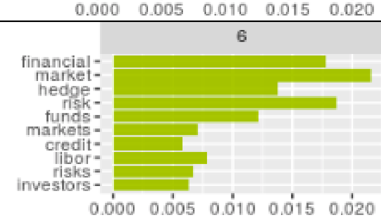
89	04 – Market Oversight (Policy & Regulation)	Topic: 11 Pass: 2 Iteration: 2	(1) Market Oversight (Regulation & Policy); (2) CFTC; (3) Data - Voice; (4) Recording Systems;	Trade Compliance (Voice Recording)	
90	04 – Market Oversight (Policy & Regulation)	Topic: 5 Pass: 2 Iteration: 2	(1) Market Oversight (Regulation & Policy); (2) Corporate Governance;	U.S. Bank Compensatio n / Pay	
91	04 – Market Oversight (Policy & Regulation)	Topic: 27 Pass: 1 Iteration: 1	(1) Market Oversight (Regulation & Policy); (2) Regulation;	Volcker Rule	
92	04 – Market Oversight (Policy & Regulation)	Topic: 48 Pass: 2 Iteration: 2	(1) Market Oversight (Regulation & Policy); (2) Trading Markets - Proprietary;	Volcker Rule (Proprietary Trading Exemptions)	
93	04 – Market Oversight (Policy & Regulation)	Topic: 23 Pass: 2 Iteration: 2	(1) Market Oversight (Regulation & Policy); (2) Trading Markets;	Wall Street Reform (i.e., Bill / DFA)	
94	04 – Market Oversight (Policy & Regulation)	Topic: 8 Pass: 1 Iteration: 1	(1) Market Oversight (Regulation & Policy); (2) Systematic Risk;	Wall Street Reform (i.e., too big to fail)	

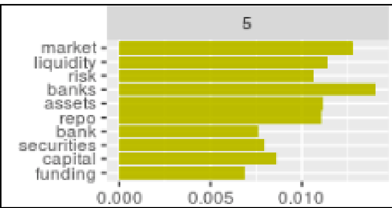
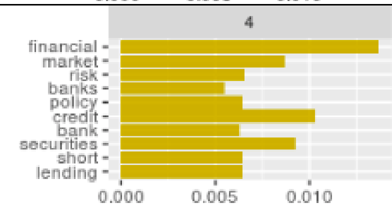
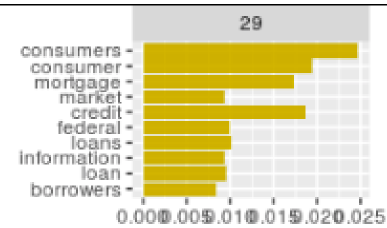
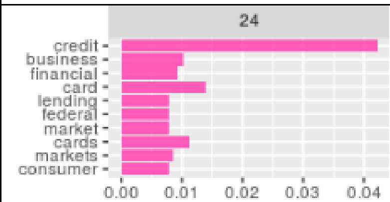
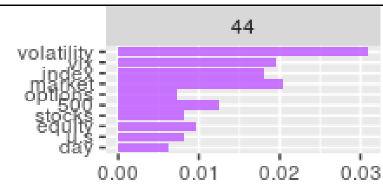
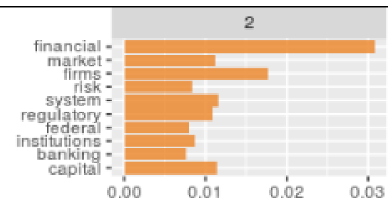
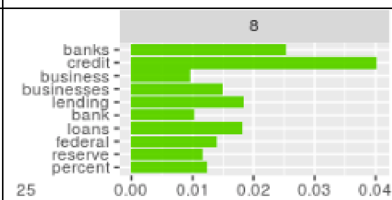
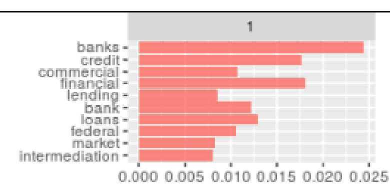
95	05 – Investing & Trading	Topic: 8 Pass: 2 Iteration: 1	(1) Market Oversight (Regulation & Policy); (2) Equity Markets; (3) Systems Technology; (4) Market Data;	Algorithmic Trading	
96	05 – Investing & Trading	Topic: 23 Pass: 2 Iteration: 1	(1) Market Oversight (Regulation & Policy); (2) Lending;	Bank Lending	
97	05 – Investing & Trading	Topic: 86 Pass: 1 Iteration: 1	(1) Market Oversight (Regulation & Policy); (2) Commercial Lending;	Bank Lending - Commercial Lending	
98	05 – Investing & Trading	Topic: 3 Pass: 2 Iteration: 1	(1) Market Oversight (Regulation & Policy); (2) Buy Side Instruction;	Buy Order	
99	05 – Investing & Trading	Topic: 87 Pass: 1 Iteration: 1	(1) Market Oversight (Regulation & Policy); (2) Financial Firm - Buy Side;	Buys Side Trading	
100	05 – Investing & Trading	Topic: 1 Pass: 2 Iteration: 1	(1) Market Oversight (Regulation & Policy); (2) Investing;	Capital Markets & Investing (stocks & derivatives)	
101	05 – Investing & Trading	Topic: 3 Pass: 2 Iteration: 1	(1) Market Oversight (Regulation & Policy); (2) Borrower;	Credit Businesses	

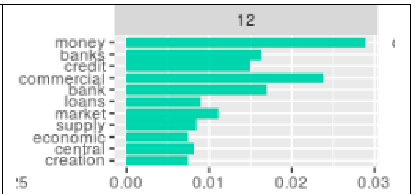
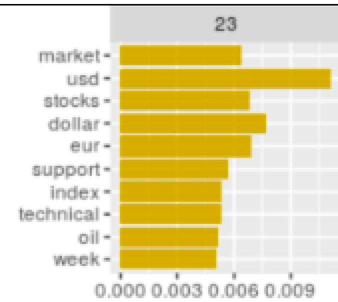
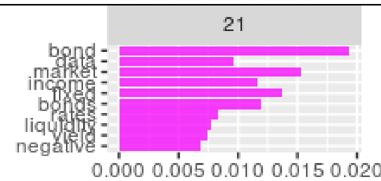
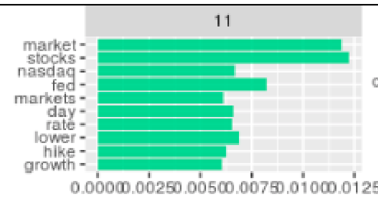
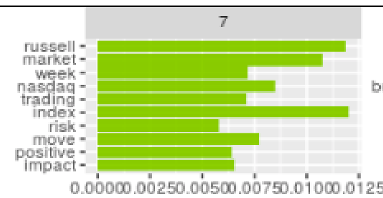
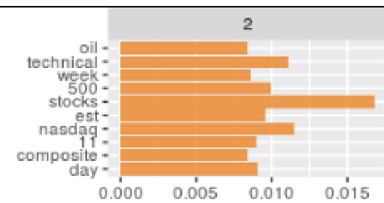
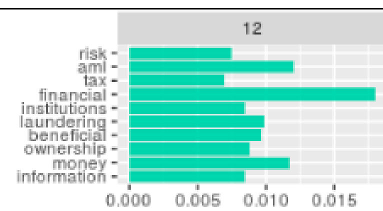
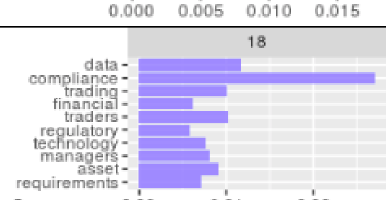
102	05 – Investing & Trading	Topic: 9 Pass: 2 Iteration: 2	(1) Market Oversight (Regulation & Policy); (2) Derivatives;	Derivatives (Futures/Swaps) Trading	
103	05 – Investing & Trading	Topic: 16 Pass: 2 Iteration: 1	(1) Market Oversight (Regulation & Policy); (2) Equity Markets; (3) Systems Technology; (4) Market Data;	Digital / Fast Trading	
104	05 – Investing & Trading	Topic: 27 Pass: 2 Iteration: 1	(1) Market Oversight (Regulation & Policy); (2) Equity Markets; (3) ETFs;	ETF Shares / Trading	
105	05 – Investing & Trading	Topic: 33 Pass: 2 Iteration: 1	(1) Market Oversight (Regulation & Policy); (2) Equity Markets; (3) Systems Technology; (4) Market Data;	High Frequency Trading	
106	05 – Investing & Trading	Topic: 94 Pass: 1 Iteration: 1	(1) Market Oversight (Regulation & Policy); (2) Equity Markets; (3) Systems Technology; (4) Market Data;	High Frequency Trading - latency	
107	05 – Investing & Trading	Topic: 45 Pass: 2 Iteration: 1	(1) Market Oversight (Regulation & Policy); (2) Equity Markets; (3) Systems Technology; (4) Market Data;	High Frequency Trading - latency (i.e., microwaves)	

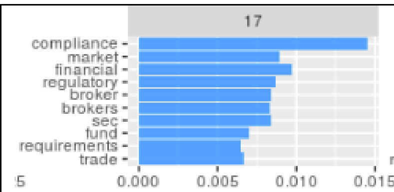
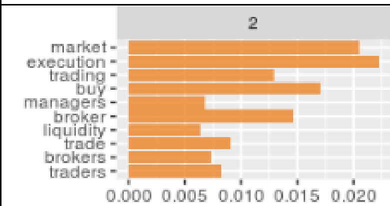
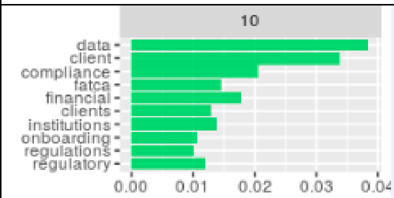
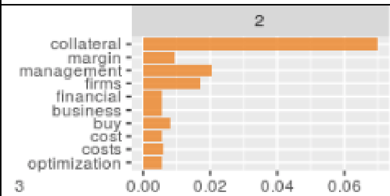
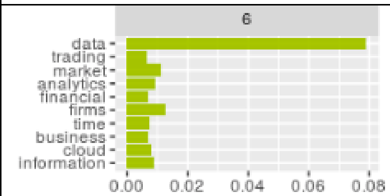
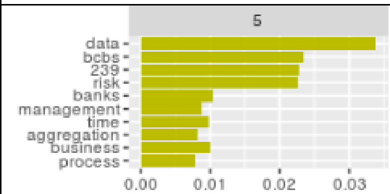
108	05 – Investing & Trading	Topic: 10 Pass: 1 Iteration: 1	(1) Market Oversight (Regulation & Policy); (2) Equity Markets; (3) Systems Technology; (4) Market Data;	High Frequency Trading - Strategies	
109	05 – Investing & Trading	Topic: 7 Pass: 2 Iteration: 1	(1) Market Oversight (Regulation & Policy); (2) Trading Markets;	Market Volatility	
110	05 – Investing & Trading	Topic: 9 Pass: 2 Iteration: 1	(1) Market Oversight (Regulation & Policy); (2) Options Markets;	Options Trading (i.e., Exchanges)	
111	05 – Investing & Trading	Topic: 30 Pass: 1 Iteration: 1	(1) Market Oversight (Regulation & Policy); (2) Options Markets;	Options/Futures Trading	
112	05 – Investing & Trading	Topic: 95 Pass: 1 Iteration: 1	(1) Market Oversight (Regulation & Policy); (2) Portfolio Management; (2) Systems & Technology;	Robo Advisors	
113	05 – Investing & Trading	Topic: 12 Pass: 2 Iteration: 1	(1) Market Oversight (Regulation & Policy); (2) Sell Instruction; (3) Equity Markets;	Short Selling	
114	05 – Investing & Trading	Topic: 24 Pass: 2 Iteration: 1	(1) Market Oversight (Regulation & Policy);	SIP (Systematic Investment Plan) Trading	

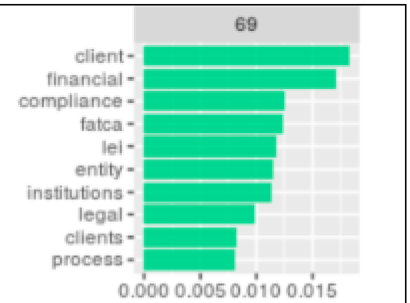
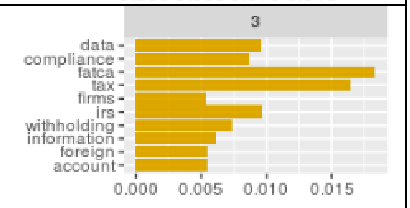
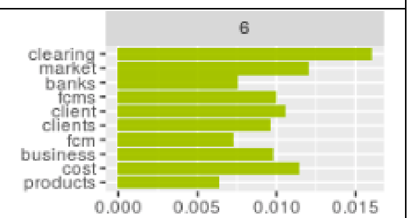
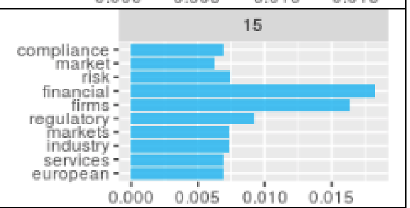
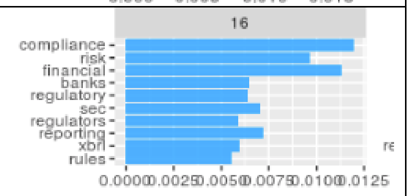
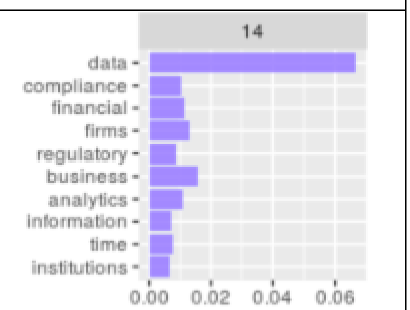
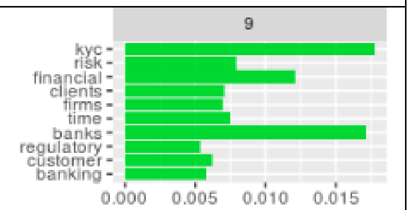
115	05 – Investing & Trading	Topic: 21 Pass: 2 Iteration: 1	(1) Market Oversight (Regulation & Policy); (2) Exchanges;	Trading - Dark Pools	
116	05 – Investing & Trading	Topic: 10 Pass: 2 Iteration: 1	(1) Market Oversight (Regulation & Policy); (2) Fixed Income Markets;	Trading - Fixed Income	
117	05 – Investing & Trading	Topic: 4 Pass: 2 Iteration: 1	(1) Market Oversight (Regulation & Policy); (2) Equity Markets; (3) Systems Technology; (4) Market Data;	Trading (High Frequency Trading)	
118	06 – Risks	Topic: 12 Pass: 1 Iteration: 1	(1) Market Oversight (Regulation & Policy); (2) Systematic Risk;	Financial Crisis / Systemic Risk	
119	06 – Risks	Topic: 1 Pass: 1 Iteration: 1	(1) Market Oversight (Regulation & Policy); (2) Risk & Regulatory Compliance	Financial Risk	
120	06 – Risks	Topic: 1 Pass: 2 Iteration: 1	(1) Credit Risk Management; (2) Market Risk Management;	Financial Risk (i.e., Market / Credit Risk)	

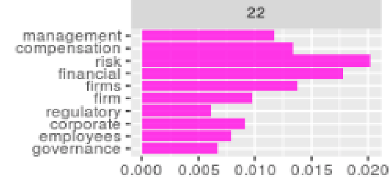
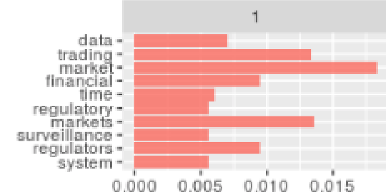
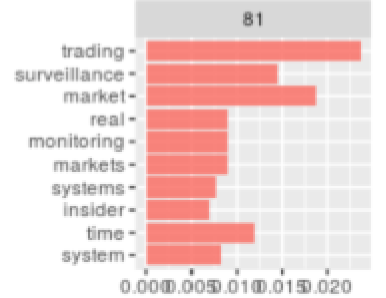
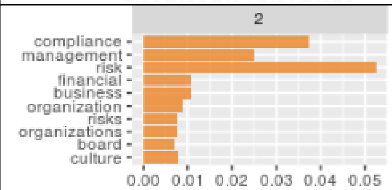
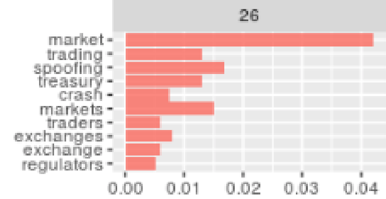
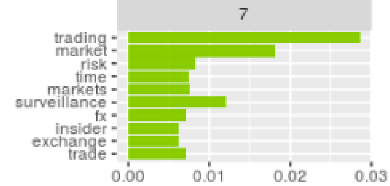
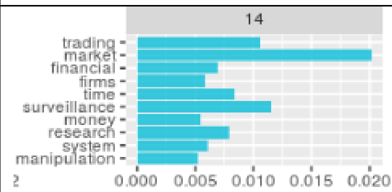
121	06 – Risks	Topic: 11 Pass: 2 Iteration: 1	(1) Market Oversight (Regulation & Policy);	Financial Stability	
122	06 – Risks	Topic: 85 Pass: 1 Iteration: 1	(1) Market Oversight (Regulation & Policy);	Inflation	
123	06 – Risks	Topic: 2 Pass: 2 Iteration: 2	(1) Market Oversight (Regulation & Policy); (2) Liquidity Risk;	Liquidity	
124	06 – Risks	Topic: 16 Pass: 2 Iteration: 1	(1) Market Oversight (Regulation & Policy); (2) Liquidity Risk;	Market Liquidity - Options	
125	06 – Risks	Topic: 68 Pass: 1 Iteration: 1	(1) Market Oversight (Regulation & Policy); (2) Liquidity Risk	Market Liquidity - Short Term / Repo	
126	06 – Risks	Topic: 3 Pass: 2 Iteration: 2	(1) Market Oversight (Regulation & Policy); (2) Market Risk Mgmt;	Market Risk	
127	06 – Risks	Topic: 6 Pass: 2 Iteration: 1	(1) Market Oversight (Regulation & Policy); (2) Market Risk Mgmt;	Market Risk - Hedge Funds	

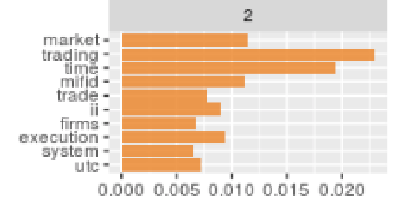
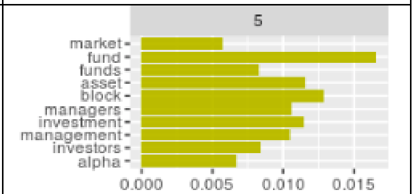
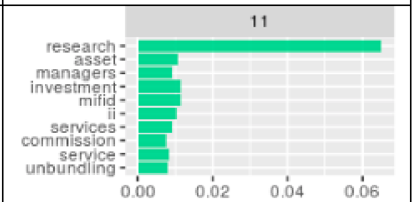
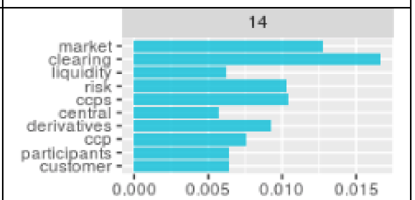
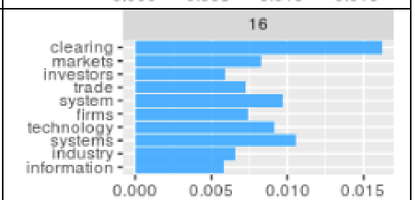
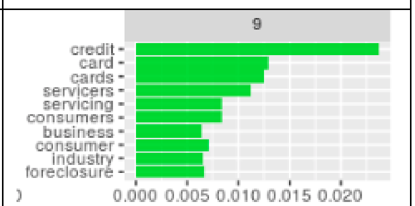
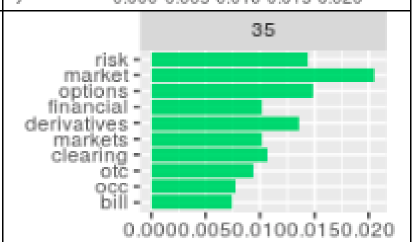
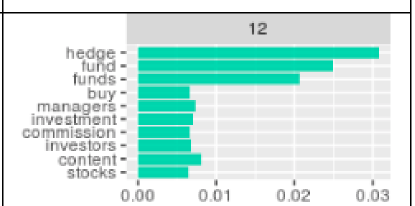
128	06 – Risks	Topic: 5 Pass: 2 Iteration: 1	(1) Market Oversight (Regulation & Policy); (2) Market Risk Mgmt;	Market Risk - Repos / Securities	
129	06 – Risks	Topic: 4 Pass: 2 Iteration: 1	(1) Market Oversight (Regulation & Policy); (2) Market Risk Mgmt;	Market Risk Policy	
130	06 – Risks	Topic: 29 Pass: 2 Iteration: 1	(1) Market Oversight (Regulation & Policy); (2) Borrower Level;	Consumer Credit	
131	06 – Risks	Topic: 24 Pass: 2 Iteration: 1	(1) Market Oversight (Regulation & Policy); (2) Borrower Level;	Credit	
132	06 – Risks	Topic: 44 Pass: 2 Iteration: 2	(1) Market Oversight (Regulation & Policy); (2) Market Risk Mgmt;	Stock Market Volatility (S&P VIX)	
133	06 – Risks	Topic: 2 Pass: 2 Iteration: 1	(1) Market Oversight (Regulation & Policy); (2) Systematic Risk;	Systemic Risk	
134	07 – Instruments	Topic: 8 Pass: 2 Iteration: 1	(1) Lending; (2) Borrower Level;	Bank Loans	
135	07 – Instruments	Topic: 1 Pass: 2 Iteration: 1	(1) Commercial Lending;	Commercial Loans	

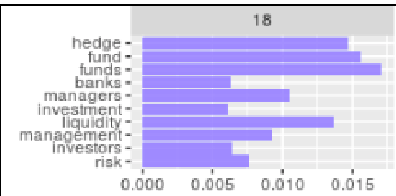
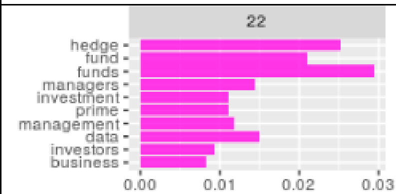
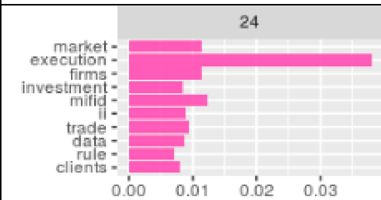
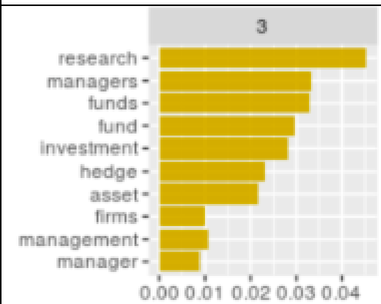
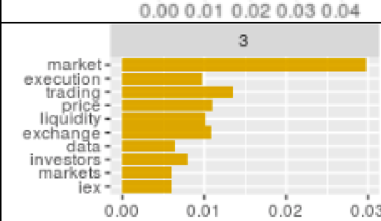
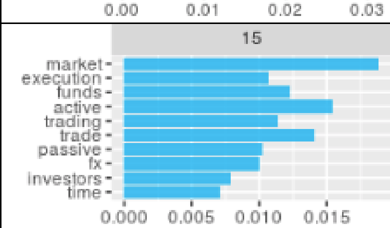
136	07 – Instruments	Topic: 12 Pass: 2 Iteration: 1	(1) Commercial Lending;	Commercial Money Markets	
137	07 – Instruments	Topic: 23 Pass: 1 Iteration: 1	(1) FX Markets;	Currencies	
138	07 – Instruments	Topic: 21 Pass: 2 Iteration: 1	(1) Fixed Income Markets;	Fixed Income (Bonds)	
139	07 – Instruments	Topic: 11 Pass: 2 Iteration: 1	(1) Equity Markets;	NASDAQ Stocks	
140	07 – Instruments	Topic: 7 Pass: 2 Iteration: 1	(1) Equity Markets;	Russell Index	
141	07 – Instruments	Topic: 2 Pass: 2 Iteration: 1	(1) Equity Markets;	Technicals - Stocks, Oil, etc.	
142	08 - Compliance	Topic: 12 Pass: 2 Iteration: 1	(1) Market Oversight (Regulation & Policy); (2) AML Management;	AML (Anti-Money Laundering) Information	
143	08 - Compliance	Topic: 18 Pass: 2 Iteration: 1	(1) Market Oversight (Regulation & Policy); (2) Regulatory Compliance;	Asset Managers Compliance	

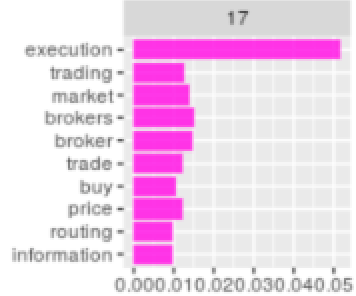
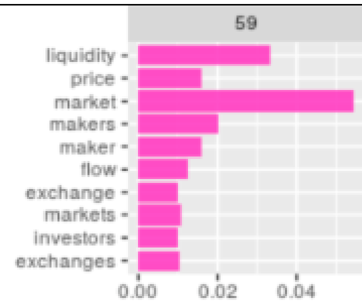
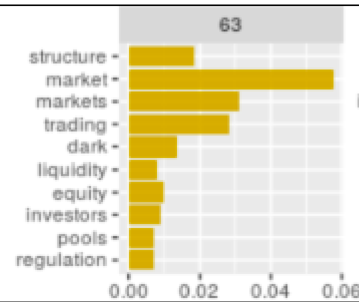
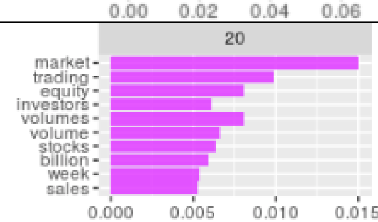
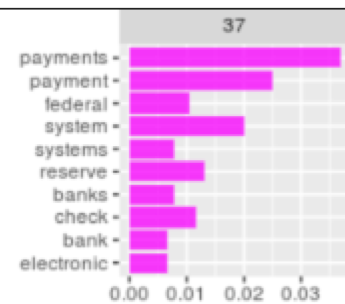
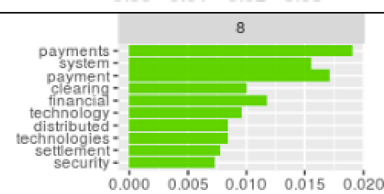
144	08 - Compliance	Topic: 17 Pass: 2 Iteration: 1	(1) Market Oversight (Regulation & Policy); (2) Regulatory Compliance;	Broker Trade Compliance	
145	08 - Compliance	Topic: 2 Pass: 2 Iteration: 1	(1) Market Oversight (Regulation & Policy); (2) Trading & Execution;	Brokers / Trade Execution	
146	08 - Compliance	Topic: 20 Pass: 2 Iteration: 1	(1) Market Oversight (Regulation & Policy); (2) Data; (3) Regulatory Compliance;	Client Data Compliance	
147	08 - Compliance	Topic: 10 Pass: 2 Iteration: 1	(1) Market Oversight (Regulation & Policy); (2) Data; (3) Regulatory Compliance; (4) Client Onboarding;	Client Onboarding - Data Regulations	
148	08 - Compliance	Topic: 2 Pass: 2 Iteration: 1	(1) Market Oversight (Regulation & Policy); (2) Collateral Optimization;	Collateral / Margin Management (and Optimization)	
149	08 - Compliance	Topic: 6 Pass: 2 Iteration: 1	(1) Market Oversight (Regulation & Policy); (2) Data;	Data / Information	
150	08 - Compliance	Topic: 5 Pass: 2 Iteration: 1	(1) Market Oversight (Regulation & Policy); (2) Data;	Data Aggregation & Reporting / BCBS 239	

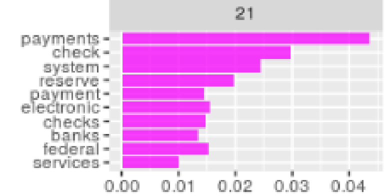
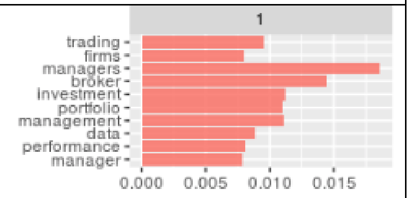
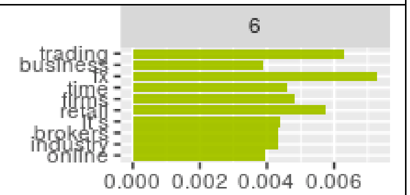
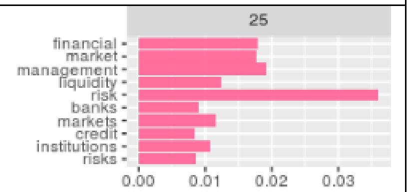
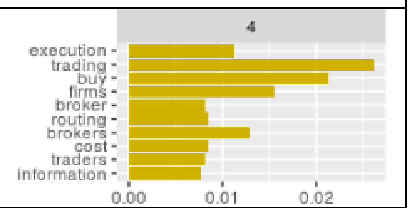
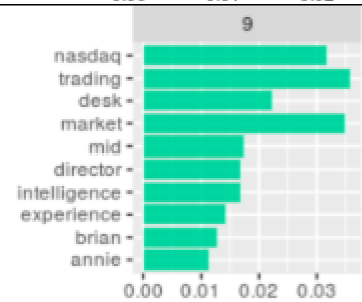
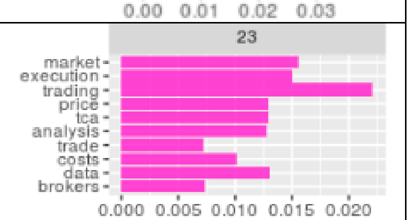
151	08 - Compliance	Topic: 69 Pass: 1 Iteration: 1	(1) Market Oversight (Regulation & Policy); (2) FATCA (Foreign Exchange Tax Compliance Act) Compliance;	FATCA / LEI compliance	
152	08 - Compliance	Topic: 3 Pass: 2 Iteration: 1	(1) Market Oversight (Regulation & Policy); (2) FATCA (Foreign Exchange Tax Compliance Act) Compliance;	FATCA Compliance - Foreign Withholding	
153	08 - Compliance	Topic: 6 Pass: 2 Iteration: 1	(1) Clearing; (2) Regulatory Compliance;	FCM (Futures Commission Merchants) Clearing Regulation	
154	08 - Compliance	Topic: 15 Pass: 2 Iteration: 1	(1) Market Oversight (Regulation & Policy); (2) Regulatory Compliance;	Financial Firms Compliance	
155	08 - Compliance	Topic: 16 Pass: 2 Iteration: 1	(1) Market Oversight (Regulation & Policy); (2) Regulatory Compliance;	Financial Compliance Risk	
156	08 - Compliance	Topic: 14 Pass: 1 Iteration: 1	(1) Market Oversight (Regulation & Policy); (2) Regulatory Compliance; (3) Systems & Technology;	Information Security Compliance	
157	08 - Compliance	Topic: 9 Pass: 2 Iteration: 1	(1) Market Oversight (Regulation & Policy); (2) KYC;	KYC (Know Your Customer) Regulation	

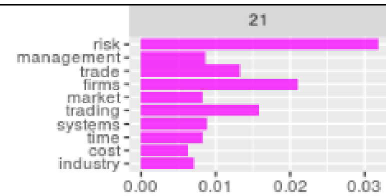
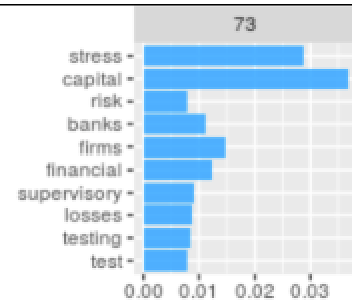
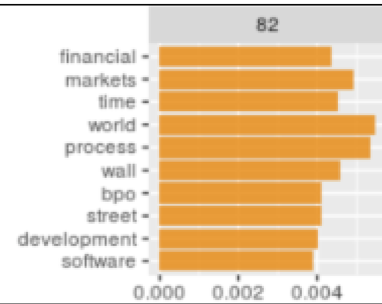
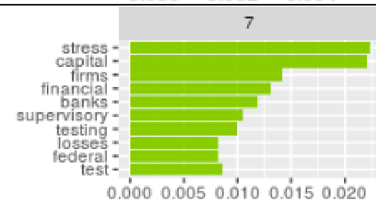
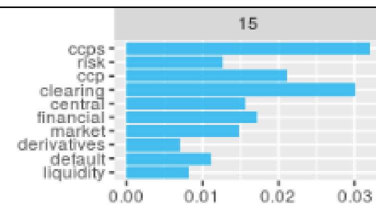
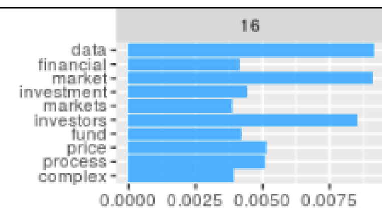
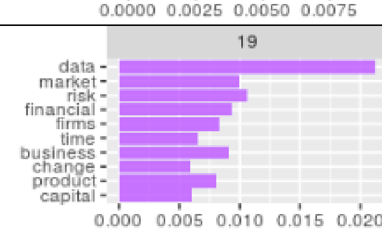
158	08 - Compliance	Topic: 22 Pass: 2 Iteration: 1	(1) Market Oversight (Regulation & Policy); (2) Corporate Governance;	Management Compensatio n Governance	
159	08 - Compliance	Topic: 1 Pass: 2 Iteration: 1	(1) Market Oversight (Regulation & Policy); (2) Regulation & Control of Securities Exchange;	Market Surveillance - Trading Data	
160	08 - Compliance	Topic: 81 Pass: 1 Iteration: 1	(1) Market Oversight (Regulation & Policy); (2) Regulation & Control of Securities Exchange;	Market Surveillance - Trading Markets	
161	08 - Compliance	Topic: 2 Pass: 2 Iteration: 1	(1) Market Oversight (Regulation & Policy); (2) Regulatory Compliance;	Risk Management / Compliance	
162	08 - Compliance	Topic: 26 Pass: 2 Iteration: 1	(1) Market Oversight (Regulation & Policy); (2) Regulatory Compliance; (3) Equity Markets;	Spoofing - Trade Compliance	
163	08 - Compliance	Topic: 7 Pass: 2 Iteration: 1	(1) Market Oversight (Regulation & Policy); (2) Regulatory Compliance; (3) Trading;	Surveillance - Trading	
164	08 - Compliance	Topic: 14 Pass: 2 Iteration: 1	(1) Market Oversight (Regulation & Policy); (2) Regulatory Compliance; (3) Trading;	Trading Surveillance	

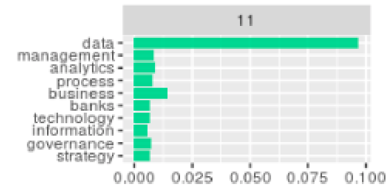
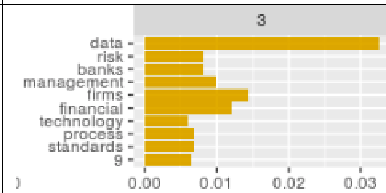
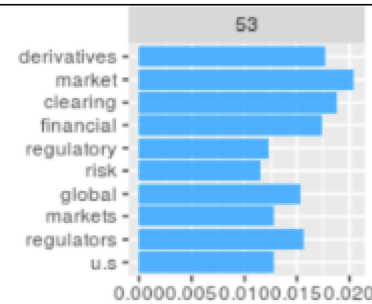
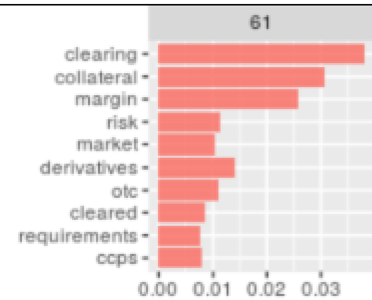
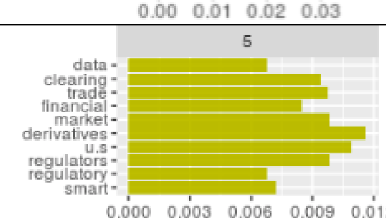
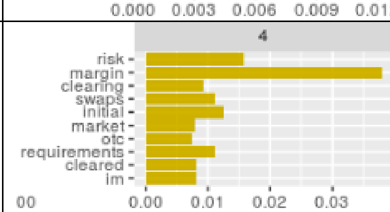
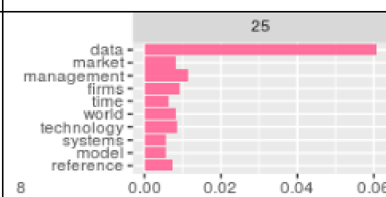
165	08 - Compliance	Topic: 2 Pass: 2 Iteration: 1	(1) Market Oversight (Regulation & Policy); (2) Regulatory Compliance; (3) Trade Execution;	Trading Time - MiFID II - Execution	
166	09 – Functions	Topic: 5 Pass: 2 Iteration: 1	(1) Asset Management; (2) Block Trades;	Asset Management - Block Trades	
167	09 – Functions	Topic: 11 Pass: 2 Iteration: 1	(1) Asset Management; (2) Research;	Asset Management - Research	
168	09 – Functions	Topic: 14 Pass: 2 Iteration: 1	(1) Exchanges; (2) Central Counterparties;	Central Counterparty Clearing	
169	09 – Functions	Topic: 16 Pass: 2 Iteration: 1	(1) Clearing; (2) Markets;	Clearing Markets	
170	09 – Functions	Topic: 9 Pass: 2 Iteration: 1	(1) Credit Cards; (2)Borrower Level;	Credit Card Servicing	
171	09 – Functions	Topic: 35 Pass: 2 Iteration: 2	(1) Derivatives Markets; (2) Clearing;	Derivatives Clearing (CDS)	
172	09 – Functions	Topic: 12 Pass: 2 Iteration: 1	(1) Hedge Funds; (2) Commissions;	Hedge Funds - Commissions	

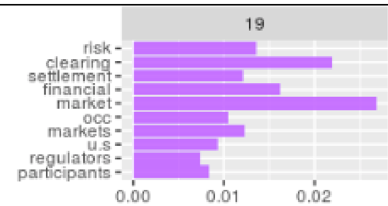
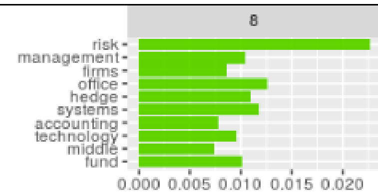
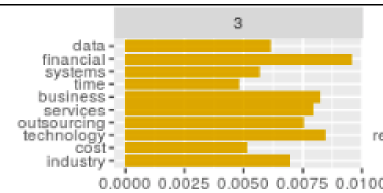
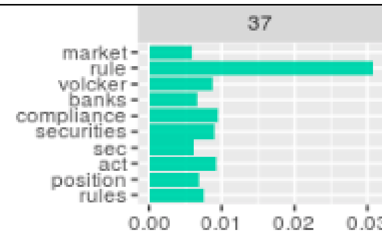
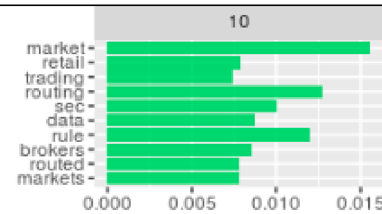
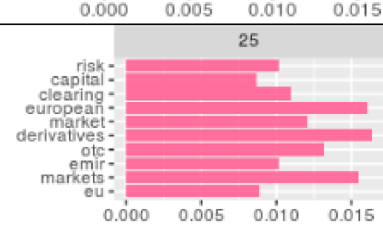
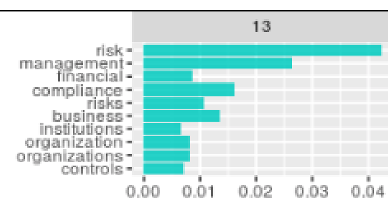
173	09 – Functions	Topic: 18 Pass: 2 Iteration: 1	(1) Hedge Funds; (2) Liquidity;	Hedge Funds - Liquidity	
174	09 – Functions	Topic: 22 Pass: 2 Iteration: 1	(1) Hedge Funds; (2) Portfolio Managers;	Hedge Funds / Investment Managers	
175	09 – Functions	Topic: 24 Pass: 2 Iteration: 1	(1) Investment Firms; (2) Execution;	Investment Firms - Market Execution	
176	09 – Functions	Topic: 3 Pass: 1 Iteration: 1	(1) Investment Firms; (2) Research;	Investment Management Research	
177	09 – Functions	Topic: 3 Pass: 2 Iteration: 1	(1) Markets; (2) Execution;	Market Execution	
178	09 – Functions	Topic: 15 Pass: 2 Iteration: 1	(1) Markets; (2) Execution;	Market Execution - Active Trading vs. Passive Trading	
179	09 – Functions	Topic: 6 Pass: 2 Iteration: 1	(1) Markets; (2) Execution; (3) Institutional Investors;	Market Execution - Institutional Investors	

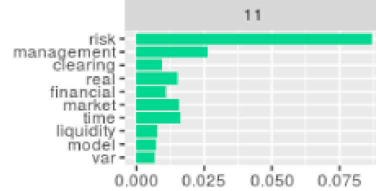
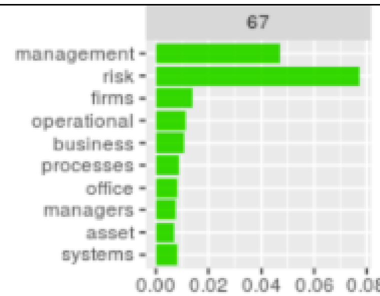
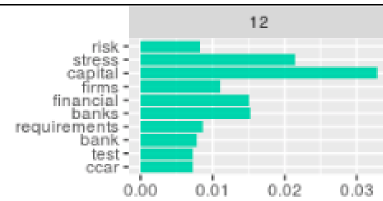
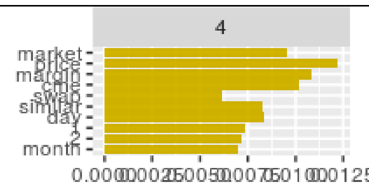
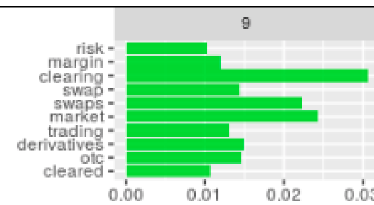
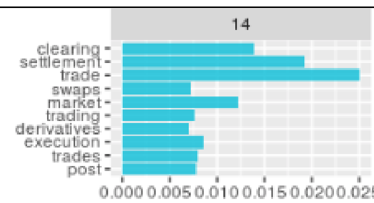
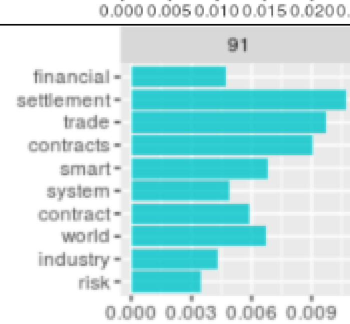
180	09 – Functions	Topic: 17 Pass: 1 Iteration: 1	(1) Markets; (2) Execution; (3) Equity Markets;	Market Execution-Equities	
181	09 – Functions	Topic: 59 Pass: 1 Iteration: 1	(1) Markets; (2) Liquidity Risk; (3) Market Makers;	Market Liquidity - Market Makers	
182	09 – Functions	Topic: 63 Pass: 1 Iteration: 1	(1) Markets; (2) Firms;	Market Makers - Desk Structure	
183	09 – Functions	Topic: 20 Pass: 2 Iteration: 1	(1) Markets; (2) Trading;	Market Trading	
184	09 – Functions	Topic: 37 Pass: 1 Iteration: 1	(1) Payments; (2) Systems & Technology;	Payment Systems	
185	09 – Functions	Topic: 8 Pass: 2 Iteration: 1	(1) Payments; (2) Systems & Technology; (3) Clearing; (4) Settlement;	Payment Systems - Clearing & Settlement	

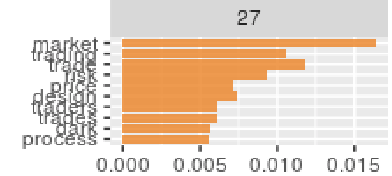
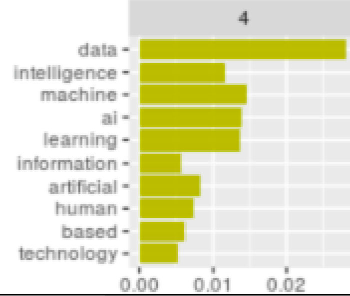
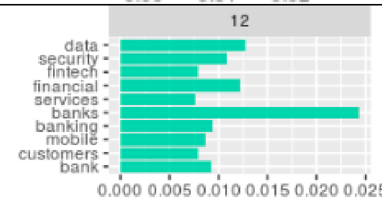
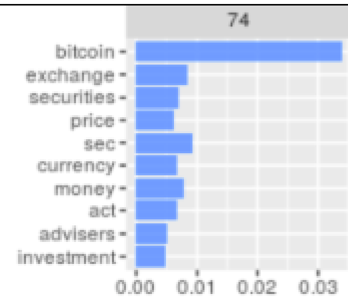
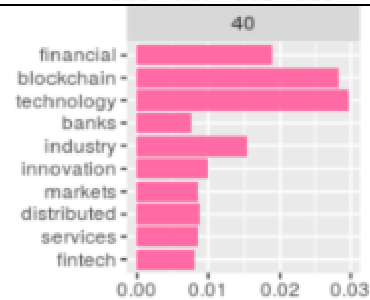
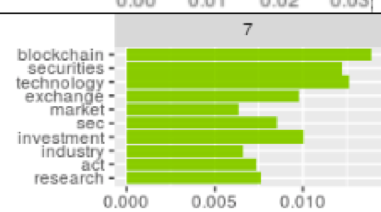
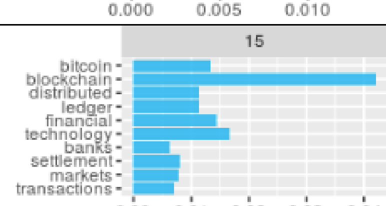
186	09 – Functions	Topic: 21 Pass: 2 Iteration: 1	(1) Payments; (2) Systems & Technology; (3) Clearing; (4) Settlement;	Payment Systems - Electronic Checks	
187	09 – Functions	Topic: 1 Pass: 2 Iteration: 1	(1) Portfolio Managers;	Portfolio Managers	
188	09 – Functions	Topic: 6 Pass: 2 Iteration: 2	(1) Retail; (2) Brokerage;	Retail Brokerage (FX)	
189	09 – Functions	Topic: 25 Pass: 2 Iteration: 1	(1) Risk Management;	Risk Management	
190	09 – Functions	Topic: 4 Pass: 2 Iteration: 1	(1) Trading; (2) Execution;	Trade Execution - Routing	
191	09 – Functions	Topic: 9 Pass: 1 Iteration: 1	(1) Trading;	Trading Desk	
192	09 – Functions	Topic: 23 Pass: 2 Iteration: 1	(1) Trading; (2) Execution;	Trading Execution	

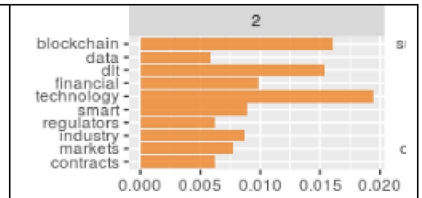
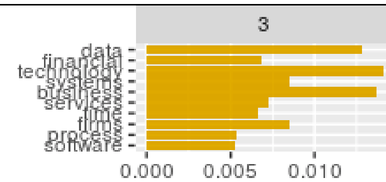
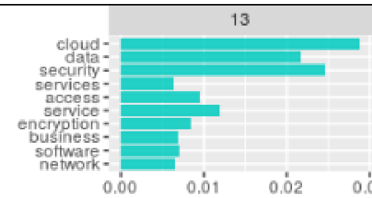
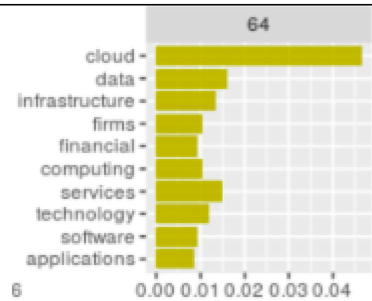
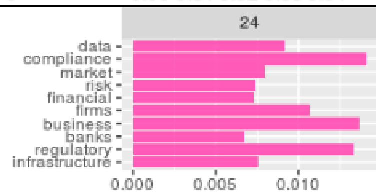
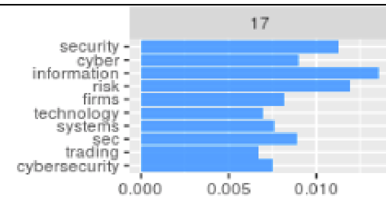
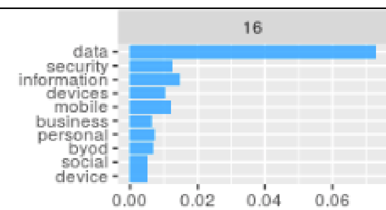
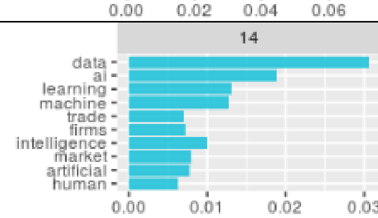
193	09 – Functions	Topic: 21 Pass: 2 Iteration: 1	(1) Trading; (2) Risk Management;	Trading Firm Risk Management	
194	09 – Functions	Topic: 73 Pass: 1 Iteration: 1	(1) Stress Testing; (2) Capital; (3) BCBS;	Bank Capital Stress Testing / Losses	
195	10 - Processes	Topic: 82 Pass: 1 Iteration: 1	(1) Process Transformation;	Business Process Transformati on	
196	10 - Processes	Topic: 7 Pass: 2 Iteration: 1	(1) Stress Testing; (2) Capital; (3) BCBS;	Capital Stress Testing / Losses	
197	10 - Processes	Topic: 15 Pass: 2 Iteration: 1	(1) Exchanges; (2) Central Counterparties; (3) Clearing;	CCPs (Central Counterparty) Clearing	
198	10 - Processes	Topic: 16 Pass: 2 Iteration: 1	(1) Processes; (2) Pricing;	Complex Process (Market Pricing / Investors)	
199	10 - Processes	Topic: 19 Pass: 2 Iteration: 1	(1) Process Transformation; (2) Data;	Data / Process Change	

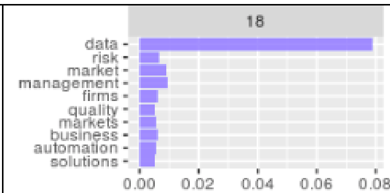
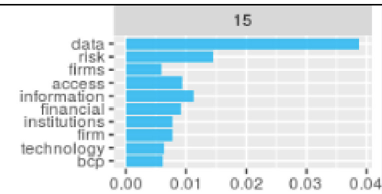
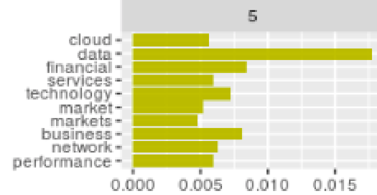
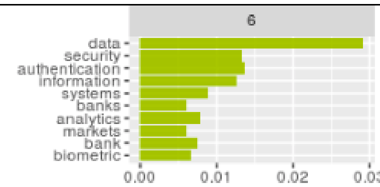
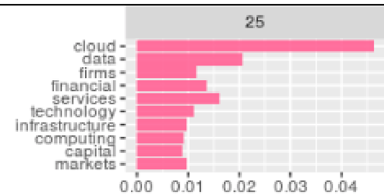
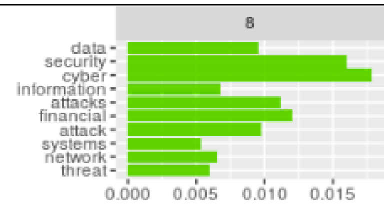
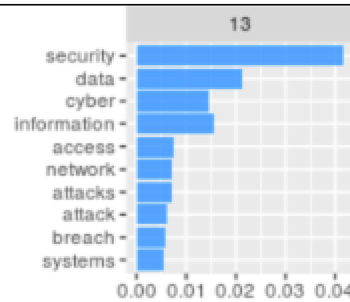
200	10 - Processes	Topic: 11 Pass: 2 Iteration: 1	(1) Data;	Data Management	
201	10 - Processes	Topic: 3 Pass: 2 Iteration: 1	(1) Data; (2) Risk Management;	Data Standards / Risk Management	
202	10 - Processes	Topic: 53 Pass: 1 Iteration: 1	(1) Derivatives Markets; (2) Clearing;	Derivatives Clearing	
203	10 - Processes	Topic: 61 Pass: 1 Iteration: 1	(1) Derivatives Markets; (2) Clearing; (3) Collateral;	Derivatives Clearing - Collateral Management	
204	10 - Processes	Topic: 5 Pass: 2 Iteration: 1	(1) Derivatives Markets; (2) Clearing; (3) Data;	Derivatives Clearing - Data	
205	10 - Processes	Topic: 4 Pass: 2 Iteration: 1	(1) Margin; (2) Swaps Markets; (3) Clearing;	Initial Margin / Swaps Clearing	
206	10 - Processes	Topic: 25 Pass: 2 Iteration: 1	(1) Markets; (2) Data;	Market Data Management	

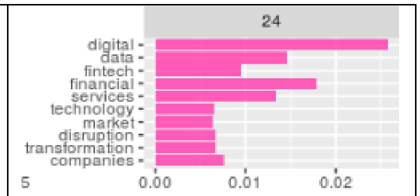
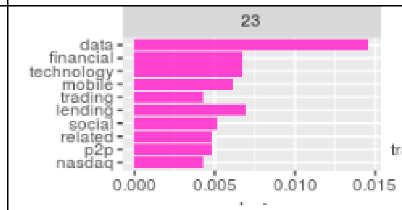
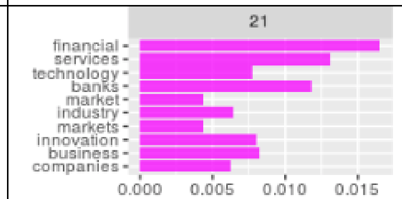
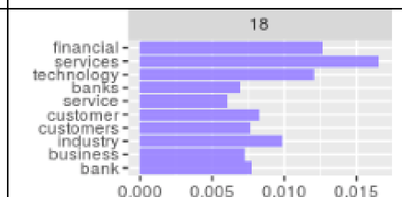
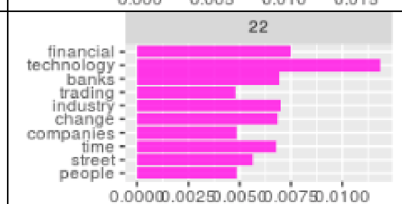
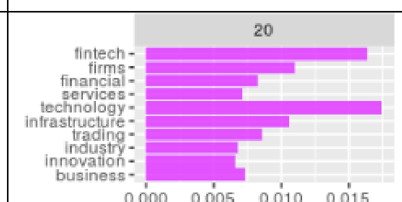
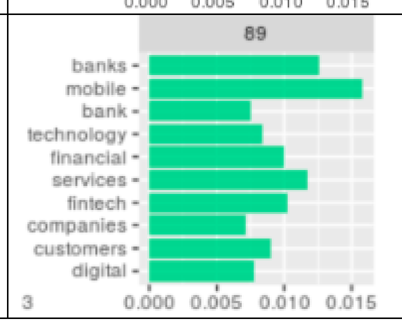
207	10 - Processes	Topic: 19 Pass: 2 Iteration: 1	(1) Markets; (2) Clearing; (3) Settlement;	Market Participants - Clearing & Settlement	
208	10 - Processes	Topic: 8 Pass: 2 Iteration: 1	(1) Middle Office; (2) Risk Management;	Middle Office - Risk Management / Accounting	
209	10 - Processes	Topic: 3 Pass: 2 Iteration: 1	(1) Data; (2) Processes;	Outsourcing Data Services	
210	10 - Processes	Topic: 37 Pass: 2 Iteration: 2	(1) Commodities Markets; (2) Position Limits;	Position Limits - Commodities	
211	10 - Processes	Topic: 10 Pass: 2 Iteration: 1	(1) Retail; (2) Brokerage; (3) Trade Routing;	Retail Trade Routing	
212	10 - Processes	Topic: 25 Pass: 2 Iteration: 1	(1) Risk Management; (2) Clearing; (3) EU Markets;	Risk - Clearing - European Markets	
213	10 - Processes	Topic: 13 Pass: 2 Iteration: 1	(1) Risk Management; (2) Controls;	Risk Management - Controls	

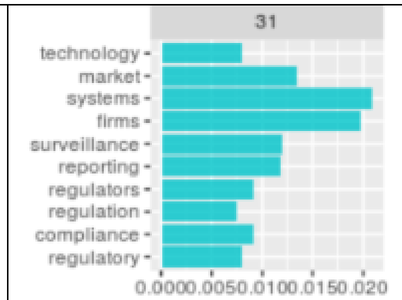
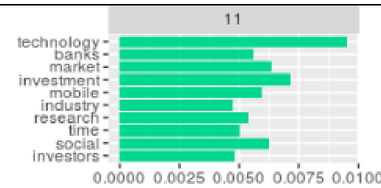
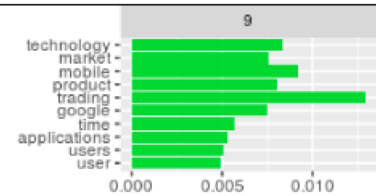
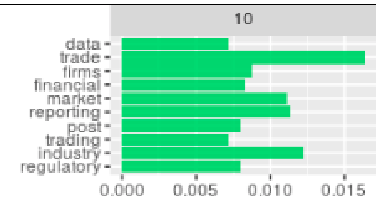
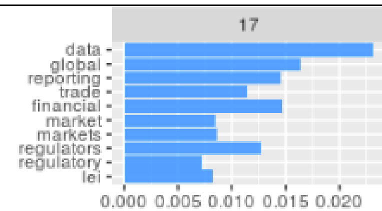
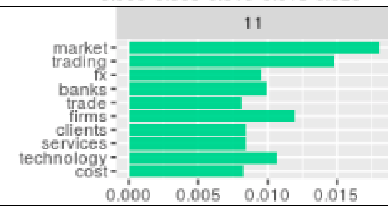
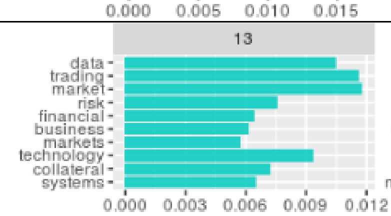
214	10 - Processes	Topic: 11 Pass: 2 Iteration: 1	(1) Risk Management; (2) Models;	Risk Management - Models (VaR)	
215	10 - Processes	Topic: 67 Pass: 1 Iteration: 1	(1) Risk Management; (2) Operating Model;	Risk Management Operating Model	
216	10 - Processes	Topic: 12 Pass: 2 Iteration: 1	(1) Stress Testing; (2) Capital; (3) BCBS;	Stress Testing / CCAR	
217	10 - Processes	Topic: 4 Pass: 2 Iteration: 2	(1) Swaps Markets; (2) Margin; (3) Pricing;	Swaps (Margin) Pricing	
218	10 - Processes	Topic: 9 Pass: 2 Iteration: 1	(1) Swaps Markets; (2) Clearing;	Swaps Clearing	
219	10 - Processes	Topic: 14 Pass: 2 Iteration: 1	(1) Swaps Markets; (2) Clearing; (3) Settlement;	Swaps Trading - Clearing & Settlement	
220	10 - Processes	Topic: 91 Pass: 1 Iteration: 1	(1) Settlement; (2) Systems & Technology;	Trade Settlement with Smart Contracts	

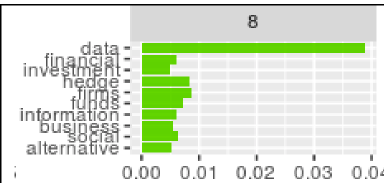
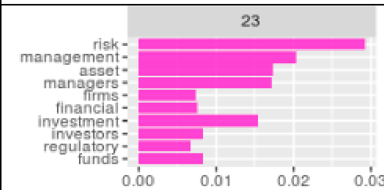
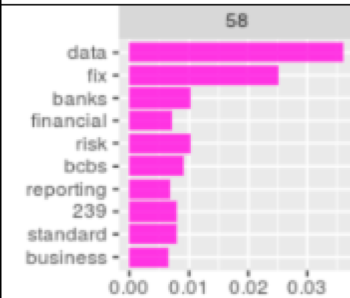
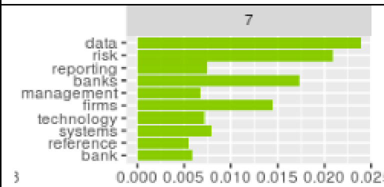
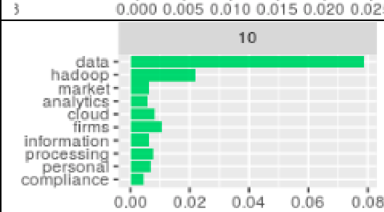
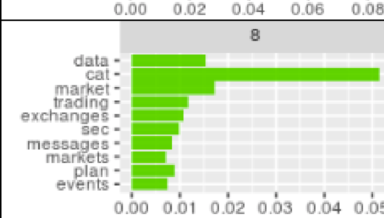
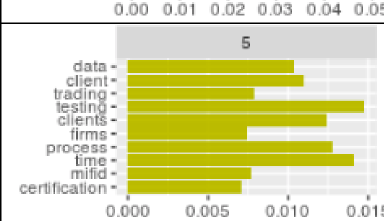
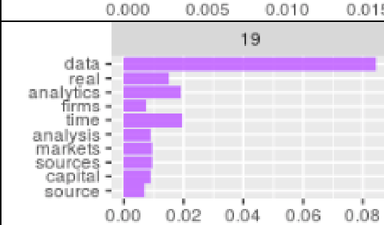
221	10 - Processes	Topic: 27 Pass: 2 Iteration: 2	(1) Trading; (2) Markets; (3) Processes	Trading Market Risk (Process Design)	
222	11 – Systems & Technology	Topic: 4 Pass: 1 Iteration: 1	(1) Systems & Technology; (2) Data;	AI & Machine Learning	
223	11 – Systems & Technology	Topic: 12 Pass: 2 Iteration: 1	(1) Systems & Technology; (2) Data; (3) Fintech;	Banks / Data / Fintech	
224	11 – Systems & Technology	Topic: 74 Pass: 1 Iteration: 1	(1) Systems & Technology; (2) Currencies;	Bitcoin	
225	11 – Systems & Technology	Topic: 40 Pass: 1 Iteration: 1	(1) Systems & Technology;	Blockchain	
226	11 – Systems & Technology	Topic: 7 Pass: 2 Iteration: 1	(1) Systems & Technology; (2) SEC;	Blockchain - Securities Exchange Commission	
227	11 – Systems & Technology	Topic: 15 Pass: 2 Iteration: 1	(1) Systems & Technology; (3) Settlement; (3) Data;	Blockchain - Settlement	

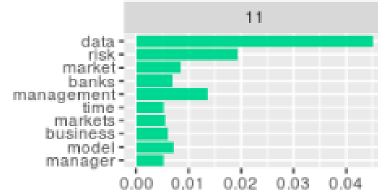
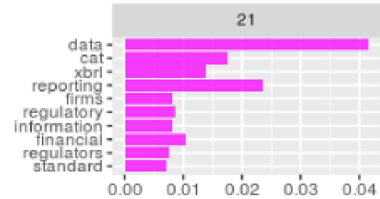
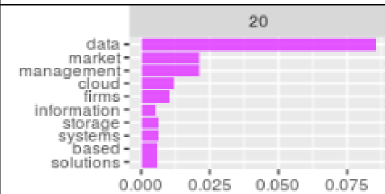
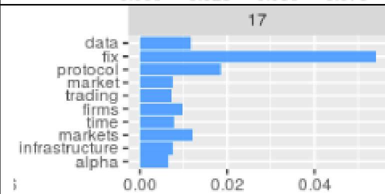
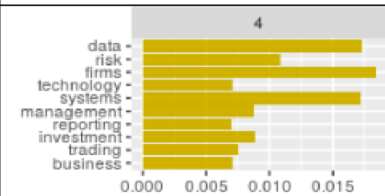
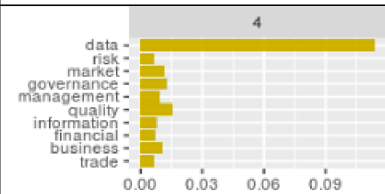
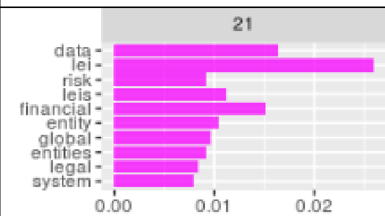
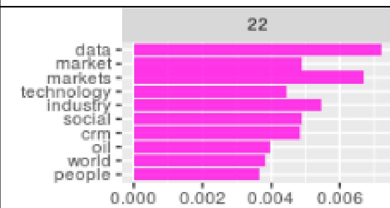
228	11 – Systems & Technology	Topic: 2 Pass: 2 Iteration: 1	(1) Systems & Technology;	Blockchain Technology	
229	11 – Systems & Technology	Topic: 3 Pass: 2 Iteration: 1	(1) Systems & Technology;	Business Financial Software	
230	11 – Systems & Technology	Topic: 13 Pass: 2 Iteration: 1	(1) Systems & Technology; (2) Data;	Cloud Data Security Services	
231	11 – Systems & Technology	Topic: 64 Pass: 1 Iteration: 1	(1) Systems & Technology; (2) Data;	Cloud Infrastructure / Data Services	
232	11 – Systems & Technology	Topic: 24 Pass: 2 Iteration: 1	(1) Systems & Technology; (2) Data;	Compliance Data Infrastructure	
233	11 – Systems & Technology	Topic: 17 Pass: 2 Iteration: 1	(1) Systems & Technology; (2) Cybersecurity;	Cybersecurity	
234	11 – Systems & Technology	Topic: 16 Pass: 2 Iteration: 1	(1) Systems & Technology; (2) Data;	Data - Mobile	
235	11 – Systems & Technology	Topic: 14 Pass: 2 Iteration: 1	(1) Systems & Technology; (2) Data; (3) Machine Learning;	Data / Artificial Intelligence / Machine Learning	

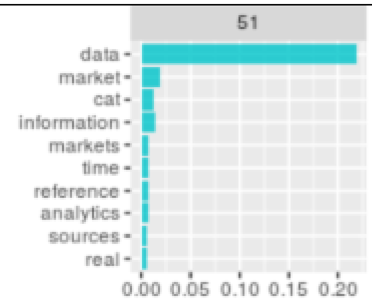
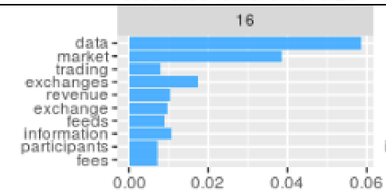
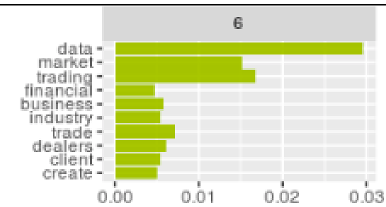
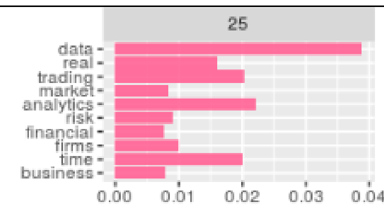
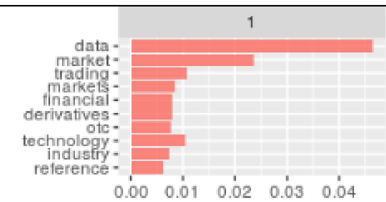
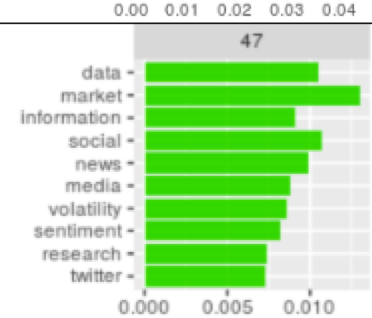
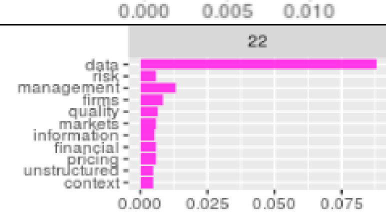
236	11 – Systems & Technology	Topic: 18 Pass: 2 Iteration: 1	(1) Systems & Technology; (2) Data;	Data / Automated Solutions	
237	11 – Systems & Technology	Topic: 15 Pass: 2 Iteration: 1	(1) Systems & Technology; (2) Data;	Data / Business Continuity	
238	11 – Systems & Technology	Topic: 5 Pass: 2 Iteration: 1	(1) Systems & Technology; (2) Data;	Data / Cloud / Network Performance	
239	11 – Systems & Technology	Topic: 6 Pass: 2 Iteration: 1	(1) Systems & Technology; (2) Data; (3) Infosec;	Data / Information Security	
240	11 – Systems & Technology	Topic: 25 Pass: 2 Iteration: 1	(1) Systems & Technology; (2) Data;	Data in the Cloud	
241	11 – Systems & Technology	Topic: 8 Pass: 2 Iteration: 1	(1) Systems & Technology; (2) Cybersecurity; (3) Data;	Data Security / Cyber Attacks	
242	11 – Systems & Technology	Topic: 13 Pass: 1 Iteration: 1	(1) Systems & Technology; (2) Cybersecurity; (3) Data;	Data Security / Cybersecurity	

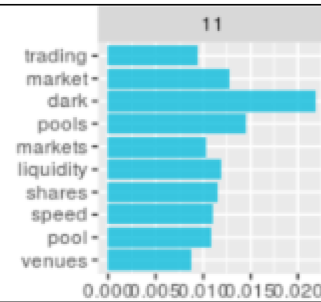
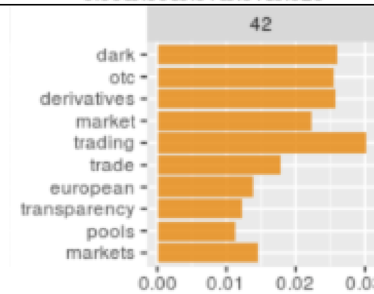
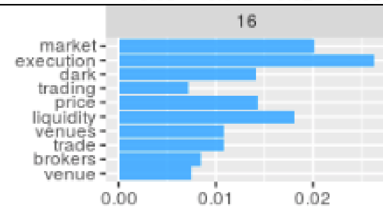
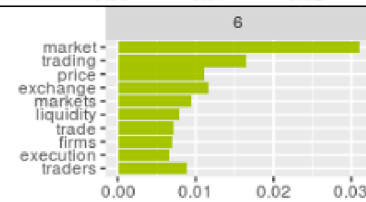
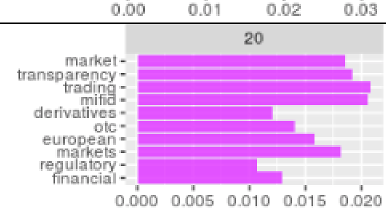
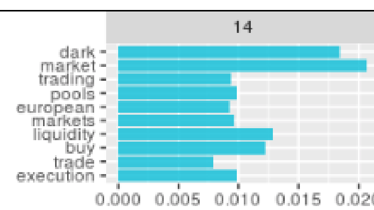
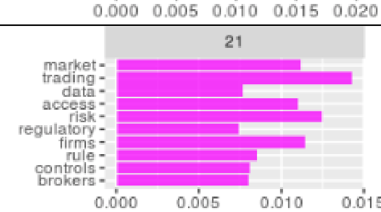
243	11 – Systems & Technology	Topic: 24 Pass: 2 Iteration: 1	(1) Systems & Technology; (2) Data;	Digital Transformati on / Data	
244	11 – Systems & Technology	Topic: 23 Pass: 2 Iteration: 1	(1) Systems & Technology; (2) Data; (3) Banks;	Financial Data - P2P / Social	
245	11 – Systems & Technology	Topic: 21 Pass: 2 Iteration: 1	(1) Systems & Technology; (2) Data; (3) Banks;	Financial Services / Bank Technology / Innovation	
246	11 – Systems & Technology	Topic: 18 Pass: 2 Iteration: 1	(1) Systems & Technology; (2) Data; (3) Banks; (3) Customers;	Financial Services Customer Technology	
247	11 – Systems & Technology	Topic: 22 Pass: 2 Iteration: 1	(1) Systems & Technology; (2) Data; (3) Banks; (4) Operating Model;	Financial Technology - Change in People	
248	11 – Systems & Technology	Topic: 20 Pass: 2 Iteration: 1	(1) Systems & Technology; (2) Data; (3) Banks; (4) Fintech;	Fintech / Technology / Innovation	
249	11 – Systems & Technology	Topic: 89 Pass: 1 Iteration: 1	(1) Systems & Technology; (2) Data; (3) Banks; (4) Fintech;	Fintech & mobile banking	

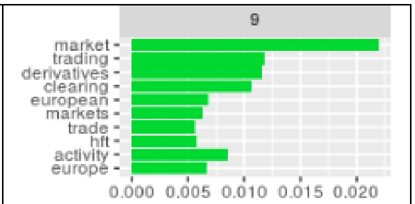
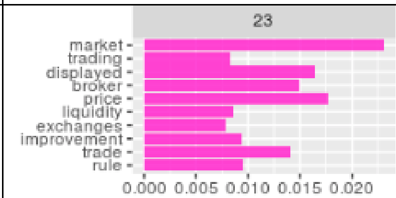
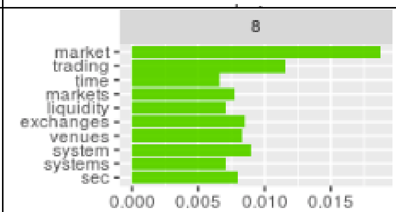
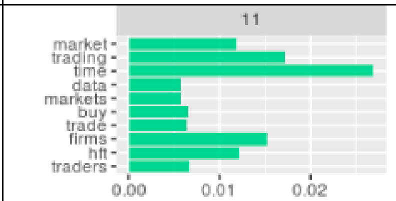
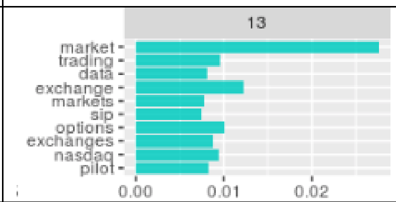
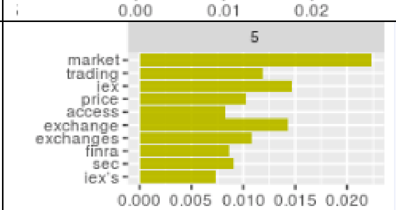
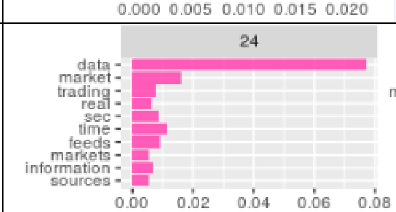
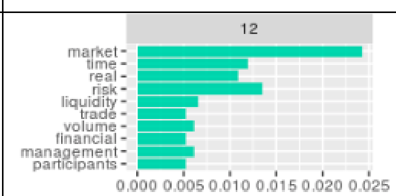
250	11 – Systems & Technology	Topic: 31 Pass: 1 Iteration: 1	(1) Systems & Technology; (2) Reporting; (3) Regulation;	Market Surveillance - Reporting Systems	
251	11 – Systems & Technology	Topic: 11 Pass: 2 Iteration: 1	(1) Systems & Technology; (2) Data; (3) Banks; (4) Fintech;	Mobile Technology / Investment	
252	11 – Systems & Technology	Topic: 9 Pass: 2 Iteration: 1	(1) Systems & Technology; (2) Data; (3) Banks; (4) Fintech;	Mobile Trading / Technology	
253	11 – Systems & Technology	Topic: 10 Pass: 2 Iteration: 1	(1) Systems & Technology; (2) Data; (3) Banks; (4) Trade Reporting;	Trade Reporting	
254	11 – Systems & Technology	Topic: 17 Pass: 2 Iteration: 1	(1) Systems & Technology; (2) Data; (3) Banks; (4) Trade Reporting; (5) LEI;	Trade Reporting (Global LEI)	
255	11 – Systems & Technology	Topic: 11 Pass: 2 Iteration: 1	(1) Systems & Technology; (2) Data; (3) Banks; (4) Trading;	Trading - Technology Costs	
256	11 – Systems & Technology	Topic: 13 Pass: 2 Iteration: 1	(1) Systems & Technology; (2) Banks; (4) Collateral;	Trading & Collateral Systems	

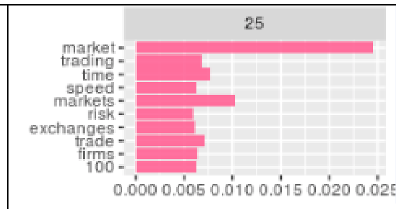
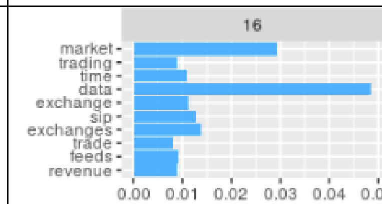
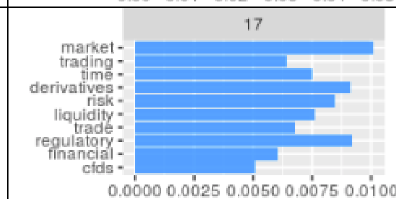
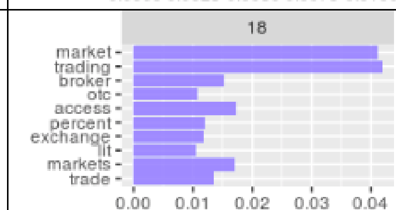
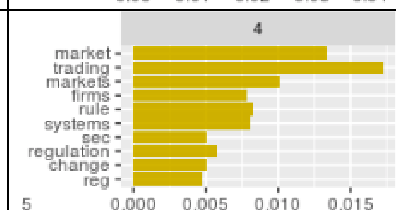
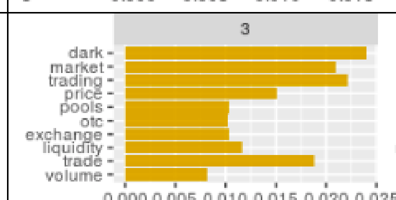
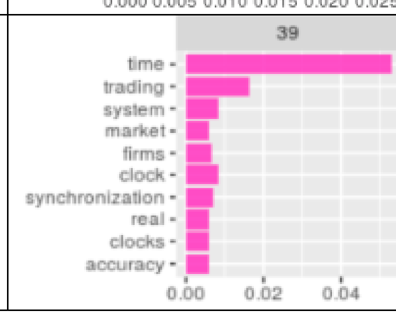
257	12 – Data	Topic: 8 Pass: 2 Iteration: 1	(1) Data;	Alternative Data (i.e., for Hedge Funds)	
258	12 – Data	Topic: 23 Pass: 2 Iteration: 1	(1) Data; (2) Asset Managers;	Asset Management Risk Data	
259	12 – Data	Topic: 58 Pass: 1 Iteration: 1	(1) Data; (2) BCBS; (3) Regulation;	Bank Data Quality for Reg Reporting	
260	12 – Data	Topic: 7 Pass: 2 Iteration: 1	(1) Data; (2) BCBS; (3) Regulation; (4) Reporting;	Bank Risk Reporting Data	
261	12 – Data	Topic: 10 Pass: 2 Iteration: 1	(1) Data; (2) Systems & Technology;	Big Data / Data Lake (i.e., Hadoop)	
262	12 – Data	Topic: 8 Pass: 2 Iteration: 1	(1) Data;	CAT / Data	
263	12 – Data	Topic: 5 Pass: 2 Iteration: 1	(1) Data;	Client Data	
264	12 – Data	Topic: 19 Pass: 2 Iteration: 1	(1) Data; (2) Systems & Technology;	Data / Real-time Analytics	

265	12 – Data	Topic: 11 Pass: 2 Iteration: 1	(1) Data; (2) Systems & Technology; (3) Risk Management;	Data / Risk Management	
266	12 – Data	Topic: 21 Pass: 2 Iteration: 1	(1) Data; (2) Reporting; (3) Accounting;	Data / Standardized Reporting - XBRL	
267	12 – Data	Topic: 20 Pass: 2 Iteration: 1	(1) Data; (2) Systems & Technology;	Data / Storage Management	
268	12 – Data	Topic: 17 Pass: 2 Iteration: 1	(1) Data; (2) Systems & Technology; (3) Trading;	Data / Trading / FIX Protocol	
269	12 – Data	Topic: 4 Pass: 2 Iteration: 1	(1) Data; (2) BCBS; (3) Regulation; (4) Reporting;	Data & Management Reporting Systems	
270	12 – Data	Topic: 4 Pass: 2 Iteration: 1	(1) Data; (2) BCBS; (3) Regulation; (4) Reporting;	Data Quality / Governance	
271	12 – Data	Topic: 21 Pass: 2 Iteration: 1	(1) Data; (2) Reference Data; (3) Counterparties;	LEI (Legal Entity Identifiers) Data	
272	12 – Data	Topic: 22 Pass: 2 Iteration: 1	(1) Data; (2) Markets;	Market Data	

273	12 – Data	Topic: 51 Pass: 1 Iteration: 1	(1) Data; (2) Markets; (3) Systems & Technology;	Market Data - Real Time	
274	12 – Data	Topic: 16 Pass: 2 Iteration: 1	(1) Data; (2) Markets; (3) Systems & Technology;	Market Trading / Data Feeds	
275	12 – Data	Topic: 6 Pass: 2 Iteration: 1	(1) Data; (2) Markets; (3) Systems & Technology; (4) Trading;	Market Trading Data	
276	12 – Data	Topic: 25 Pass: 2 Iteration: 1	(1) Data; (2) Markets; (3) Systems & Technology; (4) Trading;	Real-time Trading Data Analytics	
277	12 – Data	Topic: 1 Pass: 2 Iteration: 1	(1) Data; (2) Markets; (3) Systems & Technology; (4) Trading; (5) Reference Data;	Reference Data - Market Data	
278	12 – Data	Topic: 47 Pass: 1 Iteration: 1	(1) Data; (2) Markets;	Social Media Market News	
279	12 – Data	Topic: 22 Pass: 2 Iteration: 1	(1) Data;	Unstructured Data	

280	13 - Networks	Topic: 11 Pass: 1 Iteration: 1	(1) Trading; (2) Venues;	Dark Pools / Trading Venues	
281	13 - Networks	Topic: 42 Pass: 1 Iteration: 1	(1) Trading; (2) Derivatives;	Derivatives Trading - Dark Pools	
282	13 - Networks	Topic: 16 Pass: 2 Iteration: 1	(1) Trading; (2) Execution;	Market Execution - Venues	
283	13 - Networks	Topic: 6 Pass: 2 Iteration: 1	(1) Trading; (2) Systems & Technology; (3) Data; (4) Pricing;	Market Price - Access	
284	13 - Networks	Topic: 20 Pass: 2 Iteration: 1	(1) Trading; (2) Trade Reporting; (3) Data;	Market Trading - Market Transparency OTC Derivatives	
285	13 - Networks	Topic: 14 Pass: 2 Iteration: 1	(1) Trading; (2) Systems & Technology; (3) Dark Pools;	Market Trading - Dark Pools	
286	13 - Networks	Topic: 21 Pass: 2 Iteration: 1	(1) Trading; (2) Systems & Technology; (3) Data;	Market Trading - Data Access	

287	13 - Networks	Topic: 9 Pass: 2 Iteration: 1	(1) Trading; (2) Systems & Technology; (3) Clearing;	Market Trading - Derivatives Clearing Venues	
288	13 - Networks	Topic: 23 Pass: 2 Iteration: 1	(1) Trading; (2) Systems & Technology; (3) Data; (4) Pricing;	Market Trading - Displayed Price	
289	13 - Networks	Topic: 8 Pass: 2 Iteration: 1	(1) Trading; (2) Exchanges;	Market Trading - Exchanges / Venues	
290	13 - Networks	Topic: 11 Pass: 2 Iteration: 1	(1) Market Oversight (Regulation & Policy); (2) Trading & Execution; (3) Systems & Technology;	Market Trading - HFT	
291	13 - Networks	Topic: 13 Pass: 2 Iteration: 1	(1) Trading; (2) Options;	Market Trading - Options Exchanges	
292	13 - Networks	Topic: 5 Pass: 2 Iteration: 1	(1) Trading; (2) Systems & Technology; (3) Data; (4) Pricing;	Market Trading - Price Access - IEX	
293	13 - Networks	Topic: 24 Pass: 2 Iteration: 1	(1) Trading; (2) Systems & Technology; (3) Data;	Market Trading - Real-time Feeds	
294	13 - Networks	Topic: 12 Pass: 2 Iteration: 1	(1) Market Oversight (Regulation & Policy); (2) Risk Management; (3) Systems & Technology;	Market Trading - Real-time risk management	

295	13 - Networks	Topic: 25 Pass: 2 Iteration: 1	(1) Market Oversight (Regulation & Policy); (2) Trading & Execution; (3) Systems & Technology;	Market Trading - Speed - 100 mls	
296	13 - Networks	Topic: 16 Pass: 2 Iteration: 1	(1) Trading; (2) Data; (3) Exchanges;	Market Trading / Data Exchanges	
297	13 - Networks	Topic: 17 Pass: 2 Iteration: 1	(1) Market Oversight (Regulation & Policy); (2) Derivatives;	Market Trading / Derivatives	
298	13 - Networks	Topic: 18 Pass: 2 Iteration: 1	(1) Market Oversight (Regulation & Policy); (2) Brokerage;	Market Trading / OTC Broker	
299	13 - Networks	Topic: 4 Pass: 2 Iteration: 1	(1) Market Oversight (Regulation & Policy); (2) Trading & Execution;	Market Trading Rules	
300	13 - Networks	Topic: 3 Pass: 2 Iteration: 1	(1) Derivatives Markets; (2) Exchanges;	OTC Derivative Trading - Dark Pools	
301	13 - Networks	Topic: 39 Pass: 1 Iteration: 1	(1) Exchange; (2) Systems & Technology;	Trading Clock	

302	13 - Networks	Topic: 6 Pass: 2 Iteration: 1	(1) Exchanges;	Trading Markets - Access / Exchanges