A Tutorial on Approaching the Topic Modeling of Bank Regulation

By

Loretta Clare Ardaugh

Thesis Project
Submitted in Partial Fulfillment of the
Requirements for the Degree of

MASTER OF SCIENCE IN PREDICTIVE ANALYTICS

May, 2017

Alianna J. Maren, Ph.D., First Reader
San Cannon, Ph.D., Second Reader

Abstract

A Tutorial on Approaching the Topic Modeling of Bank Regulation

Loretta Clare Ardaugh

Regtech, a reference to the application of new technologies to bank regulation, mandates a conversation about reducing the burden of bank regulation by letting computers take over some of the handling of regulatory text. Bank regulations and the related manuals, guidance, or other supplements are mostly unstructured. Software tools and statistical models have evolved to "read" unstructured text and create actionable insights by way of "text analytics", but there are only limited cases of use and application within bank regulation. I contribute to this discussion by reviewing the text of regulatory guidance using text analytics tools. The model I employ seeks to determine the "topics" in which documents may be categorized. In this context, a "topic" division may be based on the bank activity to which the regulation applies, the regulator who authored the text, or even a time period in which the regulatory text was relevant. My objective was to appreciate whether the model could identify the first example - topics based upon the bank activity to which the regulatory text applied. I find that the model's "topics" are aligned with those of experts, plus are suggestive of a next level deeper than the experts' topics. However, the model is sensitive to changes in the formatting and word choices in the underlying text and processing choices applied. While the findings promise that there is opportunity in managing regulatory text with text analytics to create efficiency for the human implementers of regulation, they also show the importance of considering how the underlying text will affect the outcome. To that end, I recommend that creators of the text take the needs of text analytics work into account.

Abbreviations and acronyms list

Letter cases are modified in text analytics. Abbreviations and acronyms used here apply regardless of case. Some of the abbreviations and acronyms are found in the paper only in discussions of their place in lists of most probable terms.

BoW	bag of words
BSA	Bank Secrecy Act
DTM	document term matrix
FBO	Foreign Banking Organization(s)
Fed	Federal Reserve Board, Banks, or System
FFIEC	Federal Financial Institutions Examination Council
FIBO	Financial Industry Business Ontology
FinCEN	Financial Crimes Enforcement Network
Fintech	Technology in the Financial Industry, usually disruptive
FRBOG	Acronym used as an adjective referring to the Federal Reserve Board of Governor topic assignments
LDA	Latent Dirichlet Allocation
MRR	Market Risk Rule
OFAC	Office of Foreign Assets Control
Regtech	Technology in Regulation, usually disruptive
SLHC	Savings and Loan Holding Company(ies)
SRL	Supervision and Regulation Letter(s)

Table of Contents

Executive summary

Compliance with regulation is a critical function in a banking organization. The number of regulations that apply to a particular organization will be dependent upon its size, the activities in which it engages, and its geographic footprint. Its charter and structure will determine the number of regulators to which it is accountable. While regulators will collaborate to release a single regulatory document when there is shared responsibility, each may prepare a press release, FAQs, guidance or other material that accompanies the regulation. A banking organization may choose to monitor proposed regulations by monitoring any combination of Congressional discussions, media reports, proposals for comments, applicable comment letters submitted, and transcripts of speeches of regulators. Once a regulation is implemented by regulators and banking organizations, supervision of banking organizations includes testing compliance with the regulations and that may include reviewing banking organizations' policies, procedures, meeting minutes, and other text documents, often a repeated process at a specified interval of time. A current initiative of regulators is to tailor regulation by create different versions of regulation according to the activities and complexity of the banking organizations. At the same time as tailoring will reduce the regulatory burden overall, it increases the amount of regulatory text a banking organization must assess for applicability.

Text analytics offers great possibility for reducing the burden of regulation by automating the process of ingesting regulations and related text. Software tools have simplified the process of transforming text to "tokens", where tokens may be a single word, a multi-word noun phrase, or any combination of text. Readings in text analytics offer a

variety of algorithms which can be used to elicit insights from a file of tokens. Text analytics is an important tool in Regtech, a movement to improve the regulatory process through technology.

This paper shares a process of how one may approach the application of text analytics to ingest regulation by demonstrating some of the choices involved and how they impacted the outcome in this small-scale effort. The text used is a sample of the Federal Reserve (Fed) Supervision and Regulation Letters (SRLs), a series of guidance on supervisory policies and procedures. The sample includes 125 SRLs that were 1) issued inclusive of and between 2006 and 2016 and 2) were in active status as of December 31, 2016. I explored the classification of SRLs as a proxy for the myriad of regulatory communications and some of the related text submitted by banking organizations to show compliance.

For classification, I used a Latent Dirichlet Allocation (LDA) model, a probabilistic model that attempts to replicate the generation of the submitted text in order to group documents that have been generated from the same process into a topic. The LDA model requires a bag of words (BoW) as input. A BoW is a file of tokens and usually the tokens' frequency, but may make include various weighted versions of tokens. I apply LDA four times to show the effects of changes in output across changes in the BoW. I use the simplest creation of a BoW in the first run, using all of the text in an SRL PDF. I then tweak the BoW to show the effect on LDA outcomes from a couple simple string changes. In a third run of LDA, I use a BoW which includes only nouns and adjectives extracted from the letter body of the SRL. And then in a final run, the BoW is manipulated to show effects of

modifying a common practice in which tokens are “stemmed”, or replaced by the word stem.

In my work, I find that LDA based upon the simplest BoW produces strong results when I compare the model classifications to expert classifications. However, tweaking the BoW provided useful results as well in addition to more meaningful terms used to define the classifications. Unlike regressions and other traditional models, LDA does not yet have a dominant framework of measures of conceptual soundness. Human topic review most often serves as the measure of stability and sensitivity. By comparing four cases, I show sensitivity of the LDA output to choices made either as the BoW is created or as tokens are managed. I conclude that the choices in formatting are important and that text analytics of regulatory text would be facilitated by regulators’ consideration of a need for landmarks within the text and overall consistency when creating regulatory communications.

Chapter 1. Introduction

This paper considers an application of text analytics to the text of a banking regulator's supervisory guidance series. The intent is to introduce text analytics as a tool to facilitate more efficient regulatory compliance by enabling speedier ingestion of supervisory guidance. Perhaps banks and regulators will work under a common ontology, where regulatory material can speedily find its home in the bank, not just for the benefit of those responsible for its implementation, but also within the banks' technology systems to relate banks' policies to regulation or even risk processes to regulation, kicking off internal approvals and presentations and even submissions of information to regulators.

Specifically, I conducted exploratory text analysis, feature selection, and topic modeling using the text of the Federal Reserve (Fed) Supervision and Regulation Letters (SRLs). SRLs are issued by the Fed in accordance with its mission "to ensure the safety and soundness of the nation's banking and financial system and to protect the credit rights of consumers"[1]. SRLs include policies and procedures and are targeted at either or both of the Fed's staff responsible for supervising regulated entities and the regulated entities' staff responsible for complying with applicable regulations. My analysis is conducted entirely in R in an RStudio environment using a myriad of R packages which support text manipulation and analytics.

Section 1.1 Statement of opportunity

There are many opportunities to employ text analytics in the banking business as well as in the related business of regulating banks and central bank activities generally. Existing banking applications of text analytics are varied in purpose and design but, as is the case in many industries, do not yet realize the full potential of text analytics. To identify the state of text mining in the financial industry, (Kumar and Ravi 2016) conducted a survey of 89 relevant papers or conference presentations from various literature sources, including Springer and Elsevier. They found that the 89 applications of text analytics fall into the business categories of stock or foreign exchange rate prediction, cyber fraud identification, and customer relationship management and that these applications employed varied techniques, including classification, sentiment analysis, and clustering. Literature shows research is expanding in the application of text analytics beyond the areas identified in (Kumar and Ravi 2016). A particular exploration of interest for this paper is work related to a call to move beyond the text analytics conducted in just one organization to instead create a framework for shared text analytics applications which will facilitate greater efficiency in processes that are shared between banks and their regulators. I will add to that exploration by conducting text analytics on a regulatory guidance communication series. Next, I set the context for my study by providing a high-level explanation of bank regulation.

Banking is a heavily regulated industry. Regulation defines the activities in which banks may engage and how the regulation of those activities will be conducted. Regulation is often supplemented with guidance and the text I study in this paper, SRLs, represents just one series of guidance. The language of the SRLs is formal, which lends itself to an automated

text process, e.g., a misspelling would be extremely rare (the concept of formality is further explored in Appendix 1).

The text of regulation is not static. Following are a few examples of how regulations changed in response to changes in the business of banking or changes in the environment in which banks operate. In recent decades, technology and economic conditions accelerated the rate of change in how banking is conducted and banks' operations and balance sheets display a commensurate increase in complexity. Regulation responded to that increase in complexity. Regulation also must respond to pressures on the banking industry, such as occurred following the recent financial crisis when awareness of taxpayer exposure to the costs of failed banks expanded. Regulatory changes may be demanded by the financial industry. As the differences between a small bank conducting a basic banking business and a large bank providing global financial services continues to increase, a movement towards "tailored" regulation has begun. Tailored regulation results in multiple versions of supervisory text according to the characteristics of banks to which it is applicable.

The amount of text is further multiplied by existence of multiple bank regulators. Due to the evolution of the United States' banking system, there are many regulators to whom banking organizations must respond, depending on the charter and structure of the organization, its size, and the activities in which it is engaged. Prudential federal regulators of banks in the U.S. include the Fed, the Federal Deposit Insurance Corporation, and the Office of the Comptroller of the Currency.[2] In addition, there are state regulators that are defined by each state. The lines of regulatory authority are explained in two reports referenced for this paper, (GAO 2016) and (Murphy 2015). A final version of a regulation is communicated by each of the applicable

regulators, along with each regulator's versions of related material including manuals and procedural letters – all of which must be assessed by a regulated entity for applicability.

Ingesting these text communications is time-consuming for the regulated entities. In historical periods of slow regulatory changes and/or relatively smaller banking organizations, a banking organization may have had a regulatory officer with a small unit of staff who read the information and routed the text to those responsible for implementation – that model is no longer appropriate in banking organizations as the speed and quantity of text increases. A well-designed implementation of a text analytics program within a bank may increase a banks' capacity to ingest material not previously ingested, identify hidden topics in the regulatory text, provide a map from incoming regulatory text to existing bank systems, and/or facilitate the creation of derivative information for targeted operational areas. On a macro level, text analytics may be used to help build a framework through which SRLs and other regulatory text may be consistently identified, managed, and differentiated according to the entities or activities covered. That level will require consideration of the text analytics process during the state of creating regulatory text.

In describing the objective of my text analytics effort, I refer to (Miner et al. 2012). Authors describe seven practice areas of text analytics: Information Extraction, Natural Language Processing, Concept Extraction, Web Mining, Information Retrieval, Document Clustering, and Document Classification[4]. (Miner et al. 2012) provides a decision tree to determine the primary practice of interest. The tree identifies my effort as "Document Classification" although opportunities to aid ingestion of regulatory materials exist in each practice.

Section 1.2 Justification

(Gattuso and Katz 2016) on the Heritage Foundation's website tell us that between 2009 and 2015 "federal agencies issued 229 new major regulations that increased burdens, and only 26 reductions". Costs and benefits are discussed in any consideration of a new regulation. The opportunities offered by text analytics could impact that discussion as text analytics holds the potential of reducing costs of implementing regulation in the long-term.

Fintech (short for technology in finance) has become a well-known term in the banking industry and refers to disruptive technology in the financial industry. Regtech is a sibling to Fintech and the term's use has been spreading over the last couple of years. In (van Liebergen et al. 2016), the Institute of International Finance (IIF) defines Regtech as "the use of new technologies to solve regulatory and compliance requirements more effectively and efficiently" and as "a niche market, requiring collaboration between unlikely partners: regulators and regulatory experts, technology and software developers, and entrepreneurs willing to invest."

An extensive amount of text is exchanged between regulators and the regulated entities. Consider the annual stress test process that ensures banks' have adequate capital to withstand challenging economic scenarios. Text analytics offers the possibility of using the regulators' text to set up a framework for creation of the qualitative text submission required for compliance with the regulation. Perhaps it may be so simple as ensuring the language of the regulation and the banks' labels for policies and procedures match, and then creating an automated tool could read new regulation and either highlight areas of policies and procedures that may require change or extract the policies and procedures for creation of a submission packet. Perhaps upon receipt of new regulatory text, a bank may, in an automated fashion, use text analytics to identify whether the regulation requires changes to its activities through evaluation of a profile of a bank

against the regulatory text and where there is a relationship, using the text of regulation to extract related text from a database of a banks policies, procedures, meeting presentation templates, etc. I focus on using text analytics to classify the text of regulatory guidance to the end of appreciating how the text itself affects the success of the classification, just a small piece of the work of Regtech. To further support exploring the application of text analytics to bank regulatory information, I have collected quotations from several of the collaborators in Regtech.

Publication date: 2016
Publisher: Deloitte
Publication title: RegTech is the new FinTech
How agile regulatory technology is helping firms better understand and manage their risks

Relevant section heading: What Is RegTech & Why Do We Need It?
Link: Deloitte Reference

Quote

> *Kent Mackenzie (Deloitte Director, Edinburgh) sees a significant opportunity for so-called RegTech providers to bring clarity and efficiency into the way in which regulation is interpreted, how compliance is managed and most of all how reporting is and will be automated. The use of cognitive technologies and enhanced analytics is beginning to help the industry rapidly and automatically understand not just explicit meaning from regulation but also the implicit meaning or 'nuance' that is so often a challenge to digest and assess. As we all know data is meaningless unless it is organized in a way that enables people to understand it, analyse it and ultimately make decisions and act upon it i.e. by creating consumable information. In recent work Kent has undertaken for clients in deploying RegTech solution, they have been able to identify the '1 to many' relationship for the first time, i.e. where 1 control satisfies many regulations, or where a single regulatory paragraph requires many multiple controls.*

Publication date: March, 2016
Publisher: Institute of International Finance
Publication title: Regtech in Financial Services: Technology Solutions for Compliance and Reporting
Relevant section heading: "I. Regulatory and Reporting Requirements That Would Benefit From Regtech"
Link: IIIF Reference
Quote:

7. Making financial institutions more aware of regulatory developments
Identifying new regulations applying to the organization, flagging their potential implications, and allocating the accompanying reporting and compliance obligations to the right organizational units is a complex task requiring significant capacity and human resources to interpret the regulations. Large FIs operating in multiple jurisdictions are faced with local, regional and global regulations that are constantly changing. It is challenging to keep track of the different regulations being promulgated, especially since regulators publish new regulations in different formats. Analyzing how regulations compare to each other, and on which points they are consistent, and then applying obligations in a coherent way within the institution is a particular challenge.

Publication date: July, 2016
Publisher: Financial Conduct Authority
Publication title: Feedback Statement FS16/4 Call for input on supporting the development and adopters of RegTech
Relevant section heading: "What RegTech could be introduced?", subheading
Link: FCA Reference
Quote

Integrate, standardise and understand
Technology that drives efficiencies by closing the gap between intention and interpretation
Semantic tech and data point models
Technology that converts regulatory text into a programming language.

- *Machine-readable regulation would allow more automation and could significantly reduce the cost of change.*
- *It could also help ensure greater consistency between the intentions of a regulation and its implementation.*

Shared data ontology
A formal naming and definition of the types, properties, and interrelationships of entities.

- *Sharing a common understanding of the structure of regulatory data would improve efficiency, reduce costs, ease interactions and help remove ambiguity.*

...

New directions
Technology that allows regulation and compliance processes to be looked at differently (please note that this is not an exhaustive list)
Inbuilt compliance
Regulatory requirements can be coded into automated rules applied when relevant.

A system that can automatically apply the regulatory 'programme code' would improve compliance, reducing regulatory and staff costs.

Publication date: 2016
Publisher: EYGM Limited
Publication title: Innovating with RegTech – Turning regulatory compliance into a competitive advantage
Relevant section heading: RegTech in practice
Link: EY Reference
Quote

Regulatory compliance automation
Future RegTech platforms will be used to interpret regulations, including upcoming changes.

Publication date: January, 2017
Publisher: KPMG N.V.
Publication title: Will regtech save us from regulations?
Relevant section heading: There are hardly any regtech early adopters. What do you expect from a technology that has not yet proven itself?
Link: KPMG Reference
Quote

(AnneMarie) Smit: "What do all the regulations say, what do I have to comply with and how do all these various rules inter-relate? Nowadays, staff are manually comparing thousands of pages of regulations. No one's really arguing against the fact that technology can do a better job."
(Rens) Rozekrans: "We're in the process of proving it ourselves. Currently we're working on a service for customers which will connect KPMG's expertise with that of IBM Watson. Our ambition is to market a cognitive system, within a year, that will map various legislation in detail for organisations. It's a complex system that requires significant investment, but we can keep the costs down because basically we'll be able to offer it to any organisation."

Publication date: January, 2017
Publisher: Chris Skinner (from his biography: "Mr. Skinner is a regular commentator on BBC News, Sky News, CNBC and Bloomberg about banking issues; he is a Judge on many awards programs including the Asian Banker's Retail Excellence Awards, as well as working closely with leading banks such as HSBC, the Royal Bank of Scotland, Citibank and Société Générale, as well as the World Economic Forum.")
Publication title: Chris Skinner's Blog
Relevant section heading: The Semantic Regulator (#Regtech Rules)
Link: The Finanser Reference

Quote, but first please note an acronym used, "FIRO" – it is the Financial Industry Regulatory Ontology actively worked on by Ireland's Governance, Risk and Compliance Technology Centre, as described by Mr. Skinner.

> *...thus, a regulatory ontology such as FIRO can help:*
> *Financial services companies to monitor, assess, and apply a multitude of regulations within and across regulatory domains to business processes and data;*
> *Model the regulations to help simplify their consumption;*
> *Make it simpler for enterprises to map GRC policies onto regulations and perform Regulatory Change Management;*
> *Help organisations keep abreast of the ramifications of complex interacting regulatory rules and policies;*
> *Reason over regulations to identify risks and compliance issues;*
> *Contribute to the emergence of SMART Regulation.*
> *There's quite a lot more on this area if interested, as it's all about the rise of RegTech.*

In summary, I have presented the context for my study and a strong justification for the work of applying a topic modeling technique to supervisory guidance. My work will provide useful information for the implementation of a text analytics application to ingest regulatory communications on a small scale in addition to providing considerations for creators of the communication as they make formatting choices.

Chapter 2. Review of literature

Section 2.1 The application of text mining in central bank activities

In 2015, the Bank of England published *Handbook - No. 33 Text Mining for Central Banks* (Bholat et al. 2015). The focus of the publication is on "unsupervised machine learning techniques because they resonate with the Bank's evolving 'big data' ethos." As a justification of need for their handbook, the authors cite two interesting works. The first (Bank of England 2015) is a discussion paper in which authors discuss how machine learning techniques applied to social media can measure uncertainty about the economy, explain how policy communications are interpreted, and improve the collection of economic indicators. Of relevance to my paper is the Handbook's authors' discussion of using text mining to elucidate interactions between monetary and supervisory policies through measuring "textual interconnectedness" as a proposed regulation is considered and note that the measure "could help quantify ex ante the potential adverse interactions between monetary, macro-prudential, and micro-prudential changes." Another interesting example included in their report is (Li et al. 2015) in which text mining was used to measure changes in overall size and word additions and deletions over time in the United States Code. The authors conclude with a discussion of the process of text mining and a description of models utilized.

The Fed also has turned its eye to text mining. With respect to monetary policy, I identified articles in which sentiment analysis was performed. On one end, (Cannon 2015) shows text mining with and without sentiment analysis in order to better understand how monetary policy discussions may be or have been interpreted. On the other end, (Sinha 2014) and (Shapiro, Sudhof, and Wilson 2017) have published studies on how text mining news with sentiment analysis can be used to form features for use in econometric models. With respect to the Fed's supervision and regulation duties, (Goldsmith-Pinkham, Hirtle, and Lucca 2016)

published insights from text analytics of supervisory findings. Insights into the findings were gained through the application of the Latent Dirichlet Allocation (LDA) topic modeling technique and then the topic assignments were used as a feature in a regression to understand the relationship between examination findings and other characteristics of bank health.

A working paper from Bank of Canada (Hendry and Madeley 2010) uses a topic model to identify themes in various monetary policy communications. The authors found overlap among themes and used a method of regressing a theme on a following theme in an iterative process until they found a final set of orthogonal themes. Themes were then used as features in econometric models.

In response to the implementation of the Single Supervisory Mechanism in Eurozone countries in 2014, (Nopp and Hanbury 2015) consider sentiment as an aid to fill the need for expanded information about banks' risk appetites. In a first case, the authors extract CEOs' annual letters and outlook articles from banks' annual reports, apply sentiment tags to indicate whether words are positive, negative, or uncertain, and compare these outcomes to the fluctuation magnitude of the tier 1 capital ratio (a capital measure that accounts for risk via a risk-weighted asset denominator) on an aggregated basis. They also regress the fluctuation of the tier 1 capital ratio on the negativity sentiment score and find it is significant, but note that additional analysis indicated the findings do not hold at an individual bank level. In a second case, the authors label the extracted texts according to the fluctuation direction of the tier 1 capital measure, eliminate words that are too common or not informative in the model, and use the remaining information to train a model to classify text according to the future fluctuation direction. The authors do not find this technique successful when it is compared to a non-

modeled approach. The authors conclude that sentiment work may provide valuable insights into the macroprudential environment for supervisors when additional research has been done.

Section 2.2 Techniques of text analytics

(Blei 2012) explains the multitude of considerations to address before running LDA. Blei tells us we will be computing the posterior probabilities and describes them as "the conditional distribution of the topic structure given the observed documents". He explains the intractability of computing the denominator of the posterior, labeled the marginal probability of the observations and described as "the probability of seeing the observed corpus under any topic model". He explains that while other methods approximate the posterior with sampling methods, such as a Markov chain, the 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. Thus the inference problem is transformed to an optimization problem." He describes a number of extensions to LDA to address corpuses that violate assumptions, such as a correlated topic model which seeks to account for correlation between topics. Finally, he tells us about future directions, such as are fleshed out further in papers discussed below.

(Zhu, Blei and Laffterty 2006) posits that LDA as originally set out will not accommodate the inclusion of domain knowledge. They propose "tagLDA" as a potential solution, where "tag" may refer to domain-specific tags, part of speech tags, html tags, or even a tag for the section of the document from which the word was taken. They note their paper does not address tags for "higher order" terms such as bigrams and that one way of handling the tags may be to build a k-topic model for each group of terms in which we are interested. However, they then present a version of LDA which will account intra-process for the tags by accounting for the tag as the word probabilities are determined. Finally, the authors discuss the lack of advantage of tagLDA for document classification and specify future work to support the incorporation of domain knowledge into the LDA process.

(Handler et al. 2016) considers multiple tools to support extraction of noun phrases and begin with a case for moving beyond the unigram bag of words (BoW) because unigrams will not "preserve meaningful multiword phrases". The authors propose a pattern-based extraction method which relies upon existing tag applications, using a definition of noun phrase as defined in the paper. Their technique takes into account coordinating conjunctions as well as parenthetical post-modifiers (their example is a 4-gram, "401(k)") and numerical modifiers. Term sequences are tagged with pairs of start and end symbols. The authors found their method provided phrases that were "less ambiguous and more interpretable than unigrams".

(Chuang, Manning, and Heer 2012) considers the issue of keyphrase utilization versus unigram. Their study is focused on how humans perceive visualizations of text, such as word clouds. They distinguish between work done to facilitate search effectiveness/information retrieval and work done to facilitate "document understanding". They consider the risk of creating nonsense when we employ trigrams. In a study of descriptions of text formed by humans, the authors observed humans' use of primarily multiword terms, especially noun phrases. The humans who participated in their study also gravitated to mid-frequency terms, rather than most or least frequent terms. Humans' descriptions also varied according to the number of documents being reviewed and the terms humans selected were not randomly positioned. The study authors used this information as input to a logistic regression to model "keyphrase quality" and went on to build an algorithm to create word clouds according to a number of factors.

(Chang et al. 2009) discusses the lack of a standard measure when the goal of topic modeling is exploration in an unsupervised fashion rather than predictive. They tell us "While this common latent space has explored for over two decades, its interpretability remains

unmeasured." Two human-based assessments were designed to measure the quality of topics and assignments. The first task is labeled *word intrusion*, and "measures how semantically 'cohesive' the topics inferred by a model are and tests whether topics correspond to natural groupings for humans". The second task, *topic intrusion*, "measures how well a topic model's decomposition of a document as a mixture of topics agrees with human associations of topics with a document". The authors' experiment with human detection of word and topic intrusion validated the use of applying topic modeling to group documents into topics and the authors note a possibility of someday progressing to a "computational proxy that simulates human judgments."

Section 2.3 Specific applications of text analytics

(Lau, Law, and Wiederhold 2005) describe government regulations as "semi-structured text documents that are often voluminous, heavily cross-referenced between provisions and even ambiguous". In their paper, they discuss similarity analysis of government regulations in general, including both state and federal regulations, along with the handbooks and guidance related to the regulation. The authors propose a design for a tool which would be made widely available and used to identify similarities in regulations which may otherwise be "hidden" from a reader of the regulations. Their tool would take advantage of the "natural hierarchy" in regulations and support the regulated entities' need to identify material related to regulatory concepts and access linked features, such as exceptions and measurements. The authors find their system identifies "'hidden" similarities among regulations. Their future plans include using the same technology to identify conflicts among regulations.

In a novel application of the topic model, (Doyle and Elkan 2009) proposes financial topic models, in which the topics demonstrate the network relationship of publicly-traded companies. A "topic" will include identifiers for companies assigned to the topic. Unlike the data challenges that exist when measuring systemic risk via company relationships (e.g., loans), the stock price data used by the authors is more readily available. The grouping of companies as a "topic" may facilitate the identification of the "network" and improve the understanding of interconnectedness which increases systemic risk in the financial system.

Section 2.4 Ontologies dealing with bank supervision

(Bonsón-Ponte et al. 2009) discusses the increasing complexity of banking supervision in the European Union and focuses primarily on the qualitative and quantitative information required from banking organizations by supervisors. They present a case for creating a shared ontology to facilitate communication between supervisors and supervised. They posit that benefits will include simpler re-use of information to prove compliance and the creation of a record of decision processes which will facilitate the use of decision technologies. With respect to the latter, they consider the need for support of an ontology as not just a role of the organizations' technology functions but also potentially requiring overall cultural transformation in the organization.

(Bennett 2013) introduces the Financial Industry Business Ontology (FIBO). He refers to FIBO as a step to resolve the "reconciliation hell" of data management in the financial services industry. The author posits that a common vocabulary is not a solution in the problem of integrating data derived from different systems and encourages a conceptual model which supports the linkage between business requirements and deliverables. Existing web language principles will be enhanced to support the FIBO. Bennett concludes that implementation of the FIBO with the incorporate unambiguous shared meaning of concepts will be of value in the management of the systemic risk which is of concern to regulators.

(McCarthy 2013) tackles a specific case of the lack of an unambiguous shared language. He calls for technology based on language used by both market participants and regulators which will support real-time identification of high-frequency trading irregularities. As others have done, McCarthy notes the important role of regulation in the stability of the financial system and

the corresponding mounting complexity of the regulation. In his case for a shared language in regulation, he presents a visualization of the regulatory structure, copied here in Figure 1.

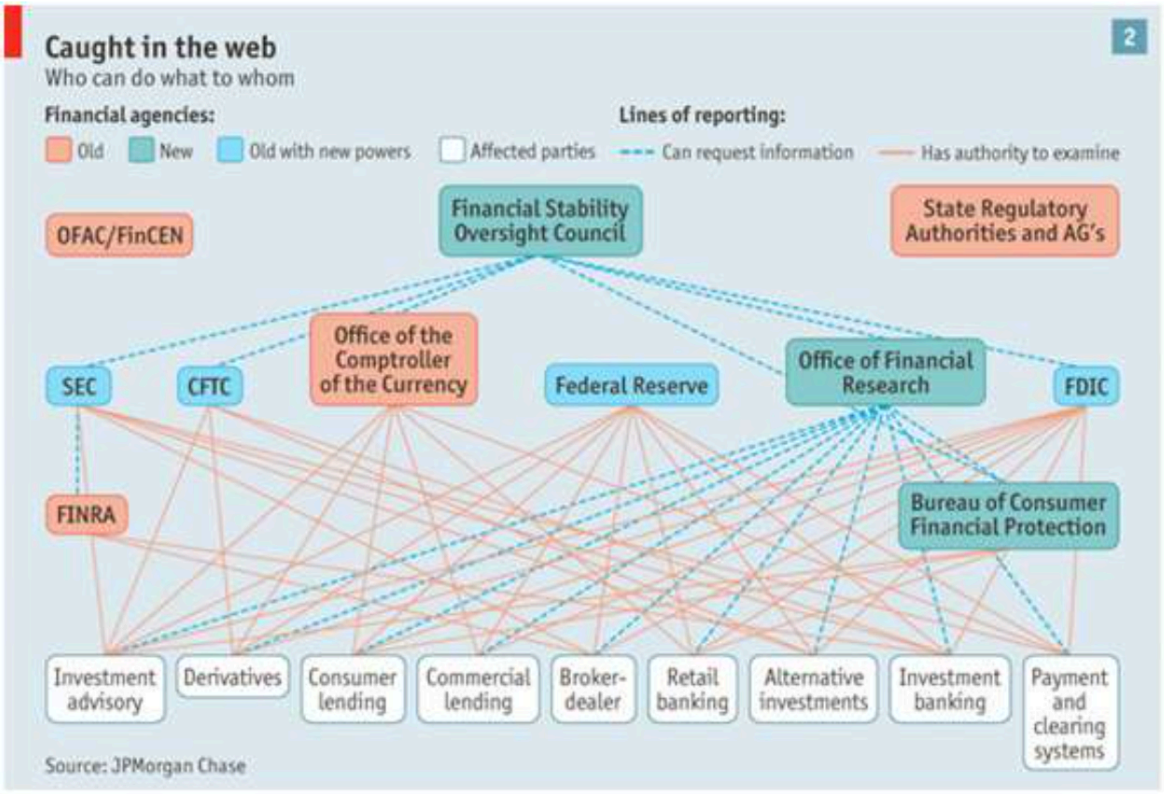
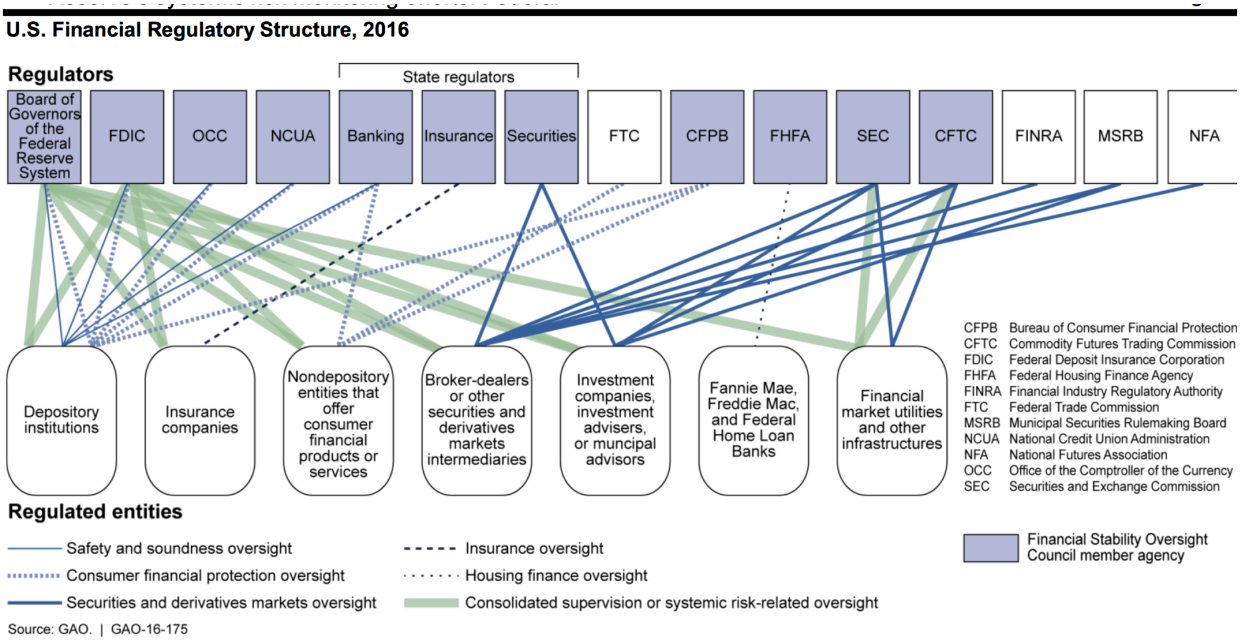


Fig. 2. The complexity that is Financial Regulation – An Example from the United States

Figure 1. The web of regulators. Source: This graphic is a copy of Figure 2 of Financial Industry Ontologies for Risk and Regulation Data (FIORD) – A Position Paper, published in Springer Conference Series and also available at ARAN Library .

For comparison to McCarthy's graphic in Figure 1, I include here a reference from *Financial Regulation: Complex and Fragmented Structure Could Be Streamlined to Improve Effectiveness* (GAO 2016). (GAO 2016) does not discuss ontologies, but fragmentation and overlapping authority which is discussed would be aided by a common ontology. Later in the paper, I develop a measure to assess the results of conducting topic modeling on the SRLs. The measure is based upon a very small part of what is a very large and complex ontology.



Note: This figure depicts the primary regulators in the U.S. financial regulatory structure, as well as their primary oversight responsibilities. "Regulators" generally refers to entities that have rulemaking, supervisory, and enforcement authorities over financial institutions or entities. There are additional agencies involved in regulating the financial markets and there may be other possible regulatory connections than those depicted in this figure.

Figure 2. GAO graphic of regulatory structure. This is graphic is a copy of a figure located on page 2 of the Highlights version of FINANCIAL REGULATION: Complex and Fragmented Structure Could Be Streamlined to Improve Effectiveness, GAO-16-175 available at GAO Report.

Chapter 3. Modeling process

Section 3.1 A text analytics technique to apply to SRLs

I offer several examples of how document classification may be conducted using text analytics and then focus on the method I adopt for this paper. A first example is the classification of documents according to sentiment. While sentiment is most often determined by identifying words according to whether they evoke a positive or negative feeling, sentiment may also tell us about other feelings relevant in finance, such as opportunity or restriction. All sentiment analysis relies upon a dictionary which assigns feelings to terms. (Loughran and McDonald 2011) found that words read as negative in the "usual" sense are not so in finance, including such words as "tax" and "cost". In another example of classification, we may look to text analytics to classify our documents according to a "named entity". For example, named entity recognition may enable classification of SRLs according to their relationship with banking laws, such as the Bank Secrecy Act (BSA) or the Dodd-Frank Wall Street Reform and Consumer Protection Act of 2010. Or classification may be made according to a lexical complexity score. This classification may identify regulatory text which perhaps is inconsistent with the plain language requirements of federal banking agencies as mandated in the Gramm-Leach-Bliley Act of 2009 and of other federal agencies as mandated in the Plain Writing Act of 2010. Another example is classification according to overall similarity or differences between documents. Perhaps text analytics would be used to classify SRLs by associated regulations even when the SRLs and regulations were not co-located or cross-referenced.

The text analytic modeling method I adopt for document classification is "topic modeling" using a probabilistic generative model named LDA. LDA seeks to identify similar documents based on the co-occurrence of words. Using the co-occurrence information, LDA will model two probability distributions – one relating topics to words and the second relating

documents to topics. In addition to the topics to which the documents are assigned, LDA will return a list of terms, by topic, with the probability that finding that term in a document will indicate the document belongs to that topic. The topic is the "hidden" information detected by LDA, while the words of each document are the observed information. A "topic" may tell us about the underlying theme of the document to enable classification, but we may also gain insights into authorship, temporal, or other factors that generated the documents included in the topic. LDA is referred to as a mixed-membership model because it is designed to deal with documents that fit into multiple topics with varying probabilities, which accommodates the structure of bank regulatory text where seldom will any one piece fit into one topic exclusively. The remainder of this paper focuses on topic modeling by LDA. Perhaps the relatively simple SRL series does not justify text analytics, as benefits of humans' superior text classification skills would outweigh the costs of an implementation of text analytics. But SRLs are just one series of regulatory communication within many and those many are escalating rapidly in quantity and complexity.

In two of the four cases of applying LDA to the SRLs, I use natural language processing to tag words for their part of speech. The work of technology to create tags for part of speech is challenged to understand the context of the word's use and if a word can play different roles in a sentence. Compare how a human versus a computer would tackle the task of inferring the meaning of "banks" in the following sentences:

> SRL guidance applies to banks.
> He banks using a piggy bank.
> There are trees on the river banks.
> The plane banks when there is snow on the ground.

I will include one last level of general description of methodology here – machine learning.

Machine learning methods run in a data-driven fashion instead of the traditional rules-based fashion. Machine learning models may be categorized as supervised or unsupervised and LDA, one of many machine learning methods, may be used as either supervised or unsupervised. Running LDA in an unsupervised fashion drives the model to look for relationships without the benefit of a predefined targeted relationship. My application of LDA to SRLs is unsupervised. While for this specific case of the SRLs, we have access to a human expert's classification system and could have added that system as the model's supervision, I chose not to share those classifications with the model. From my results, I would like to be able to draw a conclusion about the use of LDA on bank regulatory text regardless of whether it has been labeled by a human expert. Note though that I do use the classifications after the model has run to compare results of modeling the topics to show the change in results from changes in processing.

Section 3.2 An Objective Measure of model performance

Unsupervised LDA does not yet have the pervasive goodness of fit standards found in traditional models. Although I will later discuss some subjective review of LDA topic assignments, I wanted an objective measure to compare outcomes of choices made in creating the BoW as well as running the LDA. For this objective measure, I use LDA's grouping of SRLs into topics compared to the grouping done in the experts' FRBOG topics. All but two SRLs were assigned to at least one of 33 topics by FRBOG experts. To facilitate the creation of a measure, I made assignments of the two SRLs not otherwise assigned. SR0904 discusses "Applying Supervisory Guidance and Regulations on the Payment of Dividends, Stock Redemptions, and Stock Repurchases at Bank Holding Companies" and I assigned it to the Capital Adequacy topic. SR1215 discusses "Investing in Securities without Reliance on Nationally Recognized Statistical Rating Organization Ratings" and I assigned it to the Securities topic.

To familiarize ourselves with the FRBOG topics, we review some descriptive charts. The FRBOG topics represent an ontology level, as we see in Figure 3. In Figure 4, we see that more SRLs are categorized as Examination and Supervision Guidance (ExamSupGuidance) than in any other category, just one SRL is categorized in each of Fraud-Related Activities (Fraud) and Affiliate Transactions (Affiliate), and varying numbers of SRLs are categorized in the remaining FRBOG topics.

In Figure 5, we see SRL issuance by year for those SRLs in our sample. In Figure 6, we review the number of FRBOG topics into which each SRL is categorized. We see that while many SRLs are a member of just one or two FRBOG topics, some are included in more. An outlier is SR1319, which is a member of 10 FRBOG topics.

For my objective measure of LDA performance, I will count the number of SRLs in an LDA topic that fall into the FRBOG topics most represented in the LDA topic and second most represented FRBOG topic in the LDA topic. This method does not give consideration to "ties" in the representation of FRBOG topics in the LDA topics and so a tie is agnostic to the choice of FRBOG topic matching to the LDA topic. In that respect, there is some arbitrariness which prevents us from interpreting the meaning of the LDA topic assignment according to the FRBOG topic labels. Labeling LDA topics often requires human intervention.

Figure 3. A 3-level partial ontology for SRLs.

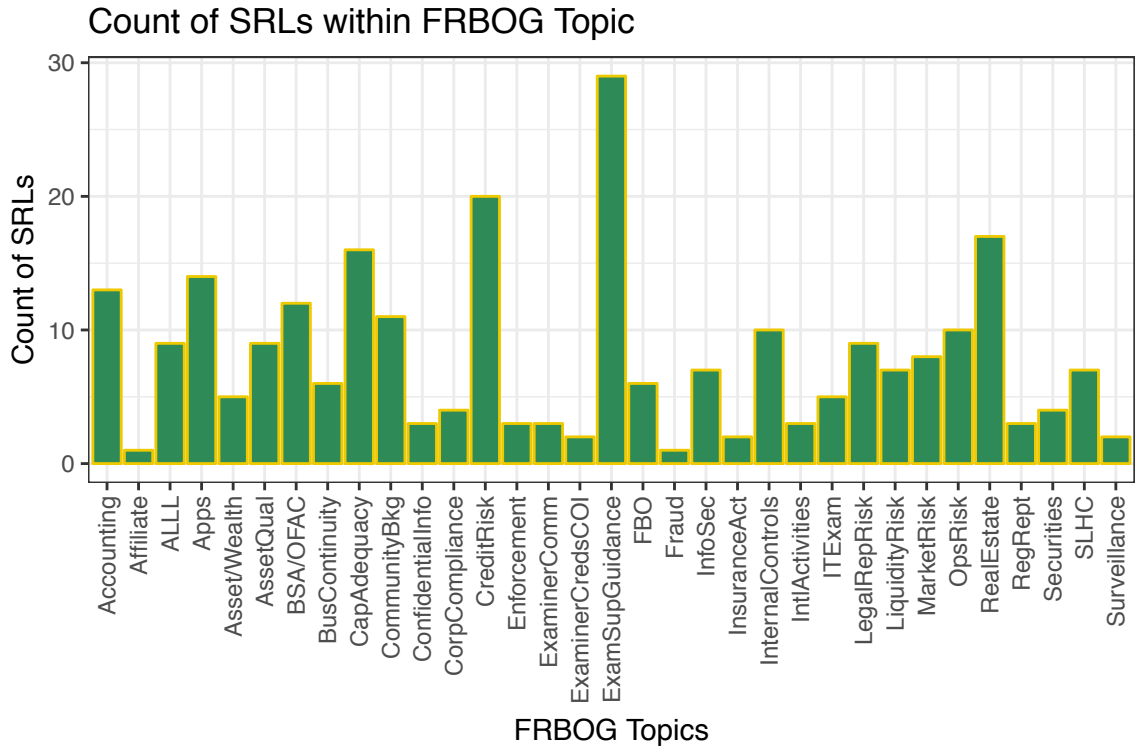


Figure 4. Histogram for counts of SRLs by FRBOG topic.

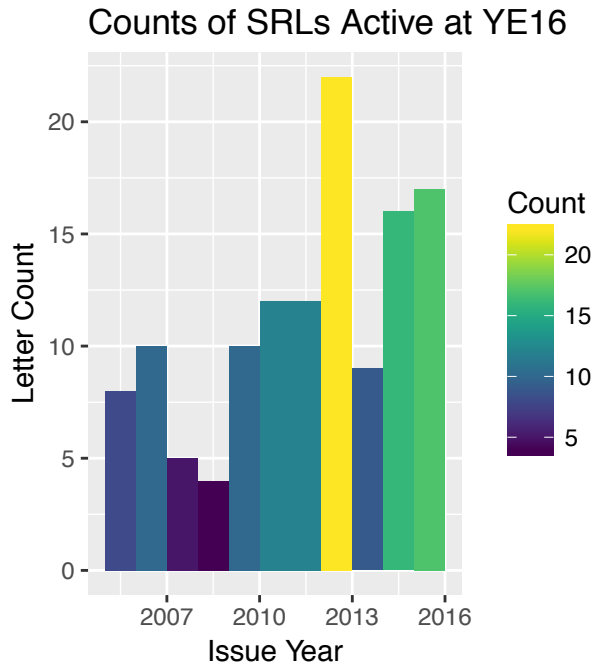
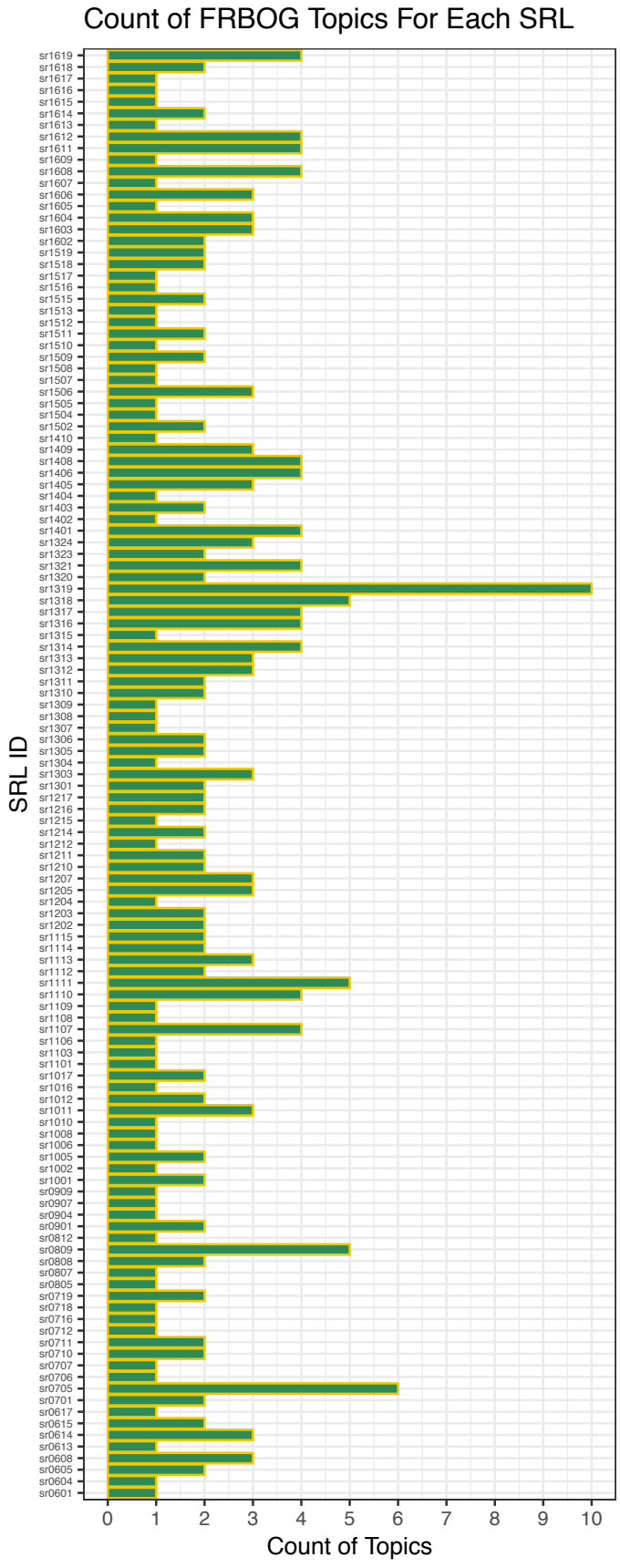


Figure 5. Histogram for counts of SRLs by issue year.

Using this method, any LDA topic made up of just one or two SRLs will automatically match to FRBOG topic grouping at 100%. Rigor of the measure would be improved by introducing a penalty term for these cases. Though I do not create a penalty term, I will note such LDA topics in the upcoming results section.



The code for this process is described in Appendix 3. From here on, I refer to this measure as the "Objective Measure".

Figure 6. Histogram for counts of FRBOG topics by SRL.

Chapter 4. The application of the LDA topic model to the SRLs

Section 4.1 The corpus

In the context of text analytics, the word "corpus" refers to the collection of documents used in the analytics. I obtained the SRLs for the corpus from a Fed website. The website allows download of the SRLs as text via a print file, a PDF, or an html file. All SRLs in effect are accessible at the website and are listed by year of issuance or by topic (hereafter those topics are referred to as FRBOG topics). Despite a description of PDFs as "often the hardest to extract from" in *Taming Text*, I chose to download the PDF versions of the SRLs and downloaded 125 SRLs. The 125 SRLs were those issued between and inclusive of 2006 to 2016, if the letter remained active as of December 31, 2016. I used the R package PDFTOOLS to transform the PDF files into text.

SRLs follow a traditional letter format. I considered their structure in order to identify "landmarks" which may enable my extraction of metadata or the elimination of text I feared may distract the model later on. Here I will briefly discuss the structure of SRLs and related possible or actual actions based upon them. I discuss the pieces as landmarks because of their impact on how text mining is conducted and to also note inconsistencies which perhaps may become an area of opportunity as regulators seek to facilitate text analytics.

- **Identifier**: The header includes the SRL identifier, which is in a format such as "SR"+YY+##, where ## is the one-up number assigned within a year. The format of the identifier differed between letters according to whether a space was placed between "SR" and YY, a hyphen was used between YY and ##, or a leading zero was used in ##. While these differences are not noticed when humans are reading one letter at a time, they prevent simple retrieval of the identifier for inclusion in metadata.

- **Addressee and subject**: These potential landmarks are in a section that extends in an abbreviated horizontal fashion. First, an addressee is provided; second, a subject is given. Letters are usually addressed to 1) the Reserve Bank head of the function responsible for the supervision area that implements the related regulation and 2) the financial institutions required to comply with the letter. The field may begin with "TO THE OFFICER" or "OFFICER", an inconsistency which limits its use as a landmark. The subject begins with "SUBJECT", which I did use as a landmark to extract the first line of the subject as metadata for my corpus. I was challenged to extract the full subject as there was no landmark to use as a stopping point for multi-line subjects.
- **Applicability**: The next landmark is a box of text defining "Applicability". The included text in this boxed section specifies to which regulated entities the guidance applies based upon size, activities, or other characteristics. I believe this represents a best practice as a text miner could test applicability to rules and, if appropriate, stop ingesting and skip modeling if the guidance was not applicable in a particular mining case. I did not include this information in my metadata as the box is included beginning in 2011 and my sample began in 2006.
- **The body of the letter**: This part is the text of interest when seeking insights.
- **The end of the letter body**: The body of the letter ends with a paragraph with question referral information. In some cases, I used the start of the question referral sentence as a landmark to stop ingesting. Identifying the start of the sentence was made time-consuming by the variety of wording used and on occasion, the lack of uniqueness of the start of the sentence. When the wording of the question referral sentence was not unique in an SRL, I instead used the signature field as a stop landmark. While the html versions

of SRLs appear to consistently use the words “signed by”, the PDFs are inconsistent in its use. The signers’ names were used then to establish a stop landmark, but in some cases, middle initials were inconsistently applied.

- **Letter attachments and lists of superseded SRLs, attachments to the current SRL, and/or cross-references**: While lists of superseded SRLs, attachments to the current SRL, and cross-references to other SRLs or regulatory information are important, they were not useful in this application of topic modeling. In addition, some SRLs include relevant attachments in their entirety while others only include the reference to the name of the attachment. In consideration of the potential effect on term frequencies of inconsistently included attachments, in some cases I removed them by stopping text extraction at the question reference paragraph or the signature line.

Section 4.2 Corpus to BoW to document term matrix (DTM)

Text analytics often requires the creation of a BoW. A BoW is created by breaking the long character strings of each document in the corpus into a file of individual "tokens". Tokens may be words from the text or combinations of the words. Moving from a corpus to a BoW follows a well-established set of steps. However, how the steps are performed – the order, the tools used, etc. – will affect the BoW and that affects the outcome of LDA. In a last step, the BoW is transformed to a DTM. A DTM is the input to text analytics models and consists of tokens for columns and documents for rows, with matrix cells filled with frequencies of tokens within documents. Here, I discuss the steps and related choices I made.

Step 1. Tokens

Sentences or phrases must be split into words. In my program, I did this by using R's string manipulation capabilities. I retained all words from each SRL PDF file, using a space as the boundary of a word.

Step 2. Text case

Depending upon the defaults for handling text case, a computer may not identify same words as such if case is different. While I wouldn't want the computer to differentiate "Thesis" from "thesis", I would want "ARMS" (when capitalized, the acronym for adjustable-rate mortgages) differentiated from the word "arms". While the former is simple to solve, the latter is not. "ARMS" is just one example of the many acronyms that introduce complexity into text analytics and while it is simple to adjust a single acronym, consistent model results will occur only if a consistent way of handling the universe of acronyms exists. Most often all text is made lowercase in preparation of the BoW, even acronyms. I chose to do that using R's TM package.

Step 3. White space

There's almost always going to be some extra white space in lines of text, such as often occurs around bullets. That space may interfere with how words are tokenized. The usual practice is to remove this space. I used R's TM package for this.

Step 4. Punctuation

My next step addressed punctuation. A computer will not recognize "regulations" as the same token as "regulations:" or "regulation's", so common practice is to strip all or almost all punctuation from the corpus in the transformation to BoW. I used R's TM package for this step as well, but included an option to preserve intra-word-dashes. The SRLs included many hyphenated terms which should be distinguished as not just additional occurrences of one component word.

Step 5. Numbers

In most applications, numbers are removed in the conversion to BoW. That was appropriate for this exercise as well, and R's TM package was used to do so. However, future work is called for as it is easy to identify an example in which removing numbers may not have been appropriate, such as if the text mining task was to identify applicability of an SRL to a particular bank using the asset size cutoff set out in the Applicability section of the SRL.

Step 6. Stop words

"Stop words" are often removed from the corpus as well. Stop words are a list of commonly used words which should be excluded from the BoW. Examples include "a", "the", "above". I used the R TM package list of English stop words. Most everyone may agree removal of "the" is appropriate, but other suggested stop words may not be universally accepted as appropriate.

Step 7. Word stems

A next step is often to stem words. When words are stemmed, they may be stemmed to remove pluralization, to find a root of a word, or something in between. In a video from Cambridge Machine Learning Summer School 2009, David Blei tells us "there's no good reason not to stem" and reminds us that consideration should be given to tolerance for stemming's effect on the interpretability of LDA output. Stemming resulted in consistent overall performance improvements in my study according to the Objective Measure. I used the R TM package stemming process.

Step 8. Too rare or too common terms (aside from stop words)

A final step is my process was too address too rare or too common terms. Consider two SRLs that both discussed "influenza", an atypical topic for bank regulation. When I observed the challenge of removing too rare terms without removing "influenza", I chose to keep all rare terms, even if they occurred just once. I expect in the context of bank regulation rare words will be insightful. On the other hand, there are many, many common words in regulation, beginning with regulators' names, "supervision", or "regulation", which are not likely to be useful in classifying documents. These too common words can prevent LDA from creating meaningful topics. In this BoW preparation, words that were present in more than 50 of the 125 SRLs in the corpus were excluded by employing the TM package.

Step 9. Transformation to DTM

LDA most commonly uses a DTM in which matrix cells contain unweighted token frequencies. In some applications, the DTM's cells may instead include weighted versions of term frequencies.

At completion of these steps, our DTM created by the R TM package has 21,889 non-sparse (non-zero) token entries. The token entries represent 63,332 tokens in total, of which there are 3,526 unique tokens.

Section 4.3 Topics in a Simple Case

LDA was applied to the DTM created in Section 3.5. LDA is commonly referred to as a "BoW model" in which tokens are interchangeable and order is not significant. (Blei, Ng, and Jordan 2003) "emphasize that an assumption of exchangeability is not equivalent to an assumption that the random variables are independent and identically distributed". They state that it is rather "conditionally independent and identically distributed, where the conditioning is with respect to an underlying latent parameter of a probability distribution." The preparation steps described in Section 3.5 will bring the SRL BoW closer to meeting this assumption, enough so that LDA can be used. Though beyond the scope of this paper, there is a large amount of research exploring ways of relaxing this assumption. For example, (Wallach 2006) describes a bigram language model which aims to predict one word based on the preceding word.

LDA requires that we set "k", the number of topics in the model. "k" is assumed to be known and fixed. The parameter is commonly set through mathematical methods or human expertise. For this case, I chose to rely upon the human expertise used to create the FRBOG topics in Figure 3. There were 33 FRBOG topics to which SRLs are assigned on a one-SRL-to-many-topic basis. LDA is a mixed-membership model in which documents may be allocated to more than one topic and this is consistent with the human FRBOG expert's classification. Based upon the overlap, the number of SRLs in my sample, and LDA performance according to the Objective Measure, I chose to use k = 23.

As a Bayesian model, LDA will begin with priors and then update them with observed information until the model converges on posterior probabilities of a document's inclusion in a topic and a topic's inclusion of a token. Choices about priors must take into consideration the generative nature of LDA as LDA seeks to explain how documents were generated from the

terms of a topic. Factors considered include expectations about the outcomes as well as whether any mathematical derivations were appropriate. α, which tells us about the distribution of documents over topics, and β, which tells us about the distribution of topics over terms, are both assumed to be "sampled once in the process of generating a corpus" unless priors are given to the LDA code (Blei, Ng, and Jordan 2003). For the α prior, I chose to set a value to initiate the process of estimating topic proportions. (Griffiths and Steyvers 2004) recommended beginning with an $\alpha = 50 / k$, which in this version of my model equaled 2.17. I applied trial and error to explore outcomes using other values, and settled at 2.7 as it produced the highest Objective Measure given the other model choices. When a symmetry assumption is met, a higher value of α implies that documents are more likely to include multiple topics; when symmetry is not met, the higher α will imply documents load mostly to one topic.

A final choice in my LDA was the method of estimation of the probabilities of topics over terms and documents over topics. I chose the variational inference method discussed earlier with an implementation using the R Topicmodels package.

Section 4.4 Results of the Simple Case

After the completion of the steps described above, LDA results are reviewed. First, we review the Objective Measure which tells us that 102 of the 125 SRLs were grouped within an LDA topic's most common two FRBOG topics.

The output of the LDA process provides probabilities that documents and terms have been generated from an LDA topic. First, I discuss the probability of a document having been generated from an LDA topic. (Grun and Horn 2017), the R Topicmodels documentation, defines gamma (γ) as a matrix of "parameters of the posterior topic distribution for each document". Since our model allows for documents to have membership in more than one topic, the LDA topic of greatest interest is the one from which an SRL was most likely to have been generated. In Figure 7, we see γ for each SRL within the LDA topic from which the SRL was most likely generated. Most SRLs have a high probability of having been generated from the terms of the LDA topic to which they were assigned, but not all. To review a couple of specific SRLs, let's begin with SR1212 in LDA topic 4 which has γ of .54. This tells us LDA was challenged in identifying if SR1212 belonged in LDA topic 4. The probability of SR1212 having been generated from the terms of LDA topic 13 was .46, not too far behind. In contrast, we see in LDA topic 8 that SR1314 and SR1604 have γ of over .99, telling us that there is a very high probability that the SRLs were generated from terms in this topic. The γ chart is useful also as a quick glance at how SRLs are spread across the LDA topics. We expect that LDA topics will be useful classifications and contain an appropriate share of SRLs, as opposed to say 22 SRLs in their own individual topic and the remaining 103 SRLs in the one remaining LDA topic. As a precursor to anticipated future work in which I will review the evolution of regulatory text relative to banks' financial statistics, I also plot γ of each SRL according to the year(s) in which

the SRL was issued in Figure 8. Using these charts, we see how some SRLs loaded mostly to one LDA topic (spikes) while others loaded to multiple LDA topics (low bounce). We note that some years' SRLs have no representation in some LDA topics, such as the 2010 SRLs with $\gamma <$ .00 across the board in LDA topic 8. We see that SRLs issued in 2013 and 2016 are included in many of the LDA topics, likely reflecting the regulatory environment which called for varied regulatory changes.

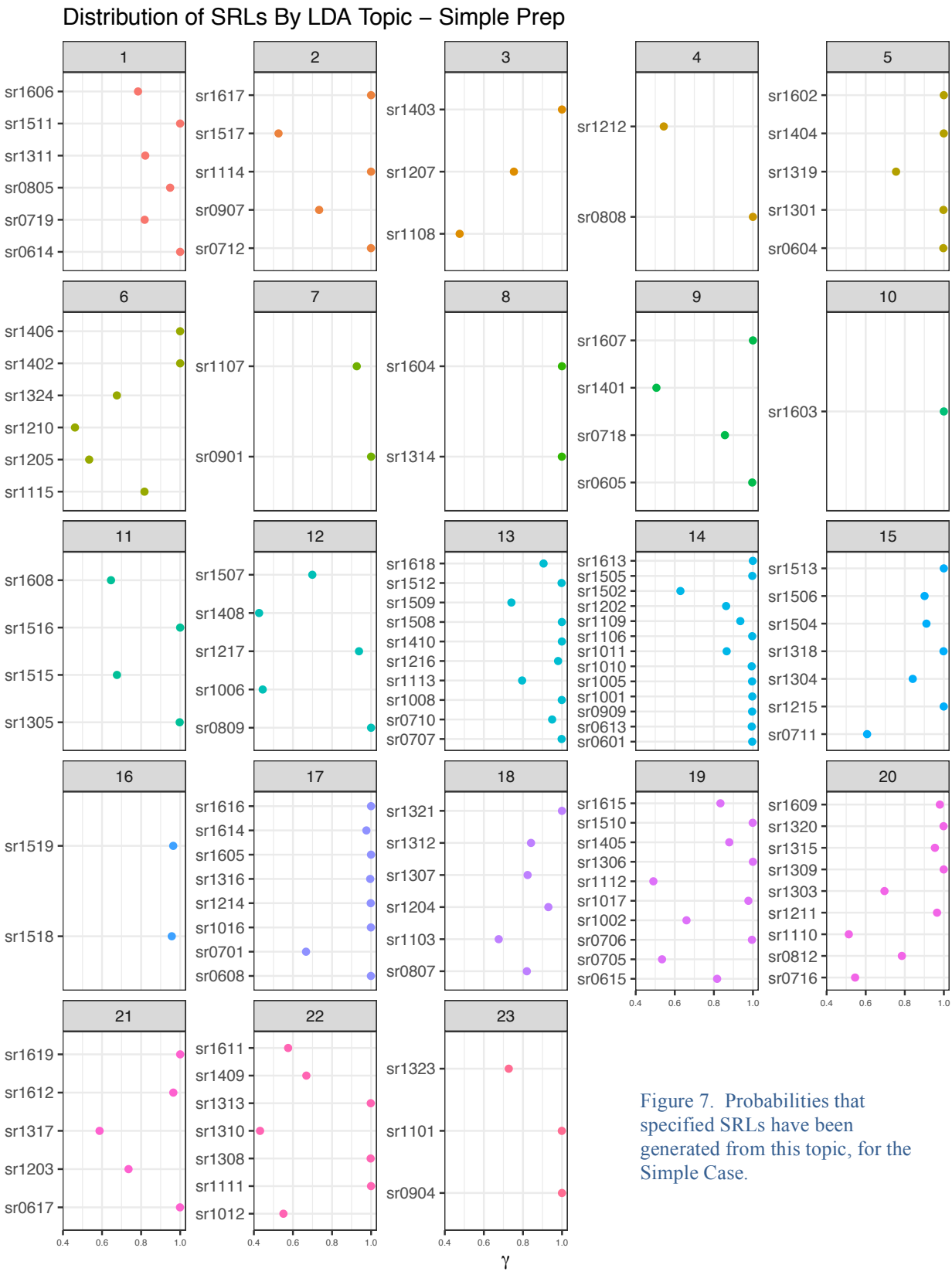


Figure 7. Probabilities that specified SRLs have been generated from this topic, for the Simple Case.

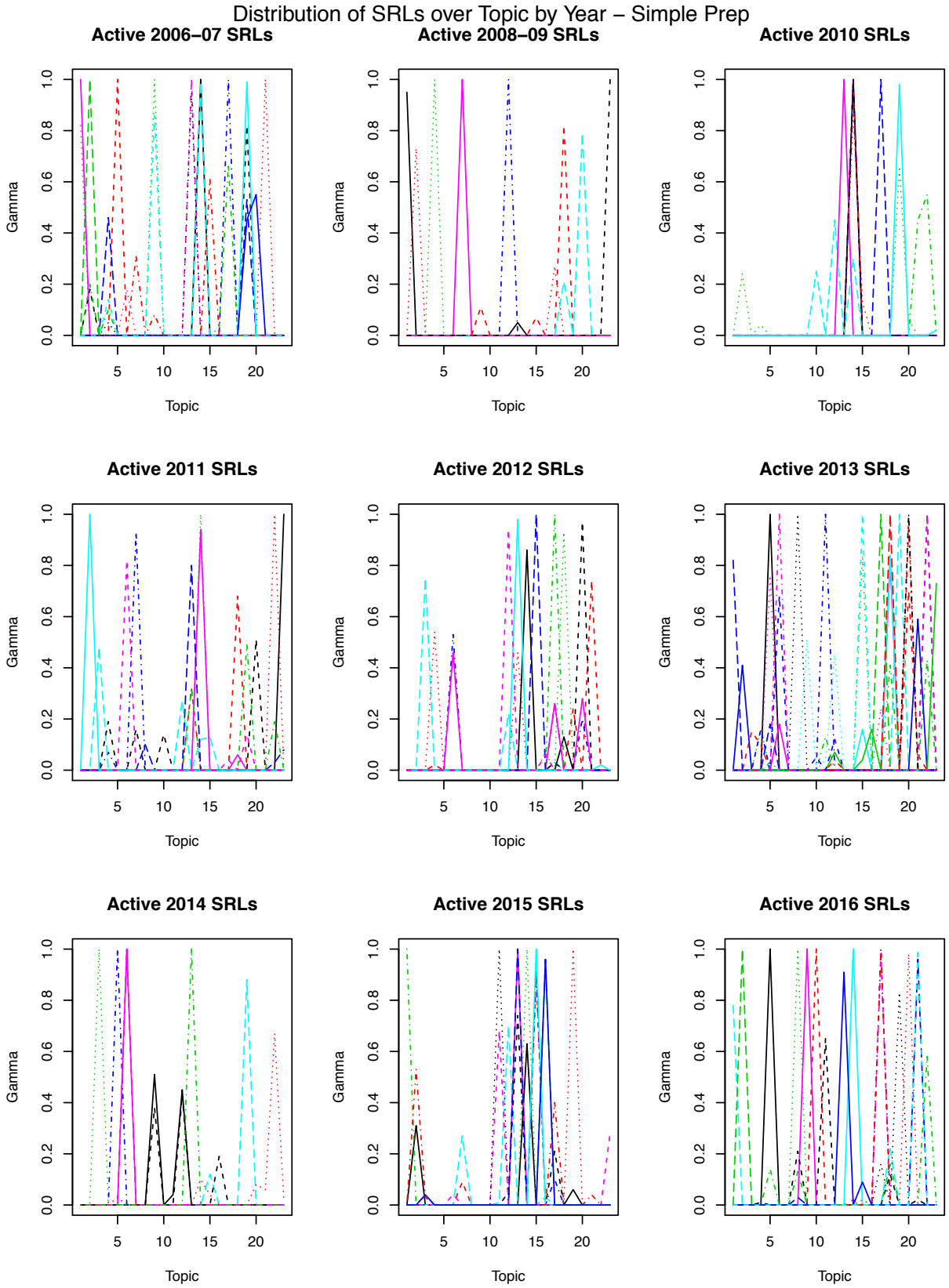


Figure 8. A version of Figure 7 divided by issue year.

Next, we review β, the probability of a term within a topic, which tells us the terms we are most likely to see in a document generated from the related topic. Because we stemmed, the top terms are returned as stems. We review the top ten most probable terms by plotting β for each LDA topic in Figure 10. In LDA topic 3, we see that "test" is somewhat defining of the LDA topic, where β tells us there is more than a 0.08 probability that an occurrence of "test" has been generated from LDA topic 3. If an SRL includes both "test" and "stress", there is a bit more than .15 probability that it will have been generated from LDA topic 3. In LDA topic 5, "audit" has a β near 0.08. We see "firm" in LDA topic 16 with a similar level β, a bit surprising since firm is not a very specific term and perhaps "firm" would be added to stop words in a future run. On the other hand, if we look at LDA topic 1, we see the highest β of terms is just a bit over 0.02 for "access". If we compare LDA topics 11 and 18, we'll see "surveil" and "rate" switch places as the most probable and least probable (among the top 10) terms, though the level of probability differs. We see also an odd term, "-site", in both LDA topics 8 and 18.

We'll take advantage of noticing the "-site" to discuss some opportunities for improvement in how we topic model regulatory text. SRLs use the terms onsite and offsite most often to describe where examination work takes place. There is inconsistent use of the hyphen in the SRLs, i.e., "offsite" might be "offsite" or it might be "off-site". No SRL includes just "-site"; instead, we are seeing in the β plot the effect of pre-processing choices. When "off" was removed during the stop word process, we were left with "-site", and since I selected to preserve intra-word hyphens, "-site" passes through the model and turns out to be an important term. What is important to note is that without further exploration we don't know whether "-site" supports the LDA process better than "onsite" and "offsite" – the term frequency of "-site" will have a higher term frequency than that of "onsite" or "offsite" individually but here we only have

"-site" when an SRL author chose to use a hyphen. While consistency in hyphenation is a good goal, we cannot say for certain the direction of the effect on results. For example, an author of SRLs that used the hyphen may be a specialist in the BSA and an author of SRLs that did not use the hyphen may be a specialist in managing market risk. Unintentionally, the inconsistency may help to differentiate the topics.

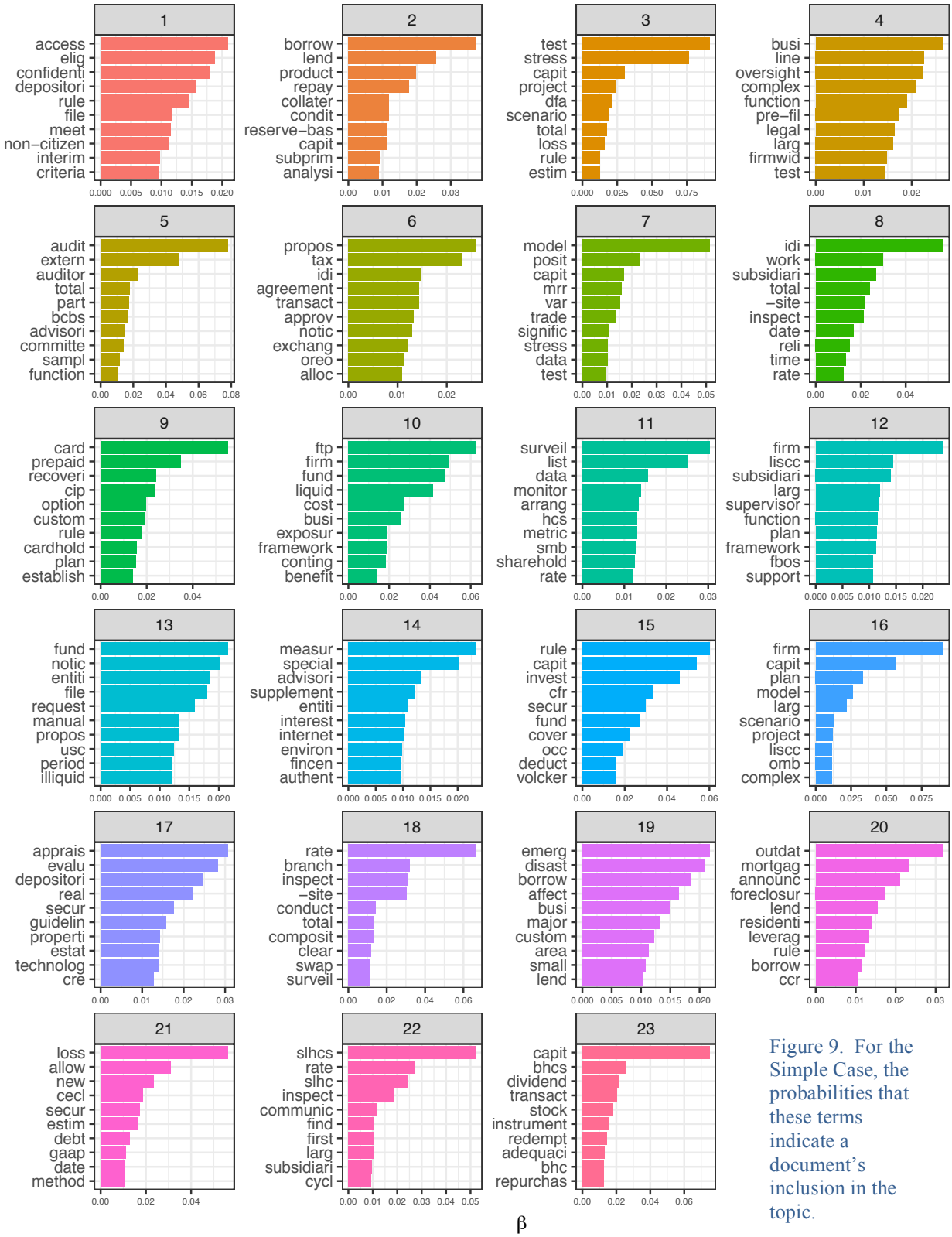


Figure 9. For the Simple Case, the probabilities that these terms indicate a document's inclusion in the topic.

Earlier I mentioned the 102 / 125 SRLs were grouped according to the Objective Measure. While that was one way of reviewing results, another is to review other classifications that this simple run of LDA found. To explore that, next I subjectively review some specific LDA topic assignments in comparison to FRBOG topic assignments.

Example 1. LDA topics where the FRBOG topic "Capital Adequacy" was prominent

In Figure 4, we see that we begin with 16 SRLs in the FRBOG topic "Capital Adequacy". Three of these 16 make up LDA topic 3 in its entirety. This LDA topic 3 had high βs for "test" and "stress". Here LDA has gone beyond "Capital Adequacy" to detect that these three SRLs are related to tests of adequate capital under stressed economic scenarios. LDA topic 7 includes just two SRLs, which both address the impact of market risk on capital adequacy. We see "stress" and "test" show up in LDA topic 7's highest probability terms, both not at the level at which they showed up in LDA topic 3. We see "model" at the topic of LDA topic 7's terms, and "mrr" (in its capitalized form, an acronym for Market Risk Rule), in the fourth from the top spot. We also see a high representation of "Capital Adequacy" in LDA topic 15, in which the top term was "rule" with $\beta = 0.06$, and the common factor seems to be these SRLs establish rules for capital calculations. In LDA topic 16, we see two more "Capital Adequacy" SRLs – these two SRLs were issued simultaneously and establish the same provisions but are tailored for different asset size and complexity of the entities to which they were applied. Finally, we see "Capital Adequacy" again for each of the three SRLs that make up LDA topic 23. These three SRLs discuss how a bank manages its capital directly and we see top terms such as "dividend" and "repurchase". We begin to appreciate that LDA may be working towards creating subsets of the FRBOG topics.

Example 2. LDA Topic 14

LDA topic 14 includes 13 SRLs, with six drawn from the FRBOG topic "BSA/OFAC" (Bank Secrecy Act / Office of Foreign Assets Control), two drawn from "Liquidity Risk", and the remainder drawn from other FRBOG topics in single occurrence. Though there is limited commonality of content, it seems the defining commonality of these SRLs is that they are issued on an interagency level, though the LDA topic does not represent the universe of SRLs with interagency guidance. Interagency guidance is usually provided as an attachment to an SRL and the authors of interagency guidance are different than those of Fed-specific guidance. Top terms include "measure", "special", and "internet", while the first occurrence of a BSA/OFAC-specific term, "fincen" (in its capitalized form, the acronym for Financial Crimes Enforcement Network), shows up in position nine of the top ten terms. Here, we must consider whether the interagency effect is strong enough to trump the effect of underlying content similarities and how that may be remedied in the future when the classification goal is to classify on content.

In summary, we see that LDA provided meaningful grouping according to the Objective Measure and is also providing some insights into subsets of SRLs within and across FRBOG topics. LDA was not set up to duplicate the mapping of each SRL into the universe of applicable topics as is done by the FRBOG experts, but LDA has certainty created topics that are aligned with the FRBOG topics. The alignment is strong enough to have shown that text analytics will enable the work of Regtech to create a framework for automating the movement of regulatory communication between regulators and the regulated. As this was the simplest BoW using standard pre-processing steps, we'll next review some tweaks to appreciate how LDA responds to deviations.

Section 4.5 Note on sentiment analysis

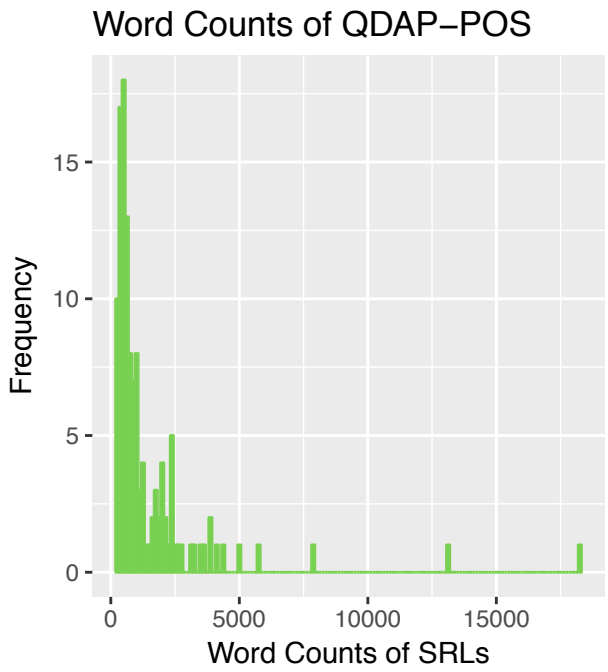


Figure 10. A display of SRLs according to a word count calculated by QDAP. Most SRLs are relatively brief.

As an area of inquiry separate from the topic modeling work, I provide a note on sentiment analysis, using the BoW created in the Simple Case before stemming and stopword elimination was applied. As noted above, sentiment analysis utilizes a dictionary that assigns feelings, or sentiments, to terms used in the corpus. So long as dictionaries allow for the use of unstemmed words, the unstemmed versions will allow for greater precision in sentiment scoring. For example, "acquirer" may not reflect sentiment in SRLs, but "acquirers" may reflect negative sentiment if it is usually the case that only failed institutions have more than one acquirer. The word counts of SRLs are displayed in Figure 10 to the extent it aids in setting a benchmark. (Note that this is an output of QDAP, and may differ from other word counts used in this paper as word count is affected by treatment of hyphens, exclusion of too common words, etc.).

Sentiment analysis of the SRL corpus is made quite challenging by the lack of a dictionary specific to bank regulation. Words that may be perceived as positive or negative in consumer reviews or even in financial disclosures may not be so in bank regulatory communications. I began an exploration of sentiment analysis to determine if it would aid in the labeling of regulatory communications as "permissive" or "restrictive" and found that it did not when using the available dictionaries which were not developed for bank regulation. I share

results of the application of two dictionaries to the SRLs here to introduce a case for the creation of a sentiment dictionary specific to bank regulators' communications.

First, I use the R QDAP package's "polarity" function which enables the comparison of the polarity in language across documents or time. I use only the function's assignment of words as "positive" or "negative". Next, we consider the most frequent of each of positive and negative words. The QDAP package assigns sentiment based on a dictionary introduced in (Hu and Liu 2004) which discussed mining opinion features in customer reviews. In Figure 11, we see that "risk" and "risks" are the most frequent negative words. In Figure 12, we see that "guidance"

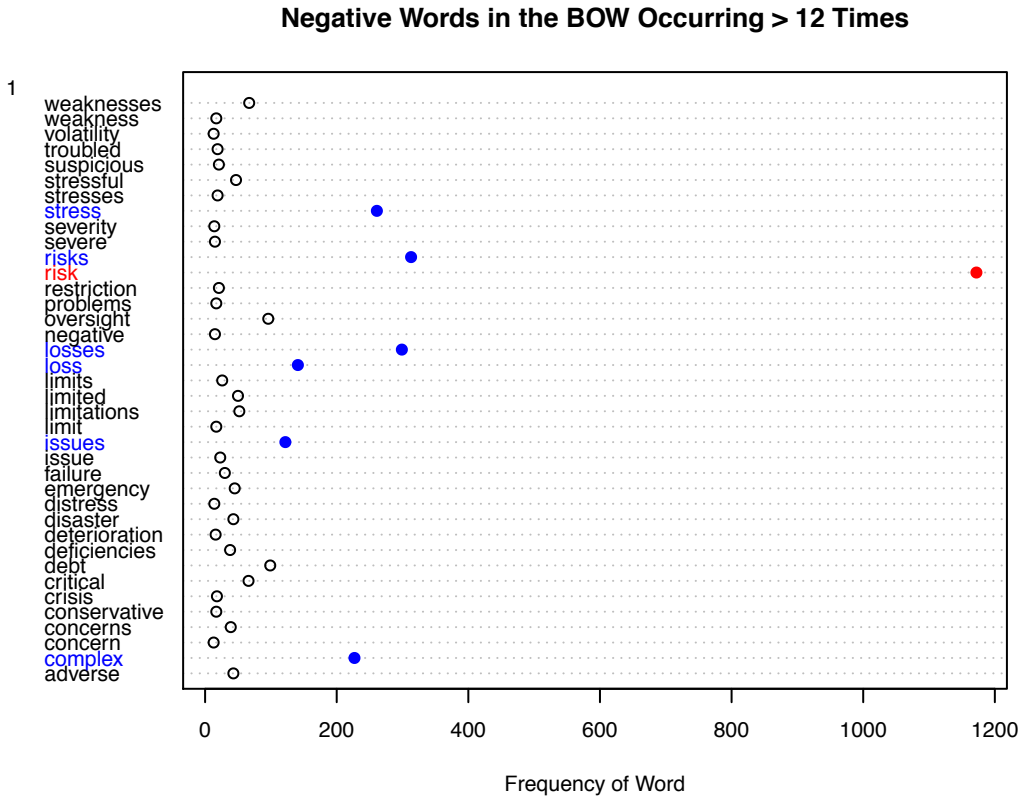


Figure 11. QDAP's identification of negative words.

is the most common positive word and in general there are more "positive" terms than "negative". The dictionary's assignments of sentiment are inconsistent with usage in SRLs.

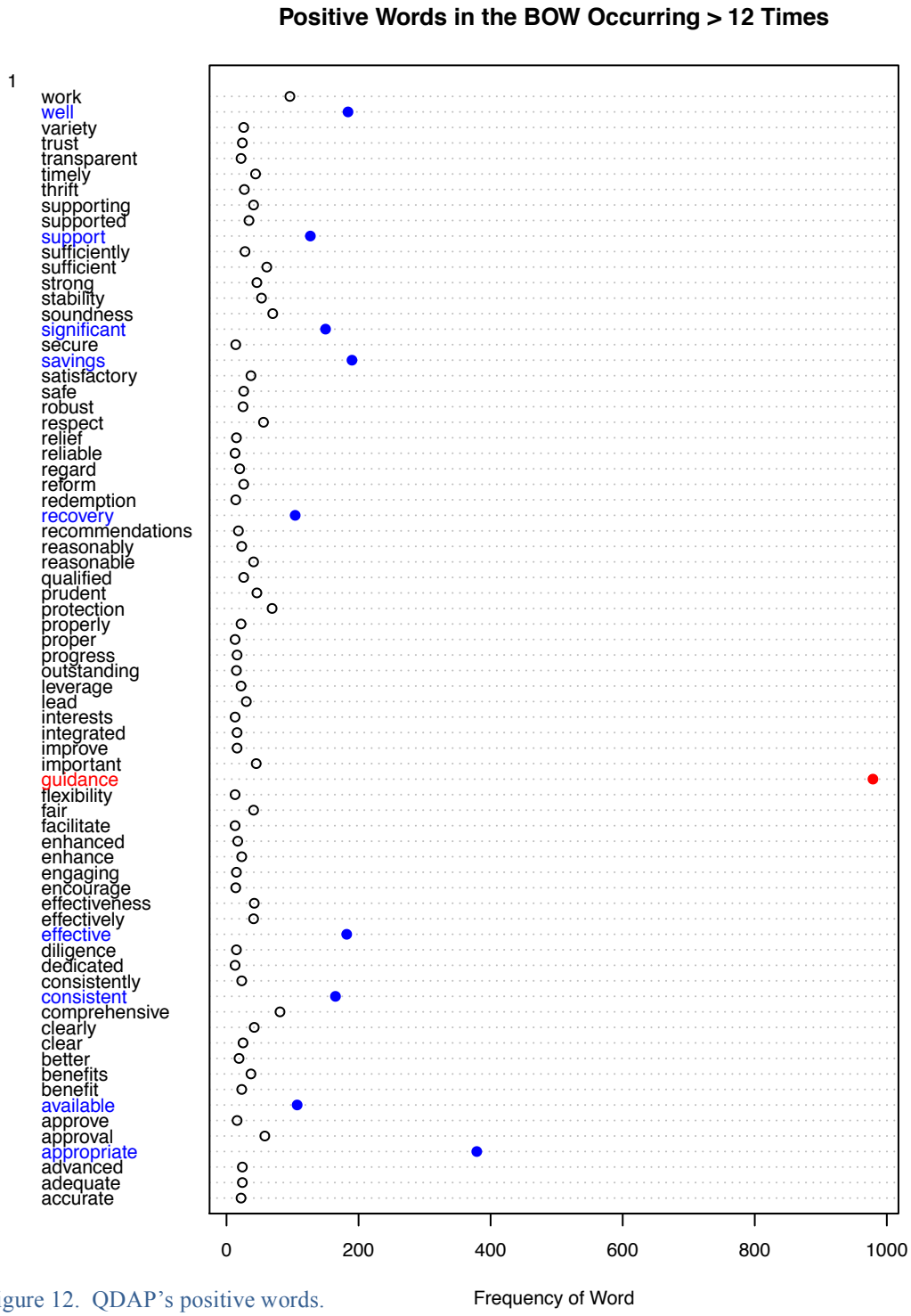
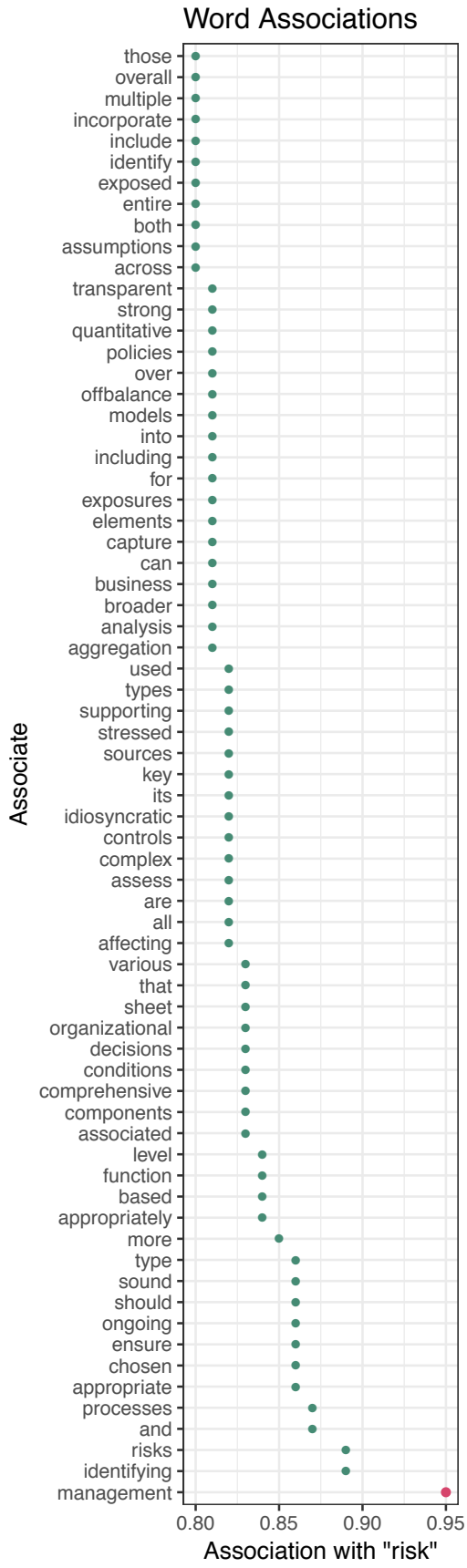


Figure 12. QDAP's positive words.



Because "risk" was such an outlier, I show its associates and dispersion, using text mining capabilities from the TM and QDAP packages. The "findAssoc" function identifies terms that co-occur with a term of interest, where co-occurrence is determined by exceeding a least amount of correlation an associate must have with the term of interest. In Figure 13, we see interesting co-occurrences, such as an association with "management". Further investigation is required to distinguish between insightful co-occurring terms and terms that naturally coincide with "risk" at a high frequency. Some expected terms that are missing from the list of associates include the adjectives most often used to describe risk, such as "credit" or "reputational".

Figure 13. Terms most associated with "risk" using the TM package function "findAssoc".

In reviewing the dispersion plot in Figure 14, we see that there is no concentration of use of "risk", it's everywhere in the SRLs. We would expect these results as both banking and bank regulation are inherently businesses of managing risk.

Figure 14. Dispersion of the term "risk" throughout the SRLs.

For the second application of sentiment analysis, I employ the R Tidytext package. Tidytext now includes the "loughran" dictionary as a sentiment tool, available only through the GitHub version for now[3]. I mentioned (Loughran and McDonald 2011) earlier in this paper. (Loughran and McDonald 2011) finds that "almost three-fourths of negative word counts in 10-K filings based on the Harvard dictionary are typically not negative in a financial context." Their work included the expansion of sentiment categories to classify words as negative, positive, uncertainty, litigious, strong modal, or weak modal. As an introduction, we review polarity charts using the "loughran" option in Tidytext. In Figure 15, we review the

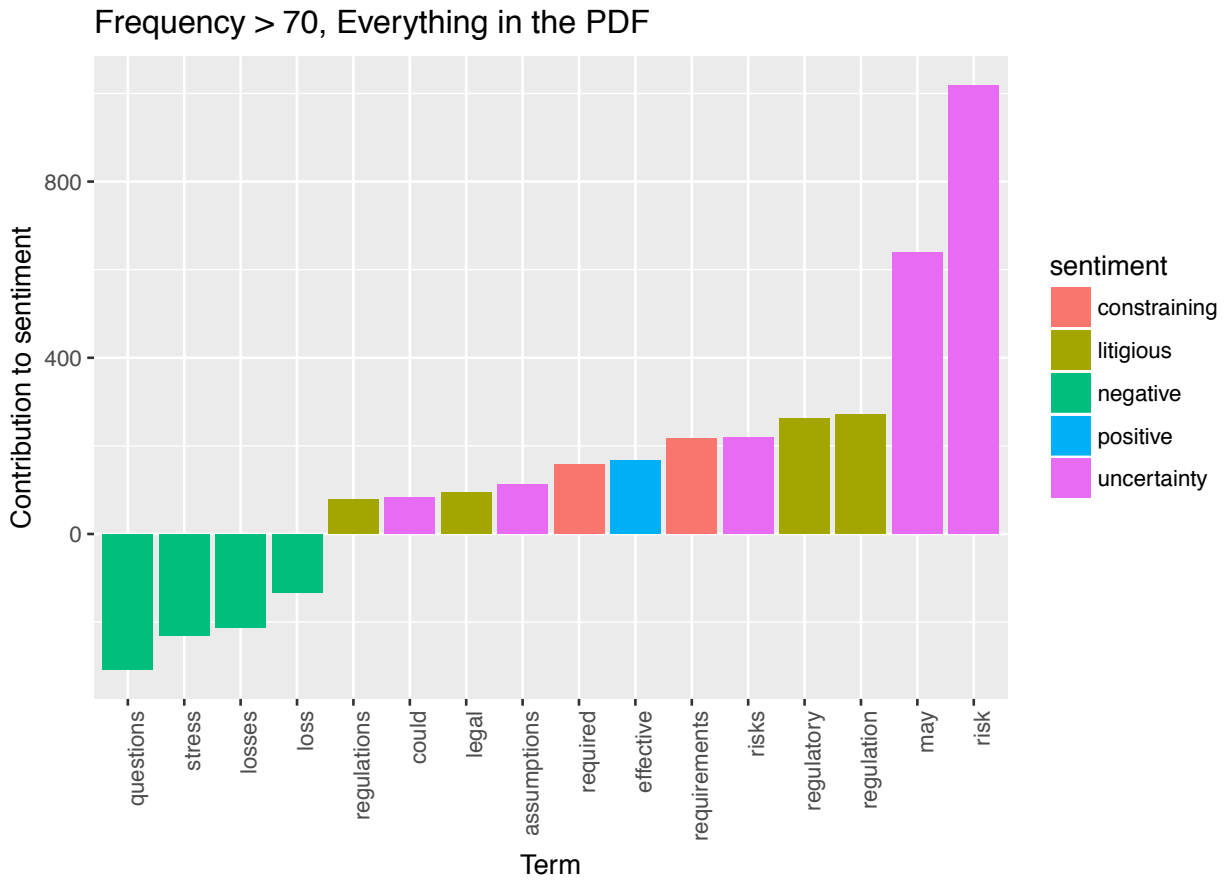


Figure 15. Tidytext's largest contributors to sentiment, using everything that reads as text in the SRL PDF files.

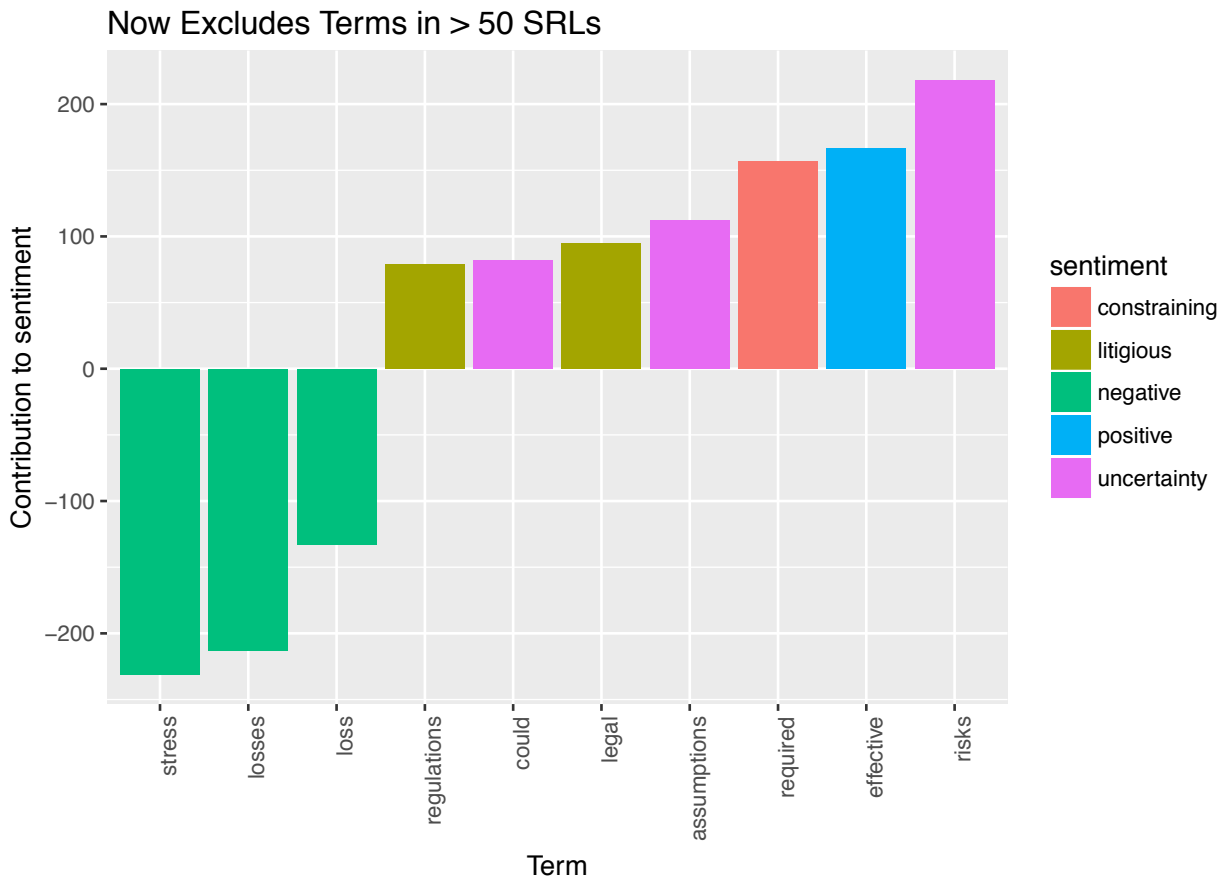


Figure 16. A new version of Figure 15 that excludes those terms that occurred in more than 50 SRLs.

sentiment of words occurring more than 70 times in the Simple Case corpus, and in Figure 16, we review the same but qualified now to exclude the terms which show up in more than 50 of the SRLs. In comparing the two charts, we see that "questions" falls off due to its inclusion in more than 50 of the SRLs. 'Questions" was labeled a "negative" word, but has no sentiment connotation in usage when it was directing a SRL reader what to do should he/she have a question. We see "may" and "risk" fall off; they were designated as indicating uncertainty. We did not "stop" month names, so we are uncertain the context of the usage of "may", and I would propose "risk" is a neutral word. The most frequent words are similar to those found with QDAP, and are not particularly telling. A red flag is the word "stress'. While the word is tagged

"negative", its use in bank regulation is somewhat neutral, especially since regulators began conducting stress tests to determine banks' capital adequacy.

While sentiment analysis provides a different lens through which to review the text of the SRLs, it did not show sentiment. There are many contradictions when using a limited scope dictionary that was not designed for the SRLs. Even when using a finance-based dictionary, we see that contradictions can result when applied specifically to bank regulation.

Chapter 5. Changes to process choices of the Simple Case

Section 5.1 Tweaks Set 1: "-site", "page", and De Novos

We'll add a little extra to the preparation of the BoW here, again using R's string manipulation capabilities, and compare results to the Simple Case. The goal of the tweaking is to demonstrate how seemingly small changes in the preparation of the BoW can dramatically affect the outcome of the LDA process. In addition to the changes I describe blow, I made other similar changes and employed several versions of each of a custom stopword and custom stemming list. In all cases, the LDA topic outcome changed but not in a consistently positive way as measured by the Objective Measure and/or subjectively. Instead of showing the results of a single best outcome, I will show the impact of select changes to introduce a case for consistent formatting and shared master stopword and stemming lists. With that in mind, I add a caveat this section to say that I chose these tweaks to create an important effect, but the development of techniques that will be used in facilitating regulatory compliance should have a well-defined method rather than one-off tweaking. Three "tweaks" are described here.

1. In relation to the discussion in Section 3.5, "off-site" and "on-site" will be made "offsite" and "onsite" throughout the BoW before any stop word removal occurs.
2. In response to "page" showing up in some top term lists in intermediate LDA results, "page" is tossed out of the corpus at the start of forming a BoW.
3. To retain the concept of "de novo" (or "De Novo" because we are doing this before we lower case) becomes "denovo". De Novo refers to a new bank starting from scratch, an important concept in regulation and the targeted entities of some SRLs.

With over 100,000 words in the corpus, it may not be expected that just these changes will make much of a difference in the outcome. It turns out that the Objective Measure increases from 102 / 125 to 106 / 125. Because LDA is a machine learning method, these small code changes resulted in quite different output. LDA is not just reassigning SRLs with these terms, it is looking at all of the co-occurrences again. We review γ and β charts in Figures 17, 18, and 19. This LDA's Topic 9 is a little crowded, and we see the top term is "ffiec" (in its capitalized form, the acronym for Federal Financial Institutions Examination Council), which tells us attachments continue to have an effect. Though some LDA topics are similar, there are many differences. We also see the surfacing of some named entities in the top terms, such as "slhc" (in its capitalized form, the acronym for Savings and Loan Holding Company), but we see the stemmer didn't know what to do with these acronyms, e.g., both "slhc" and "slhcs" made it into top terms.

I provide the same plots of LDA output for comparison to the Simple Case.

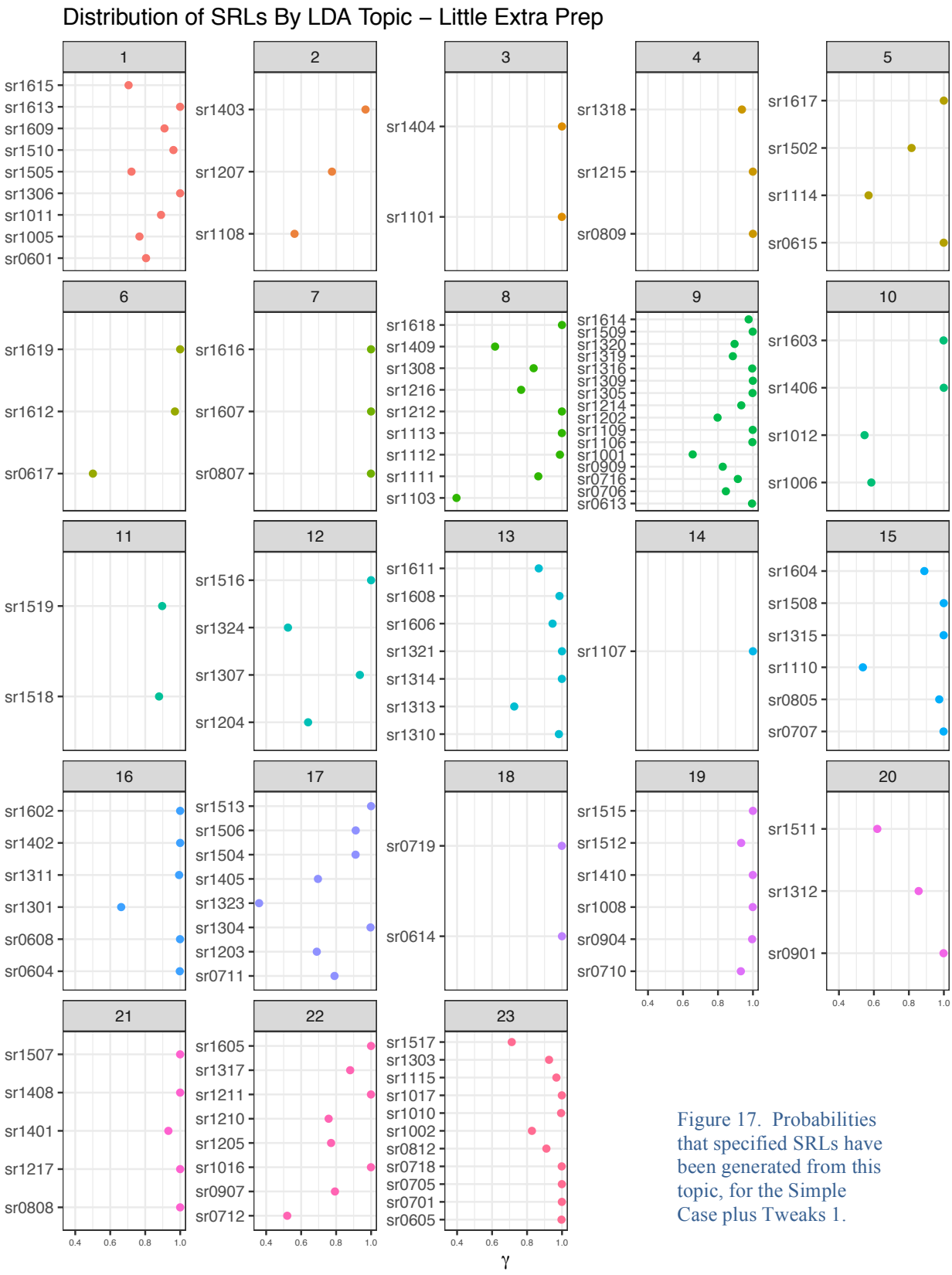


Figure 17. Probabilities that specified SRLs have been generated from this topic, for the Simple Case plus Tweaks 1.

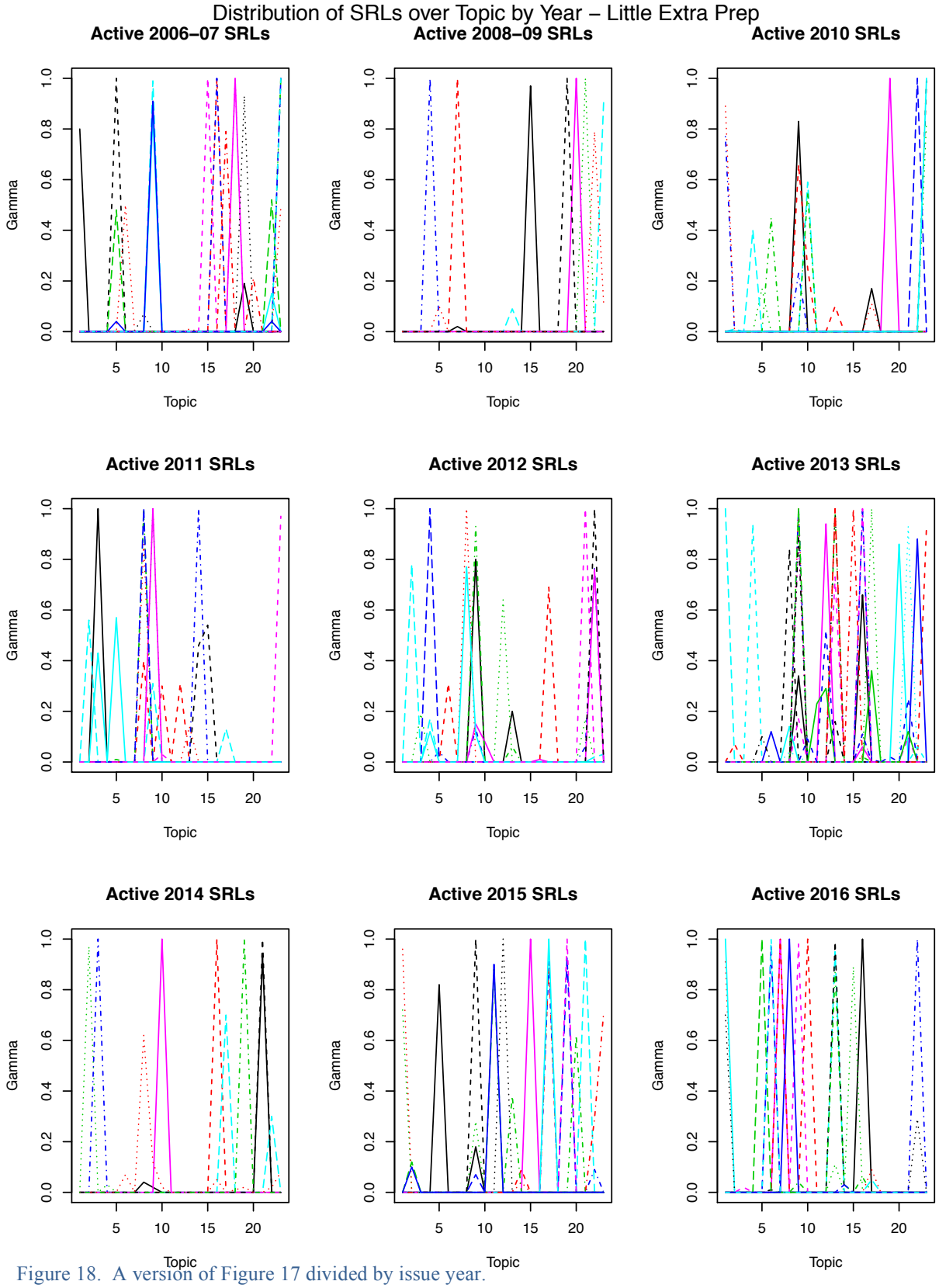


Figure 18. A version of Figure 17 divided by issue year.

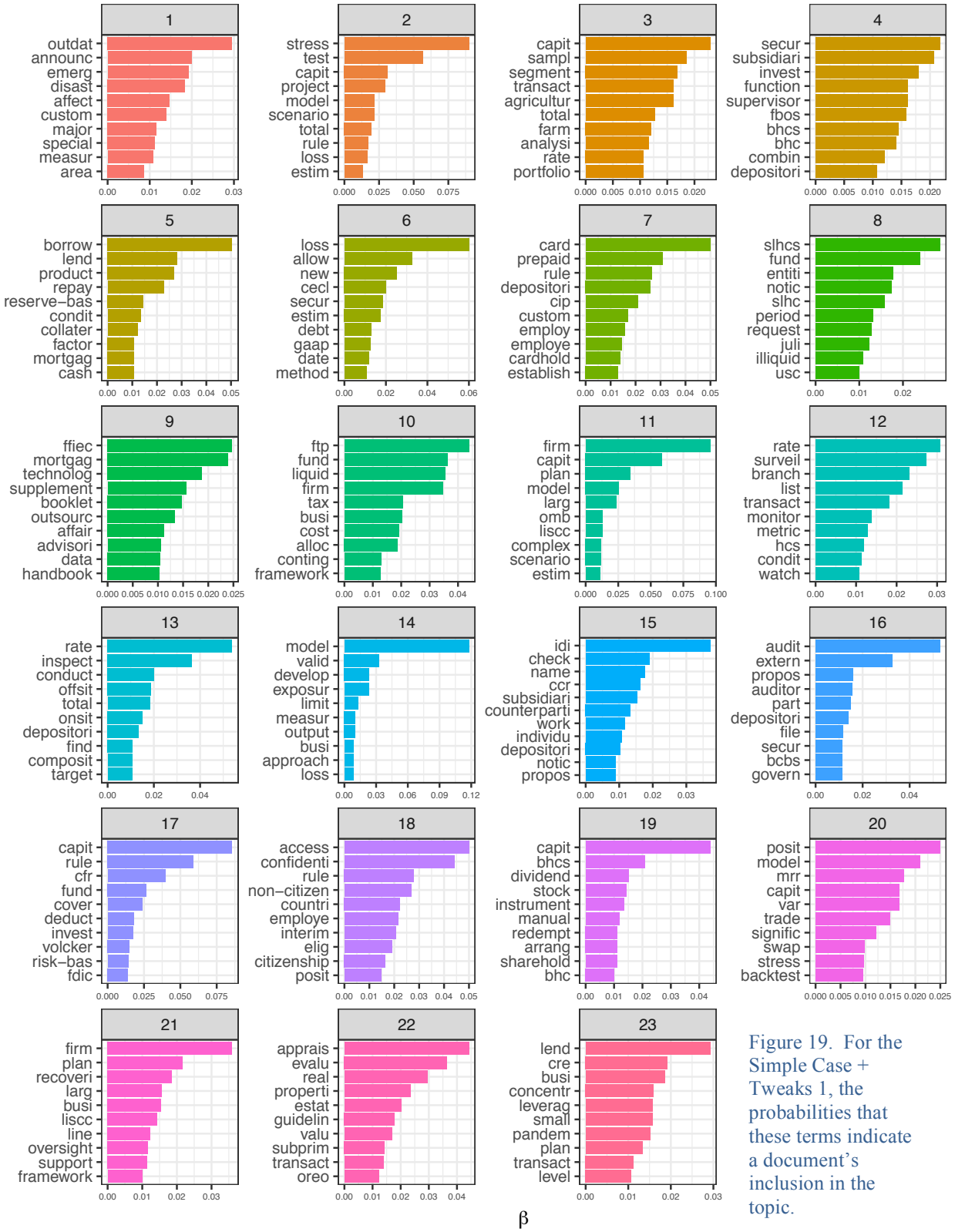


Figure 19. For the Simple Case + Tweaks 1, the probabilities that these terms indicate a document's inclusion in the topic.

Section 5.2 Tweaks Set 2: Nouns and adjectives

The inclusion of named entities is helpful in the interpretation of β plots in Section 3.6. In this set of tweaks, we'll make BoW changes expected to help LDA focus on concepts and named entities. A best way of doing this would be to feed LDA noun phrases. I describe work towards that end in Appendix 2 but that work will extend beyond this thesis. As a next best to noun phrases, I decided to focus on the letter body and extract just nouns and adjectives for inclusion in the BoW while retaining Tweaks Set 1. We will see that, for this particular study, we lose some of the Objective Measure in order to gain more meaningful top terms. My steps are described here.

Step 1. Extracting the SRL body

I use the SRL landmarks in order to extract just the text that is between the letter's subject and the letter's question referral paragraph (or the signature when the question referral is not unique). This was done with R's string processing capabilities.

Step 2. Make adjustments to the R package used for pre-processing

So far, I have been using R's TM package. In order to identify parts of speech, I'll use R's QDAP package, specifically the POS function. In addition to tagging all terms for their part of speech, this function automatically strips white space, removes punctuation, removes numbers, and lowers case so I trade the TM package code for the QDAP code.

Step 3. Extract nouns and adjectives

The POS function applied in Step 2 tags each word with a part of speech. It uses the Penn Tree Bank tags described earlier. I mentioned earlier in the paper that part of speech tagging is quite complex and in a corpus even this size is not expected to be perfect. From the tagged terms

output from the QDAP POS function, I uses R's string capabilities to extract all terms tagged as nouns or adjectives.

Step 4. Stemming

Like the primary case, I stem using the R TM package.

Step 5. Adjust the determination of terms too common to be helpful

I tested changes to the level of occurrence of a term that would determine if it were tossed due to being too common. In the Simple Case BoW, I excluded terms that were included in more than 50 SRLs. In this Noun and Adjective BoW, I exclude terms only if they are included in more than 63 SRLs. This is ½ of the SRLs.

The DTM now has 13,736 non-sparse entries (versus 21,889 in the original case) with 2,306 unique terms (versus 3,526 in the original case). In Figures 20, 21, and 22, we may review γ and β plots as before. The Objective Measure shows that only 95 of the 125 SRLs grouping within LDA topics according to the FRBOG topics. However, the subjective measure of the value of LDA improves as the list of top terms is now more insightful. The use of a BoW made up of nouns and adjectives is more likely to produce terms with intrinsic value that may be used to interpret an LDA topic. I also liked that the most "crowded" LDA topic includes less than in the Simple Case. Here, the most crowded LDA topic has just 11 SRLs. We may consider that somewhat offset by an increase of one to two LDA topics with just one SRL.

In summary, while the terms LDA works with have greater value for interpretation, we lost results according to the Objective Measure. We'll review one more set of tweaks.

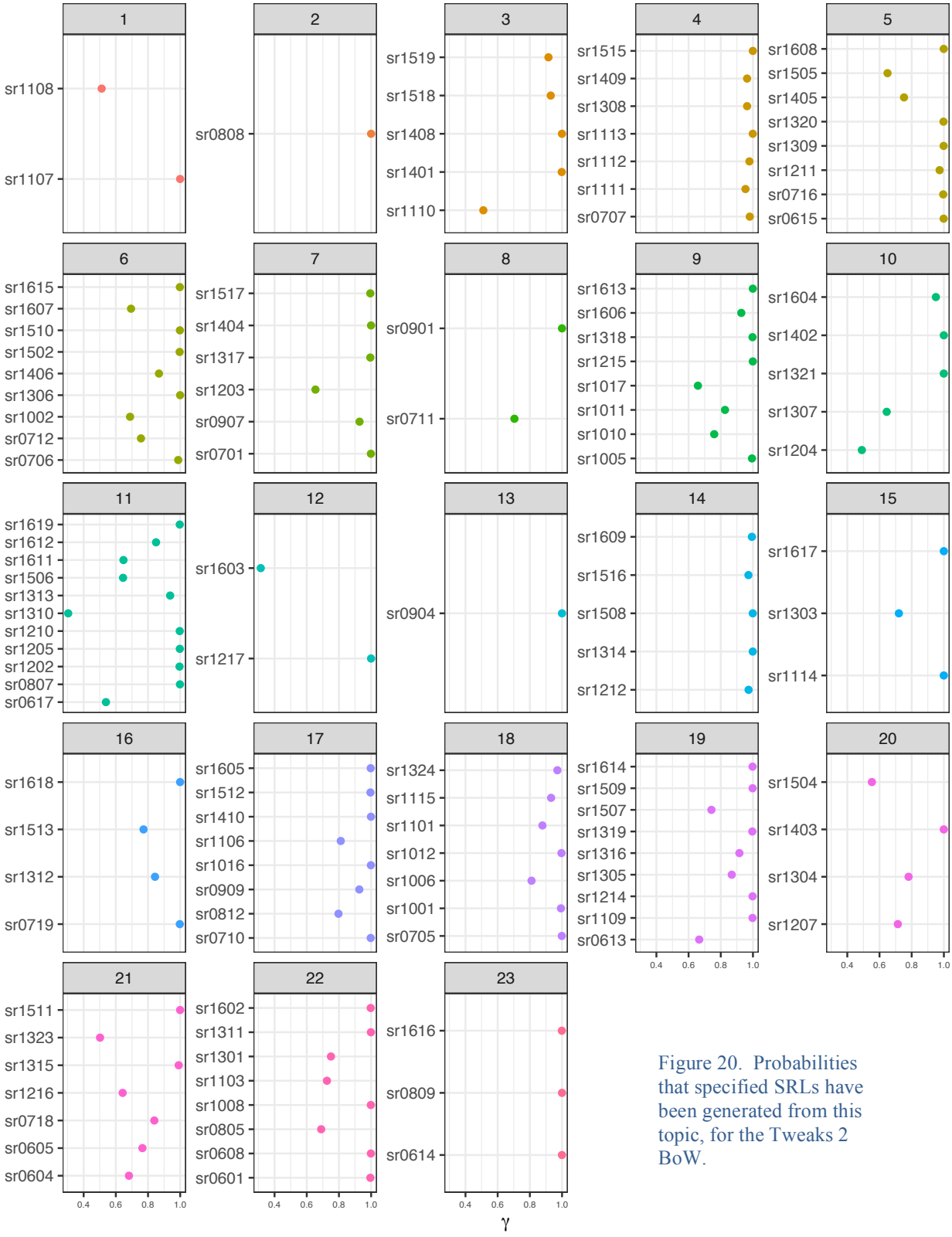


Figure 20. Probabilities that specified SRLs have been generated from this topic, for the Tweaks 2 BoW.

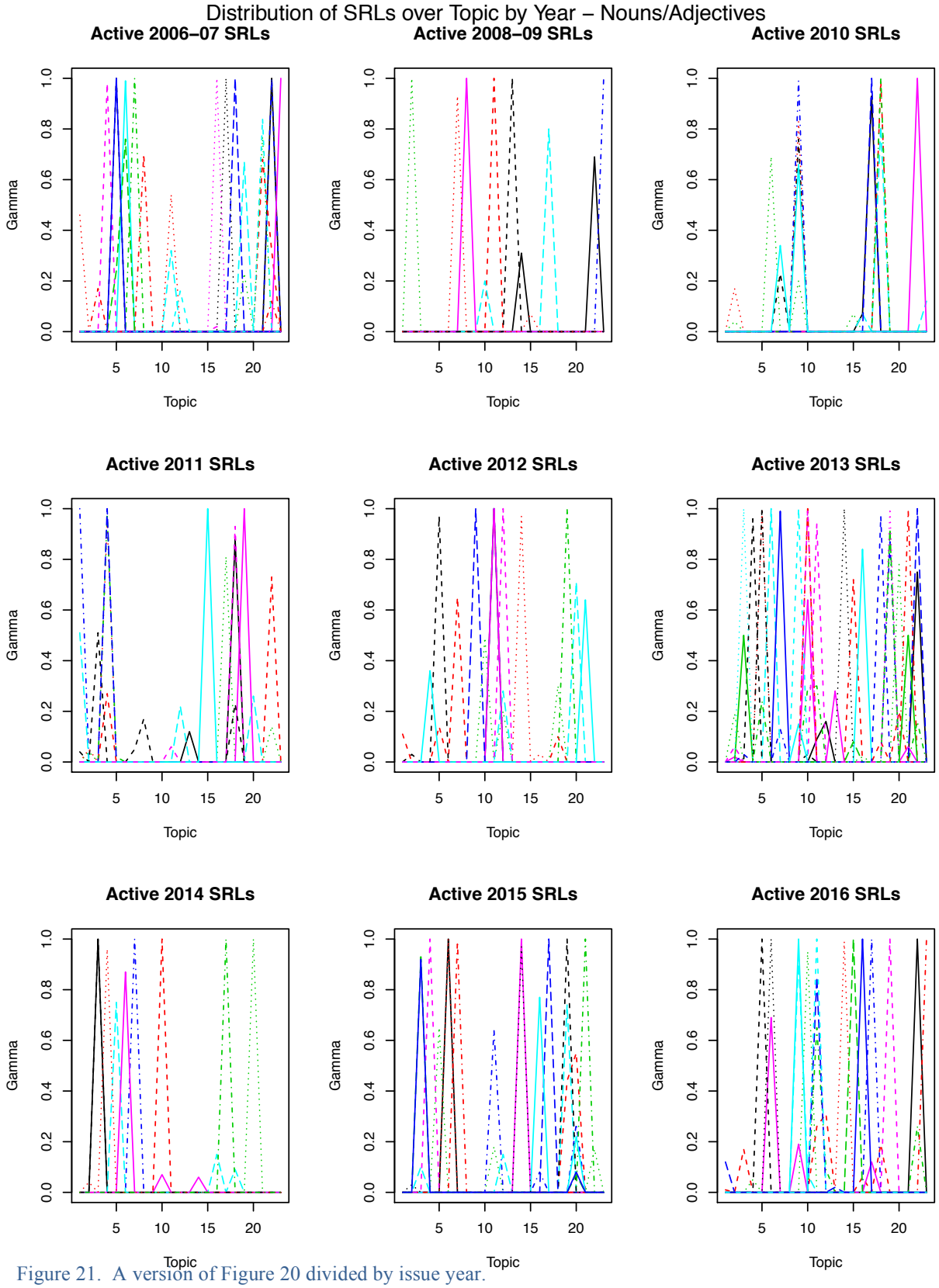


Figure 21. A version of Figure 20 divided by issue year.

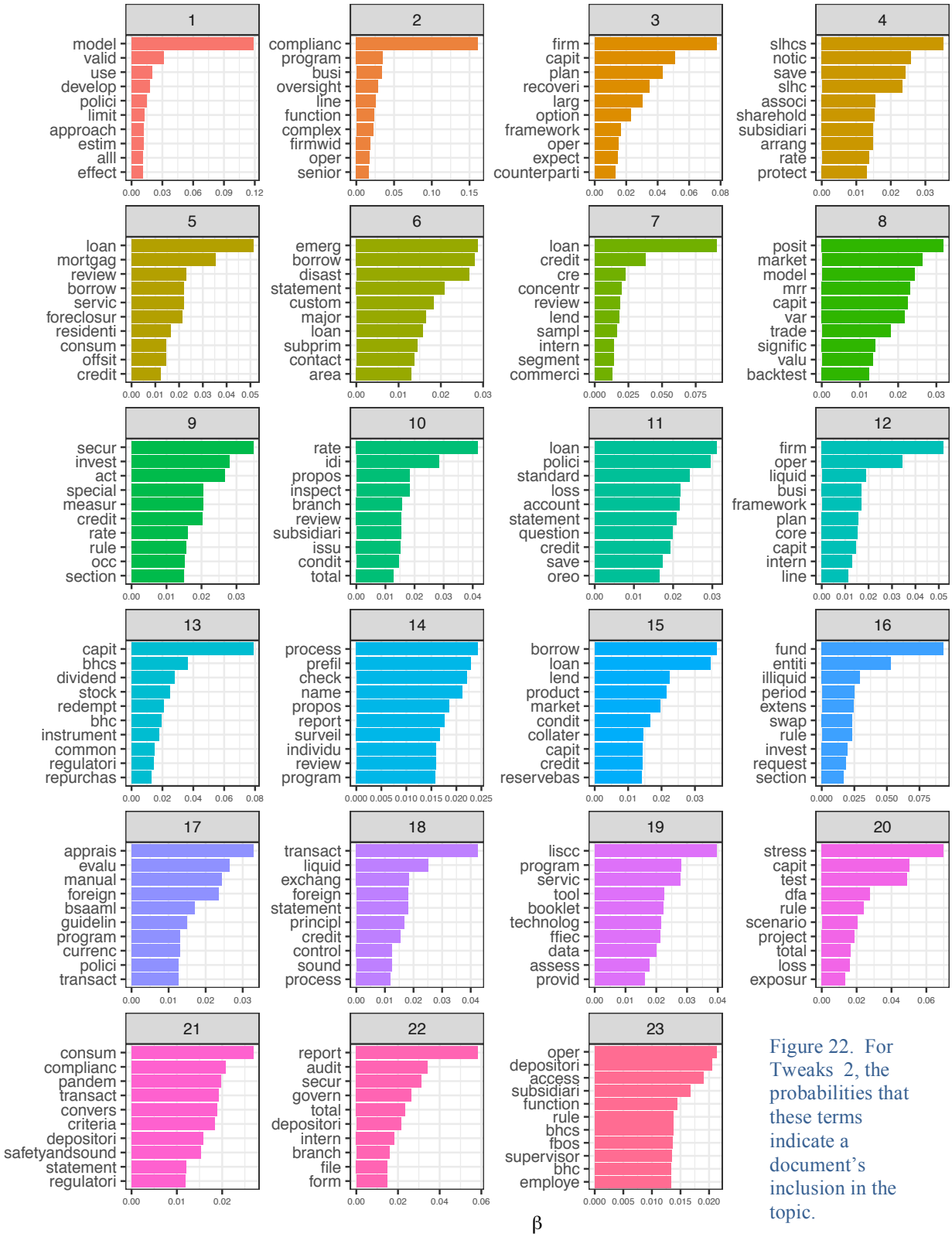


Figure 22. For Tweaks 2, the probabilities that these terms indicate a document's inclusion in the topic.

Section 5.3 Tweaks Set 3: Two examples of the power of stemming

I make two changes to the Noun and Adjective BoW formed by the work described in Section 3.7 to further demonstrate the impact of choices in BoW preparation. We will see how stemming decisions, in this case, improve our overall results.

Step 1. Stemming one acronym

We see acronyms showing up in the top term list and can expect many will be of importance in topic modeling. However, the TM package stemming function will not remove the "s" on a plural of an acronym since the acronym is not in its dictionary. The first BoW change I will make here is to make singular any plural versions of the acronym for Foreign Banking Organizations – "FBOs" will become "FBO".

Step 2. Modifying stemming for one root

The words of interest here stems to the root "consid". I will use R's string capabilities to replace each occasion of " consideration", "considerations", or "considerable", with the token "spconsideration". Consider (no pun intended!) the lexical dispersion plot in Figure 13. The lexical dispersion plot displays the SRL corpus with number of words on the horizontal axis, beginning with 2006 SRLs and ending on the right with 2016 SRLs. When "consideration", "considerations" and "considerable" stem to "consid", word dispersion in the top plot shows "consid" is quite common and, depending on setting, will likely be removed with other too common terms. In the middle plot, we view the dispersion of "consideration" and "considerations". The pattern is quite different than in the top plot. The bottom plot shows dispersion of "considerable" and we see that "considerable" may be an important distinguishing term. Modifying the consideration and considerable tokens improved the Objective Measure. Anecdotally, I expect what happened is that "consider" carries no connotation, but when SRLs

include the word “considerable”, the related noun likely is more burdensome or of concern and is co-occurring with terms that indicate such.

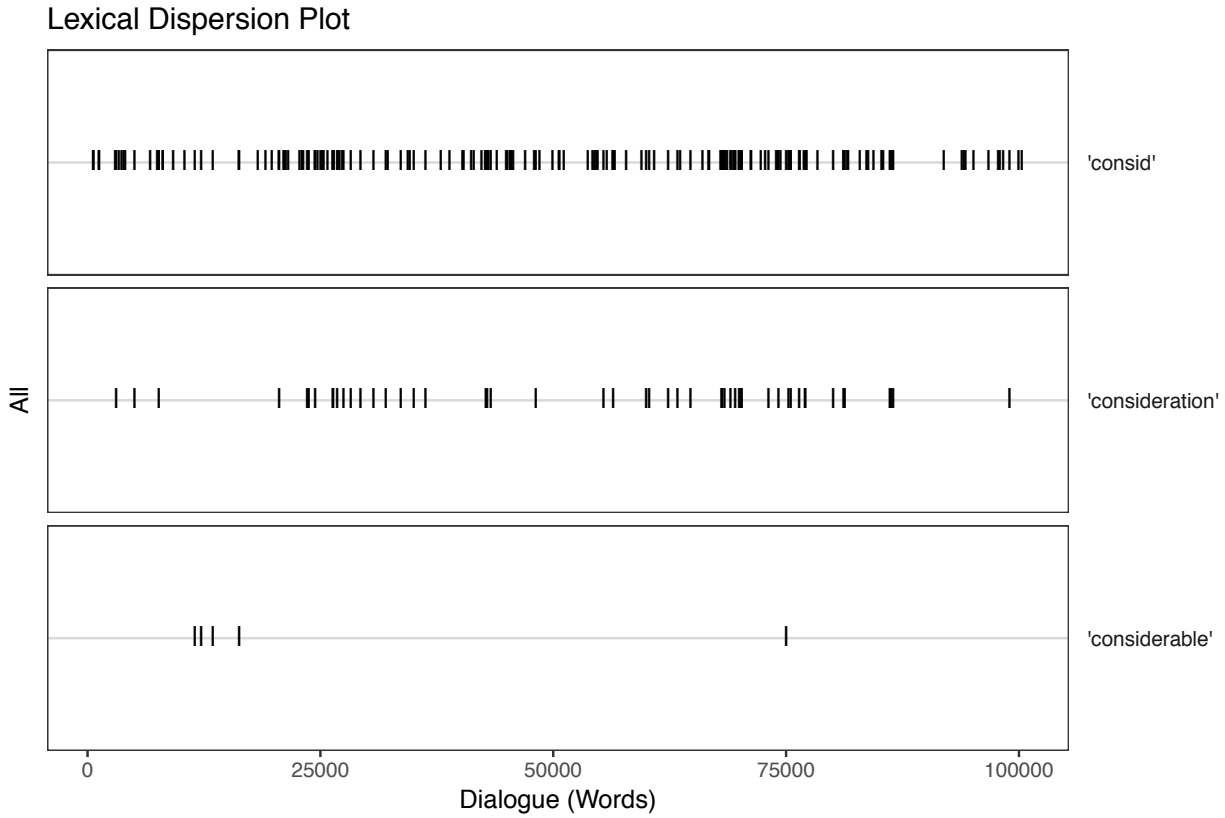


Figure 23. A display of the location of the term specified to the right, where the dialogue begins with the 2006 SRLs.

These two changes resulted in an increase of eight in the Objective Measure, from 95 / 125 after Tweaks Set 2 to 103 / 125 after making the changes. Although the Objective Measure is not at the level of the Simple Case, we see that the creation of a dictionary for use in acronym stemming and an understanding of stemming are critical to obtaining successful LDA topics. We review γ plots in Figures 24 and 25. We see that we continue to have one LDA topic to which a single SRL has been allocated but note that it is a different SRL now isolated. It is interesting to observe the counts of SRLs in each LDA topic as there is somewhat of a tendency for just a few SRLs or relatively many SRLs. We review the β plot for this scenario in Figure 26 and discuss a couple examples of latent topics as we did in the Simple Case.

Example 1 SRLs on modeling and stress tests

Recall that the Simple Case's LDA topic 3 included three SRLs related to stress testing, but one of the three related to banks' capital requirement calculations in accordance with Basel capital standards. That SRL, SR1304 is now in LDA topic 9 with another SRL on Basel guidance, and both of those SRLs are joined in an FRBOG subcategory of the FRBOG topics. LDA topic 9 has a top term of "model" with β just over 0.12. A third SRL in LDA topic 9 includes an SRL on the Fed's model risk management guidelines. This version of LDA has a topic 23 that is more focused on stress testing and includes the four SRLs that make up stress testing guidance.

Example 2 "Real Estate" in LDA topic 16

The FRBOG topic of "Real Estate" is divided into the following subtopics: Appraisals (four SRLs in my sample), Commercial Real Estate (three SRLs in my sample) and Residential Real Estate/Mortgages (ten SRLs in my sample). LDA topic 16 includes 11 SRLS made up of six of the SRLs in the FRBOG subgroup Residential Real Estate/Mortgages, two from the FRBOG subgroup Appraisals, and one from the FRBOG subgroup Commercial Real Estate which is related to loan workouts. With an improved stop word list and stemming process, I expect that LDA could arrive at a high correspondence at a subtopic level in addition to the topic level measured by the Objective Measure.

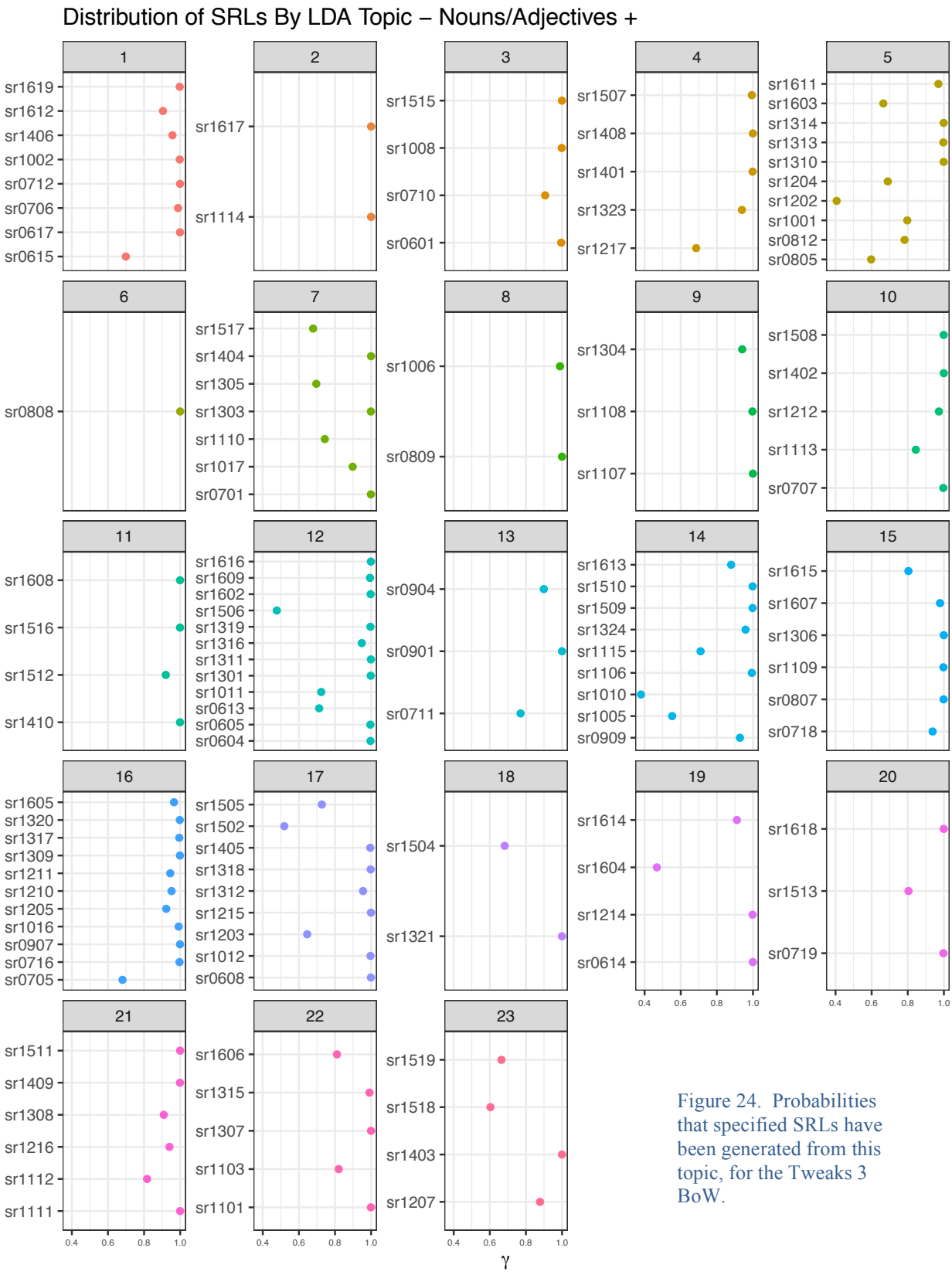


Figure 24. Probabilities that specified SRLs have been generated from this topic, for the Tweaks 3 BoW.

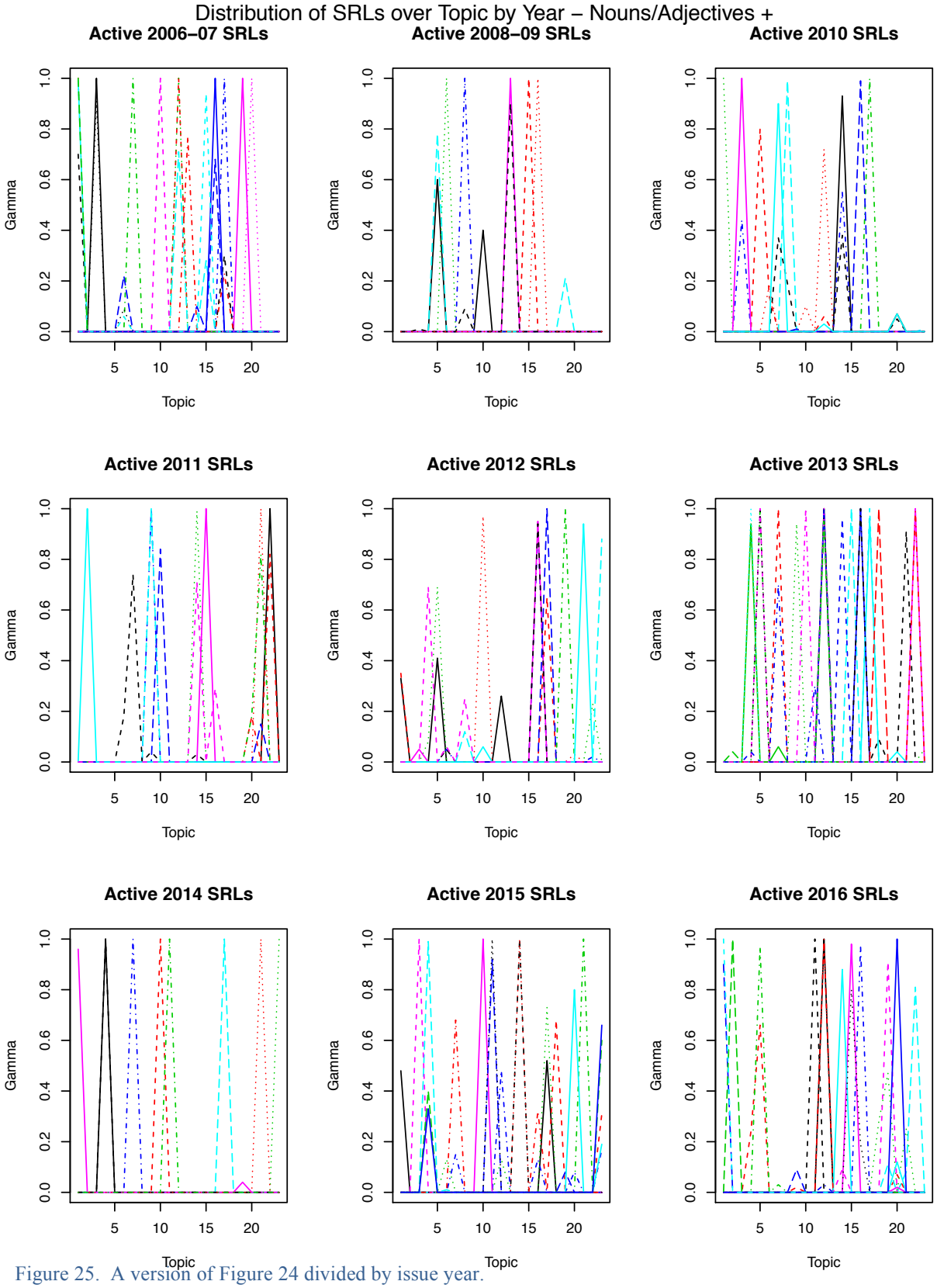


Figure 25. A version of Figure 24 divided by issue year.

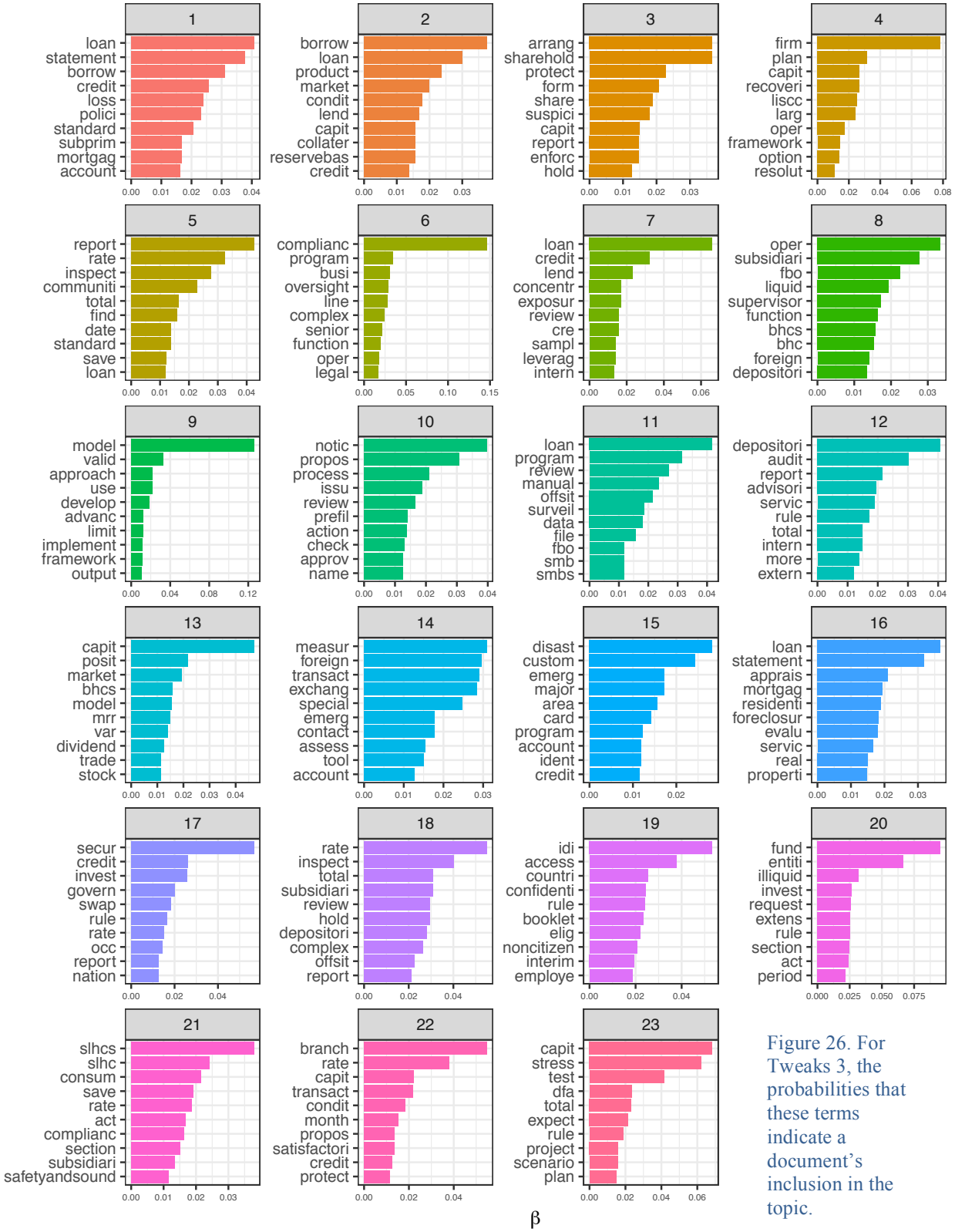


Figure 26. For Tweaks 3, the probabilities that these terms indicate a document's inclusion in the topic.

Chapter 6. Summary of results

In Table 1, I summarize the results including those discussed in Chapter 3.

Summary of Work

	Simple	**Tweaks 1**	**Tweaks 2**	**Tweaks 3**
Head of SRL	In BoW	In BoW	Eliminated	Eliminated
Subject as metadata	Retained	Retained	Retained	Retained
Tail of SRL	In BoW	In BoW	Eliminated	Eliminated
Part of speech tagging	None	None	By QDAP	By QDAP
Use of part of speech tags	NA	NA	Nouns,Adjs	Nouns,Adjs
Lower case	TM	TM	QDAP	QDAP
Strip white space	TM	TM	QDAP	QDAP
Remove punctuation	TM	TM	QDAP	QDAP
Remove stopwords	TM	TM	NA	NA
Stem	TM	TM	TM	TM
Reduction of sparse terms	None	None	None	None
Reduction of common terms	> 50 SRLs	> 50 SRLs	> 63 SRLs	> 63 SRLs
Results:				
DTM Word Count	63332	63332	33674	33671
Non-sparse entries	21889	21894	13736	13743
Corpus unique term count	3526	3525	2306	2308
Objective Measure	**102**	**106**	**95**	**103**
LDA topics where n=<2	5	4	5	4

Table 1. A summary of the results of applying LDA topic modeling to four different BoWs.

In reviewing the LDA results for the Objective Measure, we see that LDA does a good job relative to the human experts' FRBOG topics. Some of the FRBOG topics cross LDA topics and we looked at a few examples where we see that LDA may find subtopic areas of similarity.

From a review of the results of tweaking a simple BoW, we see that a word's inconsistent hyphenation, inclusion in a stop word list, or stem choice can have a significant impact on our results. We note also the special handling required for acronyms.

Figure 27 compares the LDA topic results of the Simple Case and the Tweaks, in which we can see how LDA topic members grouped according to FRBOG topics.

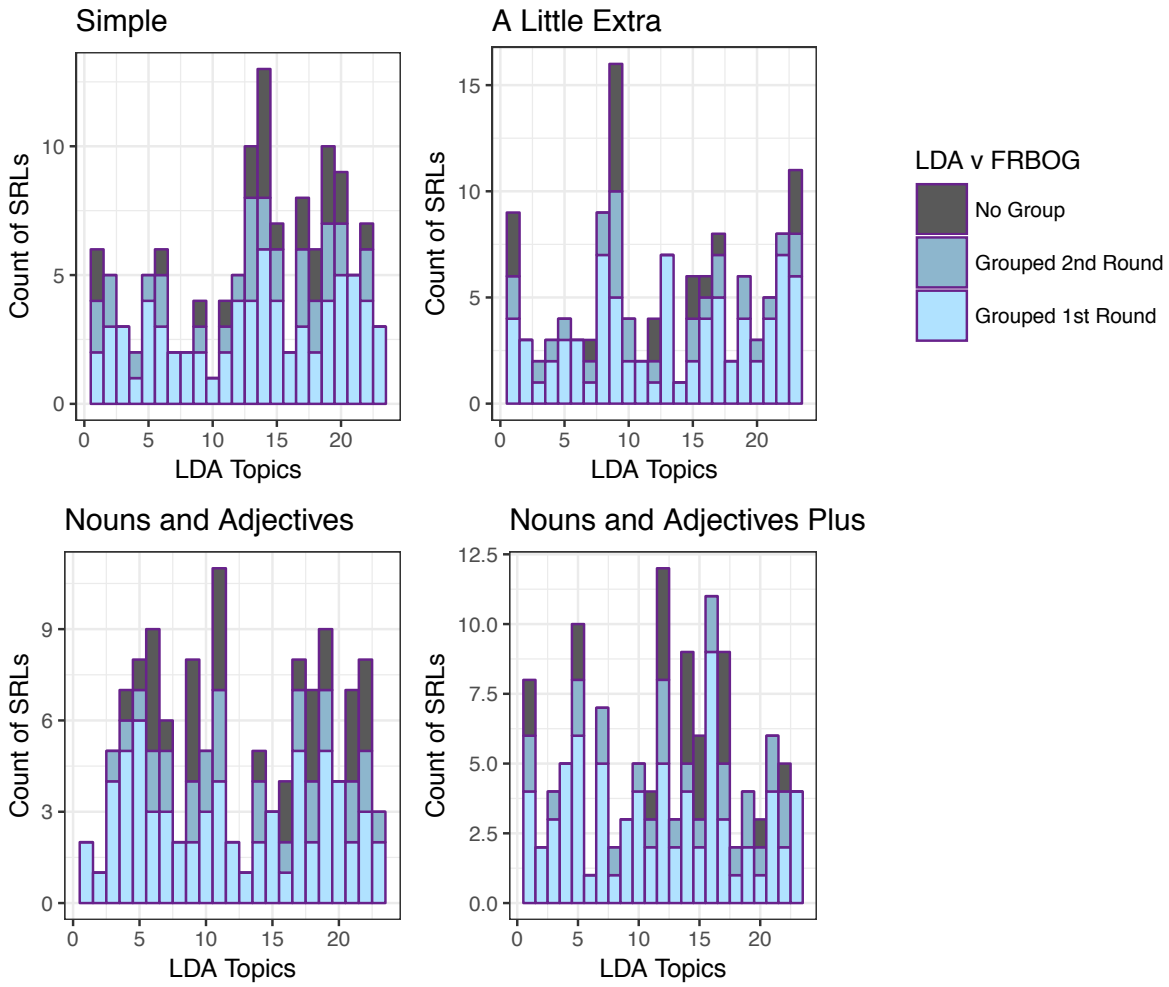


Figure 27. Reviewing the extent to which SRLs grouped by a common latent factor that corresponded to the FRBOG topic groups.

In summary, R packages were used to conduct topic modeling of SRL PDFs and find useful classifications that follow the lines of expert classifications and are suggestive of a next level of commonality. Having done this in an unsupervised fashion, i.e., without defining a desired target outcome, allows us to see the level of classification achievable without intervention. As concern over changing regulation and scarce resources escalates, this finding takes on greater significance. An application of topic modeling on regulatory text will find all pertinent guidance on a specific issue, without being subject to human intervention to pre-identify and label the issues.

Chapter 7. Conclusions

In this paper, I consider ways of forming a BoW of a specific series of bank regulatory guidance and apply LDA to model topics based on the BoWs. I expect that my findings will contribute to the work of Regtech to create automated sharing of regulatory information between regulators and regulated. We see that topic modeling produces promising results, but those results are dependent on applying best practices to BoW formation. To that end, I make the following recommendations related to the preparation of communications.

- When information is communicated in the form of a series or other designation of like documents, contents should be standardized and consistent across formats. For example, attachments should be consistently included or not included and the use of "signed by" preceding a signature should be consistent in all formats.
- Fields that may be drawn in by text analytics as metadata should be formatted consistently whenever possible.
- Landmarks in the communication may be critical to implementing effective text analytics. Consistency of text, form, and placement of landmarks is necessary to establish them as such within the text.
- Grammar rules on hyphenation may be complicated, but consistent use of hyphens is necessary. A standard may include applicable grammar rules in addition to practices specific to the industry.
- The effectiveness of stemming is related to the actions taken by the stemmer, e.g., singularizing plurals or using the root. Decisions about stemming should take into consideration how any outputted lists of terms will be used in addition to common word usage. Common word usage will enlighten identification of terms for which inference is diluted by stemming.

- A stop word list for text analytics of bank regulatory information must be agreed upon and consistently applied.
- Acronyms are troublesome in text analytics. A special stemming dictionary will help, or perhaps it will be determined that making an acronym plural or possessive could be done in some way that would not create a new "token" different than the acronym itself.
- Concepts should be referenced consistently as much as possible. Consider SR0605, in which we see "influenza pandemic" when describing "preparedness" but "pandemic influenza" when describing "outbreak". Or SR 1607, in which we see "prepaid access card" in the subject but "prepaid card" everywhere else.
- I mention earlier the laws that require the use of plain language in Federal agency proposed and final rulemakings. In addition to that requirement in the Plain Writing Act of 2010 and the Gramm-Leach-Bliley Act of 2009, we also now have the DATA Act of 2014 which addresses standards in communicating information about federal expenditures. I believe that these requirements should be extended to encompass a set of best practices for creating communications that will not unnecessarily hinder text analytics efforts.

Chapter 8. Future work

A future work will research alternative sentiment dictionaries and potentially begin building a dictionary applicable to bank regulatory text and perhaps other regulatory text. Such a dictionary would include words that are "restrictive" versus "permissive" in banking regulation.

A second work will be to explore the expansion of the use of word clouds. I particularly appreciate the promise of data visualizations to share statistically-gained insights without a requirement that the viewer understand statistics. (Castella and Sutton 2014) discusses expanding the use of word clouds to create "word storms." They discuss using word clouds as a tool to compare individual documents, to view changes in documents over time, and to establish a hierarchy among the documents. They recommend new algorithms which would increase control over the cloud content more than is provided by current word cloud tools. An example of what added control may provide is a visualization of "small multiples" of clouds that display the same word in the same place in the cloud across the storm. I used the R package wordcloud to create some comparative clouds as an initial step in this direction. I include a word storm of comparisons of the SR Letters issued in 2016 in Appendix 4.

End Notes

1. *Federal Reserve Purposes and Functions*.
2. Murphy 2015
3. Silge, J. at http://stackoverflow.com/questions/43282771/loading-loughran-finance-sentiment-into-tidytext

Bibliography

The bibliography has been divided into the following sections:

SRLs
Books and book chapters
Reports
Video
Blogs, articles, posts, and speeches
Academic, academic-style, and conference papers
R Packages

SRLs

Board of Governors of the Federal Reserve System Division of Banking Supervision and Regulation (2006 to 2016), Supervision and Regulation Letters SR0601 through SR1619. Retrieved between December 26, 2016, and January 7, 2017, from https://www.federalreserve.gov/supervisionreg/srletters/srletters.htm.

Books and book chapters

(2016). " Supervising and Regulating Financial Institutions and Activities." In The Federal Reserve System Purposes and Functions. Washington, D.C.: Federal Reserve System.

Blatt, B. (2017). "Searching for Fingerprints." In Nabokov's Favorite Word Is Mauve: What the Numbers Reveal About the Classics, Bestsellers, and Our Own Writing, Simon & Schuster Digital Sales, Inc.

Ingersoll, G. S., T. S. Morton and A. L. Farris (2013). Taming text: how to find, organize, and manipulate it, Manning Publications Co.

Jockers, M. L. (2014). Text analysis with R for students of literature, Springer.

Miner, G., J. Elder IV and T. Hill (2012). Practical text mining and statistical analysis for non-structured text data applications, Academic Press.

Sanchez, G. (2013). "Handling and processing strings in R." Trowchez Editions. Retrieved first quarter, 2017, from http://gastonsanchez.com/Handling_and_Processing_Strings_in_R.pdf.

Silge, J.A. and D. Robinson (2017). Text Mining with R: A Tidy Approach, forthcoming from O'Reilly Media, Inc. Retrieved April, 2017, from http://tidytextmining.com/index.html.

Teetor, P. (2011). "R Cookbook: Proven recipes for data analysis, statistics, and graphics, " O'Reilly Media, Inc.

Weiss, S. M., N. Indurkhya and T. Zhang (2010). Fundamentals of Predictive Text Mining, Springer.

Weiss, S. M., N. Indurkhya, T. Zhang and F. Damerau (2010). Text mining: Predictive Methods for Analyzing Unstructured Information, Springer Science & Business Media.

Reports

From regulators or other public organizations

U.S. Government Accountability Office (2016), GAO-16-175 “FINANCIAL REGULATION: Complex and Fragmented Structure Could Be Streamlined to Improve Effectiveness” (Washington, U. S. G. A. Office). Retrieved first quarter, 2017 from https://www.gao.gov/products/GAO-16-175.

(2016). "Call for input on supporting the development and adopters of RegTech FS16/4." Retrieved fist quarter, 2017, from https://www.fca.org.uk/publication/feedback/fs-16-04.pdf.

Bholat, D. M., S. Hansen, P. M. Santos and C. Schonhardt-Bailey (2015). "Handbook - No. 33 Text mining for central banks." Retrieved Janaury, 2017, from http://www.bankofengland.co.uk/education/Documents/ccbs/handbooks/pdf/ccbshb33.pdf.

Murphy, E. V. Who Regulates Whom and How? An Overview of U.S. Financial Regulatory Policy for Banking and Securities Markets. Congressional Research Service. Washington, D.C. Retrieved first quarter, 2017, from http://digitalcommons.ilr.cornell.edu/cgi/viewcontent.cgi?article=2154&context=key_workplace.

From private firms

(2015). "RegTech is the new FinTech How agile regulatory technology is helping firms better understand and manage their risks." Retrieved fist quarter, 2017, from https://www2.deloitte.com/ie/en/pages/financial-services/articles/RegTech-is-the-new-FinTech.html.

(2016). "Innovating with RegTech Turning regulatory compliance into a competitive advantage." Retrieved fist quarter, 2017, from http://www.ey.com/Publication/vwLUAssets/EY-Innovating-with-RegTech/$FILE/EY-Innovating-with-RegTech.pdf.

(2017). "Will regtech save us from regulations?" KPMG Insights. Retrieved fist quarter, 2017, from https://home.kpmg.com/nl/en/home/insights/2017/01/will-regtech-save-us-from-regulations.html.

English, S. and S. Hammond (2014). “Cost of Compliance 2016”. Retrieved first quarter, 2017, from https://risk.thomsonreuters.com/en/resources/special-report/cost-compliance-2016.html.

van Liebergen, B., Portilla, A., Silverberg K., French, C. (2016). Regtech in Financial Services: Technology Solutions for Compliance and Reporting, Institute of International Finance. Retrieved first quarter, 2017, from https://www.iif.com/publication/research-note/regtech-financial-services-solutions-compliance-and-reporting.

Video

Blei, D. (2009). Topic Models Machine Learning Summer School (MLSS) [internet video], Cambridge 2009, University of Cambridge.

Blogs, articles, posts, and speeches

(2017, March 7, 2017). "Supervisory Policy and Guidance Topics." Retrieved throughout first quarter, 2017, from https://www.federalreserve.gov/supervisionreg/topics/topics.htm.

(2017, April 22, 2016). "Supervision and Regulation Letters." Retrieved throughout first quarter, 2017, from https://www.federalreserve.gov/bankinforeg/srletters/about.htm.

Bergmann, T. (2015). "Understanding the generative nature of LDA with R." Retrieved first quarter, 2017, from http://tillbergmann.com/blog/lda-generation-R.html.

Blei, D. (2010, February 3, 2010). "assigning documents to topics when document frequency matrix is too large." Princeton Topic-models Mailing List. Retrieved first quarter, 2017 from https://lists.cs.princeton.edu/pipermail/topic-models/2010-February/000705.html.

Blei, D. (2014, November 18, 2014). "From final.gamma to \theta." Princeton Topic-models Mailing List, 2017. Retrieved from https://lists.cs.princeton.edu/pipermail/topic-models/2014-November/002990.html.

Crosman, P, M. Wisniewski, T. Macheel and others (2016). Regtech Is Real; The big takeaways from IBM's deal to acquire the consulting firm Promontory. *American Banker Magazine.* 126 No. 11: 14.

Daume III, H. (2008). "Evaluating topic models." natural language processing blog. Retrieved first quarter, 2017, from https://nlpers.blogspot.com/2008/06/evaluating-topic-models.html 2017.

Frederick, A. (2016). "10 things CIOs must know about big data text analytics platforms." Retrieved April 9, 2017 from https://www.ibm.com/blogs/watson/2016/06/10-things-cios-must-know-big-data-text-analytics-platforms/.

Gattuso, J., and Katz, D. (2016). “Red Tape Rising 2016: Obama Regs Top $100 Billion Annually”. Retrieved first quarter, 2017, from http://www.heritage.org/government-regulation/report/red-tape-rising-2016-obama-regs-top-100-billion-annually.

Goldsmith-Pinkham, P., Hirtle, B., and Lucca, D. (2016) "A Peek behind the Curtain of Bank Supervision." *Liberty Street Economics*. Retrieved first quarter, 2017, from http://libertystreeteconomics.newyorkfed.org/2016/04/a-peek-behind-the-curtain-of-bank-supervision.html.

Grimes, S. (2014). "Naming & Classifying: Text Analysis Vs. Text Analytics." *The Blog*. Retrieved first quarter, 2017, from http://www.huffingtonpost.com/seth-grimes/naming-classifying-text-a_b_4556621.html 2017.

Grün , B. (2015). "Course: "Introduction to Text Mining in R"." Retrieved February, 2017, from http://ifas.jku.at/gruen/TextMining/slides-02-topicmodels.pdf.
Hornik, K. (2016). "openNLP: Apache OpenNLP Tools Interface." R package version 0.2-6.

Jockers, M. (2011). "The LDA Buffet Is Now Open; Or, Latent Dirichlet Allocation for English Majors." Retrieved first quarter, 2017, from http://www. matthewjockers. net/2011/09/29/the-lda-buffet-is-now-open-or-latentdirichlet-allocation-for-english-majors.

Levy, M. (2016). "Playing with Twitter Data." *Thoughts on networks, environmental behavior, policy, data science, and academic life.* Retrieved first quarter, 2017, from http://michaellevy.name/blog/conference-twitter/ 2017.

Meza, D. (2015). "Topic Modeling in R." *David Meza* Retrieved first quarter, 2017, from http://davidmeza1.github.io/2015/07/20/topic-modeling-in-R.html 2017.

Matthun, R. (2016). "RegTech will change the way we regulate: The emerging field of RegTech promises an algorithmic alternative to regulation that can reduce friction and simplify compliance." Retrieved first quarter, 2017 from http://www.livemint.com/Opinion/8WcbKsj7my7ZrPiuN9mOkL/RegTech-will-change-the-way-we-regulate.html.

Nash, B. J. (2013). "To be clear: Muddy language can be costly." *Econ Focus*(1Q): 17-19. Retrieved first quarter, 2017, from https://www.richmondfed.org/-/media/richmondfedorg/publications/research/econ_focus/2013/q1/pdf/feature1.pdf.

O'Connor, B. (2016). A Little Bit of NLP Goes a Long Way: Adding Phrases to the Term-Document Matrix using Finite-State Shallow Parsing [a slide deck]. *Conference on New Directions in Text as Data*. Retrieved first quarter, 2017, from http://brenocon.com/oconnor_textasdata2016.pdf.

Rinker, T. (unavailable). "Formality Score." Retrieved first quarter, 2017, from https://trinker.github.io/qdap/formality.html.

Rinker, T. (2012). "Presidential Debates with qdap-beta." *TRinker's R Blog*. Retrieved first quarter, 2017, from https://trinkerrstuff.wordpress.com/2012/10/04/presidential-debates-with-qdap-beta/ 2017.

Rinker, T. (2013). "qdap Package Vignette." Retrieved first quarter, 2017, from https://trinker.github.io/qdap/vignettes/qdap_vignette.html.

Rinker, T. W. (2016). "qdap-tm Package Compatibility." Retrieved first quarter, 2017, from https://cran.r-project.org/web/packages/qdap/vignettes/tm_package_compatibility.pdf.

Silge, J. (2016) "The Life-Changing Magic of Tidying Text." *Data science ish.* Retrieved first quarter, 2017, from http://juliasilge.com/blog/Life-Changing-Magic/.

Silge, J. (2016). "Term Frequency and tf-idf Using Tidy Data Principles." *Data science ish.* Retrieved first quarter, 2017, from http://juliasilge.com/blog/Term-Frequency-tf-idf/.

Silge, J. (2016, August 2, 2016). *Rpubs*. "NASA Metadata: tf-idf of Description Texts and Keywords." Retrieved first quarter, 2017, from https://rpubs.com/juliasilge/200028.

Silge, J. (2016, October 21, 2016). *Rpubs*. "NASA Metadata: 64 Topics for Topic Modeling?." Retrieved first quarter, 2017, from https://rpubs.com/juliasilge/220655.

Silge, J. (2016). "NASA Metadata: Topic Modeling of Description Texts." *Rpubs*. Retrieved first quarter, 2017, from https://rstudio-pubs-static.s3.amazonaws.com/201707_d839d6be1b9c4946bcf131121b1ba4b0.html.

Silge, J., D. Robinson (2016). "Tidy Topic Modeling." [Vignette at CRAN.] Retrieved first quarter, 2017, from https://cran.r-project.org/web/packages/tidytext/vignettes/topic_modeling.html.

Sinha, N. (2014). "Using big data in finance: Example of sentiment-extraction from news articles." FEDS Notes. Retrieved first quarter, 2017, from https://www.federalreserve.gov/econresdata/notes/feds-notes/2014/using-big-data-in-finance-example-of-sentiment-extraction-from-news-articles-20140326.html 2017.

Skinner, C. (2017). "The Semantic Regulator (#RegTech Rules)." *Chris Skinner's Blog.* Retrieved first quarter, 2017, from http://thefinanser.com/2017/01/semantic-regulator-regtech-rules.html/ 2017.

Tarullo, Daniel K. (2014). "Rethinking the Aims of Prudential Regulation " speech delivered at the Federal Reserve Bank of Chicago Bank Structure Conference. Retrieved first quarter, 2017, from https://www.federalreserve.gov/newsevents/speech/tarullo20140508a.htm.

Trim, C. (2013). "The Art of Tokenization." *Language Processing*. Retrieved first quarter, 2017, from https://www.ibm.com/developerworks/community/blogs/nlp/entry/tokenization?lang=en 2017.

Underwood, T. (2012). "Topic modeling made just simple enough." *The Stone and the Shell.* Retrieved first quarter, 2017, from https://tedunderwood.com/2012/04/07/topic-modeling-made-just-simple-enough/.

Wallach, H. M. “topic models: priors, stop words and languages” [a slide deck]. Retrieved first quarter, 2017, from https://people.cs.umass.edu/~wallach/talks/priors.pdf.

Wikipedia Contributors. (March 10, 2017). "Latent Dirichlet Allocation." *Wikipedia, The Free Encyclopedia.* Retrieved fist quarter, 2017, from https://en.wikipedia.org/w/index.php?title=Special:CiteThisPage&page=Latent_Dirichlet_allocation&id=769669632.

Yazici, M. (2015). "Deriving insights from text mining and machine learning." *IBM Big Data & Analytics Hub*. Retrieved first quarter, 2017, from http://www.dataversity.net/banking-fibo-financial-institutions-turn-standard-value-compliance/.

Zaino, J. (2016) "Banking on FIBO: Financial Institutions Turn to Semantic Standard." Retrieved first quarter, 2017, from http://www.dataversity.net/banking-fibo-financial-institutions-turn-standard-value-compliance/.

Zhao, W. (unknown). “Best Practices in Building Topic Models with LDA for Mining Regulatory Textual Documents, NCTR CTP Working Group” [a slide deck]. Retrieved first quarter, 2017, from http://www.phusewiki.org/wiki/images/c/c9/Weizhong_Presentation_CDER_Nov_9th.pdf.

Academic, academic-style, and conference papers

Bank of England (2015), ‘One Bank Research Agenda Discussion Paper’. Retrieved first quarter, 2017, from http://www.bankofengland.co.uk/research/Documents/onebank/discussion.pdf.

Binkley, D., D. Heinz and D. Lawrie "Understanding LDA for Software Engineering." Retrieved March, 2017, from http://www.cs.loyola.edu/~lawrie/papers/Loy_TR_3_23.pdf.

Blei, D. M. (2012a). "Probabilistic topic models." *Communications of the ACM* 55(4): 77-84.

Blei, D. M. (2012b). "Topic modeling and digital humanities." *Journal of Digital Humanities* 2(1): 8-11.

Blei, D. and J. Lafferty (2006). "Correlated topic models." *Advances in neural information processing systems* 18: 147.
Blei, D. M. and J. D. Lafferty (2007). "A correlated topic model of science." *The Annals of Applied Statistics:* 17-35.

Blei, D. M. and J. D. Lafferty (2009). "Topic models." *Text mining: classification, clustering, and applications* 10(71): 34.

Blei, D. M., A. Y. Ng and M. I. Jordan (2003). "Latent dirichlet allocation." *Journal of Machine Learning Research* 3(Jan): 993-1022.

Brett, M. R. (2012). "Topic Modeling: A Basic Introduction." *Journal of Digital Humanities* 2(1).

Cannon, S. (2015). "Sentiment of the FOMC: Unscripted." *Economic Review-Federal Reserve Bank of Kansas City*: 5-31.

Castella, Q. and C. Sutton (2014). Word storms: Multiples of word clouds for visual comparison of documents. *Proceedings of the 23rd international conference on World wide web*, ACM.

Chandrasekaran, B., J. R. Josephson and V. R. Benjamins (1999). "What are ontologies, and why do we need them?" *IEEE Intelligent Systems and their applications* 14(1): 20-26.
Chaney, AJ-B. and D.M. Blei (2012). "Visualizing Topic Models." In *Proc. of the 6th Intern. Conf. on Weblogs and Social Media*, vol. 3.

Chang, J., J. L. Boyd-Graber, S. Gerrish, C. Wang and D. M. Blei (2009). "Reading tea leaves: How humans interpret topic models." *Nips*, vol. 31, pp. 1-9.

Chuang, J., C. D. Manning and J. Heer (2012). ""Without the Clutter of Unimportant Words": Descriptive keyphrases for text visualization." *ACM Transactions on Computer-Human Interaction (TOCHI)* 19(3): 19.

Doyle, G. and C. Elkan (2009). Financial topic models. *Working Notes of the NIPS-2009 Workshop on Applications for Topic Models: Text and Beyond Workshop*.

Du, J., J. Jiang, D. Song and L. Liao (2015). Topic modeling with document relative similarities, *IJCAI*.

Fei, G., Z. Chen, and B. Liu (2014) "Review Topic Discovery with Phrases using the Pólya Urn Model." *COLING*, pp. 667-676.

Griffiths, T. L., and M. Steyvers (2004). "Finding scientific topics." *Proceedings of the National Academy of Sciences* 101, no. suppl: 5228-5235.

Handler, A., M. J. Denny, H. Wallach and B. O'Connor (2016). "Bag of What? Simple Noun Phrase Extraction for Text Analysis." *NLP+ CSS* 2016: 114.

Hendry, S. and A. Madeley (2010). "Text mining and the information content of Bank of Canada communications." Bank of Canada-Staff Working Paper 2010-31.

Heylighen, Francis. "Advantages and limitations of formal expression." *Foundations of Science* 4, no. 1 (1999): 25-56.

Heylighen, F. and J.-M. Dewaele (2002). "Variation in the contextuality of language: An empirical measure." *Foundations of Science* 7(3): 293-340.

Hornik, K. and B. Grün (2011). "topicmodels: An R package for fitting topic models." *Journal of Statistical Software* 40(13): 1-30.

Hu, M. and B. Liu (2004). Mining opinion features in customer reviews. *AAAI*.

Ko, Y. and J. Seo (2000). Automatic text categorization by unsupervised learning. *Proceedings of the 18th conference on Computational linguistics-Volume 1, Association for Computational Linguistics.*

Lau, G., K. H. Law and G. Wiederhold (2005). "Analyzing government regulations using structural and domain information." *Computer* 38(12): 70-76.

Li, W., P. Azar, D. Larochelle, P. Hill, and A. W. Lo. "Law is code: a software engineering approach to analyzing the United States code." *J. Bus. & Tech. L.* 10 (2015): 297.

Loughran, T. and B. McDonald (2011). "When is a liability not a liability? Textual analysis, dictionaries, and 10 Ks." *The Journal of Finance* 66(1): 35-65.

Martin, F. and M. Johnson (2015). “More Efficient Topic Modelling Through a Noun Only Approach.” *Australasian Language Technology Association Workshop 2015*.

McCarthy, J., L. Vasiliu, A. D. Grody, C. Muckley, D. Lawrence, F. Zervoudakis, S. Tabet, J. van Grondelle, T. Bouras and K. Fernandes (2013). “Financial Industry Ontologies for Risk and Regulation Data (FIORD)—A Position Paper.” Working Conference on Virtual Enterprises, pp. 737-744. Springer Berlin Heidelberg.

Meyer, D., K. Hornik and I. Feinerer (2008). “Text mining infrastructure in R.” *Journal of Statistical Software* 25(5): 1-54.

Nopp, C. and A. Hanbury (2015). “Detecting Risks in the Banking System by Sentiment Analysis.” *EMNLP*.

Park, Y. and R. J. Byrd (2001). “Hybrid text mining for finding abbreviations and their definitions.” *Proceedings of the 2001 conference on empirical methods in natural language processing*, pp. 126-133.

Ramage, D., Rosen, E., Chuang, J., Manning, C. D., McFarland, D. A. (2009). “Topic Modeling for the Social Sciences.” *NIPS 2009 Workshop on Applications for Topic Models: Text and Beyond,* Whistler, Canada.

Reed, C. (2012) "Latent Dirichlet Allocation: Towards a Deeper Understanding."

Santorini, B. (1990). "Part-of-speech tagging guidelines for the Penn Treebank Project (3rd revision)." University of Pennsylvania Department of Computer and Information Science Technical Report. No. MS- CIS-90-47

Shapiro, A. H., M. Sudhof and D. Wilson (2017). “Measuring News Sentiment.” Federal Reserve Bank of San Francisco-Working Paper Series. No. 2017-1.

Stabler, R. (2013). "What We've Got Here Is Failure to Communicate: The Plain Writing Act of 2010." *J. Legis*. 40: 280.

Thomas, A., M. Kowar, S. Sharma and H. Sharma (2011). "Extracting Noun Phrases in Subject and Object Roles for Exploring Text Semantics." *International Journal on Computer Science and Engineering (IJCSE)* vol-3.

Tobback, E., H. Naudts, W. Daelemans, E. J. de Fortuny and D. Martens (2016). "Belgian economic policy uncertainty index: improvement through text mining." *International Journal of Forecasting*.

Wallach, H. M. (2006). “Topic modeling: beyond bag-of-words.” *Proceedings of the 23rd international conference on Machine learning*, pp. 977-984, ACM.

Wallach, H. M., D. M. Mimno and A. McCallum (2009). “Rethinking LDA: Why priors matter.” *Advances in neural information processing systems*, pp. 1973-1981.

Wallach, H. M., I. Murray, R. Salakhutdinov and D. Mimno (2009). “Evaluation methods for topic models.” *Proceedings of the 26th annual international conference on machine learning*, pp. 1105-1112. ACM.

Wilson, A. T. and P. A. Chew (2010). “Term weighting schemes for latent dirichlet allocation.” *human language technologies: The 2010 annual conference of the North American Chapter of the Association for Computational Linguistics, Association for Computational Linguistics*, pp. 465-473. Association for Computational Linguistics.

Zhao, W., J. J. Chen, R. Perkins, Z. Liu, W. Ge, Y. Ding and W. Zou (2015). "A heuristic approach to determine an appropriate number of topics in topic modeling." *BMC bioinformatics* 16(13): S8.

Zhu, X., D. Blei and J. Lafferty (2006). “TagLDA: bringing document structure knowledge into topic models”, Technical Report TR-1553, University of Wisconsin.

R software citations

RStudio
RStudio Team (2015). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA URL http://www.rstudio.com/.

R
R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

DPLYR
Hadley Wickham and Romain Francois (2016). dplyr: A Grammar of Data Manipulation. R package version 0.5.0. https://CRAN.R-project.org/package=dplyr

GGPLOT2
H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2009.

NLP
Kurt Hornik (2016). NLP: Natural Language Processing Infrastructure. R package version 0.1-9. https://CRAN.R-project.org/package=NLP

PLYR
Hadley Wickham (2011). The Split-Apply-Combine Strategy for Data Analysis. Journal of Statistical Software, 40(1), 1-29. URL http://www.jstatsoft.org/v40/i01/.

PDFTOOLS
Jeroen Ooms (2016). pdftools: Text Extraction and Rendering of PDF Documents. R package version 1.0. https://CRAN.R-project.org/package=pdftools

QDAP
Rinker, T. W. (2013). qdap: Quantitative Discourse Analysis Package. 2.2.5. University at Buffalo. Buffalo, New York. http://github.com/trinker/qdap

QUANTEDA
Benoit, Kenneth and Paul Nulty. (2017). "quanteda: Quantitative Analysis of Textual Data". R Package version: 0.9.9-3. URL https://github.com/kbenoit/quanteda

RWEKA
Hornik K, Buchta C and Zeileis A (2009). “Open-Source Machine Learning: R Meets Weka.” _Computational Statistics_, *24*(2), pp. 225-232. doi: 10.1007/s00180-008-0119-7 (URL: http://doi.org/10.1007/s00180-008-0119-7).

SLAM
Kurt Hornik, David Meyer and Christian Buchta (2016). slam: Sparse Lightweight Arrays and Matrices. R package version 0.1-37. https://CRAN.R-project.org/package=slam

SNOWBALLC
Milan Bouchet-Valat (2014). SnowballC: Snowball stemmers based on the C libstemmer UTF-8 library. R package version 0.5.1. https://CRAN.R-project.org/package=SnowballC

STRINGR
Hadley Wickham (2016). stringr: Simple, Consistent Wrappers for Common String Operations. R package version 1.1.0. https://CRAN.R-project.org/package=stringr

TIDYR
Hadley Wickham (2016). tidyr: Easily Tidy Data with `spread()` and `gather()` Functions. R package version 0.6.0. https://CRAN.R-project.org/package=tidyr

TIDYTEXT
Silge J and Robinson D (2016). "tidytext: Text Mining and Analysis Using Tidy Data Principles in R." _JOSS_, *1*(3). doi: 10.21105/joss.00037 (URL: http://doi.org/10.21105/joss.00037), <URL: http://dx.doi.org/10.21105/joss.00037>.

TOPICMODELS
Grün B and Hornik K (2011). "topicmodels: An R Package for Fitting Topic Models." _Journal of Statistical Software_, *40*(13), pp. 1-30. doi: 10.18637/jss.v040.i13 (URL: http://doi.org/10.18637/jss.v040.i13).

TM
Ingo Feinerer, Kurt Hornik, and David Meyer (2008). Text Mining Infrastructure in R. Journal of Statistical Software 25(5): 1-54. URL: http://www.jstatsoft.org/v25/i05/.

VIRIDIS
Simon Garnier (2016). viridis: Default Color Maps from 'matplotlib'. R package version 0.3.4. https://CRAN.R-project.org/package=viridis

WORDCLOUD
Ian Fellows (2014). wordcloud: Word Clouds. R package version 2.5. https://CRAN.R-project.org/package=wordcloud

Appendices

Appendix 1

Note on formality of the SRLs

In this appendix, I share a review of a formality measure available from the R QDAP package. First, QDAP is used to assign the Penn Treebank Project part of speech (POS) tags[1] to the extracted letter body text. For example, here is a tagged sentence from SR1619:

> the/DT agencies/NNS plan/NN to/TO issue/VB a/DT series/NN of/IN faqs/NNS between/IN now/RB and/CC the/DT implementation/NN date/NN of/IN the/DT standard/JJ to/TO address/VB questions/NNS on/IN the/DT implementation/NN of/IN cecl/NN

The POS tagger provides useful information, although like any tagger for parts of speech, is not perfect. We see that though "plan" is used in the sentence above as a verb, it is tagged as a noun. Though "standard" is used as a noun, it is tagged as an adjective. However, "cecl", has been appropriately tagged as a noun. "cecl" is an acronym for "Current Expected Credit Loss" and in this sentence it is understood to be "current expected credit loss methodology" or "current expected credit loss model".

Using the POS tagged corpus, the QDAP package allows us to easily compute a Formality Score (F-score) based upon (Heylighen and Dewaele 2002). In referencing an earlier paper by Heylighen, (Heylighen and Dewaele 2002) explores the role of context in communication and tell us "In order to minimize ambiguity and maximize the objectivity and universality of its statements, science tries to express its result as much as possible through formal languages (Heylighen, 1999)". I would argue that SRLs fall into the same category as "science" where objectivity is favored over ambiguity and only limited background information should be needed to understand.

The F-score formula as reported in QDAP documentation is

$$F = 50\left(\frac{n_f - n_c}{N} + 1\right)$$

$$where \quad f = \{noun, adjective, preposition, article\}$$

$$c = \{pronoun, verb, adverb, interjection\}$$

$$N = \sum(f + c + conjunctions)$$

where *f* and *c* categories use the Penn tree bank POS tags listed in Table A1.

To establish a benchmark for the F-score, we may refer to (Heylighen and Dewaele, 2002 p. 316) which reports scores for "Information Writing" and "Prepared Speeches" in the English language, which were 61 and 50, respectively. For the SRL corpus with SRL bodies only, QDAP output reports 103,339 words have a formality score of 77.81. For the entire corpus, QDAP output reports 183,292 words with a formality score of 78.5 (implying attachments tend towards greater formality). QDAP author Tyler Rinker in (Rinker 2013) reminds us that (Heylighen and Dewaele 2002) establish a sample size of "a few hundred words for the measure to be minimally reliable[5]" so we review the formality score of each SRL with caution.

Noun	Adjective	Preposition	Articles	Pronoun	Verb	Adverb	Interjection
NN	CD	IN	“a”	PRP	MD	RB	UH
NNS	JJ	RP	“the”	PRP$	VB	RBR	
NNP	JJR	TO		WDT	VBD	RBS	
NNPS	JJS			WP	VBG	WRB	
POS	PD*			WP$	VBN		
				EX	VBP		
*underlying code derives PD = DT - articles					VBZ		

Table A1. POS Groupings Used in QDAP Formality Score numerator. Source: thesis writer's interpretation of code that underlies QDAP function "formality".

First, I present two pairs of triplots in Figure A1. The first of each pair is the division of the corpus according part of speech, the second of the pair is according to the formality component. For the second triplot, I calculated the number of formal words and the number of contextual words according to the QDAP tag specification for those categories as used in the formality score, with an exception in how I handled articles which was made for simplicity in coding. The first pair represents just the SRL bodies, the second represents everything in the PDF. We see just slight differences in the pairs, and that overall the plots indicate high noun usage and formality.

We will also look at plots of the QDAP formality output. In Figure A2, we compare information about the components of formality for the SRL bodies versus the full PDF contents. In Figures A3 and A4, we compare the same plots but with detail for each of the most recent SRLs. The plots are useful in identifying the direction of the text analytics, e.g., consider that one SRL, SR1614, is a bit of a outlier for pronouns, due to information technology, “IT”.

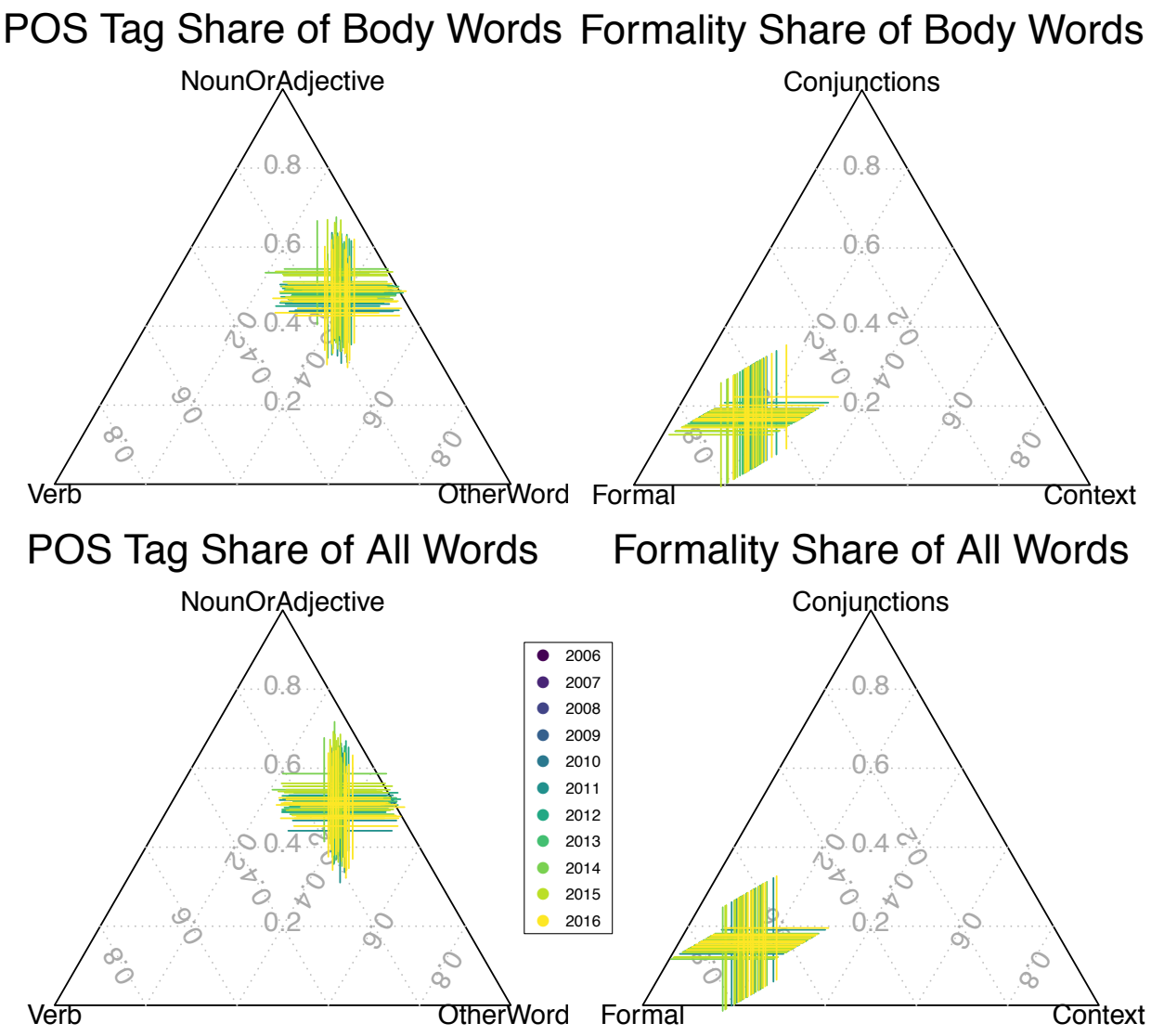


Figure A1. Triplots of the SRL BoW composition in POS and formality componenets.

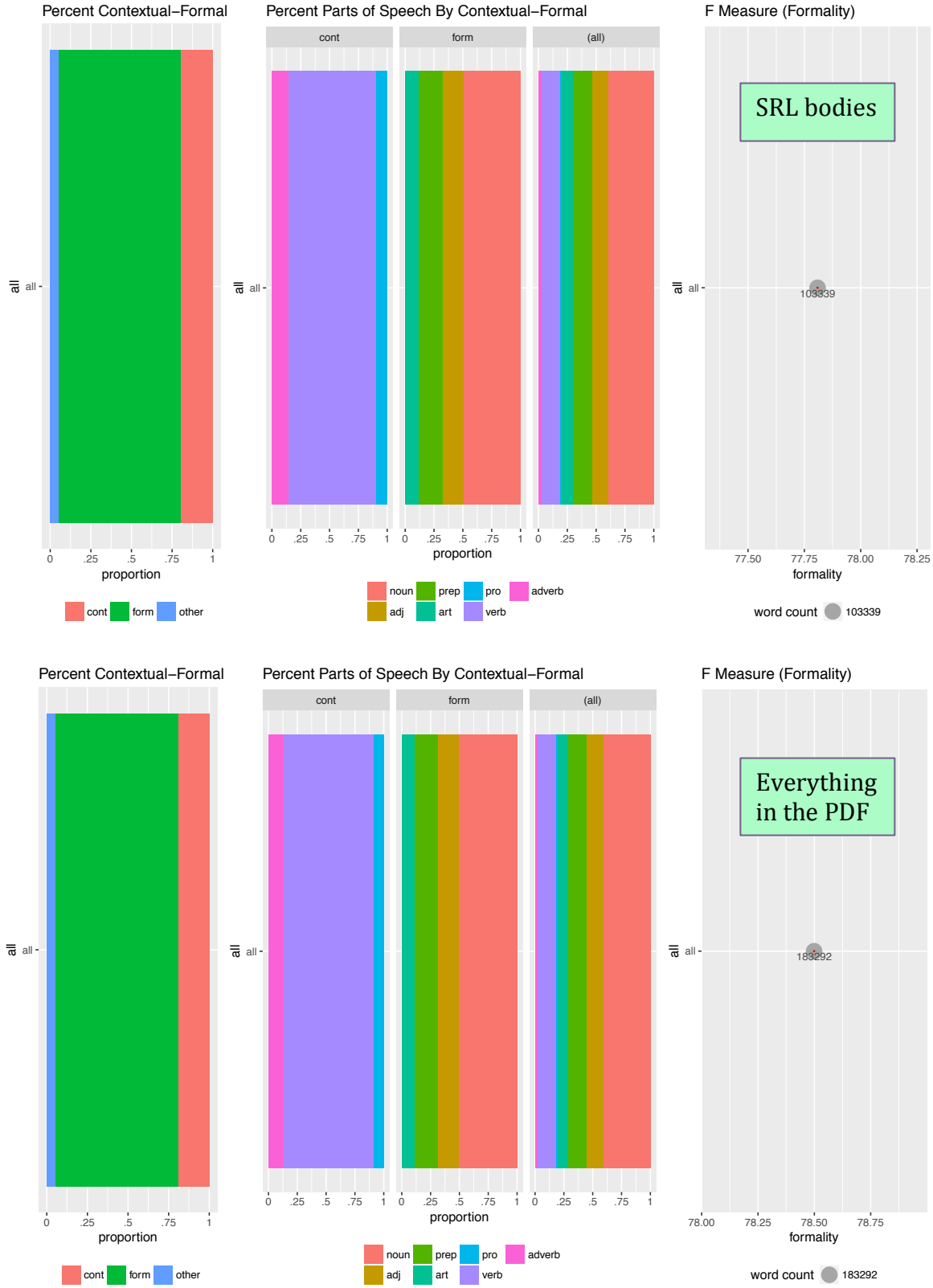


Figure A2. QDAP formality plots for two versions of the BoW.

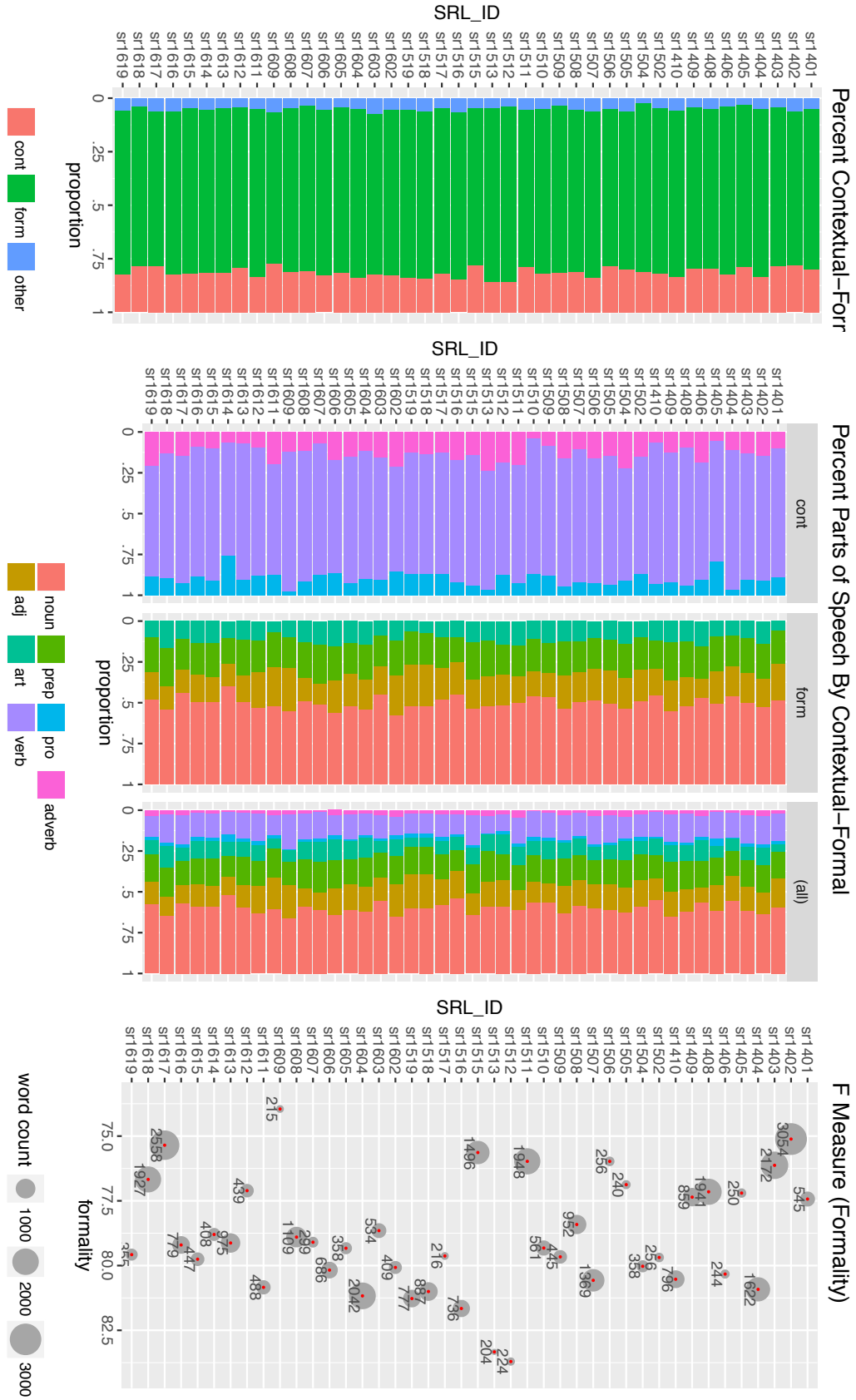


Figure A3. QDAP formality plot for selected SRLs, using the SRL body only.

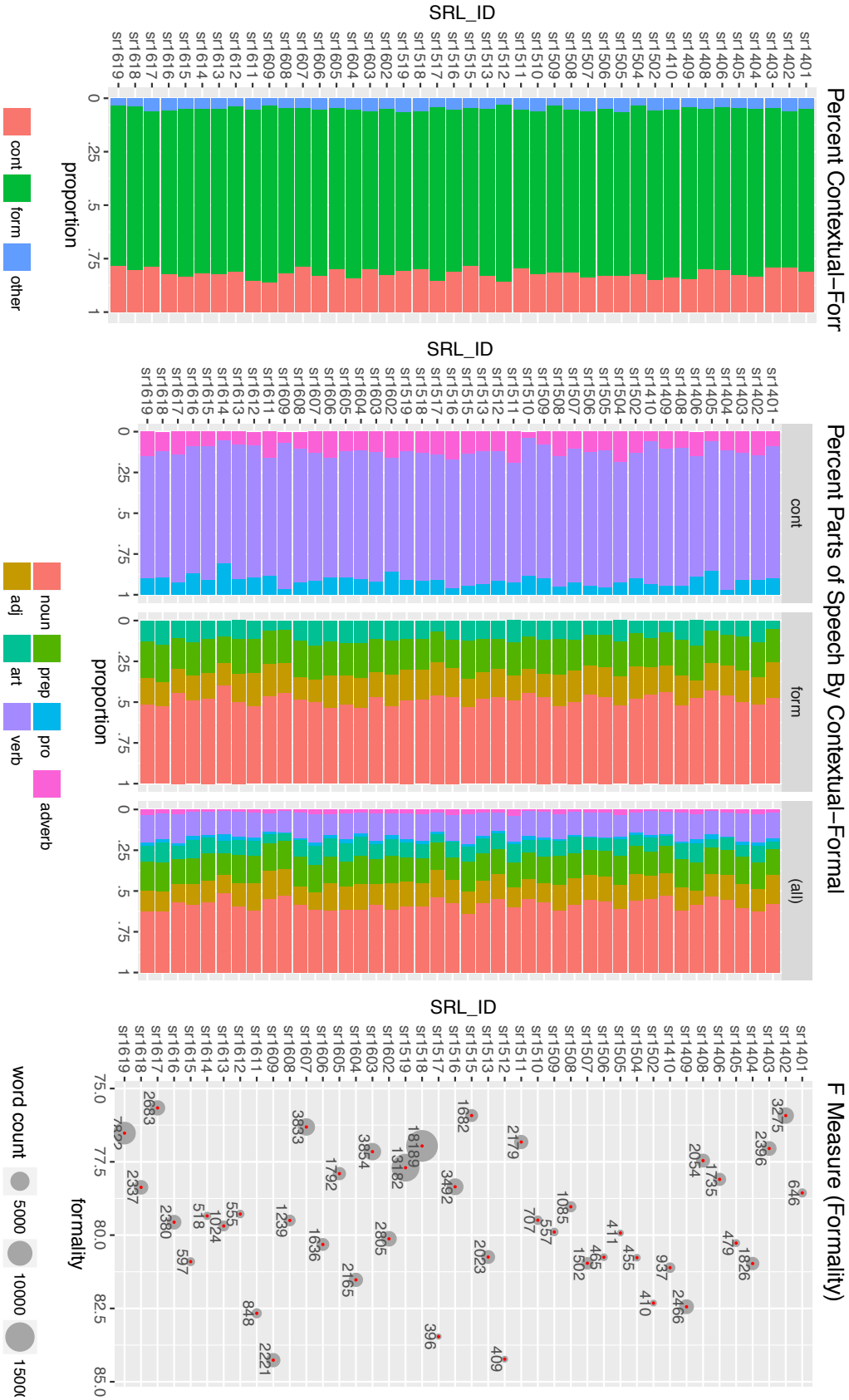


Figure A4. QDAP formality plot for selected SRLs, using everything in the SRL PDFs.

Appendix 2

Bigrams, trigrams, ngrams and noun phrases

Section 1 of Appendix 2 Overview of work on noun phrases

In the first runs of LDA models, I used a BoW that was intended to include noun phrases. At the beginning, I attempted to derive noun phrases via the use of a tokenizer, but my Objective Measure results fell significantly and meaning in LDA's highest probability terms for a topic lost interpretability. Returning meaningful terms was a secondary goal to my primary goal of determining whether LDA would provide meaningful themes in topics. Though I hoped noun phrases such as "Federal Reserve" would be pushed out as a too common phrase, I wanted to retain noun phrases such as "credit risk", "reputational risk", or "anti-money laundering". In (O'Connor 2016) discuss of the work of (Handler et al. 2016), he discusses the stand-alone nature of a noun phrase and defined BaseNP = (Adj | Noun) * Noun, where NP is an abbreviation of "noun phrase". He also recognizes two other versions of noun phrases, but my work stopped with the first one. (Handler et al. 2016) and this presentation discuss the R package PHRASEMACHINE which I have not yet tried. Here, I describe my noun phrase work as it is informational but incomplete.

For this work, I used the results of the R package QDAP POS function. I began by exploring combinations of just one adjective or noun plus a noun. To extract this information, I used an idea presented in a stack overflow post, specifically member 42-'s response on January 5, 2011 to a question "Extracting noun+noun or (adj|noun)+noun from Text". Member 42- suggested a simple solution of using R's grepl function to test for noun or adjective tags, and when found, using R's base functions to set a new variable, a "switch" if you will, to "true" or "false". He followed setting the switch by testing a relevant record's switch along with the first previous record's switch simultaneously.

Member 42-'s suggestion was the core of my noun phrase creation, and here I describe how I expanded on it. To create a detectable end of sentence marker that would persist after punctuation was removed, I replaced each period with ". meowed." (In throes of the thesis, this noun phrase formation had me up in the middle of the night coding while the TV was on. In searching for a word that is always a verb and not likely to ever show up in an SRL letter, I dismissed "hoarded", "renovate", "survived", "baked", and other words that come to mind when late-night television is in the background. After all, cash may be "hoarded", the closing institution in a merger "survived", etc.) Sometimes one or both of the members of the noun-adjective pairs were also a member in a different pairing, and in some cases strings of pairs occurred. After creating indicators of starts and stops of one or any number of pairs, I created ngrams by iterating over the BoW and combining pair members with hyphens. Last, I dropped all terms that were neither nouns nor adjectives. I describe the code further in the following section but first, here, I review selected output.

Examples of ngrams that indicated promise included:

term-funding
state-banking-agency
state-member-banks
strong-risk-management-practice

Examples of ngrams that indicated the string-combination process may have gone too far included:

traditional-stresstesting-program-banking-organization
standard-specificrisk-calculation-supervisor

An example of signs of token formation contradicting the principles upon which LDA is based in that I was creating sparse terms:

troubled-debt-restructurings-tdrs versus troubled-debt-restructurings
toptier-banking-entity versus toptier-bank

And finally, an example of code weaknesses and an incorrect POS tag:

supervisory-staff-commence

Any ngram that was moving beyond 7 terms was converted to some smaller ngram and single nouns. Overall, this effort had become over-engineered for the purpose of this thesis and non-R tools will perform this process more efficiently in future work. But it remains an educational case and I will explore it further later as the noun phrases, even as they were built here, may facilitate named entity recognition.

Section 2 of Appendix 2 A description of R Code

I will use this example sentence from an SRL throughout this appendix.

> Because the specific special measures imposed regarding 311 entities can vary, covered financial institutions should refer to FinCEN's rulemaking or order pertaining to each 311 entity for guidance regarding the nature, applicability, and scope of the imposed special measures. meowed.

The POS function I used included other text preparation steps, such as punctuation removal.

> because/IN the/DT specific/JJ special/JJ measures/NNS imposed/VBN regarding/VBG entities/NNS can/MD vary/VB covered/VBN financial/JJ institutions/NNS should/MD refer/VB to/TO fincens/NNS rulemaking/VBG or/CC order/NN pertaining/NN to/TO each/DT entity/NN for/IN guidance/NN regarding/VBG the/DT nature/NN applicability/NN and/CC scope/NN of/IN the/DT imposed/VBN special/JJ measures/NNS meowed/VBN

In next steps, I create indicators. First, an indicator is TRUE when the tag includes "JJ" for adjective or "NN" for noun. Second, an additional indicator is set to TRUE when a term with a TRUE first indicator is preceded by a term that also had a TRUE first indicator. I then query these tags to find the position of a term in a string of sort-of noun phrases. For example, if the first indicator is TRUE, and the preceding and following terms are FALSE, it is assigned "S" for a single noun or adjective. Another example is the handling of strings. When both indicators were true, the term was assigned an "L" to indicate a potentially last term in a noun or adjective string. When a term was TRUE for adjective or noun but FALSE for last member of a pair, and it was followed by an "L" item, it was assigned a "1". The program then traversed over the strings of terms labeled "L" to identify whether it was a true "L" or a second term in a three term string, and so on. Our sentence with the indicators appears next.

Term	Tag	NN, JJ?	> 1 ?	Plan
because	IN	FALSE	FALSE	out
the	DT	FALSE	FALSE	out
specific	JJ	TRUE	FALSE	1
special	JJ	TRUE	TRUE	2
measures	NNS	TRUE	TRUE	L
imposed	VBN	FALSE	FALSE	out
regarding	VBG	FALSE	FALSE	out
entities	NNS	TRUE	FALSE	S
can	MD	FALSE	FALSE	out
vary	VB	FALSE	FALSE	out
covered	VBN	FALSE	FALSE	out
financial	JJ	TRUE	FALSE	1

institutions	NNS	TRUE	TRUE	L
should	MD	FALSE	FALSE	out
refer	VB	FALSE	FALSE	out
to	TO	FALSE	FALSE	out
fincens	NNS	TRUE	FALSE	S
rulemaking	VBG	FALSE	FALSE	out
or	CC	FALSE	FALSE	out
order	NN	TRUE	FALSE	1
pertaining	NN	TRUE	TRUE	L
to	TO	FALSE	FALSE	out
each	DT	FALSE	FALSE	out
entity	NN	TRUE	FALSE	S
for	IN	FALSE	FALSE	out
guidance	NN	TRUE	FALSE	S
regarding	VBG	FALSE	FALSE	out
the	DT	FALSE	FALSE	out
nature	NN	TRUE	FALSE	1
applicability	NN	TRUE	TRUE	L
and	CC	FALSE	FALSE	out
scope	NN	TRUE	FALSE	S
of	IN	FALSE	FALSE	out
the	DT	FALSE	FALSE	out
imposed	VBN	FALSE	FALSE	out
special	JJ	TRUE	FALSE	1
measures	NNS	TRUE	TRUE	L
meowed	VBN	FALSE	FALSE	out

Table A2. Tag indicators.

Only NN and JJ tagged items are retained for the next step. A better process would have addressed coordinating conjunctions, determiners, and articles.

Term	Tag	NN, JJ?	> 1 ?	Plan
specific	JJ	TRUE	FALSE	1
special	JJ	TRUE	TRUE	2
measures	NNS	TRUE	TRUE	L
entities	NNS	TRUE	FALSE	S
financial	JJ	TRUE	FALSE	1
institutions	NNS	TRUE	TRUE	L
fincens	NNS	TRUE	FALSE	S
order	NN	TRUE	FALSE	1
pertaining	NN	TRUE	TRUE	L
entity	NN	TRUE	FALSE	S

guidance	NN	TRUE	FALSE	S
nature	NN	TRUE	FALSE	1
applicability	NN	TRUE	TRUE	L
scope	NN	TRUE	FALSE	S
special	JJ	TRUE	FALSE	1
measures	NNS	TRUE	TRUE	L

Table A3. Nouns and adjectives.

Code then traverses over the NN or JJ terms to decide if they are output as a single term ("S") or should be concatenated. I add a hyphen between terms when they are concatenated. From this effort, the BoW is left with the following list. I chose this sentence as it included results that will act deceptively in the LDA process. The LDA process will not recognize the linkage of specific-special-measures and special-measures. The LDA process is also unlikely to encounter another "nature-applicability".

ngrams
specific-special-measures
entities
financial-institutions
fincens
order-pertaining
entity
guidance
nature-applicability
scope
special-measures

Table A4. Output.

I believe that my technique may prove useful in a more thorough application but in its current form we see how easily the ambiguity in created ngrams may work against the machinations of LDA.

Appendix 3

A description of the R code which creates the success measure

Output from the LDA model includes a file of the topic assignments of each SRL. Among other information, the file includes the topic assignment – "VEMtopics" and the SRL_ID. A subset is displayed in Table A5.

VEMtopics	SRL_ID
1	sr1108
1	sr1207
1	sr1403
2	sr1001
2	sr1006
2	sr1010
2	sr1106
2	sr1107
2	sr1324
2	sr1410
2	sr1603

Table A5. VEM Topics

A "long" file is manually created based upon information from the FRBOG website. The FRBOG website categorizes SRLs into broad topics, such as "Accounting". SRLs may be categorized into more than one topic. A subset of this file is displayed in Table A6.

SRL_ID	BOGTopicL
sr0601	BSA/OFAC
sr0604	InternalControls
sr0605	Opsrisk
sr0605	BusContinuity
sr0608	ExamSupGuidance
sr0608	Securities
sr0608	Asset/Wealth
sr0613	InfoSec
sr0614	ExaminerCredsCOI
sr0614	ConfidentialInfo
sr0614	InfoSec
sr0615	ALLL
sr0615	RealEstate
sr0617	ALLL
sr0701	RealEstate
sr0701	ALLL

Table A6. FRBOG topics

SRL_ID	VEMtopics	BOGTopicL
sr0601	3	BSA/OFAC
sr0604	7	InternalControls
sr0605	14	OpsRisk
sr0605	14	BusContinuity
sr0608	19	ExamSupGuidance
sr0608	19	Securities
sr0608	19	Asset/Wealth
sr0613	4	InfoSec
sr0614	18	ExaminerCredsCOI
sr0614	18	ConfidentialInfo
sr0614	18	InfoSec
sr0615	7	ALLL
sr0615	7	RealEstate
sr0617	7	ALLL
sr0701	8	RealEstate
sr0701	8	ALLL
sr0705	8	Accounting
sr0705	8	Asset/Wealth
sr0705	8	CreditRisk
sr0705	8	LegalRepRisk

Table A7. LDA topics with FRBOG topics

Merging on SRL_ID, a new "long" dataset is created in which we can now see all of the "BOGTopicL" that have been grouped into the LDA topic. A subset of this file is displayed in Table A7. We see in the yellow and peach highlighted rows that SRLs may be associated with 1 or more FRBOG topics. The blue highlighted rows display some of the SRLs assigned to topic "7" by LDA and those FRBOG topics associated with the SRLs.

Our goal here is to see if the SRLs have been grouped into LDA topics similarly to the grouping into FRBOG topics by an expert human. We do this without regard to the FRBOG topic label, and instead look for the most frequent FRBOG topic label among the SRLs assigned to LDA topic, and then look again for a most frequent FRBOG topic label among the remaining SRLs assigned to the LDA topic. We'll follow the evolution of the remaining process using examples from the LDA topic assignment "1".

In this example run of LDA, topic 1 included SRLs 1108, 1207, and 1403. In Table A8, we see that multiple FRBOG topics were associated with 2/3 SRLs and that each of the SRLs in LDA topic 1 is associated with FRBOG topic "CapAdequacy".

SRL_ID	VEMtopics	BOGTopicL
sr1108	1	CapAdequacy
sr1207	1	CapAdequacy
sr1207	1	ExamSupGuidance
sr1207	1	LiquidityRisk
sr1403	1	LiquidityRisk
sr1403	1	CapAdequacy

Table A8. FRBOG topics within LDA topic

VEMtopics	max	maxBOGt
1	3	CapAdequacy
2	4	LiquidityRisk
3	3	CommunityBkg
4	5	Accounting
5	2	Accounting
6	4	ExamSupGuidance
7	4	RealEstate
8	4	CreditRisk
9	3	CapAdequacy
10	3	Apps
11	4	SLHC
12	1	Asset/Wealth
13	1	Accounting
14	6	RealEstate
15	3	Apps
16	1	
17	2	CreditRisk
18	3	BSA/OFAC
19	2	ExamSupGuidance
20	1	Apps

Table A9. Max FRBOG topic, round 1

VEMtopics	BOGTopicL	freq
1	ExamSupGuidance	1
1	LiquidityRisk	2

Table A10. LDA topic 2, round 2.

Using "ddply" from the R package plyr, I find the maximum frequency among FRBOG topics of SRLs assigned to LDA topics. We see "CapAdequacy" for LDA topic 1. This is displayed in Table A9.

These "maximums" are then merged back to the larger dataset and the associated SRLs are extracted from the dataset. The remaining rows are then submitted for a repeat of the process. In Table A10, we see the new frequencies of the FRBOG topics found for the remaining SRLs of each LDA topic. In Table A11, we see the most frequent FRBOG topic for each LDA topic, as we did in Figure 5. We see the LDA topic 1 highlighted in yellow. An aside: Two rows are highlighted in blue to call out the lack of a second most frequent FRBOG topic – in LDA topic 10, all 3 SRLs had a FRBOG topic of "Apps" and only "Apps". In Figure 8, we see that the code identified the common occurrence of "CapAdequacy" among our LDA topic 1 SRLs.

For understanding, it is critical to note at this point that the concept of "CapAdequacy" should not be used to draw conclusions. The code extracted SRLs based upon the most frequent first-found common FRBOG topic. For example, in LDA topic 13, all FRBOG topics associated with the SRLs within the LDA topic 13 were unique. Most frequent topic was the first-found topic with a frequency of 1, next most frequent topic was the next occurring topic with a frequency of 1 among the remaining SRLs in topic 13. This introduces arbitrariness in the identification of the top term. Our goal was to identify if LDA topics grouped SRLs in the manner of an expert human, not whether the descriptive word or phrase for topic labeling was appropriate.

VEMtopics	max	maxBOGt
1	2	LiquidityRisk
2	2	BSA/OFAC
3	3	ExamSupGuidance
4	4	ALLL
5	2	AssetQual
6	3	LegalRepRisk
7	3	Apps
8	3	Accounting
9	2	ExamSupGuidance
11	3	ExamSupGuidance
12	1	CorpCompliance
13	1	CapAdequacy
14	4	CreditRisk
15	3	ExamSupGuidance
16	1	ExamSupGuidance
17	1	CommunityBkg
18	2	ConfidentialInfo
19	1	Asset/Wealth
20	1	CapAdequacy

Table A11. Max FRBOG topic, round 1

At this point in this process, a dataset includes a variable "toptop" which is 1 if an SRL was matched to the most frequent FRBOG topic within the LDA topic, 2 if an SRL was matched to the next most frequent FRBOG topic within the LDA topic, and NA if it didn't match to either. As R reads NA as less than 1, "SO" was created to hold 1 or 2 if there had been a match, or 3 if not. Our sample included 125 SRLs and transforming to a long dataset with one row per SRL/FRBOG topic resulted in 260 rows for this example run so we must transfom to 1 row per SRL. An SRL within an LDA topic which matched to either of the 2 most frequent FRBOG topics is a match. Of course this is more meaningful in an LDA topic with 15 SRLs than an LDA topic with 2 SRLS, but nevertheless it provides a comparative measures for various runs of LDA modeling of SRLs. Matches to FRBOG topics within a LDA topic tell us of how LDA may have duplicated some of the effort of an expert human. In Table A13, we see match status within the first 2 LDA topics.

VEMtopics	BOGTopicL	SRL_ID
1	CapAdequacy	sr1403
1	CapAdequacy	sr1108
1	CapAdequacy	sr1207

Table A12. Completed LDA topic 1.

VEMtopics	SRL_ID	BOG_TOPIC	SO	countkey
1	sr1403	CapAdequacy & LiquidityRisk	1	MatchTo2
1	sr1108	CapAdequacy	1	MatchTo2
1	sr1207	CapAdequacy & ExamSupGuidance & LiquidityRisk	1	MatchTo2
2	sr1107	CapAdequacy & LiquidityRisk & InternalControls & MarketRisk	1	MatchTo2
2	sr1006	LiquidityRisk	1	MatchTo2
2	sr1603	ExamSupGuidance & FBO & LiquidityRisk	1	MatchTo2
2	sr1001	LiquidityRisk & MarketRisk	1	MatchTo2
2	sr1410	BSA/OFAC	2	MatchTo2
2	sr1106	BSA/OFAC	2	MatchTo2
2	sr1010	ExamSupGuidance	3	NoMatch
2	sr1324	LegalRepRisk & OpsRisk & IntlActivities	3	NoMatch

Table A13. Identified matches, by SRL, of FRBOG top 2 topics to LDA topic combinations.

fVEMtopics	MatchTo2	NoMatch
1	3	0
2	6	2
3	4	5
4	6	11
5	2	4
6	5	1
7	7	8
8	4	6
9	3	0
10	3	0
11	5	0
12	1	0
13	2	0
14	8	5
15	5	1
16	2	0
17	2	0
18	5	1
19	2	4
20	2	0
	77	48

Table A14. Summary matches.

Recall we have results of LDA based upon different BoWs. Table A14 is summarized to a count of matches in a single case, which was just one example.

Appendix 4

Comparison clouds

Figure A.5. The next four pages of 2016 SRLs,
in comparison cloud format, with comparison on the left and commonality on the right.

sr1619

sr1618

sr1618

sr1617

sr1617

sr1616

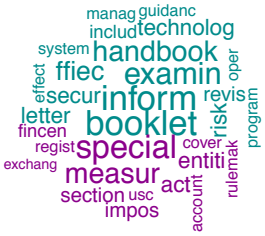
sr1616

sr1615

sr1615

sr1614

sr1614



sr1613

sr1613

sr1612

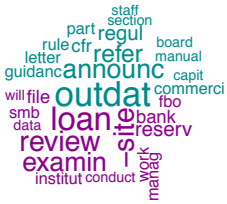
sr1612

sr1611

sr1611

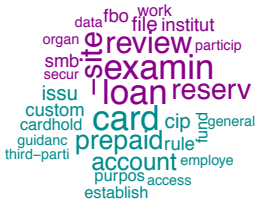
sr1609

sr1609



sr1608

sr1608

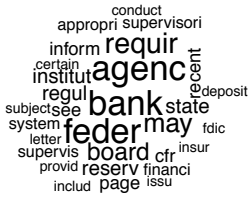


sr1607

sr1607

sr1606

sr1606



sr1605

sr1605

sr1604

sr1604

sr1603

sr1603

sr1602