Structured Insights or Preprocessing Artifacts? “Breaking Down” the Impact of Text Chunking Strategies on Topic Model Interpretability

By Becky Marcus

Computational text analysis (CTA) has become an essential tool for sociologists seeking to extract cultural meaning from texts, particularly with the increased digitization of historical text corpora. Structural Topic Modeling (STM) is a popular exploratory tool for highlighting latent themes in text that warrant further investigation, but with new tools come new challenges for reliability and validity. Most researchers will adhere to the text preprocessing methods suggested in prominent CTA literature, though with some variability in the size of the text chunk, word tokenization, and vocabulary simplifying and filtering. These decisions are often made somewhat uncritically as part of an established methodological procedure yet have the potential to yield markedly different topics and interpretations. Supplementing existing literature on preprocessing strategies for topic models, this study takes a mixed-methods data science approach to evaluate the effect that text unit size has on topic content and prevalence. It analyzes a corpus of presidential addresses from meetings of the American Economic Association from 1888-2022 and includes meeting year as a topic prevalence covariate. Contributing to a larger research effort on social science presidential addresses, this study not only uncovers preliminary patterns in institutional and academic discourse over time but also discusses how text chunking can be aligned with different empirical questions.

Structured Insights or Preprocessing Artifacts? “Breaking Down” the Impact of Text Chunking Strategies on Topic Model Interpretability

By Becky Marcus

Mount Holyoke College

Department of Mathematics, Statistics, and Data Science

30 June, 2025

Acknowledgements

The biggest thanks goes to my advisor, Ben Gebre-Medhin, for supporting me all throughout this process: cheering me on and laughing with me during the ups and providing kind and honest support during the downs. I really feel so lucky to have had you in my corner over the past few years. Other people who helped me on this emotional roller coaster include my friends, my parents, and my late-night study buddies Sophia, Amy, Tabitha, Grace, Anya, Fatma, and Méabhan (to name a few). I certainly would not have been able to get this done without the Data Science Team of the METRICS Lab (Robin, Hema, and Sanskriti) and my research assistants Alex and Gus who put in the tedious work of fixing character and formatting errors in the text. Many thanks to my co-advisees; Coco, who always brings super smart thoughts and wonderful vibes to our thesis meeting; and Katherine, our fun and full of light lab manager and a great friend of mine. I want to acknowledge the other lab seniors: Dening, who I've loved getting to know this year; and Estelle for bringing so much kind and fun energy into every interaction. Much love as well to Lela and Hanh who will be carrying forward the wonderful culture of the lab next year. I'd like to also thank Arie Shaus for being on my committee and always so excited about and supportive of my work. Finally, I would be remiss not to acknowledge Eleanor Townsley for the science fiction book chats and recommendations, as well as for always believing in me and pushing my thinking. I will certainly miss sitting in between you and Ben at lab meetings and feeling my brain expand *in real time* during our lively discussions about computational methods and epistemological culture.

Table of Contents

Figures and Tables

CHAPTER I - INTRODUCTION

Computational text analysis methods have been of great interest to sociologists who are working on extracting cultural meaning from text. This has become increasingly true with digitization of historical texts. Specially, unsupervised methods of clustering texts can reveal latent patterns in corpora and highlight key concepts and constructs that warrant further investigation. Such methods allow for greater breadth, speed, and reliability when coding than close reading or qualitative coding approaches, though with the drawbacks of decontextualization of words and loss of nuance. Many sociologists prefer Structural Topic Modeling (STM) by Roberts et al. (2013) which has the advantage of incorporating document-level metadata into estimations of topic prevalence and content. The meaning of words is also easier to deduce when in the context of other words in a given topic.

Still, like any form of computational text analysis, text preprocessing is necessary to get meaningful results: It is common practice to remove stop words (a, of, it, the, and, etc.) and combine words with similar meanings but different parts of speech using lemmatization or stemming. There is also the question of whether to only measure the frequency and co-occurrence of singular words or try to capture greater meaning using n-grams, noun phrases, or named entities. Historical texts also may present the challenge of word or character error rates from imperfect digitalization and text extraction that add noise to the data and could distort or obscure valuable insights.

Researchers therefore must make decisions about these preprocessing steps, as well as hyperparameters in the models themselves—most crucially the number of topics. However, it is possible that different preprocessing configurations will recommend models with different numbers of topics that highlight different meanings. Even if these insights were all relevant and interpretable, this would call into question the reliability of STM as a tool for identifying underlying patterns. Since computational methods run the risk of imitating objectivity even as they only abstractly represent complicated meanings, great care must be taken to understand the values and limitations of such methods.

That is why this study will take a data science approach to model optimization, tuning, and visualization to investigate one of the earliest decisions in the topic modeling process: the unit of text that will be passed to the model. While this is not a concern for studies of smaller text objects, there is little guidance on how to break up larger texts such as novels and speeches. To determine the effect that text chunking has on STM results, models will be created from text split to different units of analysis, then evaluated for interpretability and potential empirical applications.

The text being modeled is Presidential Addresses given at annual meetings of the American Economic Association (AEA). Previously unstudied, this corpus is composed of 128 speeches over 132 years and ranging from approximately 730 to 19,000 words in length. The question of how to chunk the text is especially pertinent given the variance in address length, so the results of this investigation can inform future studies of similar corpora. This corpus is also ideal because of the consistency of

the speeches' context and the relevant temporal metadata to be incorporated into the model.

STM will therefore be applied to the AEA Presidential Address corpus chunked at the whole document, page, and paragraph level with meeting year as a topic prevalence covariate. The main questions of this analysis are as follows:

1) What is the relationship between the text unit of analysis and the optimal number of topics (K), as determined by semantic coherence and exclusivity?
2) How consistent is topic content at different values of K and between models of differently chunked addresses?
3) What can be learned about the AEA by measuring topic prevalence across the years? To what extent do the distributions vary with chunking strategy?

While the differently chunked models provided a similar range of potential Ks, *it was ultimately found that models with a higher value of K and more granular topics are the most cohesive with smaller text chunks*. This paper details the considerations and decisions made when topic modeling the AEA addresses that led to that conclusion. In doing so, it not only provides a starting point for further exploration of this corpus but also contributes to understanding of best practices for CTA in the social sciences by supporting researchers in making choices that align the modeling process with their research question.

CHAPTER II - LITERATURE REVIEW

This section describes the existing literature relevant to this investigation of how text chunking strategies affect STM results. It first outlines the domain of computational social sciences and some of the field's methodological considerations and controversies. Next is an overview of various topic modeling strategies and their applications, strengths, and weaknesses. Finally, a discussion of the dominant preprocessing methods and existing literature on the effect of preprocessing and text chunking on topic models.

2.1 Computational Text Analysis in the Social Sciences

Technological advancements in the twenty-first century have made it significantly easier to process large volumes of text with great speed and relatively low computational power. There are many applications across domains: business reviews can be summarized (Pal et al. 2023), potential spam emails can be flagged (Tusher et al. 2024), and new Large Language Models (LLMs) can write papers for students and then detect the plagiarism in them (Pudasaini et al. 2024). In academia, the field of Digital Humanities has evolved to incorporate these methods for information retrieval (Manning, Raghavan, and Schütze 2008), authorship attribution (Savoy 2020), and new methods of text analysis (Carter 2013).

In the social sciences, computational methods can supplement qualitative coding practices with more speed and reliability, though it is harder to ensure validation. Grimmer and Steward (2013) explain the challenges of qualitative text analysis for political science due to large volumes of texts and the expense of hiring coders. Providing a survey of text analysis methods that are relevant to the political science domain, the authors stress the challenges of validation and inherent incompleteness of text as data. There have also been advancements in CTA methods for psychology (Tausczik and Pennebaker 2010; Kennedy et al. 2022), sociology (Nelson 2020; DiMaggio, Nag, and Blei 2013; Gao, Wang and Liu 2024), and widely applicable procedures and models for Natural Language Processing (NLP) in general (Blei, Ng, and Jordan 2003; Sarkar 2016). Additionally, with new neural network architecture for processing texts, some claim that sequentially trained models by experts may eventually be able to replace manual coding (Do, Ollion, and Shen 2024).

However, there are still objections to using computer assisted methods in the study of culture from texts, something that is of great interest to sociologists (DiMaggio et al. 2013; Biernacki 2012; Lee and Martin 2015). In *Reinventing Evidence in Social Inquiry*, Biernacki critiques text coding practices in sociology—computational or not—arguing that coding practices lack “empirical substance” as they decontextualize the initial meaning of words with abstract mathematical transformations, only to recontextualize the results in a “ritual arena” such as graphs and tables (2012:11). Lee and Martin respond to Biernacki’s critique of coding, conceding that qualitative coding practices are intended to provide objectivity, but, as Biernacki’s failed replication of qualitative coding studies demonstrates, “no science can build objectivity where

replication is so implausible" (2015:6). Lee and Martin thus argue for a greater formalization of computational methods, explaining that quantitative text analysis still depends on the researcher's interpretation but with greater transparency and reliability than just reading. In *Patterns: The Theory of Digital Society*, Nassehi takes the stance that qualitative and quantitative research do not differ as much as people think, as both are "about supraindividual patterns and about the methodically controllable recombination of meaning" and therefore cannot be simplified to questions of "fact" versus "interpretation" (2024:36).

Responding to this ongoing debate, lack of guidelines, and the "potential [for] haphazard and undisciplined use of text analysis methods," Nelson proposes Computational Grounded Theory as a new methodological framework for sociology (2020:4). With a moderate approach similar to Grimmer and Steward (2013) and contrasting the extreme positions of Lee and Martin (2015) and Biernacki (2012), Nelson advocates for "preserving the superior abilities to interpret text holistically provided by humans" while still "incorporating the formal rigor, reliability, and reproducibility of computer assisted methods" (2020:8). It is expected that sociologists using STM will be working in a similarly balanced framework, as topic modeling methods rely heavily on qualitative interpretation.

2.2 Topic Modeling

Topic modeling is generally considered useful as an inductive and explanatory tool for text corpora. The idea is to model a structure of latent topics based on word

co-occurrence and frequencies over a specific probabilistic distribution. Latent Dirichlet Allocation (LDA) by Blei et al. (2003) is the most popular topic modeling algorithm. Its advantage over its predecessor, Probabilistic Latent Semantic Indexing (pLSI; Hofmann 1999) is its mixed-membership structure; where topics are defined by mixtures of words and documents are composed of mixtures of topics. By looking at the top words in each topic, researchers can deduce general themes of the corpus before applying their theoretical priors. Additionally, LDA accounts for the multiple meanings of words by distributing them across different topics and allowing context—that is, the other words in the topic—to determine their meaning. As such, topic models support sociological axioms of the relationality of meaning, and still require subject matter expertise to evaluate the relevance of topics for a given research question (DiMaggio et al. 2013).

Another probabilistic generative model, the Structural Topic Model (STM) further "accommodates corpus structure through document-level covariates affecting topic prevalence and/or topical content" (Roberts et al. 2013:1). In other words, specified metadata (covariates) and their interactions with topics are incorporated into the modeled distribution of each word or document. STM is widely used in the social sciences on survey response data (Roberts et al. 2014), academic literature (Lindstedt 2019; Erikson, Yao, and Karell 2023), newspapers and other media (Wachen 2018; Chandelier et al. 2018), and more.

When working with historical texts, STM can be used to investigate discourse over time with temporal data as a covariate (Jo 2019). Other temporally-focused topic models include Dynamic Topic Models (Blei and Lafferty 2006) which map the evolution of topics over discrete time steps; and incremental Hierarchical Dirichlet Processes for

continuous text streams that allow users to “observe how topics evolve over time, including its strength, content, and splitting/merging relationship” (Gao et al. 2011:1). There are also several methods for creating more detailed representations of words and topics, including the Topical N-gram Model (Wang, McCallum, and Wei 2007) and the Embedded Topic Model (Dieng, Ruiz, and Blei 2020).

Alternatively, Gerlach, Peixoto, and Altmann (2018) critique the constraint of assuming probabilistically structured priors and instead propose a “network approach to topic models” that uses community detection algorithms for modularity and a mixed-membership stochastic block model to represent non-Dirichlet topic mixtures. Similarly, seeking to “reconstruct the flow of political discourse” in State of the Union Addresses across time, yet unsatisfied with the opacity of Gao et al.’s (2011) method for continuous text streams, Rule, Cointet, and Bearman (2015) create paragraph-level word co-occurrence networks in overlapping time periods. Even with these extensive variations, rarely do social scientists employ anything other than STM and LDA given their dominance in the existing literature.

2.3 Preprocessing Decision-Making

The primary resources for computational social science researchers using Structural Topic Modeling are *Text as Data* by Grimmer, Roberts, and Stewart (2022) and the documentation of the R package *stm* (Roberts, Stewart, and Tingley 2019). These contain a canonical procedure for text preprocessing, including word standardization methods (making lowercase, stemming) and the removal of stop words

and unwanted characters. Discussion of preprocessing decisions is generally minimal—as shown in studies such as Lindstedt (2019), Margherita et al. (2023), and Popa (2025)—even though, as Nour (2024) asserts, “the inclusion of this information is essential to allow readers to gauge the robustness and generalizability of study conclusions.” There also is minimal literature on the impact of these preprocessing decisions, though some studies have suggested that the default stemming and custom stop word removal steps are not necessary and potentially detrimental to topic model performance (Schofield and Mimno, 2016; Schofield, Magnusson, and Mimno 2017).

The rarely-discussed preprocessing decision being evaluated in this paper is what counts as one unit of analysis; that is, how many documents a corpus should be split into to ensure an adequate sample size without losing meaningful word mixtures. Sbalchiero and Eder (2020) investigate text chunking by breaking up a corpus of 100 novels into chunks of 500, 1000, 5000, 10,000, 20,000, and 50,000 words and comparing the recommended number of topics based on the log likelihood. They find an inverse relationship between the size of the text chunk and number of topics but explain that their method does not always produce interpretable topics. This highlights a limitation of the log-likelihood method of selecting the number of topics and demonstrates the need for further research into the subject. Sbalchiero and Eder (2020) also refer to studies that suggest 1000-word chunks as a rule of thumb (Jockers and Mimno 2013) but without explicit reasoning. Alternatively, in their topic modeling of early economics texts, Erikson et al. (2023) use chunks of 500 words with each overlapping by 250 words, yet it is unclear how the overlapping chunks impacted the results.

Finally, Guo, Menglin, and Wei (2021) propose “An Improved LDA Topic Modeling Method Based on Partition for Medium and Long Texts” (LDAP) that consists of partitioning long texts into paragraphs, fitting an LDA model, and calculating a weighted summation of topic prevalences based on paragraph length. However, since this model was only validated by using its output matrices as features in a random forest classifier, it is unclear how it would perform in sociology research.

While these studies have initiated a conversation about the impact of text chunking on topic modeling results, few systematic inquiries have been conducted that engage directly with the topic modeling results and evaluate their empirical applications. This study will therefore contribute to the existing CTA and NLP literature by providing a formal comparison of text chunking units in Structural Topic Modeling, yet with the more grounded and interpretation-focused approach of social science researchers: The first part of this comparison is quantitative and focuses on semantic coherence and exclusivity when choosing the ideal number of topics. This is followed by a more qualitative assessment of the relevance and consistency of topic content and prevalence, thus providing a detailed estimation of the impact of text chunking strategies on model insights and highlighting the greater coherence of small text chunks in models with a higher number of topics. These sections combine to improve applications of computational text analysis methods in the social sciences by advancing understanding of the impact of chunking strategies on model fit and interpretability.

CHAPTER III - INTRODUCTION TO THE DATA

3.1 Presidential Addresses at Meetings of the American Economic Association

The primary corpus of this study consists of addresses given by presidents of the American Economic Association at their annual meeting. The AEA was founded in 1885 with the intention of encouraging free academic discussion and publication in a nonpartisan fashion (Bernstein 2008). Members gather at the AEA annual meeting to present papers, attend lectures, and discuss research, with one of the featured events being the Presidential Address (American Economic Association, n.d.). According to the association's bylaws, the leadership of the AEA includes a President, President-elect, two Vice Presidents, and six Executive Committee members. Each year, the President-elect creates a Nominating Committee who selects candidates for President-elect and forms an electoral college with the Executive Committee to choose a single nominee. Members can petition to add additional candidates, and the final election is conducted by a membership-wide vote. The winner serves as President-elect for one year before becoming President.

Annual meeting reports and papers were published by the AEA then digitized by JSTOR as PDFs. It is important to note that the content of these papers differ from what was orally delivered at the meeting, recordings of which are very limited in availability.

While the papers lack the performative context of a live audience at a large meeting, they were still distributed as a presidential communication, giving them greater authority than just another research paper. To ensure that all the addresses were collected for analysis, the names and terms of the past presidents from the American Economic Association website were cross-referenced with the PDFs. Efforts were also undertaken to ensure the use of the term-year instead of the publication-year of the address.

3.2 Data Cleaning

One of the biggest challenges of a novel data science project is the data preparation portion, especially when working with text data that are meant to be read by humans, not machines: PDFs are optimal for document reading, sharing, and printing yet they do not have the digital text representation of raw text or markdown file. There are open-source packages that extract text from PDFs such as PyPDF2, yet the quality of results is greatly varied. This is because the text embedded in the PDFs are based on various Optical Character Recognition (OCR) algorithms that find the most likely characters given the pixels of the document. As a result, text from older and lower-quality documents are more prone to character-level error rates such as mixing up “e” and “c” or “I” (the capital letter “i”), “l” (the lowercase letter “L”), and “1” (the number one). To overcome these issues, Google Cloud’s Document AI API was used. This tool leverages Large Language Models to enhance traditional OCR by considering context and spelling when determining characters and words, rather than just the shape.

Not all the text was meant to be analyzed, however: Most pages had some kind of header and footer, and quite a few had footnotes as well. Though some footnotes were several paragraphs long with some important context, there was much variety in form, so they were excluded. All of these speeches were published in the primary AEA journal of the time[1], so there were large groups of similarly formatted documents, allowing for the implementation of tailored text cleaning pipelines at scale. This was primarily achieved using regular expression pattern matching for removing headers, footers, and references in a set of documents with a given formatting. With this modular approach, it was much easier to adapt generalized functions to specific patterns and add in logic for edge cases.

Hired research assistants then went through the text files in Google Docs to fix spelling errors, remove outstanding non-body text, and split the text into paragraphs the way that the author did. Because they were able to lay eyes on every page of the text and use automatic grammar and spell checking to identify blatant errors, it is highly unlikely that any page of the addresses (about 500 words) would have more than one error on average. Therefore, the upper bound on estimated word error rate is 0.2%, which is well above standards presented in published literature.

[1] Journals include: Publications of the American Economic Association (1886-1907); American Economic Association Quarterly (1908-1910); and The American Economic Review (1911-2022)

CHAPTER IV - METHODS: TEXT CHUNKING AND PREPROCESSING DECISION-MAKING FOR TOPIC MODELS

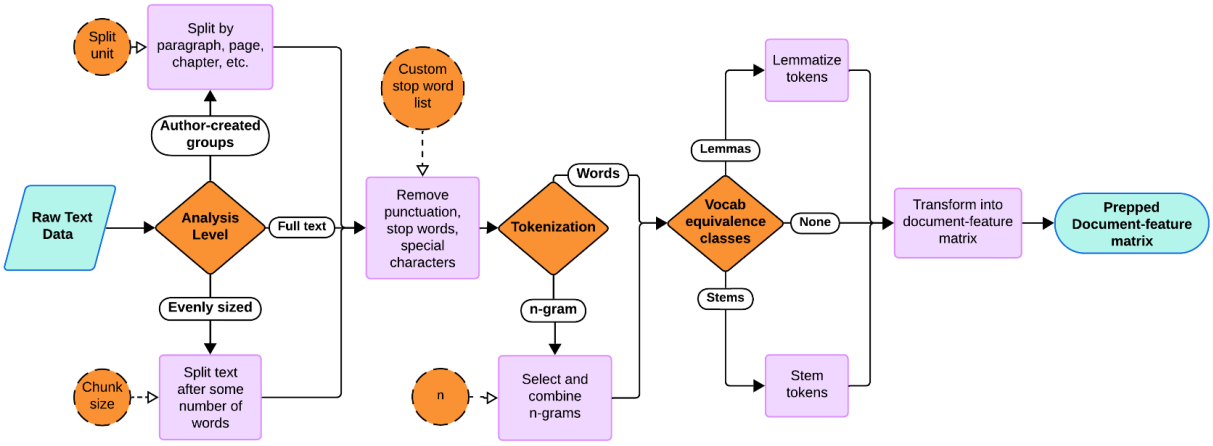
4.1 Text Preprocessing Steps and Decisions

Computational text analysis methods rely on the transformation of language into machine-readable format. Words can already be represented by computers—that is how this digital text is available in the first place to be analyzed—but it is much more challenging to digitize the *meaning* of words. That is why it is common to represent corpora in a “bag-of-words” (BoW) format, where word order is ignored and only the counts of each word in each document are recorded. This information is structured into a sparse and high-dimensional “document-feature matrix” with the potential for summarization and comparison of documents using probabilistic and algorithmic techniques (Grimmer et al. 2022). The relative “simplicity” of BoW models makes them suitable for tasks beyond text representation such as image processing and texture recognition (Qader, Ameen, and Ahmed 2019).

Still, researchers must make a wide range of potentially consequential choices when transforming raw text into document-feature matrices. Figure 4.1 illustrates the common text preprocessing steps for topic modeling and the necessary decisions.

These decisions are based on Grimmer et al.'s chapter on bag-of-word models in *Text as Data* (2022).

Figure 4.1: Bag-of-words Preprocessing Steps and Decision Points



Starting with the "raw text data" (assumed to be spell-checked and only including body text), the first branching is at *analysis level*, that is, what will be considered one unit of text in the model. For corpora of shorter texts such as journal abstracts (Anupriya and Karpagavalli 2015), essays (Alvero et al. 2021), and survey responses (Roberts et al. 2014), this step is not necessary. However, longer texts such as novels (Jockers and Mimno 2013) and journals (Lüdering and Winker 2016) are expected to have multiple meaningful themes which could be diluted if the text stays whole. The chunks may be evenly sized or split into more meaningful units such as paragraphs, sections, or chapters. After that, noise such as stop words, special characters, and punctuation are removed. Besides the default stop words such as "and," "it," and "the," researchers may define additional, corpus-specific words that they expect to be very frequent yet unhelpful.

Next, the documents are *tokenized*, usually just into words using white space, but sometimes with the addition of prevalent n-grams such as “United_states” or “labor_market.” Then, the complexity of the vocabulary is reduced by combining tokens: The words are first made lowercase, and then can be either stemmed, lemmatized, or kept as they are. Lemmatization algorithms map words to their “canonical” forms without changes in tense or other modifications, for example changing *families* to *family*, *saw* to *see* and *best* to *good*. This has more complexity than stemming the words using an algorithm that removes suffixes to get to an approximate common root (Grimmer et al. 2022). Finally, with the vocabulary set, the corpus can be transformed into a document–feature matrix.

This analysis will only be focusing on *one* of these key decisions—the unit of analysis—also referred to as chunking or text splitting strategy. As such, the rest of the parameters must be standardized across models. The next section will describe the specific configurations of text chunking that are being compared as well as the rest of the preprocessing pipeline.

4.2 Preprocessing Parameters for STM of AEA Addresses

4.2.1 Text Chunking Variations

There are four levels of text chunking that will be compared in this analysis. The first has each document in the document-feature matrix being a full address and will be abbreviated to “*Doc*” or “*Full Address*” when labeling models. Like other statistical models, STM can better represent corpus structure with a larger sample size, meaning

that there will likely be less nuance in the emergent topic structure. Other units of analysis will be uniformly sized text chunks: The addresses will be split every 500 words into chunks labeled as "*Pages*" and every 200 words into "*Paragraphs*" (sometimes abbreviated to "*Para*"). It's important to note that the word counts will not be uniform across text units passed to the document-feature matrix due to the removal of stop words and low-frequency words. The final unit of analysis being compared uses the paragraph splits created by the author and are expected to be more semantically meaningful groupings.[2] However, paragraph length varied between addresses, and some "paragraphs" were single line headers. That is why the original paragraphs from the text were combined or split into sections of 100 to 300 words.[3] Again, the word counts for each text were measured before any preprocessing. This unit will be labeled as "*Bounded Paragraphs,*" "*Paragraphs Bounded,*" or "*ParaB.*" Table 4.1 displays summary statistics for each text chunking configuration. Corpora of these differently split addresses were then passed to the same preprocessing pipeline.

[2] This configuration had the drawback of being very time-consuming. Document AI attempted to separate paragraphs when parsing the PDFs, and some inferences could be made about paragraph breaks by very short lines, but the validation was tedious and manual.

[3] To get a sense for this size range, the paragraph describing text units contains approximately 240 words.

Table 4.1: Text Unit Counts and Vocabulary Size for Address Splitting Configurations

Text Chunking Configuration	Abbreviation	Initial Word Count Bounds	Number of Text Units	Vocabulary*	Token Count† Mean (sd)
Full Addresses	*Doc*	*n/a*	126	4900	3732 (1584)
"Page" splits	*Page*	500 words	2039	5700	236 (38)
"Paragraph" splits	*Para*	200 words	5009	5861	96 (13)
"Bounded Paragraph" splits	*ParaB*	100-300 words	5772	5996	87 (37)

* The vocabulary size varies with differently split texts because the lower threshold is based on the number of documents a given token appears in.

† This token count is measured *after* removing stop words and low frequency words. It is calculated from the values of the document-feature matrix, aggregating the total number of tokens in each document (row) then calculating the mean and standard deviation.

4.2.2 Standardized Preprocessing Pipeline

Since only the impact of text chunking is being analyzed, the rest of the preprocessing steps must stay constant. As previously discussed, it's important that researchers be thoughtful when making these decisions and be able to justify them. This section will give an overview of the decisions made and values held constant, and a deeper discussion can be found in Appendix A. The first step in this process is removing stop words, punctuation, and special characters. Though some researchers choose to curate a corpus-specific stopword list, this step is time-consuming and does not yield significantly better results (Schofield et al. 2017). Most of this preprocessing was done using the R package *quanteda* (Benoit et al. 2018), so their default stop word list was used.

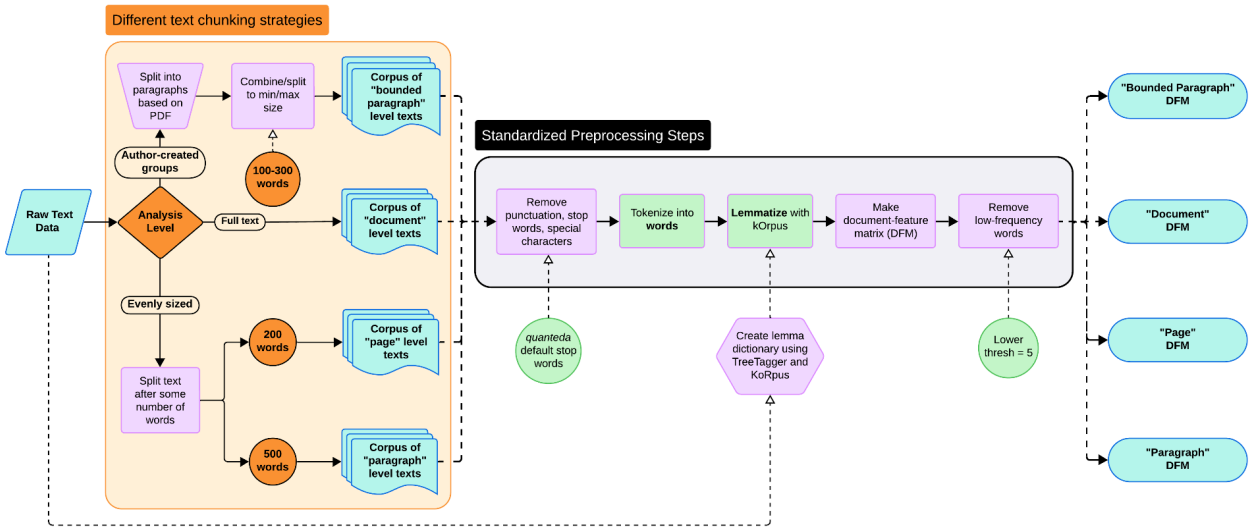
Next was the question of how the texts should be tokenized: While 2 or 3-grams could be useful for maintaining meeting of texts in bag-of-words representations, topic

modeling already allows for the meaning of words to be determined by other words in the topic (DiMaggio et al. 2013). For simplicity and readability on graphics, the tokens will only be words—or 1-grams. The vocabulary was further simplified by creating "equivalence classes" and lemmatizing words using the TreeTagger software (Schmid 1997) through the *koRpus* package (Michalke 2021). Though lemmatization has the added step of creating a dictionary mapping of tokens, it was preferred over rule-based stemming to allow for more comprehensible tokens and a greater breadth of meanings.[4]

The final step is to filter out lower-frequency tokens from the vocabulary based on a lower threshold. This decision has been investigated by Bystrov et al. in "Analysing the Impact of Removing Infrequent Words on Topic Quality in LDA Models" (2023): Using randomly generated corpora with known topic distributions, they found that topic quality only significantly diminished when removing terms that were in more than 3% of the documents, making it a worthwhile step for improving runtime. Using the *plotRemoved* function in the R package *stm* (Roberts et al. 2019), a lower threshold of 5 was set for all models. A summary of text chunking strategies and the established preprocessing decisions are in Figure 4.2.

[4] See Appendix A for an illustration of the advantages of lemmatization over stemming.

Figure 4.2: Outline of Text Chunking Variations and Standardized Preprocessing Steps



4.3 Structural Topic Model Hyperparameter Tuning

Once the data is prepared, the researcher must set hyperparameters for the structural topic model. Guides for social scientists put the most emphasis on selecting the optimal number of topics, emphasizing the importance of qualitative validation (Lindstedt 2019; Roberts et al. 2013; Roberts et al. 2019; Ulstein 2024). This will be addressed in depth later in the paper. The other important hyperparameters are initiation state (*init.type*) and the *prevalence* and *content* covariate formulas. If runtime is a concern, may also specify the maximum number of iterations (*max.em.its*) and convergence tolerance (*em.tol*).

What separates STM from other topic modeling strategies are the incorporation of covariates into the corpus structure. They can be used in the estimation of topic prevalence—the distribution of *topics* across *documents*; or topic content—the

distribution of *words* across *topics*. The model also outputs estimated covariate effects which are “analogous to GLM coefficients familiar to social scientists” (Roberts et al. 2019:2). In this analysis, “meeting year” will be the prevalence covariate, transformed with a natural cubic spline with 10 degrees of freedom, as it is not expected to have a linear relationship with the topics. Visualizing the changing prevalence of topics in addresses across time will be an integral part of model interpretation and evaluation.

Due to the added complexity of these covariates, the output of STM is especially sensitive to different random seeds. That is why it is common practice to use Arora et al.’s (2013) spectral initiation which produces deterministic results. Researchers may also fit models with a variety of seeds and select the most semantically meaningful one.

In this analysis, spectral initiation was used for fitting the models with different numbers of topics, which is further explained in the next chapter, followed by a discussion of methods for selecting the “optimal” number of topics given the metrics of exclusivity and semantic coherence.

CHAPTER V - SELECTING AND COMPARING THE OPTIMAL NUMBER OF TOPICS

5.1 Introduction

The first part of this analysis investigates the effect that preprocessing steps have on the recommended number of topics (K) for the final model. This approach assumes that similar meanings and categories will emerge when the words are distributed across a similar number of topics, or with the topics in some models being the condensed forms of those in others (e.g. 20 vs. 40 topics). However, as emphasized in topic modeling literature and guides, there is no definitively optimal number of topics for a given corpus—it depends on what meanings are deemed important by the researchers. The common practice is to fit models over a range of K and evaluate the results quantitatively to narrow it down, followed by a qualitative assessment of the topics.

The evaluation metrics on which this paper focuses are *semantic coherence* (Mimno et al. 2011) and *exclusivity* (Bischof and Airoldi 2012). Semantic coherence is a measure of the internal consistency of topics using co-occurrence of most prevalent words for each topic, and exclusivity refers to how distinct the topics are from one another. These measures have an inverse relationship to one another, but, for interpretability, the best model is one that balances exclusivity and semantic coherence

(Roberts et al. 2019). Topic models can also be evaluated using *held out-likelihood*—that is, their predictive ability for out-of-sample documents—yet prioritizing accuracy in this regard may decrease interpretability (Lindstedt 2019).

As previously discussed, the focus of this analysis is determining the effect that different text splitting strategies affect the topics produced by a structural topic model—both the number of topics and the meanings derived from them. Table 5.1 provides a summary of the splitting configurations being compared. The analysis will proceed as follows: For each text chunking configuration, structural topic models will be trained with 4 to 75 topics. Next, the mean exclusivity and semantic coherence across topics for each model will be calculated and graphed in faceted line plots. Then, the plots will be evaluated and a few values of K with an adequate balance of exclusivity and semantic coherence will be recorded. Finally, the analysis will discuss some trends in K values in models with differently sized units of analysis.

Table 5.1: Text Chunking Variations, Abbreviations, and Parameters

Text Chunking Configuration	Abbreviation	Initial Word Count Bounds
Full Addresses	*Doc*	*n/a*
"Page" splits	*Page*	500 words
"Paragraph" splits	*Para*	200 words
"Bounded Paragraph" splits	*ParaB*	100-300 words

5.2 Methodology

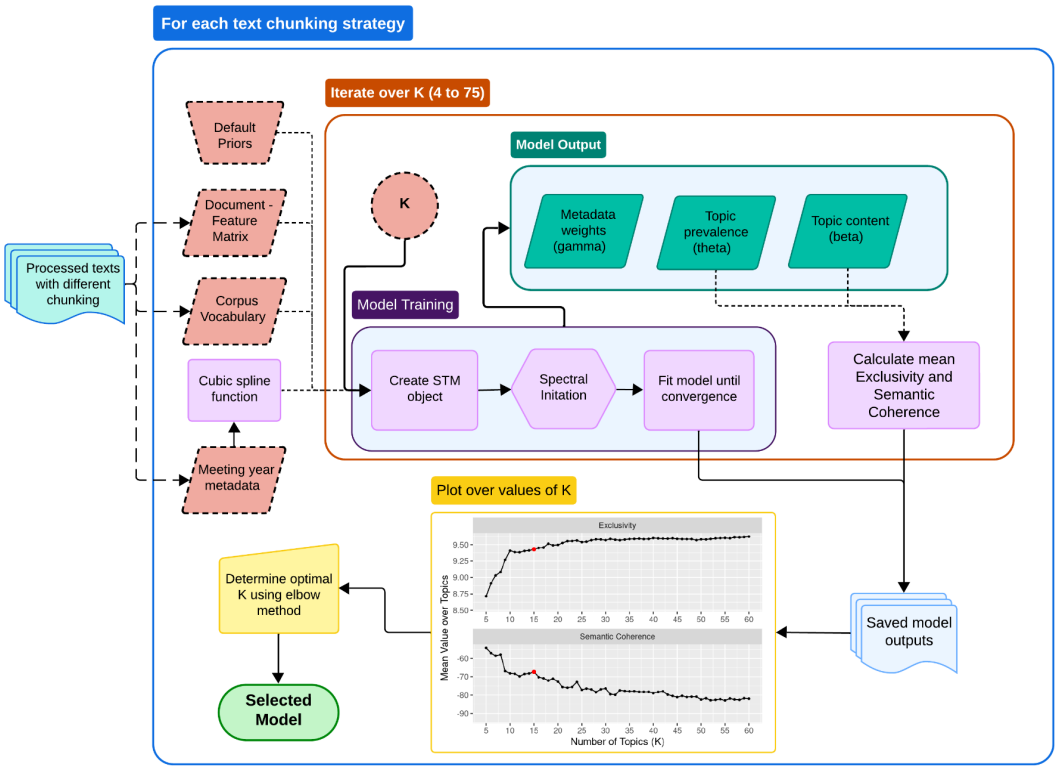
5.2.1 Fitting Models

Figure 4.2 illustrates the process of selecting the optimal number of topics across the differently split document configurations. Each configuration consists of a document-feature matrix, a vocabulary vector, and a dataframe mapping each document to its corresponding meeting year. These components, along with a specified number of topics (K), are passed to the *stm* function. A natural cubic spline is applied to the meeting year data within the topic prevalence formula, while model priors remain at their defaults.

The structural topic model is then fitted using the “spectral” initialization method, iterating until convergence.[5] If the model does not converge within 500 iterations, it is discarded. The resulting topic prevalence and topic content matrices are used to compute exclusivity and semantic coherence for each topic. This process is repeated for K values ranging from 4 to 75 and the values of average exclusivity and semantic coherence are plotted across K. Once all preprocessing configurations have been processed, the results are reviewed and one or more values of K for each model are selected that best optimize these metrics.

[5] Convergence is defined as when the change between steps is less than 0.0001

Figure 5.1: Model Selection Process Applied to Each Text Chunking Strategy

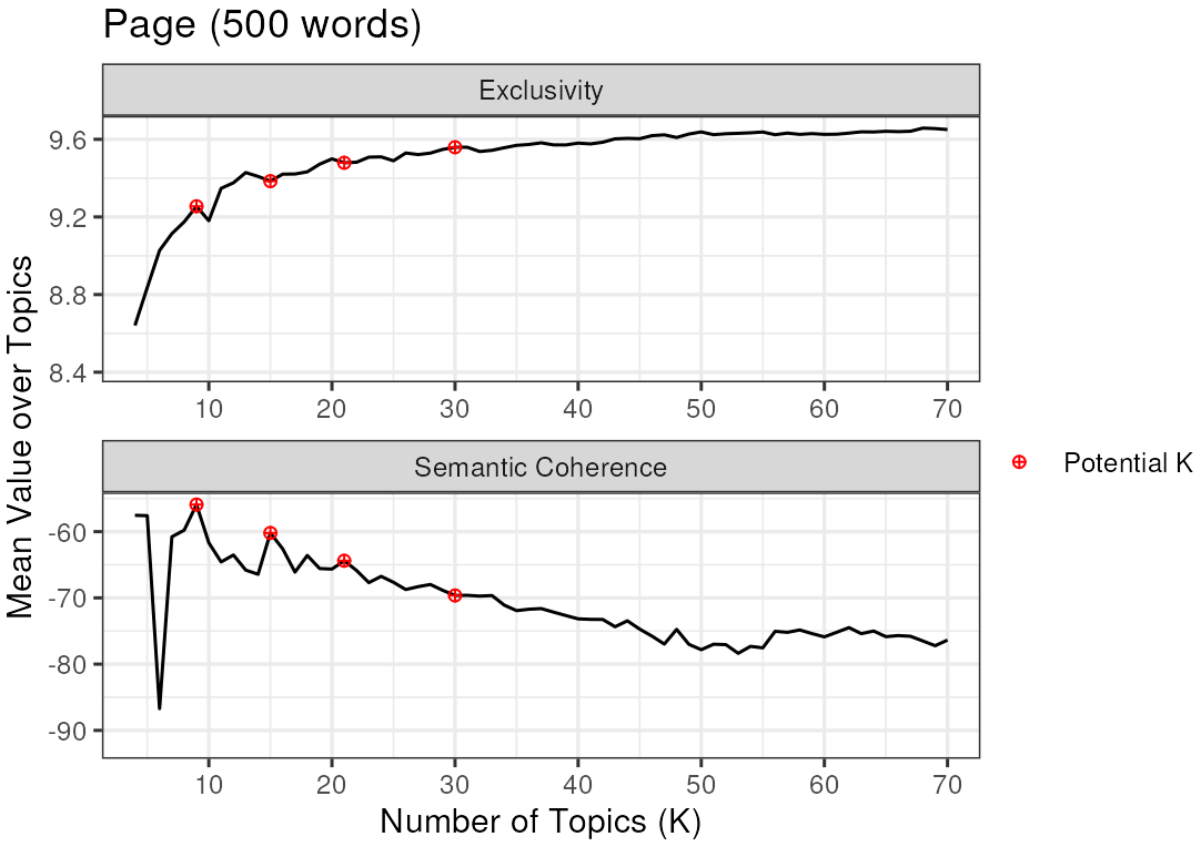


5.2.2 Selection of “Optimal” K

Though there are many metrics with which one can determine a good number of topics for a Structural Topic Model, only semantic coherence and exclusivity were calculated and then plotted across values of K. Generally, efforts are made to identify the “elbow effect”—the point at which increasing the number of topics has a diminishing return on exclusivity or corresponds with a significant drop-off in semantic coherence (Roberts et al. 2013). Values of note are also at local maxima, though a peak in one metric at a given value of K often corresponds with a drop in the other. Figure 5.2 is an example of a semantic coherence and exclusivity plot used to select the ideal number of topics for the *Page* model, with potential values of K marked in red: The values of $K = 9$ and $K = 15$ are distinguished by the significant peak in semantic coherence

without a significant decrease in exclusivity. The values $K = 21$ and $K = 30$ were also selected due to plateauing of exclusivity values there.

Figure 5.2: Mean Exclusivity and Semantic Coherence Values for the Topics of Models Trained across Different Values of K.



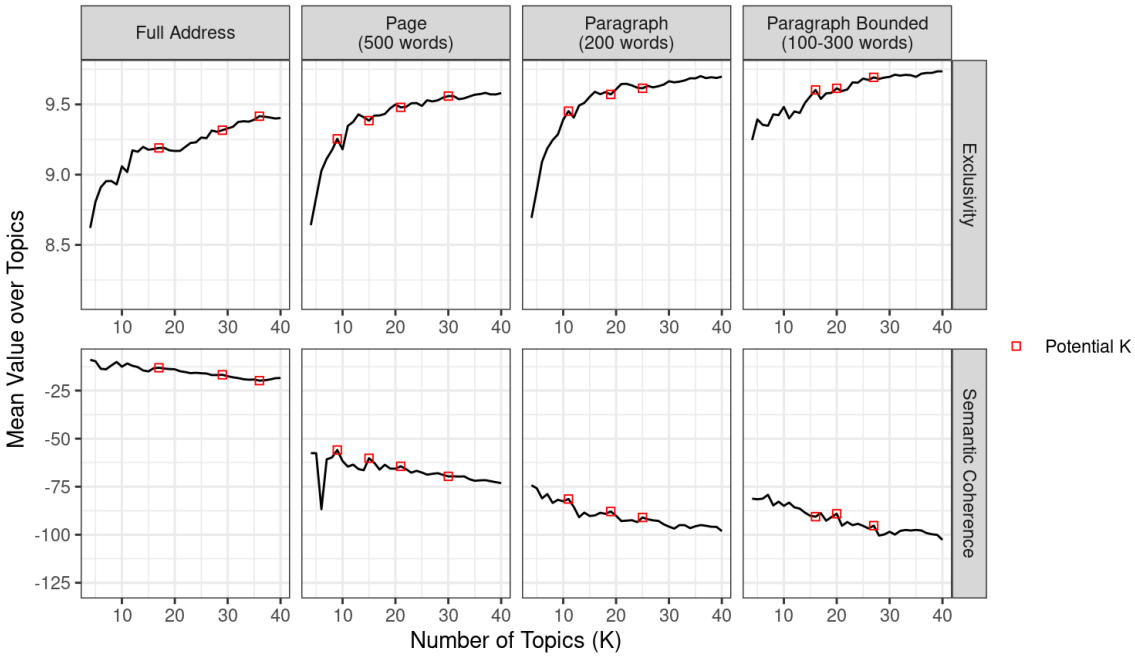
Selected values of K marked in red

5.3 Comparing K across Text Chunking Strategies

Figure 5.3 summarizes the semantic coherence and exclusivity for the models of each variation of text splitting. There are a few trends of note. Firstly, the model created from text kept as full addresses have a much greater semantic coherence than all of the other ones. This makes sense given that semantic coherence depends on word

co-occurrence, and words are more likely to co-occur when the documents are so long. Following that pattern, the “page” model has a greater semantic coherence than the paragraph one, and, inverting that relationship, the models made with paragraph chunks have a higher exclusivity than those with longer chunks.

Figure 5.3: Mean Exclusivity and Semantic Coherence across K for Each Text Splitting Configuration



Selected values of K marked in red

However, there are no strong patterns in the suggested values of K: All values are between 10 and 30, save the Full Address model at $K = 36$ and the Page model at $K = 9$. Since these ranges are overlapping, no conclusions can be made about the relationship between text chunking and optimal number of topics. So, the next step in this analysis is comparing topic content across the candidate values of K and text chunking strategies to further narrow down values of K and identify prevalent themes in the corpus.

CHAPTER VI - TOPIC CONTENT EVALUATION AND ALIGNMENT

6.1 Introduction

The next chapters will focus on comparing the results of the topic modeling and evaluating their interpretability. In this stage of a sociological analysis, the researcher would inspect the output of models with different values of K for meaningful insights given their prior domain knowledge. Both topic content and topic prevalence are relevant to this interpretation: **Topic content** refers to the distribution of words over topics, revealing main themes of the corpus; and **topic prevalence** is the distribution of topics over documents (with reference to the time covariate in this case).

For models with the same vocabulary such as these, topics content (β) distributions can be compared mathematically using entropy metrics like Pointwise Mutual Information, Jaccard similarity, and Kullback-Leiber distance; as well as coordinate space comparisons with cosine distance (Ballester and Penner 2022). However, since social scientists are more likely to label topics by the most prevalent words rather than intricacies of the β distribution, the similarity metric selected for this analysis is the number of words that overlap in the models' top 10 most prevalent words. If the meaning revealed by structural topic modeling is robust to text splitting, then the models will have several overlapping topics with similar temporal trends.

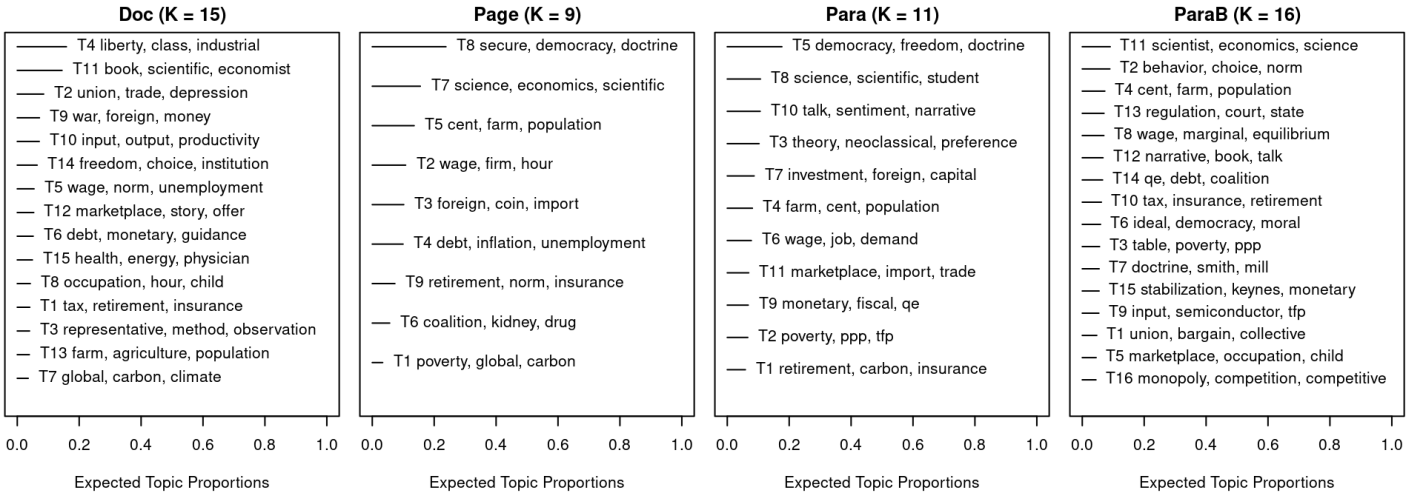
6.2 Methods: Comparing Topic Content

Topic content is generally represented qualitatively as a list of the most prevalent words in each topic. The *stm* package's *labelTopics* function calculates the top n words based on metrics such as the distribution of topic-word frequencies (β); weighted proportions of word-topic frequency vs. word-corpus frequency; or FrEx, the weighted harmonic mean of frequency and exclusivity (Roberts et al. 2017). FrEx (Bischof and Airoldi 2012) was selected for topic labeling as it produced the most differentiated results when testing, so topic similarity will be measured as the number of overlapping words in the 10 tokens with the highest FrEx values.

Figure 6.1 is the output of the standard Structural Topic Model plotting function that lists the top words in each topic and an estimation of prevalence.[6] It is expected that a researcher would use this visualization to get an initial sense of the interpretability of a given model using their subject matter expertise. This overview presents some immediately comprehensible groupings from just the top three words such as "freedom," "liberty," and "democracy" in Topic Doc-4 and "wage," "demand," and "job" in Topic Para-6.[7] However, there are some topics that the model produced that are less coherent such as Para-1 where "carbon" (likely referring to climate emissions) is grouped with "retirement" and "insurance." This case of word intrusion (Chang et al. 2009) indicates that the topics in the *Para* model are not *exclusive* enough, thereby suggesting that a higher value of K would better represent corpus structure.

[6] Models with the lowest *K* for each text unit were selected for this first comparison in order to make the figure more readable.
[7] Specific topics will be referred to in the text with the text group's abbreviation followed by a dash and the topic number such as "Doc-T4."

Figure 6.1: Basic View of Topic Content and Prevalence across Models



This visualization was created from simply calling the plot *function on the fitted Structural Topic Model. The top three words are determined by FrEx score (Bischof and Airoldi 2012) and topic prevalence is derived from the θ distribution of topics over documents of the corpus*

The next section will first illustrate the change in topic content across different values of K for each chunking strategy, then evaluate the extent to which topic content and themes stay constant, and finally compare topic content across models with differently processed text. This investigation assumes that if the text chunking strategy does not fundamentally change the emergent structure and insights of the corpus, then there will be significant overlap in the prevalent themes. Additionally, the sankey-based visualization strategy (inspired by Rule et al. 2015) helps to visualize these relationships.

6.3 Comparing Topic Content across K with Fixed Chunk Size

This analysis will first look at the topic content of the *Paragraph* STMs due to the clearly incongruent words in its Topic 1 (see Figure 6.1). Figure 6.1 only shows the top 3 words of the model, so there may be other topics that would be more coherent if split up. To investigate the hierarchical relationships and content overlap between topics at different values of K, Figure 6.2, maps changes in topic content across the Paragraph-level models when K is increased from 11 to 19 to 25.[8] In the diagram, nodes represent a single topic in a model and are linked to topics in other models that share at least one of their top 10 words. The links are weighted by the number of tokens in common, which can be represented mathematically as the cardinality of the intersection between the sets of tokens, or:

$$w_{ij} = |T_i \cap T_j|$$

where T_i and T_j are the sets of top 10 FrEx-scored tokens for topics i and j from different models. It's also important to note that the node labels only include the words that overlap with adjacent nodes. As a result, non-overlapping words and disjoint topics are not represented in this diagram.

[8] See Figure 5.2 in in Section 5.3 for exclusivity and semantic coherence graphs that recommend these values

Figure 6.2: Sankey Diagram Showing the Flow of Similar and Diverging Topics as K Increases in *Paragraph* Models.

Link weights are determined by the size of the intersection between the sets of the top 10 tokens (ranked by FrEx score) for each model, and all three structural topic models are made from addresses split at the Paragraph level (~200-word chunks).

Note that there is a pink node labeled with a 13 in the column furthest to the left, despite there only being 11 topics in that model there. This is because topic 13 is in the 19-topic model but has no incoming links from topic content overlap with the 11-topic model. As a result, the node was placed on the far-left side.

There are several key topics that persist despite differences in K, represented by a thick link between nodes. For example, the content of Topic 6 (in hot pink, second/third from the bottom) across all three models has different words relating to employment and the labor market such as “wage,” “job,” and “worker.”

There are also cases of distinct topic divisions—when increasing the number of topics makes it so that somewhat unrelated words in one topic are separated into more meaningful topics of the new model: As shown in Table 6.1, Topic 1 with the incongruent combination of “retirement” and “carbon” as top words was split two topics with a larger model. Those topics were distinct enough to persist from the model with 19 topics to the one with 25, with 9 out of the 10 words staying the same. The contents of Topic 2 are split from the 11-topic model to the 19-topic one, and then split in two *again* when there are 25 topics (see Table 6.2). The resulting groupings are much more distinct and meaningful than those when there are fewer topics.

When replicating this investigation with models created from *Full Addresses, Pages*, and *Bounded Paragraphs*, similar relationships surfaced: Of note is the splitting of some topic content in the *Full Address* model when the number of topics increases from 17 to 29 to 36. Presented in Table 6.3, these splits do not separate incongruent topics but rather delineate more specific subthemes. The full Sankey diagrams and notes on these relationships are found in Appendix D.

Table 6.1: Topic Splitting into Semantically Meaningful Groupings when Increasing the Number of Topics from 11 to 19 for *Paragraph* STMs

K = 11	K = 19	K = 25
Topic 1: retirement, carbon, insurance, coalition, emission, climate, benefit, annuity, abatement, club	Topic 19: coalition, emission, carbon, abatement, climate, tariff, club, regime, region, ton	Topic 19: coalition, carbon, emission, global, climate, club, region, abatement, tariff, regime
	Topic 17: tax, insurance, retirement, benefit, security, revenue, annuity, program, fund, taxation	Topic 17: tax, retirement, security, revenue, annuity, benefit, taxation, fund, save, taxable

Table 6.2: Example of Repeated Splitting of Topic Content as K Increases from 11 to 19 to 25 for *Paragraph* STMs

K = 11	K = 19	K = 25
Topic 2: poverty, ppp, tfp*, occupation, index, computer, icp*, hour, gender, datum	Topic 2: poverty, survey, ppp*, index, revision, icp*, poor, india, count, line	Topic 2: measure, datum, index, statistical, estimate, measurement, audience, statistic, survey, parameter
		Topic 1: reduction, poverty, revision, poor, percent, tfp*, gdp*, rich, real, inequality
	Topic 18: input, productivity, occupation, hour, output, semiconductor, fertility, gender, computer, earning	Topic 18: input, semiconductor, productivity, computer, software, capital, output, equipment, household, stock
		Topic 22: occupation, hour, gender, earning, woman, representative, gap, fraction, deviation, valuation

* Glossary of abbreviations found in Appendix C

Table 6.3: Hierarchical Splitting Behavior with Full Address-Level Topic Models		
K = 17	K = 29	K = 36
Topic 4: liberty, industrial, class, secure, competition, democracy, court, power, doctrine, property	Topic 4: liberty, industrial, class, doctrine, laborer, english, competition, mass, spirit, struggle	Topic 4: laborer, class, mass, competition, industrial, spirit, industry, workman, doctrine, secure
		Topic 30: liberty, industrial, class, modern, english, society, struggle, man, evolution, achievement
	Topic 27: court, power, property, private, sentiment, motive, public, secure, state, protect	Topic 27: property, court, private, power, public, corporation, protect, state, government, sentiment

Recognition of these dynamic groupings can ultimately help researchers choose the final model based on the granularity required for their research question. Additionally, less optimal values of K will not align as well with the other models when visualized in this way.[9] That being said, the quality of a given model cannot be determined from topic content alone, as topic prevalence is the other integral part of corpus structure.

For this analysis, visualizing the connections between topics of different models also serves to align topics across models made from different text units. While Roberts et al. (2017) uses the “Hungarian method” created by Kuhn (1955) to match topics in a way that minimizes the sum of their inner products, that method only performs one-to-one matches and therefore does not capture the splitting and merging relationships, nor does it reveal topics that are meaningfully unique to only one of the models. Therefore, this approach to topic visualization proves to be a valuable addition to STM tuning and interpretation methods.

[9] An example of this can be found in Appendix E

6.4 Overlaps in Topic Content across Text Chunking Strategies

The previous section explored the dynamics of topic content across values of K for models with the same text unit, and now that method will be used across the models created from differently sized text chunks. These will be labeled as:

- **Doc** – *Full addresses*
- **Page** – *500-word chunks*
- **Paragraph** or **Para** – *200-word chunks*
- **Bounded Paragraph**, **Paragraph Bounded** or **ParaB** – *author-created paragraphs split or combined to have between 100 and 300 words.*

Figure 6.3 visualizes the alignment and flow of similar topics across the models, highlighting some consistent streams as well as splitting/merging relationships.

From the Sankey diagram, meaningfully overlapping topic sets were recorded and structured into a spreadsheet with all 10 of the top words. These sets did not always contain all 4 of the models however—there were several topic sets that only came from two or three of the models. Then, the grouped topic content was passed to a large language model to produce informal topic labels,[10] the output of which is in Table 6.4. These themes are quite varied in nature, covering a range of methods (*Poverty Metrics*, *Monetary Policy Tools*); central economic concepts (*Labor & Industry*, *Productivity & Growth*); theories (*Behavioral Economics*, *Economic Modeling*); applications (*Climate Governance*; *Retirement Systems*); and abstract principles (*Nature of Economics*; *Decision & Ethics*).

[10] This application of LLMs was explored by Li, Zhang, and Zhou (2023) and was found to be generally credible, though the authors caution that it should not be the *only* grounds for labelling topics. However, their LLM prompting process was much more complex and dynamic than in this paper. See Appendix B for more details.

Figure 6.3: Sankey Diagram of Topic Content Similarities and Relationships across Text Chunking Strategies

†Link weights are determined by the size of the intersection between the sets of the top 10 tokens (ranked by FrEx score) for each model.
‡ Models with similar values of K and therefore the same degree of granularity were chosen to be visualized here. Each of them was also validated from their own Sankey diagrams to ensure that the topics were coherent. They are ordered based on the size of the text chunk. One limitation of this Sankey configuration is that it does not show similarities between topics in non-adjacent models.
*** Note that there is a red-orange node 11 in the column furthest to the left with links that connect to nodes in both the second and third columns. This is because topic 11 in the Page model had no incoming links from topic content overlap with the Doc model. It was therefore placed on the far-left side and brought Topic 16 from the ParaB model from the third to the second column.*

Table 6.4: Overview and Informal Labels of Consistent Topic Themes in Differently Chunked Text		
Topic Sets	**Example words**	**Informal Label**
Doc-14, Page-9, ParaB-2, Para-9	choice, preference, agent, empirical, expectation	Behavioral Economics
Doc-11, Page-14, ParaB-12, Para-13	book, mill, economist, professor, scientific	Scholarly Thought §
Doc-2, Page-3, ParaB-15, Para-7	depression, export, policy, gold, recovery	Economic Cycles §
Doc-7, Page-13, ParaB-14, Para-19	carbon, climate, emission, coalition, region	Climate Governance
Doc-14, Page-9, ParaB-2, Para-3	choice, preference, moral, social, freedom	Decision & Ethics
Doc-17, Page-2, ParaB-8, Para-6	wage, equilibrium, model, inflation, expectation	Economic Modeling
Doc-11, Page-14, ParaB-12, Para-10	book, economist, story, narrative, talk	Economic Storytelling
Doc-16, Page-15, ParaB-4, Para-4	farm, agriculture, food, population, century	Historical Agriculture
Doc-8, Page-10, ParaB-5, Para-18	occupation, gender, earning, college, hour	Labor & Gender Gaps
Doc-5, Page-2, ParaB-8, Para-6	**wage, unemployment, supply, price, employment**	**Labor Markets**
Doc-6, Page-4, ParaB-14, Para-12	debt, bank, qe*, monetary, guidance	Monetary Policy Tools
Doc-7, Page-13, ParaB-3, Para-2	poverty, ppp*, icp*, global, line	Poverty Metrics
Doc-10, Page-10, ParaB-9, Para-1	productivity, input, capital, tfp*, growth	Productivity & Growth
Doc-1, Page-5, ParaB-10, Para-17	**retirement, tax, benefit, annuity, insurance**	**Retirement Systems**
Doc-10, Page-10, ParaB-9, Para-18	productivity, computer, output, occupation, gender	Work & Tech Productivity
Page-7, ParaB-11, Para-14	economist, theory, truth, solution, sound	Economic Thought †
Page-3, ParaB-15, Para-15	depression, policy, budget, union, business	Fiscal History †
Page-11, ParaB-1, Para-15	union, organization, management, corporation	Labor & Industry †
Page-3, ParaB-15, Para-12	stabilization, monetary, policy, budget, debt	Macroeconomic Policy †
Page-11, ParaB-16, Para-15	monopoly, competition, corporation, industry	Market Structure †
Page-7, ParaB-11, Para-8	science, economics, theory, discipline, knowledge	Nature of Economics †
Page-7, ParaB-11, Para-8	economics, freedom, regulation, association	Political Economy †
Page-6, ParaB-1, Para-15	union, contract, employer, organization, company	Union Relations †

Table 6.4: Overview and Informal Labels of Consistent Topic Themes in Differently Chunked Text		
Topic Sets	**Example words**	**Informal Label**
Doc-13, Page-1	inequality, income, sector, share, underdeveloped	Income Distribution
ParaB-7, Para-13	**liberty, class, ricardo, smith, mill**	**Classical Thinkers §**

** Abbreviations defined in Appendix C*
† Themes not represented in the Full Address (Doc) model - generally more conceptual than applied in nature
§ Labeled in a separate ChatGPT instance, see Appendix B for details
***In bold** are topics that will be further explored in later chapters*

6.5 Conclusion

Identifying and labeling the persistent themes across text chunking strategies contributes to understanding the structure of the corpus—one of the main objectives of this analysis. However, the themes produced by overlaps between only *some* of the text unit configurations provide more insight into the effect of text chunking on model output: There are several labels that apply to all of the models except for the document-level one, and these generally relate to more abstract concepts such as *Economic Thought* and *Macroeconomic policy* rather than more concrete applications (see notes of Table 6.4 for exact topics).

Additionally, this summary of topic content in Table 6.4 could also inform researchers as to what the content of an AEA presidential address *could be*, but not *when* or *by whom* these topics were discussed. As mentioned before, it is necessary to interpret topic prevalence as well as topic content to get a more comprehensive estimation of corpus structure. That is why the next chapter will compare topic prevalence over time of topic groups representing similar meanings, highlighting differences in interpretive value based on the shape of the distribution

CHAPTER VII - MODEL INSIGHTS AND COMPARISONS FROM TOPIC PREVALENCE GRAPHS

7.1 Introduction

The previous chapter investigated the extent to which Structural Topic Models of the same text broken up into different lengths produce similar topics. Using a simple metric of the number of overlapping words in the topics' top 10 tokens by FrEx score, several main themes emerged which are summarized in Table 6.4. However, topic content only represents the distribution of *words* over *topics* while topic prevalence describes the distribution of *topics* over *documents*. These STMs were implemented with the year of the address as a topic prevalence covariate and transformed with a natural cubic spline for greater flexibility. For this corpus of AEA Presidential Addresses, investigating topic prevalence over time can be used to highlight changing priorities, interests, and discourses of the discipline.

So, this chapter will take a closer look at a few of the persistent themes found in the previous chapter, plot the topic prevalence of the different models, and discuss interpretive value. These analyses will also include discussions of the ways that the text chunking may have contributed to the differences in results. Ultimately, the results suggest that there are few significant differences between the *Page*, *Paragraph,* and *Bounded Paragraph* models when there are fewer topics, yet the smaller chunks

allowed for smoother trends when K was higher. Additionally, as expected, the lower document count and longer texts in the *Full Address* model resulted in limited representation of temporal patterns.

7.2 Comparing Topic Prevalence of Aligned Topics across Text Chunking Strategies

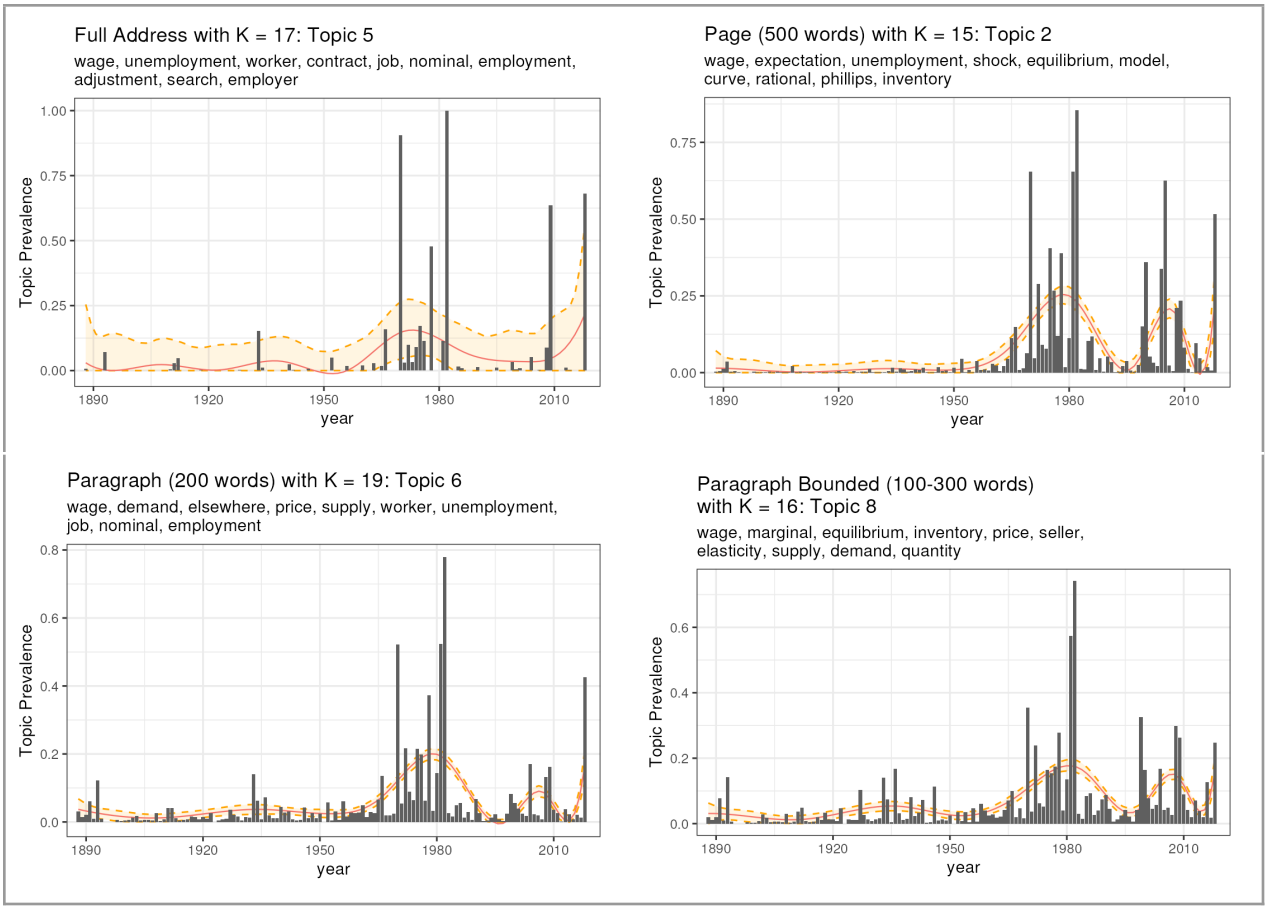
Structural topic modeling distinguishes itself from other topic modeling methods by its incorporation of document-level metadata. The output being analyzed for these comparisons leverages the *stm* package's *estimateEffect* function (Roberts et al. 2019) as well as the raw topic prevalence data of the θ distribution. Then, by plotting topic prevalence over time, researchers of this corpus can determine the changing focuses and priorities of presidents of the AEA, identify connections to historical events, and evaluate the relevance of a given topic and its model to their research question.

In the figures below, the output of *estimateEffect* is plotted as a regression spline with a 95% confidence interval. The bars display the average topic prevalence of the text chunks that make up the address of each year, weighted by chunk size. The top 10 words by FrEx score for each are also listed, allowing the researcher to extract meaning from both topic content and prevalence.

7.2.1 Topic Interpretation: “Labor Markets”

The first persistent theme being discussed is “Labor Markets,” represented by words such as “wage,” “unemployment,” “supply,” and “price." The prevalence over time for topics with this theme across each chunking strategy is presented in Figure 7.1 and can be used to evaluate the topics’ reliability and validity. The consistently shaped topic prevalence distributions across the models with peaks in the 1970s-1980s and the 2000s demonstrate substantial reliability, yet there are some differences. Notably, the bars of the *Full Document* graph are relatively sparse, and the estimated effect line has a very large confidence interval. All topics created from *Full Addresses* exhibit this behavior due to the decreased statistical robustness when only working with 128 documents. Additionally, when entire documents are being modeled, brief references to these themes are less likely to be captured, resulting in more extremes of topic presence or absence rather than smooth trends across time. Still, the peaks in topic prevalence across models align with known economic downturns 1975 and 1982 as well as the “Great Recession” of 2008 (Kalleberg and Wachter 2018), providing evidence for the topics’ validity. This example thus illustrates how the similar temporal trends were represented across text chunking strategies, even as the *Full Address* model failed to capture the same level of detail as the others.

Figure 7.1 Topic Prevalence Over Time of Labor Market–Themed Topics across Chunking Strategies



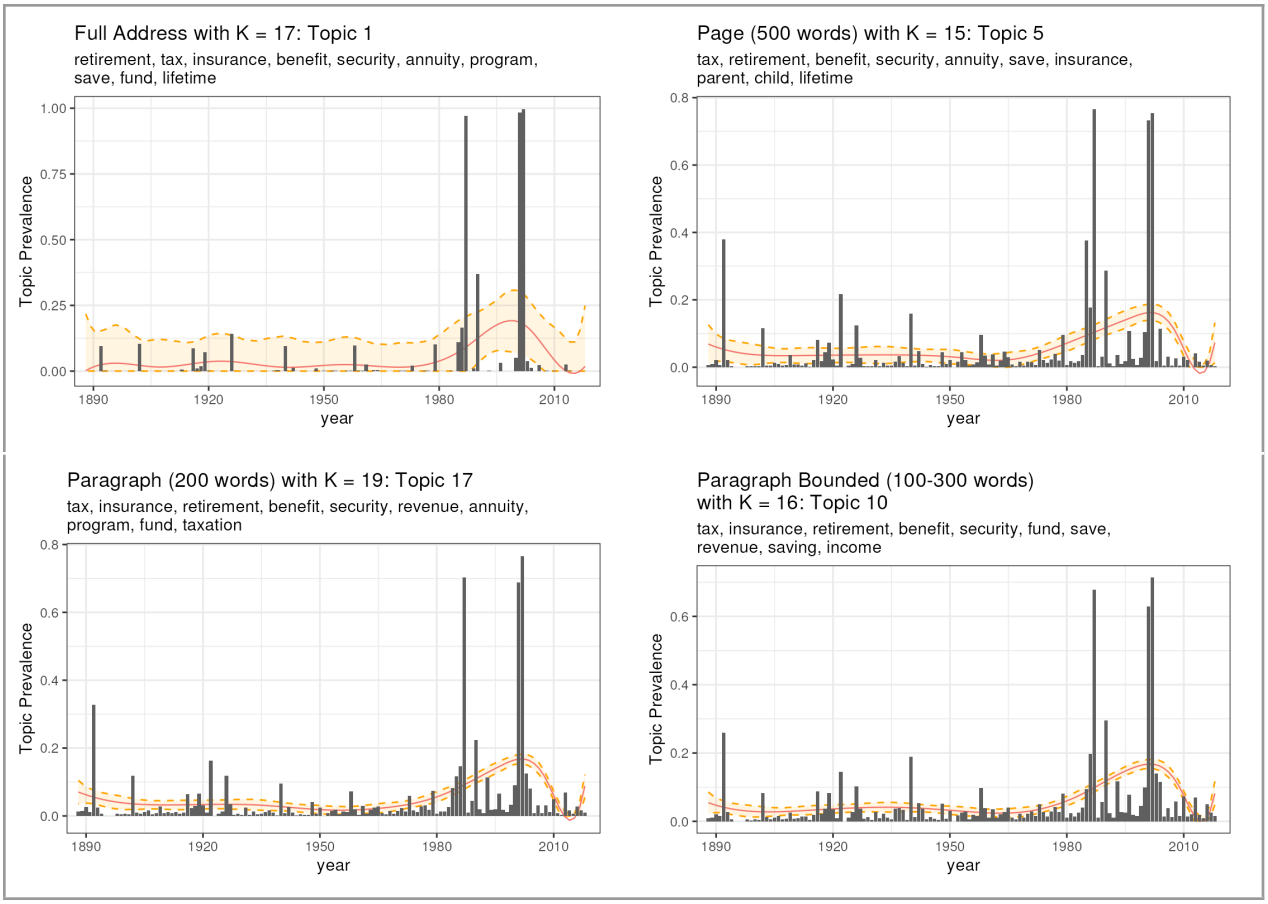
In these topic prevalence graphs, column height represents weighted average of topic prevalence for documents from each years' address, as well as a fitted regression spline with uncertainty shown in red/orange. All models have similarly shaped peaks around the 1970s and 2000s, demonstrating both reliability and validity. However, the Full Address model displays a more sparse and jagged distribution.

7.2.2 Topic Interpretation: “Retirement Systems” and Examples of Sparse Topic Prevalences

The next theme being discussed is “Retirement Systems” and includes words like “fund,” “tax,” “annuity,” and “benefit." This topic theme was very robust and consistent across values of K and text units alike (see Figures 6.2 and 6.3). As shown in Figure 7.2, the prevalence of this topic is largely concentrated in just a few addresses around 1985 and 2005. Though this allows a researcher to infer the main themes of those addresses, it doesn’t reveal as much about the overall structure of the corpus.

The fitted lines appear mostly flat, with a small bump due to the high-leverage points in the later years. Since topic modeling depends on mathematical optimization, highly influential points may lead to overfitting and end up eclipsing other valuable patterns of meaning. In this case, alternative contexts of the token “tax” were not incorporated into the structural topic model due to the dominance of the “tax” / “retirement” mixture. Even though STM is a mixed membership model and therefore could produce multiple topics with “tax” in the label, such a model would have a low exclusivity and therefore not be selected. Researchers can address this limitation by investigating the frequency of specific n-grams of “tax” such as “income tax” or “tax bracket” and rerunning the model with those included. Overfitting may also be a consequence of transforming the meeting year covariate into a b-spline with 10 degrees of freedom rather than keeping it linear, which may have surfaced more general patterns. This decision will be further examined in the discussion section.

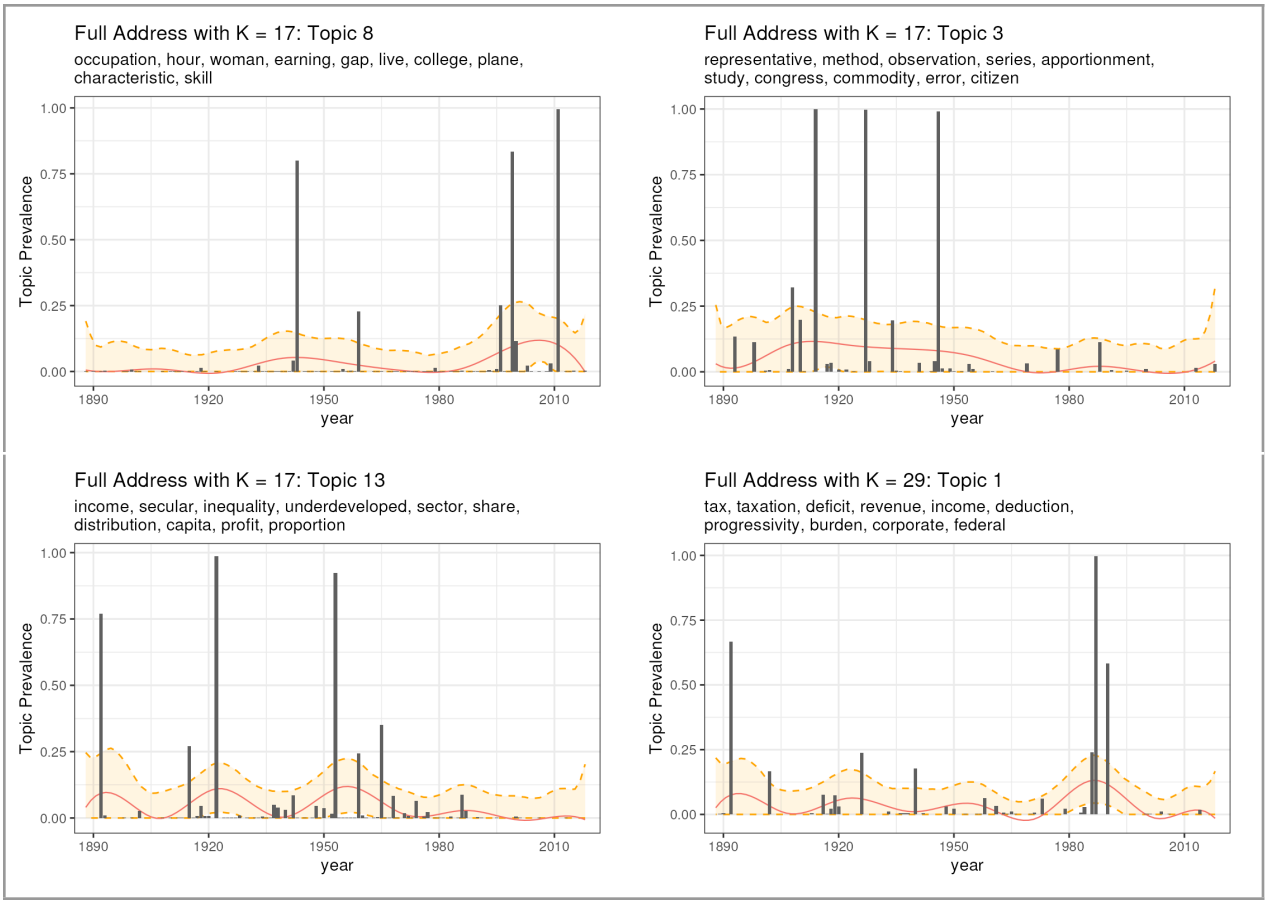
Figure 7.2: Topic Prevalence Over Time for "Retirement Systems"



This figure illustrates a slightly less coherent topic that appears to be overfitted to a few addresses with very specific themes, resulting in a low prevalence across other years. There is still interpretive value in these topics: specifically the focus on retirement and insurance in the 1990s and 2000s, yet the topic mixture is still fairly exclusive to only a few addresses.

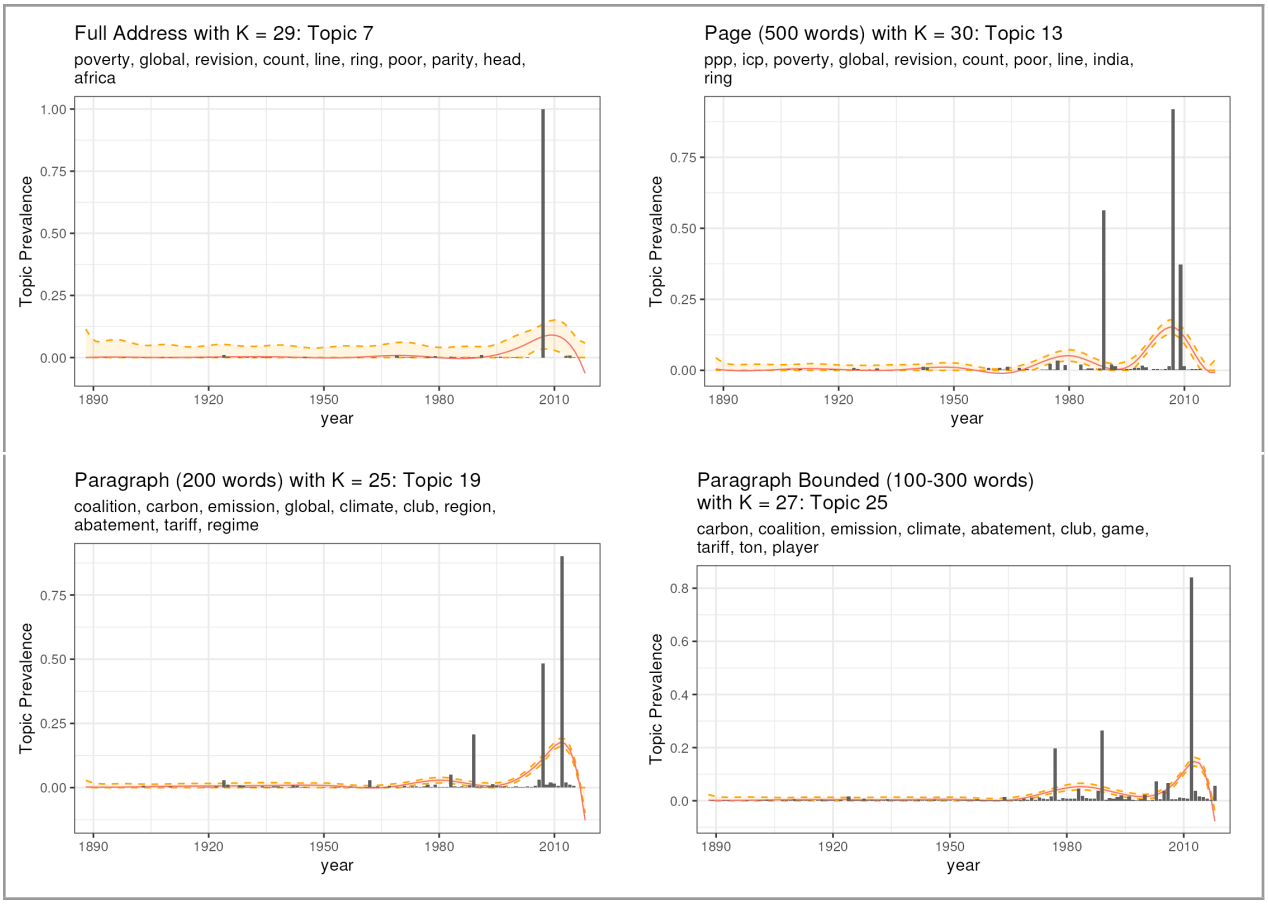
Some other examples of topics with fragmented and extreme topic prevalences resulting from overfitting are presented in the figures below: Figure 7.3 illustrates how the *Full Address* topics often show fewer temporal trends and sometimes have only a few prevalent addresses, and Figure 7.4 highlights how these sparse and potentially overfitted topics become more common and extreme with higher values of K.

Figure 7.3: Sparsity of *Full Address* Topic Prevalences



This figure shows how the Full Address models generally show less detail in temporal trends and have high levels of uncertainty in the estimation of covariate effects due to the smaller sample size. Rather than presenting patterns of discourse in the AEA across time, these graphs only show the extent to which individual addresses match the theme of the topic.

Figure 7.4: Overly Specific Topics with Higher K



Displayed here are prevalence graphs with less interpretive value due to only covering a few addresses. The most detailed model is the Paragraph Bounded one, though it is unclear if all of the topics with nonzero prevalence contain references to "carbon" and other environmental concerns or the more general concepts of "tariffs" and "games."

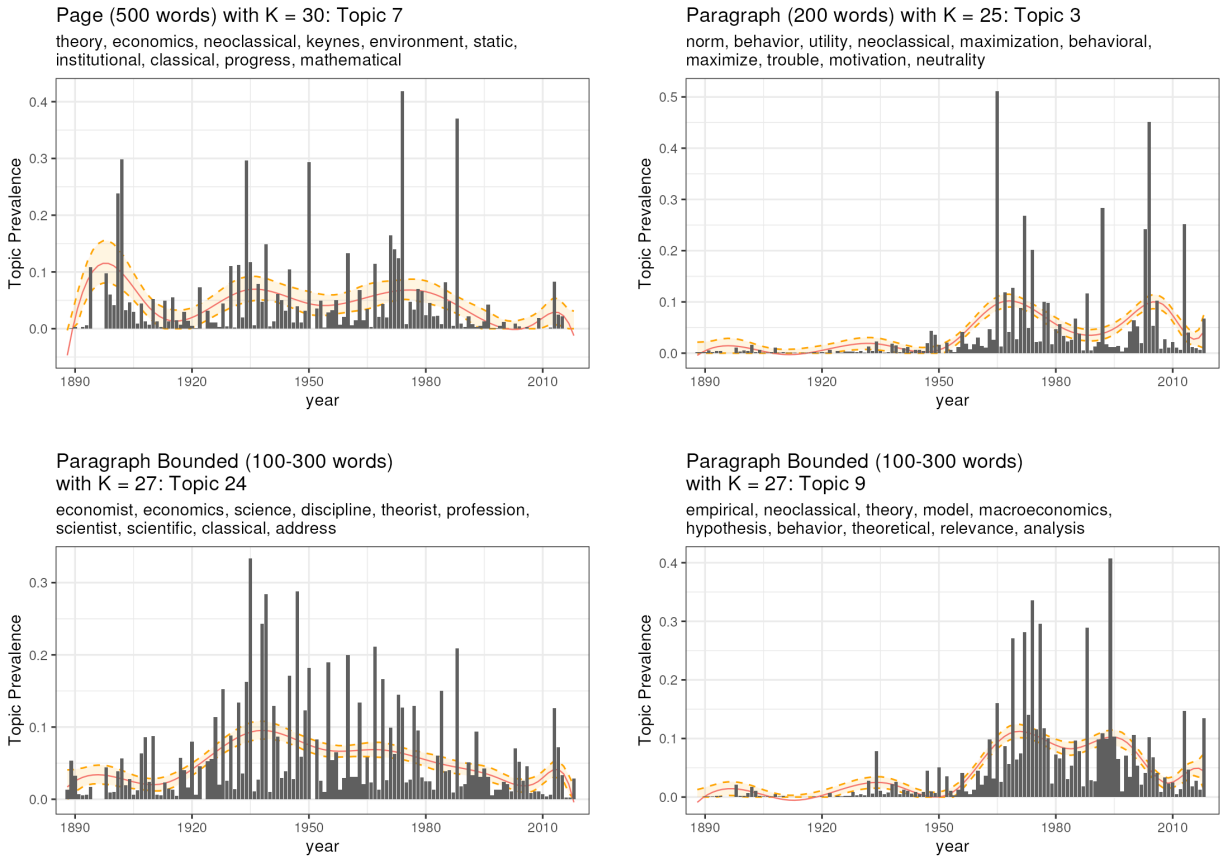
Also of note is the changing meaning of "global" in the different topics: The Full Address and Page models relate to global poverty, while the Paragraph and Paragraph Bounded ones are more related to climate issues. Had the texts been tokenized into 2-grams, "global warming" may have been a top token in this topic and allow for greater distinction between the meanings, though the topics are still very interpretable as is. This observation serves more to highlight the continued effects of preprocessing decisions overall.

7.3 Examples of Smoother Trends with Smaller Text Units

The previous section discussed how some topics will not have the same interpretive potential when only one address is prevalent for the topic, especially as K increases, but this is not always the case: Figure 7.5 presents a few examples of meaningful prevalence patterns even with a greater total number of topics for the *Page, Paragraph,* and *Bounded Paragraph* models. However, with these more granular themes, there are greater differences in topic content and prevalence. Despite the number of overlapping words between the models, they all have different temporal trends and meanings:

The *Page* model refers to different schools of thought in the economics discipline, from “classical” to “keynes(ian)” to “neoclassical,” while the *Paragraph* refers solely to “neoclassical” theory and principles of “maximization” and “behavior." The *Bounded Paragraph* model shows an even smoother prevalence trend and concepts relating to the neoclassical paradigm shift of the 1980s with its greater focus on “models,” “macroeconomics,” and “empirical” studies (Derber 2016). The *Bounded Paragraph* model also has a topic that overlaps with the *Page* model more generally referring to the development of the “discipline” across time as a “science” and a “profession." This great specificity from the *Bounded Paragraph* model is evidence of payoff from the added work of separating the semantically meaningful chunks in the text cleaning phase.

Figure 7.5: Comparison of Topic Prevalence Trends with Higher K Models



Shown here are examples of topics that had meaningful trends in prevalence even in larger models. Note the more filled-in shape of the Paragraph Bounded model.

Although smaller chunks tend to produce more cohesive and interpretable topics than larger ones at higher values of K, this does not imply that they are *less* optimal for lower Ks. While Sbalchiero and Eder (2020) found an inverse relationship between text size and optimal number of topics, this analysis illustrates a more nuanced association. Additionally, these examples show the greater deviation between topics as K increases. However, it is unclear whether this variation is due to the text chunking differences or the variance and flexibility of the models themselves. It is therefore recommended that researchers perform robustness checks on their chosen model by

fitting it with different seeds and initiations to ensure that no contradictory results appear.

7.4 Conclusion

This chapter discussed interpretations of some of the aligned topics across models given their topic content and prevalence. Graphing topic prevalence across time revealed some distinct patterns of discourse and meaning that could help shape understanding of this corpus and the AEA as a whole. Similar meanings could be inferred from the different graphs, yet with some variations: Notably, the *Full Address* level prevalences were more likely to be concentrated on a couple of addresses rather than showing trends over time, while the smaller text units were able to model smoother trends of prevalence, even with higher values of K.

Therefore, if a researcher were interested in uncovering more granular themes and expected to have many topics modeled, this would be best achieved with smaller text chunks. However, when the number of topics were in the 10-20 range, larger chunks were able to model meaningful trends as well. Additionally, the author-created paragraphs (*Bounded Paragraphs)* resulted in more semantically coherent topics when there was a greater number of them, but that could also be because some of the text units were as small as 100 words and therefore achieved more granularity.

CHAPTER VIII - CONCLUSION

8.1 Summary of Results

This study addressed the extent to which text chunking size affects the output and interpretation of Structural Topic Models. While some corpora are made up of shorter documents that can be passed individually to STMs, it is common to break up larger ones to find more granular meanings and increase the sample size (Sbalchiero and Eder 2020). However, there is little guidance on how to perform this, with some researchers using pages as they appear in the text (Nelson 2021), fixed-size chunks (Erikson et al. 2023; Jockers and Mimno 2013), or paragraphs (Guo et al. 2021). So, an unstudied corpus of presidential addresses given at AEA annual meetings was chunked at the document, page, paragraph, and bounded paragraph levels then passed to a Structural Topic Model.

For each of these chunk levels, this study followed the standard steps for preprocessing and model tuning: tokenizing and lemmatizing; removing infrequent words; fitting models over a range of values of K; and selecting candidate models based on exclusivity and semantic coherence of the topics. Then, Sankey diagrams were used to visualize the splitting and merging of topic content across K and aligning topics of models made from differently chunked texts. These also showed select

groupings of tokens that persisted across the text chunking strategies, suggesting that some of the topics produced by STM are robust to text unit size.

The models were further evaluated by graphing topic prevalence over time and comparing with historical events such as wars and recessions: All the models reflected known shifts from theoretical to empirical and applied economics with greater use of quantitative methods (Backhouse and Cherrier 2017). The biggest differences were between the *Full Address* topics and the others, as the small sample size (only one document for each year) resulted in more sporadic and disjoint peaks in prevalence over time. It was also found that models with higher values of K worked better with smaller text chunks and that there was more variance in the topic content between models in those.

8.2 Discussion and Future Research Directions

A common theme throughout this study is the joining of qualitative and quantitative interpretation required to get meaningful results when utilizing topic modeling methods. This is even true when choosing the model hyperparameters: Semantic coherence and exclusivity were calculated for the topics of the fitting models across K, but the numbers themselves were insufficient to determine the ideal value of K. Rather, they needed to be plotted and searched over for significant peaks and plateaus. Alternatively, in some less computationally “grounded” (Nelson 2020) topic modeling implementations, the “ideal” model is selected based purely on evaluation metrics (Sbalchiero and Eder 2020).

It is important to note that it is also common to select the number of topics using the *chooseK* function in the *stm* R package (Roberts, Stewart, and Tingley 2023) which holds out a test sample with which fitted models are evaluated. This is not as recommended for small corpora, however (Weston et al. 2023). Additionally, there are options to create multiple models with the same K and different initiations and then select the ones with the highest semantic coherence and exclusivity or perform robustness checks (Roberts et al. 2019).

In some ways, this flexibility in model selection calls into question STM's reliability in accurately representing the "real" structure of a corpus—some may even consider this "shopping" for models to be a form of data dredging. However, STM is not meant to be used for hypothesis testing but rather as an exploratory tool in conjunction with others. It can even prevent confirmation bias by surfacing prominent themes before researchers apply their own theories to the corpus (DiMaggio et al. 2013).

This analysis uncovered some preliminary themes across the addresses in the form of topics with similar content and prevalence regardless of text chunking or value of K. The next step for a researcher would be to read the documents with high topic prevalences to see if they are consistent with the labels created,[11] and adjust the labels as needed. If the models created from differently chunked addresses referenced generally the same excerpts, then this would provide further evidence of the robustness of STM to text chunk size and would be a valuable topic of future study.

The robustness of the observed similarities across text chunking strategies could also be tested by creating and interpreting models with different initiations and seeds, or even a different corpus: The AEA presidential address corpus was initially chosen due

[11] This can be achieved by using the *findThoughts* function in *stm* (Roberts et al. 2019)

to its heterogeneous formatting (in terms of page and paragraph size) while still having a consistent context and potential for meaningful insights. That being said, further analysis of this corpus would benefit from greater subject matter expertise. Additionally, overfitting ended up being a concern when topic prevalence would only peak for one or two addresses and be otherwise close to zero. An alternative corpus could be the State of the Union addresses, as the addresses vary less from year to year yet there are still well studied topic structures and developments across time (see Rule et al. 2015; Bilh and Bauer 2017). Another option would be to apply this methodology to presidential addresses given at other social science conferences with institutions such as the American Psychological Association and the American Sociological Association.

The overfitting to specific addresses was also caused by the flexibility of the topic prevalence estimator which had 10 degrees of freedom. This estimator was originally chosen based on the recommended use for continuous variables in the *stm* documentation (Roberts et al. 2019) as well as the default values for the spline function. It would therefore be valuable to perform this analysis using a purely linear estimator or a less flexible spline.[12] Though a limitation of this study, the unsuitability of this originally chosen hyperparameter emphasizes the hazards of always adhering to the default method.

Considering the original research question of the effect of text chunking on topic content and prevalence, the results of this study are mixed: The methodology generally focused on finding similarities between the models and aligning topics, though the presence of similarity does not equate the absence of difference. There was limited

[12] It's important to note that the nonlinearity of the estimator did show interesting patterns such as the peaks in discussions of the labor market corresponding with recessions in 1980 and 2008 (see Figure 7.1).

evidence of the *Bounded Paragraph* models producing significantly different results from the fixed-size chunks, suggesting that it is not always necessary to put in the additional work of paragraph splitting. However, they did produce more coherent topics with higher levels of K, so further inquiry is needed. Additionally, as expected, the *Full Document* models did not capture as much nuance in the corpus structure and had a higher degree of uncertainty on the metadata effects due to the smaller sample size. Further research could investigate the limits of chunk size to get meaningful results, as well as the extent to which variability in document length across the corpus can affect this.

It is also unclear whether certain text unit sizes are more suited to answer different empirical questions, though perhaps that can be determined on a case-by-case basis: Assuming computational resources and time are not a problem, a researcher could automatically split the longer documents into chunks of 200 or 500 or even 1000 words and determine which results best suit the research question. In essence, this is no different from the process used to select K. Still, the prominence of qualitative interpretation in this process may cause some to question the reliability of topic modeling in social science research—something that is supposed to be an advantage of computational methods (Grimmer and Stewart 2013; Nelson 2020). This invites future research opportunities on reliability of hyperparameter selection; for example, the extent to which researchers will choose the same values of optimal Ks given exclusivity and semantic coherence graphs.

All in all, this investigation of the effect of a single preprocessing decision ultimately discussed the range of choices a researcher must make when using topic

modeling for social science research. It was found that these decisions rarely have a clearly optimal answer and more so depends on the research question. As part of a larger conversation about reliability and validity of topic modeling and other computational methods in the social sciences, this study above all demonstrates thoughtful and synergistic integration of qualitative and quantitative analyses. This paper introduces novel visualization strategies that can aid this process and discusses assumptions, considerations, and implications of possible results. In the larger field of data science, structured investigations of computational methods such as this can not only support but also challenge established conventions of research, encouraging more deliberate and reflective analytical practices across a wide range of domains.

Appendix A - Discussion of Preprocessing Decisions

A.1 Stemming Versus Lemmatization

This section will provide additional justification for some of the standardized preprocessing decisions. The first one being discussed is why lemmatization rather than stemming was used to simplify the vocabulary. This is best illustrated in Table A1 which illustrates how many different tokens may be combined into a single stem: *“Gener”* is not only short for “general,” and “generally,” but also “generate” and “generators,” which have completely different meanings. Lemmatization is better able to capture these nuances by separating “general” from “generate” and “generic” while still providing simplification to reduce noise. However, common homonyms such as “general” (both an adjective and a rank) still can obscure meaning in STMs. Similarly, the stem “*commun”* abstracts the distinct and salient words “community,” “communication,” and “communism."

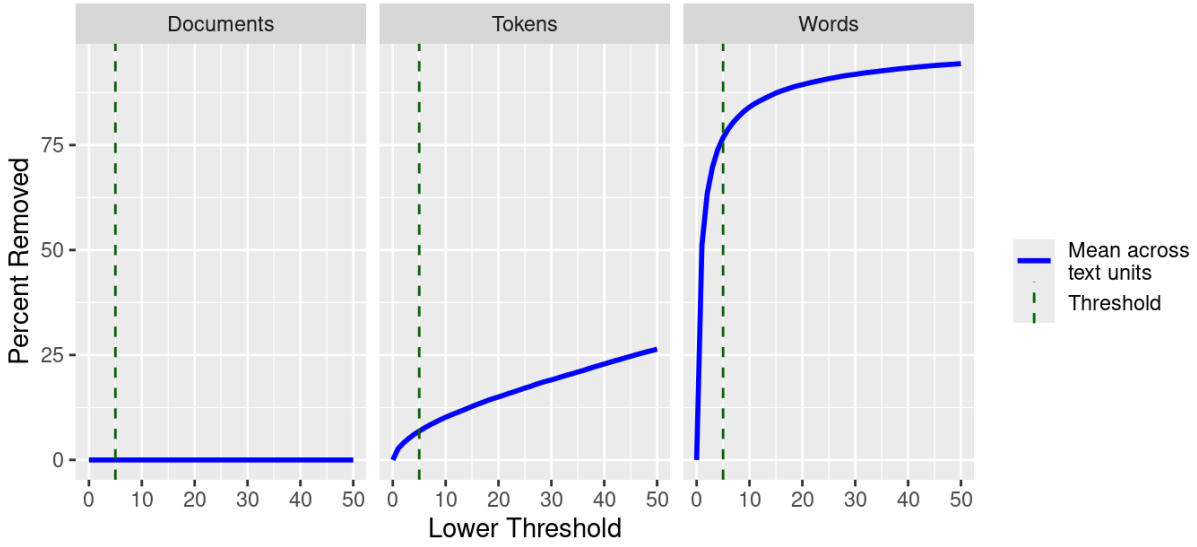
Table A1: Breakdown of Stems vs. Lemmas of "*gener*" and "*commun*"

stem	lemma	token
gener	generate	generated
		generates
		generating
	general	general
		generals
	generalize	generalized
		generalizes
	generalization	generalizations
	generally	generally
	generation	generations
	generator	generators
	generic	generics
commun	communicate	communicated
		communicates
		communicating
	communication	communication
		communications
	community	communities
		community
	commune	communes
	communism	communism

A.2 Vocabulary Lower Threshold

The next decision being discussed is the lower threshold for token frequency to be included in the STM vocabulary. The *stm* package's *plotRemoved* function calculates how many vocabulary items, overall tokens, and documents will be excluded from the corpus for a range of values (Roberts et al. 2019). It was applied to each document feature matrix and the returned values were normalized into a percentage and averaged across the text chunking units then plotted in Figure A1. The vertical green line represents the chosen threshold of 5. Less than 10% of the tokens remaining after removing stop words in the corpus were removed, but those made up around 75% of the dimensions in the document-feature matrix. As a result, their removal greatly increased the efficiency of the modeling with very little meaning lost.

Figure A1: Number of Tokens and Vocabulary Words Removed for Different Frequency Lower Thresholds, Averaged across Text Chunking Strategies



"Lower threshold" refers to the minimum number of documents a token must be in to be included in the vocabulary. Different document splitting strategies therefore result in slightly varied vocabularies.

Appendix B - Prompt for LLM Topic Labeling

Open AI's ChatGPT 4o model was used to create informal labels for some of the persistent topics across models. The prompt was engineered over several iterations with a new chat each time to avoid confounding from prior topic labels. The LLM was also used to develop the prompt itself after reaching adequate results and wanting them to be reproducible in a fresh chat. This process also surfaced some assumptions the model was making from the initial prompt. The final prompt is as follows:

I will provide groups of related topic word sets, separated by line breaks.

Each line will be formatted like: Topic X: word1, word2, ..., word10

Each group of related topics will be visually grouped together (e.g., 2–3 lines at a time, separated by line breaks), and with empty values being a "-."

Your task is to:
Treat each group as already related — do not try to regroup or merge topics.
For each group, generate:
A single unique 1–3 word informal label that captures the theme across all topic sets in that group.
A representative sample of 4-6 top words pulled from across the topics in the group.
The topic numbers as a comma-separated list with a prefix to each topic number that is either "Doc", "Page", "ParaB", or "Para," depending on its position in the list.

For example, the named topic numbers for the following input:
Topic 1: retirement, tax, insurance, benefit, security, annuity, program, save, fund, lifetime Topic 5: tax, retirement, benefit, security, annuity, save, insurance, parent, child, lifetime Topic 10: tax, insurance, retirement, benefit, security, fund, save, revenue, saving, income Topic 17: tax, insurance, retirement, benefit, security, revenue, annuity, program, fund, taxation
would be: Doc-1, Page-5, ParaB-10, Para-17

and

- Topic 13: income, secular, inequality, underdeveloped, sector, share, distribution, capita, profit, proportion Topic 1: inequality, capita, top, distribution, share, sector, tfp, income, secular, underdeveloped -
would have the named topics: Page-13, ParaB-17

Do not consider any topic groups other than the ones given.

Structure the output in to be pastable into a with the column labels "Topic Sets", "Example words," and "Informal Label."

{Topic sets}

Only 3 topic sets were passed at first, and when the model returned the proper structure, most of the rest followed. However, when asked to label a few more, the original context was lost and the model did not return the proper structure, even when prompted to “structure it like the others” and “put it in a table.” So, a new chat was opened with the same prompt. The first output separated the topic names into their own columns, so it was prompted with:

Can you separate it into a markdown so the columns have "Doc-9, Page-8, ParaB-6, Para-11" "peace, liberty, nation, democracy, property" and "Political Ideals", for example

The topics marked in this second chat are labeled with a § in Table 6.

Appendix C - Glossary of Abbreviations in Economics

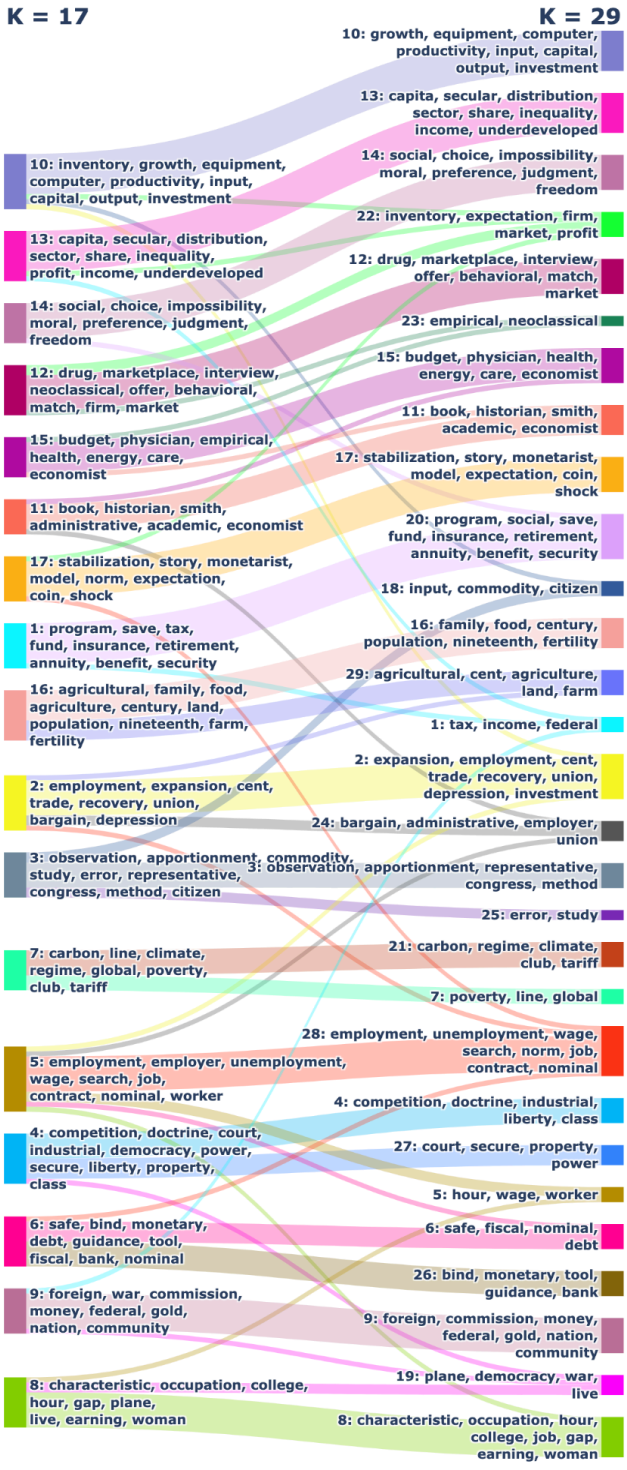
Below is a selection of economic jargon that were in the top 10 words of various models. The majority are defined using "The A to Z of Economics" by *The Economist* (n.d.), though some relate to specific organizations, definitions of which come from their websites. These are meant to provide an overview in common English and therefore do not convey the depth of meaning that a seasoned economist would be able to gather.

Table C1: Economics Abbreviations

Acronym	Meaning	Source
GDP	Gross Domestic Product	"The main measure of an economy's size. GDP is calculated from the market value of all the finished goods and services within a country's borders over a set period of time" (*The Economist*, n.d.).
ICP	International Comparison Program	"A worldwide statistical initiative to collect comparative price data and detailed GDP expenditures to produce purchasing power parities (PPPs) for the world's economies" (*World Bank, n.d.).*
PPP	Purchasing-power Parity	"A method of adjusting exchange rates to take account of the different levels of prices in different countries…. [Economists] calculate PPP exchange rates as a way of assessing whether currencies are under- or over-valued" (*The Economist*, n.d.)
QE	Quantitative Easing	"A policy introduced to alleviate the effects of the 2007-09 financial crisis. Central banks slashed interest rates but the effect of such cuts seemed to diminish as they approached the zero lower bound. Quantitative easing (QE) involved banks buying government bonds… in the secondary market and creating new money to pay the sellers …. This had the double effect of injecting liquidity into the economy and pushing down bond yields, cutting the cost of borrowing for the corporate sector" (*The Economist*, n.d.).
TFP	Total Factor Productivity	"Output relative to inputs, measured … by dividing an index of output by a combined index of labour and capita…. comes from greater efficiency or the adoption of new technology" (*The Economist n.d.).*

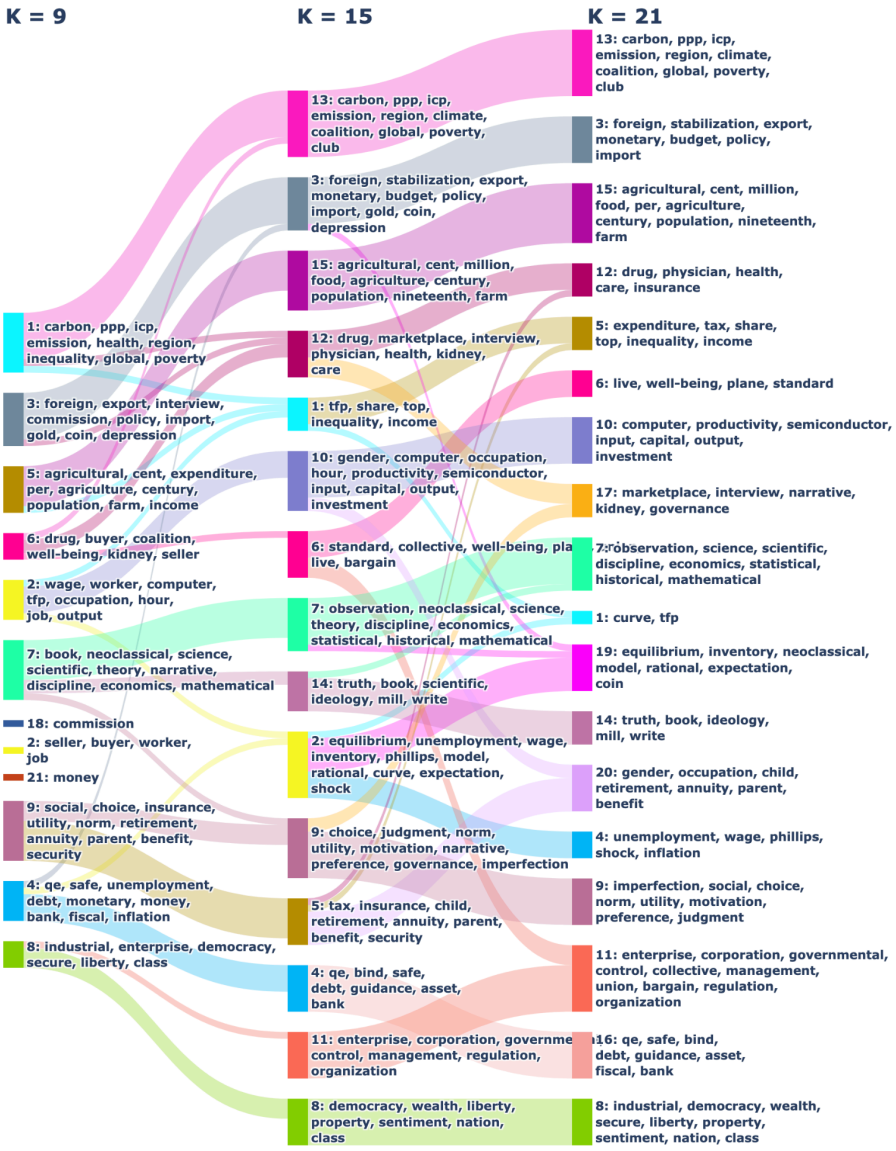
Appendix D - Full Sankey Diagrams

Figure D1: Topic Splitting Sankey Diagram with Full Address Models Going from 17 to 29 Topics



Note that a model with 36 topics was also made, as suggested by the semantic coherence and exclusivity graphs, however Sankey diagrams with more than 25 topics were extremely hard to read and therefore not included here.

Figure D2: Topic Splitting Sankey Diagram with *Page* Models Going from 9 to 15 to 21 Topics



There are a few extra nodes on the right hand side that came from the $K = 21$ model with no links coming to or from them despite the label. Time did not permit investigating and fixing this bug in the Sankey-making code.

Figure D3: Topic Splitting Sankey Diagram with *Bounded Paragraph* Models Going from 15 to 20 to 27 Topics

As shown in some other Sankey diagrams, Topic 20 of the $K = 20$ model (light purple with a dark green link) does not have any incoming nodes and therefore was pushed all the way to the left.

Appendix E - Splitting/Merging of Topics with Non-Optimal Number of Topics

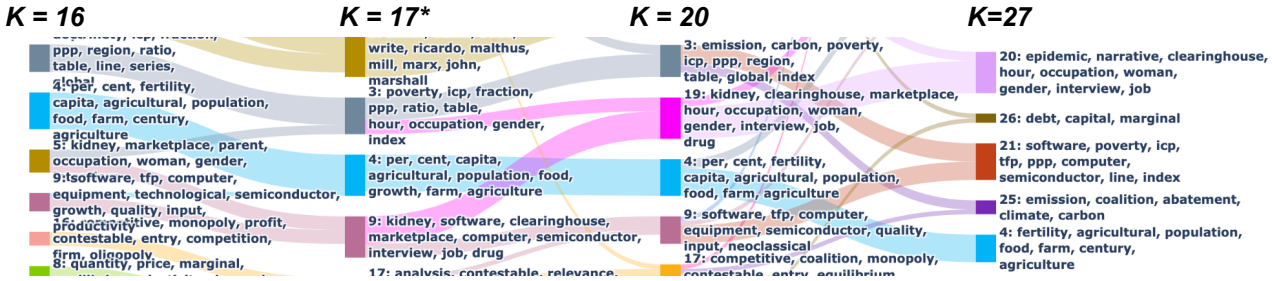
Table E1 shows the topic splitting relationships as K is increased from 16 to 20 to 27 in the *Paragraph Bounded* model. While there is merging into Topic 21 of the 27-topic model that is less semantically coherent, the relationships between topics are well-delineated. However, changing the number of topics from 16 to 17 severely disrupts this alignment. As shown in Table E2, Topic 9 of the 17-topic model splits into both Topic 9 and Topic 19 of the 20-topic one, and do not merge back together again. Indeed, it is impossible to structure the table in a way that does not split up one of the topics given the staggered relationship. This is also shown in Figure E1, a cropped Sankey diagram.

Table E1: Aligned Topic Splitting Relationship With Optimal K for *ParaB*

K=16	K=20	K=27
Topic 9: input, semiconductor, tfp, software, computer, productivity, equipment, growth, quality, technological	Topic 9: input, semiconductor, computer, software, tfp, equipment, neoclassical, contribution, quality, measurement	Topic 21: ppp, index, icp, semiconductor, poverty, computer, software, tfp, line, revision
Topic 3: table, poverty, ppp, icp, global, line, region, series, ratio, fraction	Topic 3: table, global, ppp, carbon, poverty, emission, figure, icp, index, region	Topic 25: carbon, coalition, emission, climate, abatement, club, game, tariff, ton, player
Topic 5: marketplace, occupation, child, gender, woman, parent, school, college, interview, kidney	Topic 19: marketplace, job, occupation, kidney, gender, drug, hour, interview, woman, clearinghouse	Topic 20: narrative, gender, hour, job, interview, woman, epidemic, occupation, clearinghouse, female
		Topic 19: health, kidney, drug, governance, option, physician, care, medicare, transplant, risk

Table E2: Staggered Topic Splitting Relationship With Non-Optimal K for *ParaB*		
K=17	K=20	K=27
Topic 9: marketplace, computer, semiconductor, software, kidney, job, drug, interview, clearinghouse, externality †	Topic 9: input, semiconductor, computer, software, tfp, equipment, neoclassical, contribution, quality, measurement	Topic 21: ppp, index, icp, semiconductor, poverty, computer, software, tfp, line, revision
Topic 3: table, poverty, ppp, icp, occupation, fraction, index, ratio, hour, gender	Topic 3: table, global, ppp, carbon, poverty, emission, figure, icp, index, region	Topic 25: carbon, coalition, emission, climate, abatement, club, game, tariff, ton, player
Topic 9: marketplace, computer, semiconductor, software, kidney, job, drug, interview, clearinghouse, externality †	Topic 19: marketplace, job, occupation, kidney, gender, drug, hour, interview, woman, clearinghouse	Topic 20: narrative, gender, hour, job, interview, woman, epidemic, occupation, clearinghouse, female
		Topic 19: health, kidney, drug, governance, option, physician, care, medicare, transplant, risk
† There is no way to align the cells in this table so that one of them is not split like this, demonstrating how disrupted the alignment is, especially compared to the 16-topic model.		

Figure E1: Sankey Diagram of Staggered Splitting with Non-Optimal K Value



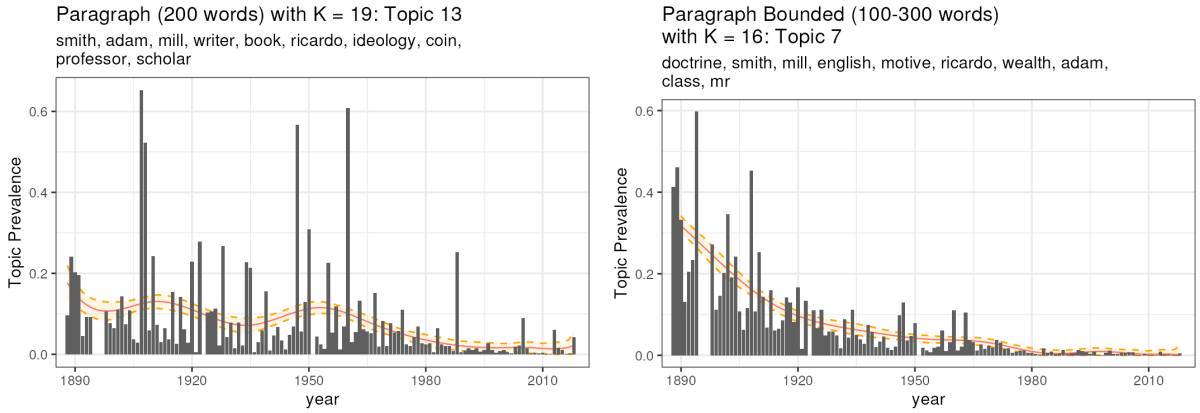
** Value not selected as a potential K*

Note how several topics in the $K = 16$ model split and merge into topics 3 and 9 of the 17 topic one, only to split back again in a staggered fashion.

Appendix F - Additional Insights from Topic Prevalence Graphs

Topic 7 of the *Bounded Paragraph* model and Topic 13 of the *Paragraph* model both refer to classical thinkers of the Economics discipline such as “smith,” “mill,” and “ricardo” (Jamieson 2023). As shown in Figure F1, the *Paragraph* topic has slight differences in topic content and prevalence with more of a bimodal shape as well as additional top words that refer to economic thought more broadly such as “professor,” “scholar,” and “book.” So, if a researcher were studying prominent economic thinkers specifically, they might consider a model with a greater number of topics to be able to better delineate themes of interest.[13]

Figure F1: Prevalence of “Classical Thinker” topics for Paragraph and Bounded Paragraph models



The topics illustrated here make explicit references to the classical economists Adam Smith, David Ricardo, and John Stewart Mill, yet with some differences in content that affect the overall shape of the distributions.

[13] Notably, the 20-topic model with Bounded Paragraphs has a topic that includes even more “big names” in economics: Mill, Ricardo, Smith, Malthus, Keynes, and Marx

It’s also important to note that a few of the top words are not as meaningful: Firstly, not much can be inferred from “mr” besides perhaps some language patterns of the time or references to key figures, though the latter is already conveyed by the rest of the topic content. Additionally, it is redundant to have “adam” as well as “smith” in the top words, as the meaning of “smith” (as a person, not a profession) can be inferred in the context of the other labelling words, or by looking at the addresses. This may prompt a researcher to adjust the vocabulary of the text by combining instances of “Smith” and “Adam Smith” to be one token or adding “mr” to the stop word list. However, according to Schofield et al. (2017), the iterative task of creating a custom stop word list is often time consuming and does not produce significantly better results. When labeling topics, a researcher can instead just replace less helpful tokens with the next highest ranked ones.

Bibliography

"The A to Z of Economics" n.d. *The Economist*. Retrieved April 7, 2025 (https://economist.com/economics-a-to-z).

"About the AEA Annual Meeting" n.d. *American Economic Association*. Retrieved April 15, 2025 (https://www.aeaweb.org/conference/about).

Alvero, Aj, Sonia Giebel, Ben Gebre-Medhin, Anthony Lising Antonio, Mitchell L. Stevens, and Benjamin W. Domingue. 2021. "Essay Content and Style Are Strongly Related to Household Income and SAT Scores: Evidence from 60,000 Undergraduate Applications." *Science Advances* 7(42):eabi9031. doi: 10.1126/sciadv.abi9031.

Anupriya, P., and S. Karpagavalli. 2015. "LDA Based Topic Modeling of Journal Abstracts." Pp. 1–5 in *2015 International Conference on Advanced Computing and Communication Systems*. Coimbatore, India: IEEE.

Arora, Sanjeev, Rong Ge, Yonatan Halpern, David Mimno, Ankur Moitra, David Sontag, Yichen Wu, and Michael Zhu. 2013. "A Practical Algorithm for Topic Modeling with Provable Guarantees." Pp. 280–88 in *Proceedings of the 30th International Conference on Machine Learning*. Vol. 28, *Proceedings of Machine Learning Research*, edited by S. Dasgupta and D. McAllester. Atlanta, Georgia, USA: PMLR.

Backhouse, Roger E., and Béatrice Cherrier. 2017. "The Age of the Applied Economist: The Transformation of Economics since the 1970s." *History of Political Economy* 49(Supplement):1–33. doi: 10.1215/00182702-4166239.

Ballester, Omar, and Orion Penner. 2022. "Robustness, Replicability and Scalability in Topic Modelling." *Journal of Informetrics* 16(1):101224. doi: 10.1016/j.joi.2021.101224.

Benoit, Kenneth, Kohei Watanabe, Haiyan Wang, Paul Nulty, Adam Obeng, Stefan Müller, and Akitaka Matsuo. 2018. "Quanteda: An R Package for the

Quantitative Analysis of Textual Data." *Journal of Open Source Software* 3(30):774. doi: 10.21105/joss.00774.

Bernstein, Michael A. 2008. "A Brief History of the American Economic Association." *The American Journal of Economics and Sociology* 67(5):1007–23. doi: 10.1111/j.1536-7150.2008.00608.x.

Biernacki, Richard. 2012. *Reinventing Evidence in Social Inquiry*. New York: Palgrave Macmillan US.

Bihl, Trevor J., and Kenneth W. Bauer Jr. 2017. "Statistical Analysis of High-Level Features from State of the Union Addresses:" *International Journal of Information Systems and Social Change* 8(2):50–73. doi: 10.4018/IJISSC.2017040103.

Bischof, Jonathan M., and Edoardo M. Airoldi. 2012. "Summarizing Topical Content with Word Frequency and Exclusivity." in *Proceedings of the 29th International Conference on Machine Learning*. Edinburgh, UK.

Blei, David M., and John D. Lafferty. 2006. "Dynamic Topic Models." Pp. 113–20 in *Proceedings of the 23rd international conference on Machine learning - ICML '06*. Pittsburgh, Pennsylvania: ACM Press.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3(Jan):993–1022.

Bystrov, Victor, Viktoriia Naboka-Krell, Anna Staszewska-Bystrova, and Peter Winker. 2023. "Analysing the Impact of Removing Infrequent Words on Topic Quality in LDA Models." 10.48550/arXiv.2311.14505.

Carter, Bryan. 2013. *Digital Humanities*. Bingley, United Kingdom: Emerald Publishing Limited.

Chandelier, Marie, Agnès Steuckardt, Raphaël Mathevet, Sascha Diwersy, and Olivier Gimenez. 2018. "Content Analysis of Newspaper Coverage of Wolf

Recolonization in France Using Structural Topic Modeling." *Biological Conservation* 220:254–61. doi: 10.1016/j.biocon.2018.01.029.

Chang, Jonathan, Jordan Boyd-Graber, Sean Gerrish, Chong Wang, and David M. Blei. 2009. "Reading Tea Leaves: How Humans Interpret Topic Models." Pp. 288–96 in *Proceedings of the 22nd International Conference on Neural Information Processing Systems*, *NIPS'09*. Red Hook, NY, USA: Curran Associates Inc.

Derber, Charles. 2016. "The Neoclassical Paradigm." Pp. 31–38 in *Capitalism: Should You Buy It? An Invitation to Political Economy*. London New York: Routledge.

Dieng, Adji B., Francisco J. R. Ruiz, and David M. Blei. 2020. "Topic Modeling in Embedding Spaces" edited by M. Johnson, B. Roark, and A. Nenkova. *Transactions of the Association for Computational Linguistics* 8:439–53. doi:10.1162/tacl_a_00325.

DiMaggio, Paul, Manish Nag, and David Blei. 2013. "Exploiting Affinities between Topic Modeling and the Sociological Perspective on Culture: Application to Newspaper Coverage of U.S. Government Arts Funding." *Poetics* 41(6):570–606. doi: 10.1016/j.poetic.2013.08.004.

Do, Salomé, Étienne Ollion, and Rubing Shen. 2024. "The Augmented Social Scientist: Using Sequential Transfer Learning to Annotate Millions of Texts with Human-Level Accuracy." *Sociological Methods & Research* 53(3):1167–1200. doi: 10.1177/00491241221134526.

Erikson, Emily, Keniel Yao, and Daniel Karell. 2023. "Salvation into Nation: Topic Modeling Early Modern Economic Writings." *Œconomia. History, Methodology, Philosophy* (13–2):357–92. doi: 10.4000/oeconomia.15688.

Gao, Yuan, Xuechun Wang, and Xu Liu. 2024. "Mapping Higher Education Internationalisation as a Research Space via Natural Language Processing

(NLP) Techniques.” *Journal of Studies in International Education* 28(5):687–710. doi: 10.1177/10283153241251924.

Gao, Zekai J., Yangqiu Song, Shixia Liu, Haixun Wang, Hao Wei, Yang Chen, and Weiwei Cui. 2011. “Tracking and Connecting Topics via Incremental Hierarchical Dirichlet Processes.” Pp. 1056–61 in *2011 IEEE 11th International Conference on Data Mining*. Vancouver, BC, Canada: IEEE.

Gerlach, Martin, Tiago P. Peixoto, and Eduardo G. Altmann. 2018. “A Network Approach to Topic Models.” *Science Advances* 4(7):eaaq1360. doi: 10.1126/sciadv.aaq1360.

Grimmer, Justin, Margaret E. Roberts, and Brandon Stewart. 2022. *Text as Data: A New Framework for Machine Learning and the Social Sciences*. Princeton University Press.

Grimmer, Justin, and Brandon M. Stewart. 2013. “Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts.” *Political Analysis* 21(3):267–97. doi: 10.1093/pan/mps028.

Guo, Chonghui, Menglin Lu, and Wei Wei. 2021. “An Improved LDA Topic Modeling Method Based on Partition for Medium and Long Texts.” *Annals of Data Science* 8(2):331–44. doi: 10.1007/s40745-019-00218-3.

Hofmann, Thomas. 1999. “Probabilistic Latent Semantic Indexing.” Pp. 50–57 in *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Berkeley California USA: ACM.

“International Comparison Program (ICP)” n.d. *World Bank*. Retrieved April 8, 2025 (https://www.worldbank.org/en/programs/icp).

Jamieson, Duncan R. 2023.“Ricardo’s Seven Key Economic Principles | EBSCO Research Starters.” https://www.ebsco.com/research-starters/history/ricardos-seven-key-economic-principles.

Jo, Wonkwang. 2019. “Possibility of Discourse Analysis Using Topic Modeling.” *Journal of Asian Sociology* 48(3):321–42.

Jockers, Matthew, and David Mimno. 2013. “Significant Themes in 19th-Century Literature.” *Poetics* 41:750–69. doi: 10.1016/j.poetic.2013.08.005.

Kalleberg, Arne L., and Till M. Von Wachter. 2017. “The U.S. Labor Market During and After the Great Recession: Continuities and Transformations.” *The Russell Sage Foundation Journal of the Social Sciences : RSF* 3(3):1–19. doi: 10.7758/rsf.2017.3.3.01.

Kennedy, Brendan, Ashwini Ashokkumar, Ryan L. Boyd, and Morteza Dehghani. 2022. “Text Analysis for Psychology: Methods, Principles, and Practices.” Pp. 3–62 in *Handbook of Language Analysis in Psychology*. New York, NY, US: The Guilford Press.

Kuhn, H. W. 1955. “The Hungarian Method for the Assignment Problem.” *Naval Research Logistics Quarterly* 2(1–2):83–97. doi: 10.1002/nav.3800020109.

Lee, Monica, and John Levi Martin. 2015. “Coding, Counting and Cultural Cartography.” *American Journal of Cultural Sociology* 3(1):1–33. doi: 10.1057/ajcs.2014.13.

Li, Dai, Bolun Zhang, and Yimang Zhou. 2023. “Can Large Language Models (LLM) Label Topics from a Topic Model?.” Retrieved April 11, 2025 (osf.io/preprints/socarxiv/23x4m_v1).

Lindstedt, Nathan C. 2019. “Structural Topic Modeling For Social Scientists: A Brief Case Study with Social Movement Studies Literature, 2005–2017.” *Social Currents* 6(4):307–18. doi: 10.1177/2329496519846505.

Lüdering, Jochen, and Peter Winker. 2016. “Forward or Backward Looking? The Economic Discourse and the Observed Reality.” *Jahrbücher Für Nationalökonomie Und Statistik* 236(4):483–515. doi: 10.1515/jbnst-2015-1026.

Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval.* Cambridge University Press.

Margherita, Emanuele Gabriel, Stefania Denise Escobar, Giovanni Esposito, and Nathalie Crutzen. 2023. “Exploring the Potential Impact of Smart Urban Technologies on Urban Sustainability Using Structural Topic Modelling: Evidence from Belgium.” *Cities* 141:104475. doi: 10.1016/j.cities.2023.104475.

Michalke, Meik. 2021. “koRpus: Text Analysis with Emphasis on POS Tagging, Readability, and Lexical Diversity.” Version 0.13-8.

Mimno, David, Hanna M. Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. “Optimizing Semantic Coherence in Topic Models.” Pp. 262–72 in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, *EMNLP ’11*. USA: Association for Computational Linguistics.

Nassehi, Armin. 2024. “The Reference Problem of Digitalization.” Pp. 15–43 in *Patterns: Theory of the Digital Society*. Cambridge Hoboken, NJ: Polity.

Nelson, Laura K. 2020. “Computational Grounded Theory: A Methodological Framework.” *Sociological Methods & Research* 49(1):3–42. doi: 10.1177/0049124117729703.

Nelson, Laura K. 2021. “Cycles of Conflict, a Century of Continuity: The Impact of Persistent Place-Based Political Logics on Social Movement Strategy.” *American Journal of Sociology* 127(1):1–59. doi: 10.1086/714915.

Nour, Matthew M. 2024. “Adequate Methodological Reporting and Sensitivity Analyses Are Essential to Allow Reproducibility and Interpretability in NLP Studies.” *Journal of Affective Disorders* 356:436–37. doi: 10.1016/j.jad.2024.04.052.

Pal, Shounak, Baidyanath Biswas, Rohit Gupta, Ajay Kumar, and Shivam Gupta. 2023. “Exploring the Factors That Affect User Experience in Mobile-Health

Applications: A Text-Mining and Machine-Learning Approach." *Journal of Business Research* 156:113484. doi: 10.1016/j.jbusres.2022.113484.

Popa, Mircea. 2025. "Modelling Policy Action Using Natural Language Processing: Evidence for a Long-Run Increase in Policy Activism in the UK." *Journal of Computational Social Science* 8(2):47. doi: 10.1007/s42001-024-00353-9.

Pudasaini, Shushanta, Luis Miralles-Pechuán, David Lillis, and Marisa Llorens Salvador. 2024. "Survey on AI-Generated Plagiarism Detection: The Impact of Large Language Models on Academic Integrity." *Journal of Academic Ethics*. doi: 10.1007/s10805-024-09576-x.

Qader, Wisam A., Musa M. Ameen, and Bilal I. Ahmed. 2019. "An Overview of Bag of Words;Importance, Implementation, Applications, and Challenges." Pp. 200–204 in *2019 International Engineering Conference (IEC)*. Erbil, Iraq: IEEE.

Roberts, Margaret E., Brandon M. Stewart, and Dustin Tingley. 2019. "Stm: An *R* Package for Structural Topic Models." *Journal of Statistical Software* 91(2). doi: 10.18637/jss.v091.i02.

Roberts, Margaret E., Brandon M. Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G. Rand. 2014. "Structural Topic Models for Open-Ended Survey Responses." *American Journal of Political Science* 58(4):1064–82. doi: 10.1111/ajps.12103.

Roberts, Margaret E., Dustin Tingley, Brandon M. Stewart, and Edoardo M. Airoldi. 2013. "The Structural Topic Model and Applied Social Science." in *Advances in Neural Information Processing Systems Workshop on Topic Models: Computation, Application, and Evaluation*. Lake Tahoe, Nevada: NIPS.

Roberts, Margaret, Brandon Stewart, and Dustin Tingley. 2023. “Stm: Estimation of the Structural Topic Model.” Version 1.3.7.

Rule, Alix, Jean-Philippe Cointet, and Peter S. Bearman. 2015. “Lexical Shifts, Substantive Changes, and Continuity in State of the Union Discourse, 1790–2014.” *Proceedings of the National Academy of Sciences* 112(35):10837–44. doi:10.1073/pnas.1512221112.

Sarkar, Dipanjan. 2016. *Text Analytics with Python*. Berkeley, CA: Apress.

Savoy, Jacques. 2020. *Machine Learning Methods for Stylometry: Authorship Attribution and Author Profiling*. Cham: Springer International Publishing.

Sbalchiero, Stefano, and Maciej Eder. 2020. “Topic Modeling, Long Texts and the Best Number of Topics. Some Problems and Solutions.” *Quality & Quantity* 54(4):1095–1108. doi: 10.1007/s11135-020-00976-w.

Schmid, Helmut. 1997. “Probabilistic Part-of-Speech Tagging Using Decision Trees.” in *New Methods In Language Processing*. Routledge.

Schofield, Alexandra, Måns Magnusson, and David Mimno. 2017. “Pulling Out the Stops: Rethinking Stopword Removal for Topic Models.” Pp. 432–36 in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, edited by M. Lapata, P. Blunsom, and A. Koller. Valencia, Spain: Association for Computational Linguistics.

Schofield, Alexandra, and David Mimno. 2016. “Comparing Apples to Apple: The Effects of Stemmers on Topic Models.” *Transactions of the Association for Computational Linguistics* 4:287–300. doi: 10.1162/tacl_a_00099.

Tausczik, Yla R., and James W. Pennebaker. 2010. “The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods.” *Journal of Language and Social Psychology* 29(1):24–54. doi: 10.1177/0261927X09351676.

Tusher, E. H., M. A. Ismail, M. A. Rahman, A. H. Alenezi, and M. Uddin. 2024. “Email Spam: A Comprehensive Review of Optimize Detection Methods, Challenges, and Open Research Problems.” *IEEE Access* 12:143627–57. doi: 10.1109/ACCESS.2024.3467996.

Ulstein, Julie. 2024. “Structural Topic Modeling as a Mixed Methods Research Design: A Study on Employer Size and Labor Market Outcomes for Vulnerable Groups.” *Quality & Quantity* 58(5):4331–51. doi: 10.1007/s11135-024-01857-2.

Wachen, John. 2018. “Media Coverage of Educational Testing: Understanding Issue Dimensions Using Topic Modeling.” The University of North Carolina at Chapel Hill University Libraries.

Wang, Xuerui, Andrew McCallum, and Xing Wei. 2007. “Topical N-Grams: Phrase and Topic Discovery, with an Application to Information Retrieval.” Pp. 697–702 in *Seventh IEEE International Conference on Data Mining (ICDM 2007)*. Omaha, NE, USA: IEEE.

Weston, Sara J., Ian Shryock, Ryan Light, and Phillip A. Fisher. 2023. “Selecting the Number and Labels of Topics in Topic Modeling: A Tutorial.” *Advances in Methods and Practices in Psychological Science* 6(2):251524592311601. doi: 10.1177/25152459231160105.