What are you talking about Mr. President?

Topic Modeling for the Economic Reports of the President

George L. Sorg-Langhans
Princeton University

Maximilian J. Vogler
Princeton University

September 10, 2020

Abstract

Despite their power, modern natural language processing techniques have not found widespread use in the economic literature. In this paper we demonstrate their potential in the context of a specific economic application, namely the analysis of the Economic Reports of the President. Specifically, we use both Non-Negative Matrix Factorization and Latent Dirichlet Allocation to extract and study the main topics of each presidential report. Whilst both approaches broadly agree on the topics, the NMF proves more versatile. Overall, the topics we identify are well-defined and display remarkable time series patterns, documenting both long-run economic trends and highlighting specific policy events. Based on these findings, the Economic Reports in combination with natural language processing techniques thus present a fertile ground for future research.

1 Introduction

Over the past two decades natural language processing (NLP), a sub-discipline of artificial intelligence concerned with the automated analysis of human language, has made sweeping advances. Examples of recent breakthroughs include the automated categorization of millions of scientific articles, the categorization of emails in user accounts or the detection of structured genetic variation in human DNA.[1] In contrast to this, economic research analyzing large textual bodies still overwhelmingly relies on the following two basic approaches. First, the researcher manually reads through all of the documents involved, a process both highly subjective and not easily scalable. Due to these shortcomings, a common alternative, the use of dictionary-based methods is common, in which the researcher counts the frequency of certain pre-selected words. As we will illustrate below, this simple approach is limited in its ability to exploit the underlying structure of both the documents itself and the English language so that its performance falls short of more modern approaches to textual analysis.

In this paper we apply natural language processing algorithms, which we deem highly relevant and interesting for economics, to the analysis of the Economic Reports of the President (ERP) - one of the most prominent economic texts. Specifically, we use two state-of-the-art approaches to extract the main topics of each report. This allows us to construct time series of the most relevant topics discussed in the reports to find interesting trends and pattern. Overall, we therefore intent to contribute to the nascent literature on natural language processing in economics in two ways. First, we demonstrate the success of these modern natural language processing approaches in an economic setting. We therefore provide a basic roadmap that future researchers can follow when studying economic text corpora. Second, we show that the Economic Report of the President is a first-order economic data source, that can help with the analysis of both overall trends and specific economic events.

In order to extract meaningful topics from the ERPs we rely on two topic extraction algorithms and contrast their performance. Both of these algorithms use the underlying cross-correlation structure of related words in the texts to form related word groups. In particular, they extract word groups that are often co-occurring and therefore hint at the

[1] For a detailed discussion of recent successes and applications see Boyd-Graber et al. (2017).

underlying text structure. These groupings form the basis for our topics. We then follow the common approach in the NLP literature of labeling these word groups with succinct topic labels based on the most frequent words in each topic.

The two approaches that we use and contrast in this paper are the Latent Dirichlet Allocation (LDA), proposed by Blei et al. (2003) and discussed in detail in Blei (2012) and Non-negative Matrix Factorization (NMF) as proposed by Lee & Seung (1999) and Lee & Seung (2001). Hansen & McMahon (2016) and Hansen et al. (2018) use the LDA to study the notes of the Federal Open Market Committee (FOMC). They are therefore the only papers we are aware of that use the same general NLP approach in economics as this paper. In contrast to them, we introduce the NMF as an additional NLP tool and study a completely different application.

Interestingly, we find that while the LDA gives us results that align broadly with those of the NMF, the NMF outperforms the LDA on one important dimension. Whilst both approaches yield word groups that are sensible and remarkably similar, the NMF provides us with further important topics that the LDA does not discover. For this reason we agree with the findings of O'Callaghan et al. (2015) who demonstrate that on medium-sized text corpora such as the ERP the NMF outperforms the LDA. Our recommendation for future research in economics is therefore to primarily rely on the NMF. The LDA is still of great import as a robustness device, which we will also demonstrate below.

Regarding the Economic Reports of the President it is not clear ex ante that they represent a valuable data source for economists as they might be overly tainted with party ideology. To investigate this issue, we embark on an in-depth study of the time-series behavior of the main policy topics revealed by our algorithms. Through careful analysis we show that the topics in the economic reports closely mimic the main policy topics of the time. In particular, if party ideology were to dominate the reports, one would expect a degenerate time series of the topic, with republican topics being uniformly important during their presidencies and equally unimportant under democratic administrations and vice versa. Noticeably, this is not what we find. Instead the time series exhibit interesting behaviors that encompass both cross-party time trends and specific well-documented administration priorities. In addition, we document spikes in topic frequency around specific policy events such as topic-related legislation. Overall, we thus conclude that we can use the Economic Reports and our extracted topics to analyze both overall economic trends and individual policy events.

We start our analysis by discussing our primary data source - the Economic Reports of the President in section 2. Next, in section 3 we introduce the two NLP methods in detail, discussing first the Latent Dirichlet Allocation and next the Non-Negative Matrix Factorization. Section 4 presents the main results of these algorithms. First, we discuss the topics identified and next we analyze the time series of these topics. Finally section 5 concludes.

2 Data

In order to investigate the topics of the Economic Reports of the President (ERP) we construct a data set of the annual reports starting in 1960.[2] This section describes the details of the data extraction and cleaning necessary to prepare our corpus of documents for the different topic extraction algorithms.

The economic reports from 1960 to 2020 are available from the Fraser Federal Reserve System.[3]. For each report, we perform the following data cleaning steps. First, we exclude all irrelevant sections from the document, such as the table of content or the bibliography that have no value for our topic modeling algorithms. As we elaborate below the natural unit for topic extraction in our context is at the paragraph level. To this end we next split our documents into their component paragraphs. We drop paragraphs that are shorter than 50 characters, the typical length of a line. These can arise for numerous reasons, such as a sub headline or a label of cut figure, and their exclusion is required since they do not contain any paragraph-level information. This yields a corpus of paragraphs on which we will run our analysis.

Considering each of these paragraphs, we will turn them into so called "bags of words".

[2]Note that due to a different encryption format the 2015 report is not yet available to us, but will be included in future versions of this paper.

[3]The reports can be found at: https://fraser.stlouisfed.org/title/economic-report-president-45?browse=1960s

This is in line with the "bag of words approach", or "tokenization" commonly used in Natural Language Processing (NLP). This means that we will not consider the sequences of words in a paragraph and instead count the occurrence of specific words. The reason for this approach is the topic-modeling algorithms we will employ only take into account the joint occurrence of words and not their relative order.

To arrive at these word bags, we go through the following cleaning steps. In order to only focus on informative words, i.e. those that are highly indicative of topics, we drop three types of words. First, we drop any non-alphabetic tokens from the corpus. Next, we drop all "stop words", which are common language particles such as "the", "as" or "in". Note that this is a form of general English language cleaning and not yet take the structure of the corpus into account. To this end we also exclude words that pervade the paragraphs and are therefore uninformative as well. In particular we drop all words that appear in more than 95% of all paragraphs. Finally, we drop words that appear less than 2 times in an entire report, since those will not guide topic selection due to their rarity.

To arrive at the final bags of words, we generate tokens, which pool words that are closely related and share a common meaning. To this end, we first transform all letters to lower case, which ensures that "Tax" at the beginning of a sentence and "tax" in the middle of a sentence are treated as the same token. Second, we want to reduce different forms of the same word into the same token. For example, we want to reduce "growth" and "growing" to a common form, e.g. "growth". This process is called stemming and is necessary since topic-specific words arise in a variety of forms. This, however, is not necessarily a lossless compression. In some contexts two words can be stemmed to the same stem even though their meaning is different. A relevant example would be "universal" which could indicate a health care topic, vs "university" pointing towards education. Conflating these two words would be problematic. For this reason it is important to judiciously pick the stemming algorithm. An example of an aggressive, i.e. overly conflating, stemmer is the "Lancaster stemmer". For this reason, we follow the most common approach in the NPL literature and opt for a less aggressive alternative, the "Porter stemmer" developed by Porter (1980). Finally, we restrict our attention to the 1000 most frequent words in the corpus. We tested this hyperparameter for robustness and found little impact on our topics when increasing the cutoff.

The final output of our data creation step is a $P \times T$ paragraph-token matrix X, where each row represents one of P paragraphs and each column represents one of the $T = 1000$ token counts. Entry (p, t) therefore represents the number of times that token t appears in paragraph p. Similarly, we have a dictionary that reports for each paragraph p the ERP that it originates from, this dictionary will be important for aggregating the paragraph-level analysis to the overall economic reports.

3 Methodology

At the heart of our inquiry is to understand what topics are emphasized in the ERPs. There is a number of ways that lend themselves to extract topics and their relative importance from a corpus of texts. As Hansen et al. (2018) note, the most common approaches for text based analyses in economics so far have been based on dictionary methods. The approach here is to associate a list of pre-chosen words for each pre-specified topic and for example count their frequency in each paragraph. While this technique is straightforward it runs into the constraint that it does not take the context of the words into account. For instance, if the researcher specifies "expansion" as an indication of health care, she might erroneously classify a paragraph as health care, even though it is actually referring to fiscal expansion. Since it is unfeasible (or at least error-prone) to write an exhaustive list of words indicating a topic, it is essential that an algorithm identifies groups of words indicating a topic. More advanced algorithms therefore analyze the underlying cross-correlation structures to endogenously generate word groupings that indicate specific topics. Thus, if "expansion" co-occurs with "coverage", it is likely to be about health care. Therefore, we will not rely on dictionary methods in this paper and instead work with algorithms that focus on revealing hidden structures in the text and provide us with groupings of words which define a topic. In the following, we will therefore treat "word groupings" and "topics" as synonyms.

In particular, we will rely on two modern algorithms widely used in NLP for topic extraction, namely the Latent Dirichlet Allocation (LDA) and the Non-Negative Matrix Factorization (NMF). These approaches will take the pargraph-token matrix X as input and output groups of co-occurring words. Additionally and importantly these algorithms also output the proportions of these word groupings for each input document. Crucially, these groupings of words often take the shape of well-identified topics. We therefore carefully analyze the word groupings outputted by the algorithms and identify if they indicate specific economic topics. We will compare the results from both algorithms in order to understand which one is better suited for our application.

Before we study the inner workings of these approaches, let us carefully discuss some further details of model inputs and model outputs.

Regarding the model inputs, the interested reader might wonder why we use a pargraph-token matrix instead of a report-token matrix, since this will require us to aggregate the topics back up to the report level. This turns out to be crucial for a successful application. A paragraph is the natural unit for a topic, since usually it is highly unusual to find a paragraph covering multiple topics. In contrast, an entire ERP always encompasses a large number of topics. This is important since both the LDA and the NMF favor a sparse distribution of topics, i.e. they usually assign only a low number (below 3) topics to each document. We cross-check our results by running our algorithms at the document level rather than the paragraph level and report our results in the Appendix. Whilst these robustness tests reveal generally the same broad topics, using the paragraph level data defines them significantly better, meaning that the words associated with a topic have a closer topical connection. When extracting topics from the entire reports at once we find that some of the resulting topics are conflated.

Turning to the model outputs one has to make a judicious trade-off when specifying the number of word groups K. This number is a crucial hyperparameter that governs how coarse the groupings are. For instance, you might have two quite similar clusters, one talking about "unemployment" and the other one about "job market uncertainty". Depending on the context we might want to keep these two topics separate or combine them. With a small K the algorithm is likely to combine these close topics whilst a large K splits them. This is in a way natural since the two topics are fundamentally linked, but depending on the context we might want to have a more detailed view of each. Consequently, g is an important parameter whose impact on our analysis we check by varying its level.

In addition, since our input documents are paragraphs as explained above, the output will also be in terms of paragraphs. That is, the algorithms will yield topic proportions for each individual paragraph. Since we are ultimately interested in the ERPs overall, we need to aggregate these paragraph-level results to ERP-level results. To do so, we use the mean over the paragraphs constituting a ERP. For example, if an ERP consists of 10 paragraphs and the algorithm predict that 5 of these paragraphs contain topic A with 50% each, and the other 5 paragraphs with 0% each, we say that the ERP contains topic A with 25%.

Finally, both of these algorithms share the same fundamental approach. They build an underlying statistical model of the text-generating process to infer the hidden cross-correlation structure of words to generate word groupings.[4]

Having thus noted the similarities between the algorithms, let us turn to describing the specific algorithms. We start by discussing the LDA as developed by Blei et al. (2003) and then move to the NMF initially proposed by Lee & Seung (1999).

3.1 Latent Dirichlet Allocation

The LDA is a hierarchical Bayesian model that belongs to the class of mixed membership models. Below we will give an outline of the most important elements and intuitions for the procedure, while we refer readers who are interested in the mathematical details of LDAs to the excellent exposiBlei (2012). Our focus here is to give an overview of the method to readers who are not familiar with the approach so that they can develop an intuitive understanding of the procedure and understand the implications for our further analysis.

At the heart of the LDA lies a generative model, i.e. an assumed structure that generates the observed documents. The individual building blocks for a document are as follows. First, we assume that there are a fixed number of topics K that the ERPs cover. A document gets created by drawing a distribution over its topics. Here a topic is a distribution over the vocabulary. A document is then a sequence of draws from its distribution of topics and their conditional distribution over words. Note however, that we do not observe either what

[4]In essence, the problem of extracting topics from our corpus is thus an example of an unsupervised learning task. Unsupervised learning refers to algorithms that extract patterns or clusters from data without explicit labels which organize the data. An excellent introductory discussion of the differences between supervised and unsupervised learning can be found in Hastie et al. (2001).

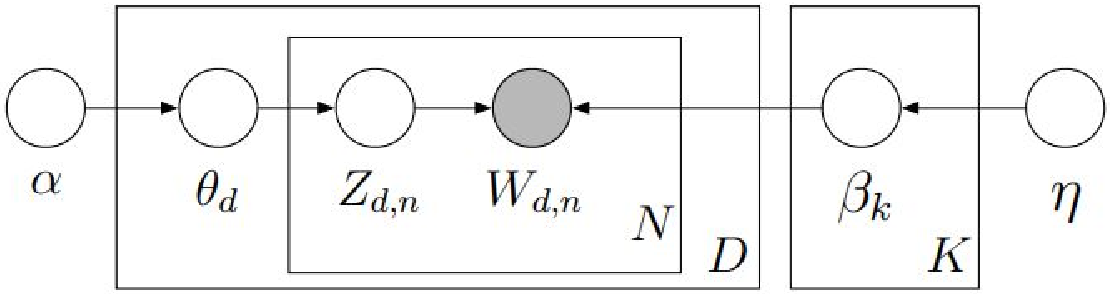


Figure 1: Plate model of LDA (Source: Blei & Lafferty (2006)

distribution a topic has over the vocabulary nor do we know what shares of topics each document has. These variables are ex ante hidden from us. Consequently, LDA is a hidden-variable model. Our task is to infer these hidden variables from the observed realizations of words that we find in our documents. Essentially, we are forming a posterior distribution over the hidden variables.

To illustrate this clearly let us formalize the assumed generative process for the creation of our documents.[5] We have K topics that each have a distribution $\vec{\beta}_k$ over the vocabulary V (comprising all 1000 words). These K distributions are drawn from a symmetric Dirichlet distribution parameterized by the scalar η. Similarly, for each of D documents the distribution of document $d = 1, ..., D$ over the K topics is denoted by $\vec{\theta}_d$. These $\vec{\theta}_d$ are drawn from a Dirichlet distribution parameterized by the vector $\vec{\alpha}$.

Let us next discuss the generative process as assumed by the LDA. First, draw the $\vec{\beta}_k$ for all K topics. Next, for a given document first draw a distribution over the topics, $\vec{\theta}_d$. Then draw N realizations from this distribution, where N is the number of words per document.[6] Having thus drawn the distributions, we now make draws from those distributions to form the specific document d. For each of the $n = 1, ..., N$ words in document d draw a topic assignment $Z_{d,n}$ according to θ_d. Having thus specified a topic for this word, draw the specific realization of the word $W_{d,n}$ from the distribution $\beta_{Z_{d,n}}$, formally $W_{dn} \sim \text{Mult}(\beta_{Z_{d,n}})$.

Figure 1 illustrates this process using a graphical representation of the LDA in plate

[5]We mostly follow the conventional notation of Blei & Lafferty (2006)
[6]To clarify, a document here in our context is a paragraph of an ERP.

notation (see Figure 1). Starting from the right, the rightmost node η represents the topic Dirichlet distribution. From this distribution we draw K topic-word distributions, $beta_k$. Turning to the leftmost node of the graph, α represents the Dirichlet distribution which assigns document-topic distribution to each document. Our corpus consists of D documents each of which is assigned a topic distribution, θ_d. The innermost plate represents a specific document d, which consists of N words. For each of these words $Z_{d,n}$ is a realization of θ_d. The eventual realization of the nth word is denoted by the grey filled central node, $W_{d,n}$. It depends on $\beta_{Z_{d,n}}$.

Note that our variables of interest are θ_d and β_k. By studying the β_k, one can infer whether topic k is interesting and coherent and assign it a real-world topic name for further analysis. The θ_d on the other hand are then crucial to determine to what extent these topics appear in the different documents and therefore eventually in the different EPRs, thus allowing us to infer what topics each EPR stresses.

The above model structure allows us to form a posterior over the desired but unobserved variables in the model, namely $\{\theta_d\}_{d=1}^D$ and $\{\beta_k\}_{k=1}^K$, given the observed variables. Here the observed variables are the word realizations in our corpus over the D documents, $\{\vec{W}_d\}_{d=1}^D$. The posterior distribution of interest is:

$$p(\{\vec{\theta}_d\}_{d=1}^D, \{\vec{Z}_d\}_{d=1}^D, \{\vec{\beta}_k\}_{k=1}^K | \{\vec{W}_d\}_{d=1}^D) = \frac{p(\{\vec{\theta}_d\}_{d=1}^D, \{\vec{Z}_d\}_{d=1}^D, \{\vec{\beta}_k\}_{k=1}^K, \{\vec{W}_d\}_{d=1}^D)}{\int_\beta \int_\theta \sum_z p(\{\vec{\theta}_d\}_{d=1}^D, \{\vec{Z}_d\}_{d=1}^D, \{\vec{\beta}_k\}_{k=1}^K, \{\vec{W}_d\}_{d=1}^D)}$$

, where $p(\cdot)$ is a general stand-in for distributions, specified by its arguments. The left-hand side denotes the posterior, i.e. the distribution over $\{\theta_d\}_{d=1}^D$ and $\{\beta_k\}_{k=1}^K$ conditional on the observed corpus, $\{\vec{W}_d\}_{d=1}^D$. This equals the joint distribution divided by the marginal distribution.

While the integral in the posterior is intractable as shown in Dickey (1983), there are a number of approximation procedures available to calculate the posterior as in Wainwright & Jordan (2007). A common approximation approach, mean field variational inference, is given by Blei et al. (2003). The idea of this approximation is to assume independence among the hidden variables and fit this simplified posterior as close as possible (with closeness measured by Kullback-Leibler divergence) to the unconstrained true model. Since the exact procedure is less relevant for the purpose of this paper and understanding our results, we refer interested readers, who want to get a detailed introduction into mean field variational inference, to either Blei et al. (2003) or WWainwright & Jordan (2007).

To summarize, the LDA allows us to extract topics as specific word groupings. We follow the LDA literature and label each topic with a phrase that summarizes the connection of the most frequent 20 words in the topic. Moreover, this approach yields a distribution over the topics for each document. As discussed above, we aggregate these paragraph-level topics to ERP-level topics to understand the focus of each report.

3.2 Non-Negative Matrix Factorization

Non-Negative Matrix Factorization (NMF) as proposed by Lee & Seung (1999) is a technique similar in spirit to principal component analysis (PCA). The underlying idea is that we can approximately decompose our overall document token matrix X into two components. Specifically, a paragraph is viewed as a linear combination of key components, which are viewed as topcis in NLP applications. In matrix notation, we can thus express this decomposition as

$$X \approx WH$$

where X is our $P \times T$ document-token matrix, W is a $T \times K$ word-topic matrix and H is a $K \times D$ topic-document matrix. In this way, it can be viewed as decomposing every document in our corpus as a combination of the topics captured in W, where the weights are provided by H. Crucially, these two matrices are exactly our desired outputs. As before in the LDA the H matrix allows us to study the components and assign them topic names based on the 20 most important words in each.

The reason why non-negative matrix factorization is particularly well suited for NLP-applications compared to other versions of PCA is that X is entirely non-negative. Therefore, we can guarantee that positive W and H exist. With positive W and H the interpretation of the components as topics becomes natural.

In order to perform the decomposition we follow the algorithm proposed by Lee & Seung (2001). Note that this algorithm ensures that at all steps W and H are non-negative as desired for our application.

In addition to the non-negativity, O'Callaghan et al. (2015) demonstrate that NMF is well-suited for NLP applications since it exhibits good topic cohesion, especially in smaller corpora. As we will show below, we find this to be true in the context of our application.

On the other hand, NMF has one disadvantage, namely that the decomposition is not necessarily unique. Therefore, different initializations might lead to different results. For this reason, we check our results with different initialization, but find that for our applications, they do not lead to significant differences.[7]

Finally, in order to further improve the power of the NMF, we follow Salton & Buckley (1988), in applying Term Frequency-Inverse Document Frequency (TF-IDF) weighting as a further pre-processing step on our corpus. The basic idea is to build weights that capture the informativeness of words. In particular, we weight each token t in document d by

$$tfidf(t, d, D) = tf(t, d) \times idf(t, D)$$

where D is the number of documents, N is the number of distinct tokens, $tf(t, d)$ is the number of occurrence of token t in document d and $idf(t, D)$ is given by

$$idf(t, D) = \log \frac{N}{|\{d \in D : t \in d\}|}$$

We include the token frequency $tf(t, d)$ to capture the fact that more frequent words are more important. On the other hand, we want to penalize words that are pervading the corpus and are therefore less informative. This is captured by $idf(t, D)$. It is given by the log of the inverse proportion of paragraphs that contain the word.

4 Results

In this section, we will present our main results. First, we will discuss the main topics emerging from the presidential reports. In particular, we will contrast the topics emerging

[7]In particular, we use random initial non-negative matrices, as well as various versions of Non-Negative Double Singular Value Decompositions as initial matrices, where the individual versions are differently suited to the sparsity of the problem.

from the LDA and those from the NMF approaches. For the most interesting of these topics we will then turn to the time series to determine what topics each president emphasizes and their presidential reports. We find that whilst both approaches broadly align for both the topic content and their time-series behavior, we find that the NMF is the preferred method for our application. Finally, we analyze to whether democratic and republican presidents are focusing on different or similar issues. Interestingly, we find that on the majority of topics the topic time trend is more indicative for whether a topic appears in a presidential report than the party alignment of the sitting president.

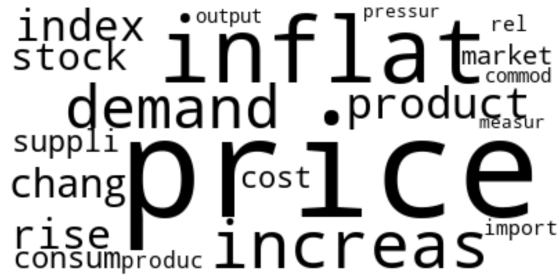
4.1 Topic Content

As discussed in the above methodology section our two approaches, LDA and NMF, yield distributions of topics over words. We follow the common practice in the NLP literature of studying the top 20 words in each of these topics to interpret them. For the convenience of exposition we furthermore assign a name to each topic that summarizes it succinctly.We now turn to the discussion of these individual topics. They fall into three categories, which we will discuss in turn. First, key policy topics that are sharply identified by both algorithms. These first key policy topics we further split up into "primary" and "secondary" topics, due to the nature of their time series, which we will discuss below. Second, key policy topics that only emerge in the NMF approach. Finally, we illustrate that some topics, independent of the algorithm, are not relevant for further analysis, since they identify parts of the report that are not policy related.

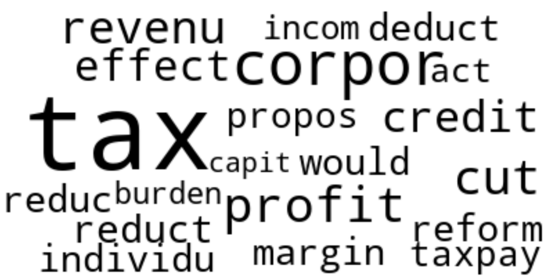
4.1.1 Topics Uncovered by both NMF and LDA

Figure 2 displays six key policy topics which will be crucial for our further analysis. Here, we contrast the results of the NMF and the LDA by displaying closely related topics in the same figure. In particular, whilst the left column shows the NMF topics, the right column displays their LDA cousins. The individual topics are represented as word clouds, where we use the aforementioned 20 most frequent words in each topic. Note that the words depicted are the stemmed version (see above details). To illustrate relative word importance each

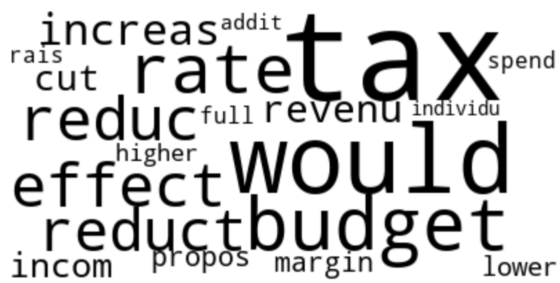
(a) Inflation/ Monetary Policy Topic NMF



(b) Inflation/Monetary Policy Topic LDA

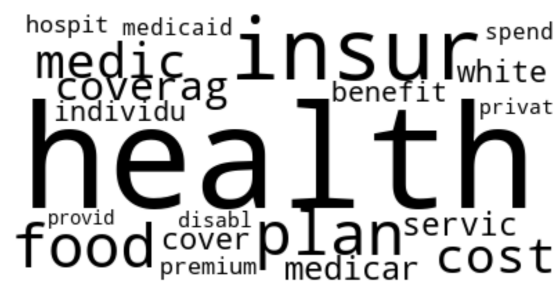


(c) Taxation Topic NMF



(d) Taxation Topic LDA

(e) Health Care Topic NMF



(f) Health Care Topic LDA

(g) Trade Topic NMF

(h) Trade Topic LDA

Figure 2: Primary policy topics: the left column depicts four topics identified by the NMF, and the right column closely related topics identified by the LDA.

word is scaled by its relative frequency, meaning that most important word has the largest font.[8]

First, panels (a) and (b) represent a topic that we have termed inflation/monetary policy. The most common words in the NMF topic are "rate", "interest", "inflat" and "increase". For the LDA those words are "price", "inflat" and "increas". Note that in both cases, all other words are also closely related to our topic name. Interestingly, while the top words are clearly about the same topic, the different algorithms stress slightly different aspects of this topic. In particular, the NMF topic is more orientated towards monetary policy, whilst the LDA topic seems slightly tilted towards the inflation part of the topic. For instance, note that the NMF does not feature "price" as a top word, whilst the LDA does not feature "rate". Therefore, while both algorithms identify the same topic, their emphasis differs slightly.

Second, panels (c) and (d) give us the next word groups which we label as "taxation". The top words of the NMF results are "tax", "corpor" and "profit". For the LDA we find "tax" also in the top position, but it is followed by a filler word "would" and the topically relevant "rate".[9] Observe that while the topics are closely related there are differences in the connotation of the top words. While both cover words about taxation, the NMF is slightly more oriented towards corporate taxation. This is evident in terms such as "corpor" (short for corporation) and "profit". On the other hand, the LDA does not incorporate these terms in its top words.

Third, we identify a health care topic for both the NMF (see panel e) and the LDA (see panel f). The top words of the NMF are "health", "care" and "insur". The LDA shares "insur" and "health" in its top words and furthermore includes "plan". The two word groups are very coherent and appear to be concerned with the exact same topic. This is in contrast to the topics described above, in which we identified a difference in topic lean between the two methods.

Finally, panels (g) and (h) are labeled as trade topics. Here, the NMF's top three words

[8]Note that importance of a word in the context of the LDA is measured by the probability mass in the topic-word distribution. On the other hand in the NMF word importance refers to the value of the word in the word-topic matrix W.

[9]Note that one can do further cleaning of these filler words and rerun the algorithm. However, we intentionally left this word in here to illustrate that even if certain artifacts survive, they do not render the topics nonsensical.

include "trade", "intern" and "agreement". The LDA has "countri", "unit" and "trade" in its top three words. Notice that the NMF in this context provides us with a slightly sharper word group, since it does not give weight to words like "unit" or "countri". While these words are clearly related, they do not seem to share the same strong intuitive connection as the words presented in the NMF results.

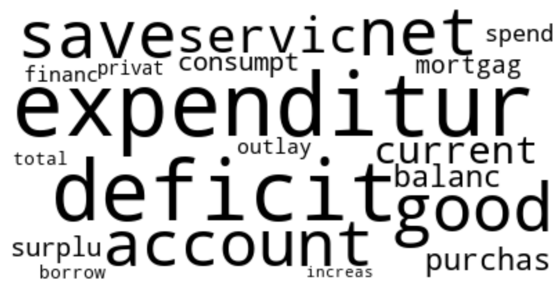
Figure 3 displays six more policy topics. The interpretation of 3 is the same as above, with the left column showing word groups based on the NMF and the right column their LDA equivalents.

First, panels (a) and (b) showcase the word groups for a topic that we call "Government Expenditure". This title is an obvious choice, given the most frequent words in the topics. For the NMF those are "govern", "expenditur", "local" and "feder". In the LDA the respective words are "expenditur", "deficit", "good" and "save". Whilst both word groups therefore clearly fall into the overall topic "Government Expenditure", the NMF grouping emphasizes mostly the implementation of government spending, stressing for example that government spending exists both at the local, the state and the federal level, that is the administrative level of government spending. The LDA topic, on the other hand, seems to be more oriented towards the politics of government spending, as indicated by the words "good" or "deficit".

Next, panels (c) and (d) display groupings that constitute a topic we term "GDP/Growth". The NMF topic's most common words are "growth", "real", "gdp" and "economi", whilst for the LDA those are "rate", "growth", "year" and "increas". While both algorithms again describe the same topic, it is also worthwhile studying the word groupings in detail. The NMF puts a very high weight on the term "growth" which is by far the most common word. In contrast to this, the LDA distributes weight more evenly across the most common words. Thus, the NMF group seems to stress growth and GDP in general, whilst the LDA group tends to describe the behavior of growth in more detail as witnessed by "increas", "rise" or "decline" - a similar but distinct focus.

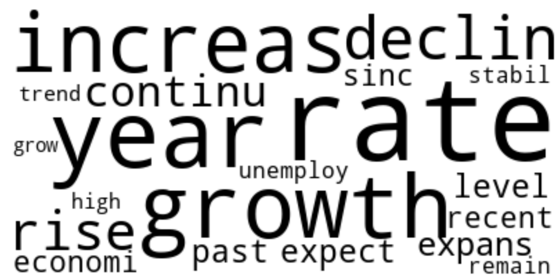
Finally, panels (e) and (f) describe a topic called "Investment". For the NMF the most frequent words are "invest", "capit", "busi" and "foreign", whilst for the LDA those terms are "invest", "capit", "credit" and "corpor". Overall, these are clearly topics that are very

(a) Government Expenditure Topic NMF



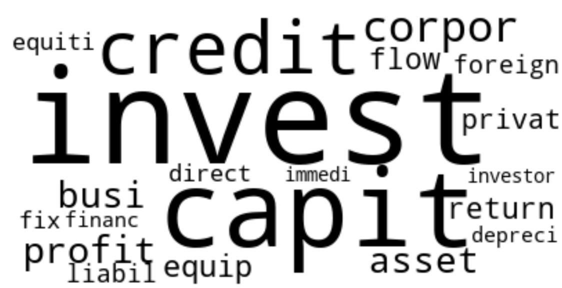
(b) Government Expenditure Topic LDA

(c) GDP/Growth Topic NMF



(d) GDP/Growth Topic LDA

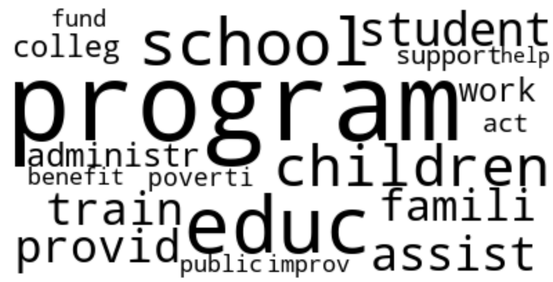
(e) Investment Topic NMF



(f) Investment Topic LDA

Figure 3: Secondary policy topics: the left column depicts four topics identified by the NMF, and the right column closely related topics identified by the LDA.

(a) Inequality Topic NMF



(b) Education Topic NMF

Figure 4: Key Policy Topics only identified by the NMF

similar and describe investment. If one considers not just the top four words but the top 20 words as represented by the word groups in the figure, one can see that also here NMF and LDA subtly differ in their orientation. The NMF topic seems more concerned with investment at the macroeconomic level, whilst the LDA topic focuses on the business side. This can be seen through words such as "infrastructur" for the NMF and "asset", "profit" or "equiti" for the LDA.

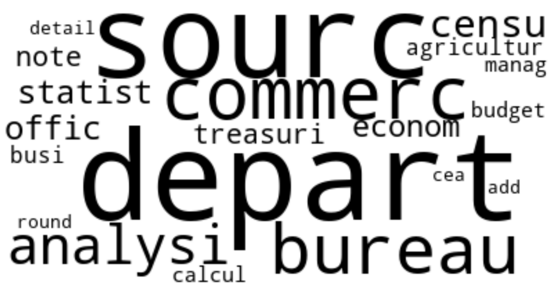
Overall, we have seen that all these topics appear in both the NMF and the LDA. Whilst the word groupings clearly describe the same overall topic, we also emphasize that they can differ in their detail and therefore may have slightly different orientations. We will discuss to what extent these subtle differences matter for the topic occurrences in presidential reports below.

4.1.2 Topics Uncovered Only by the NMF

One significant difference is that the NMF reveals more policy-relevant topics than the LDA. Examples of those are the subject of this section, and are a good reason for the use of the NMF as the primary tool of topic detection in the current context. Importantly, note that the NMF and the LDA output the same number of topics, as this is a hyperparameter in both approaches which we set to the same value. Differences in the number of policy-relevant topics are thus solely due to the different approaches. This is in line with the findings of O'Callaghan et al. (2015), who observe that the NMF is better suited for corpora of middling size.

Two topics only uncovered by the NMF are the following:

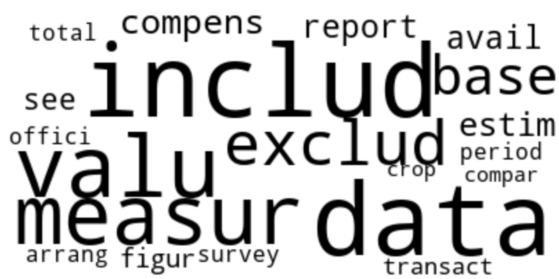
First, the NMF finds a word grouping which we label as an "income inequality" topic



(a) Government Agencies Topic NMF

(b) Months Topic NMF

(c) Units LDA



(d) Data Topic LDA

Figure 5: Non-policy topics: The left column displays topics identified by the NMF, and the right column topics identified by the LDA

(see panel (a) of figure 4). The top three words are "incom", "famili" and "person". Other noteworthy words are "poverti", "transfer" and "distribut". With the increased focus on inequality of both policy makers and academic economists, this is a topic of great interest for our analysis.

Second, we find a word group produced by the NMF that we label "Education", see panel (b) of figure 4. Here the top words include "program", "educ", "school" and "children". Since educational policies are a key yardstick for all administrations, it is noteworthy that only the NMF was able to identify it.

These additional topics that were only identified by the NMF are of first-order interest for our further analysis. For this reason, we will consider the NMF as our primary approach and use the LDA as a robustness tool. Based on these results we therefore strongly agree with O'Callaghan et al. (2015) and advocate the use of the NMF in further economic text processing applications. As the overall text corpus increases, the LDA might gain in competitiveness, but for most economic applications, the NMF seems to be the better starting point.

4.1.3 Auxiliary topics

In this section, we demonstrate that not all topics arising from the NMF and the LDA are necessarily policy relevant and thus interesting for further analysis. However, they still provide evidence for the smooth functioning of the topic-finding algorithms, since they also represent well-defined and coherent topics, albeit not policy related.

Figure 5 shows four such topics, the first row displaying two from the NMF and the second row two from the LDA. Panel (a) is a topic we term "Government Agencies", which consists of word describing government agencies mentioned in the report. The most common words are "depart", "source", "commerc", and "bureau" which can be recognized as part of agency names, e.g. the department of commerce.

Panel (b) is another very clearly defined topic "Months", since the most common terms are the name of the 12 months. Panel (c) is a topic that we term "Units", as it represents units frequently mentioned in the report. Note that it contains both time units such as "year", "quarter" or "month", but also counting units as in "first", "third" or "last". Finally, panel (d) shows a word grouping representing a "Data" topic, since it includes a lot of terms concerned with the data underlying statements in the reports. The most frequent words are "data", "includ", "valu" and "measur" and are all strongly related.

Summarizing, the above four word groupings exhibit strong internal consistency and are clearly identifiable as topics. However, these topics are not concerned with policy itself but aid the expression of them, for example by providing units, time and data terms required for the description of policy. Clearly, these auxiliary topics are not of first-order interest for our further analysis. It is nevertheless helpful to discuss them at this stage, to provide a full overview of the nature of all topics emerging from the analysis and to further stress the consistency of the topics generated by both the NMF and the LDA approaches. We also hope that by presenting these non-policy topics, readers unfamiliar with the approaches gain intuition for their inner workings.

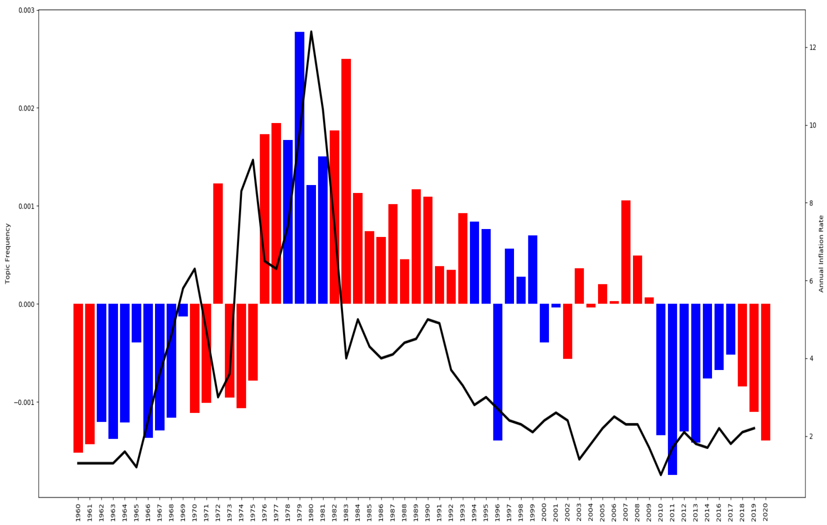
4.2 Analysis of Topic Time Series

In this section, we will analyze the time-series behavior of the topics uncovered above. This is essential, since the time series reveals to what extent each presidential report mentions each topic. It is thus one of the central pieces of our analysis.

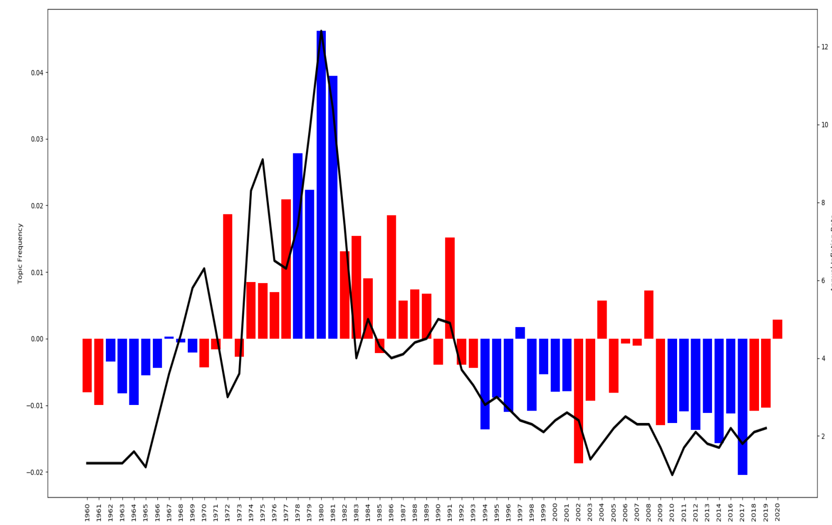
We report our results in the following format. For each topic we show the NMF time series and if possible, i.e. if the same topic is also uncovered by the LDA, the LDA time series for robustness. We demean each topic and show deviations from the average. That is, take the mean of the frequency of this topic over all presidential reports. If a presidential report mentions the topic at exactly that frequency, we assign it a score of 0, if it is mentioned at higher frequency a score higher than 0 and vice versa. We also indicate years with republican presidents with red bars and those with democratic presidents as blue ones.[10] Finally, note that we highlight legislation mentioned in the descriptions below through black vertical lines in the corresponding years of the time series graphs.

We start our analysis with panel (a) in Figure 6, which displays the "Inflation/Monetary Policy" topic. In addition to the topic frequency, we overlay the graphic with the yearly inflation rate for comparison. We see that in the early 60s the inflation topic is significantly below its average. Since in those years inflation was low and of no major policy concern, its scant mentioning in the report makes sense. As inflationary pressure increased throughout the early 1970s the topic frequency increases. It reaches its peak in the late 1970s and early 1980s when the inflation rate spiked and remained at a sustained high level. Following the Volcker interest rate hikes in the early 1980s, inflation was brought back under control. The inflation topic stays important throughout the rest of the 1980s as well as the 1990s, indicating that awareness of the issue remained high. Only in the 2000s does the topic cease its relevance. This indicates that whilst awareness of inflation spiked in line with the inflation rate, it remained a central policy concern even long after the inflationary bouts of the 1970s and early 1980s. Note that the time series of the equivalent LDA topic in panel (b) broadly confirms the results, though there the inflation topic starts slightly earlier and ends slightly sooner.

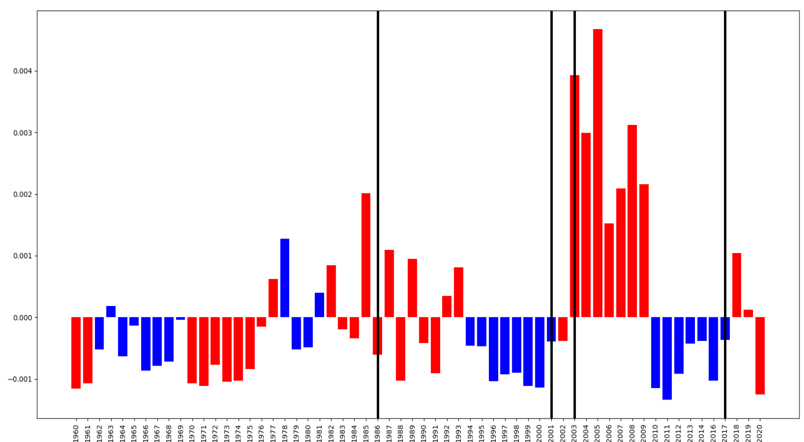
[10] For the observant reader, the reports are published in January of each year. In particular if a new president takes office in a year the report for this year will still be published by his predecessor.



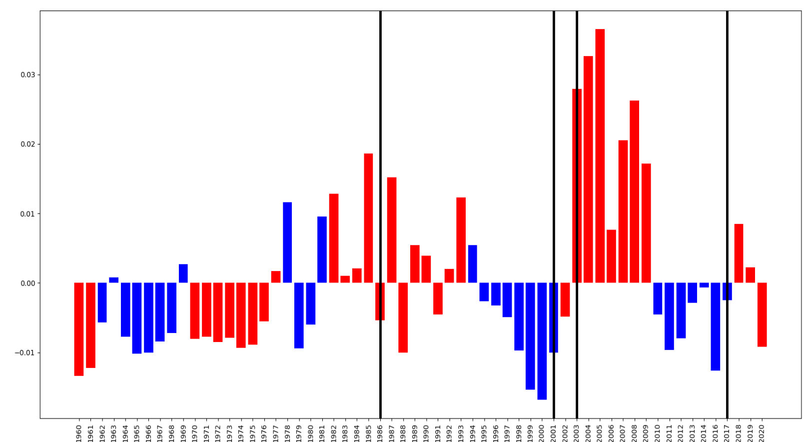
(a) Inflation/ Monetary Policy Topic NMF



(b) Inflation/Monetary Policy Topic LDA



(c) Taxation Topic NMF

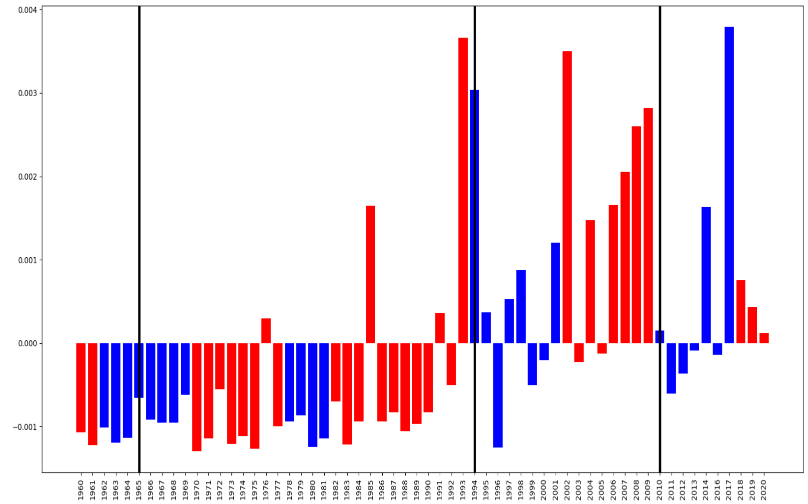


(d) Taxation Topic LDA

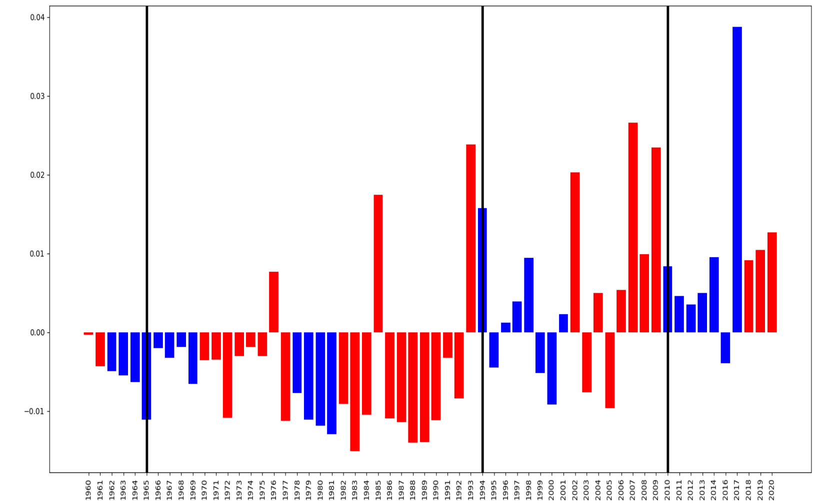
Figure 6: Time series: the left column depicts the time series behavior of the "Inflation/Monetary Policy" and "Taxation" topics identified by the NMF, and the right column the time series of their LDA equivalents.

Panels (c) in Figure 6 shows the time series for the "Taxation" topic. We see that it is below its average until the late 1970s. Only with the Reagan administration does it gain persistent prominence. In particular, we can see that taxation was a high-priority issue both in the lead-up to and in the aftermath of the "Tax Reform Act" of 1986. Taxation then ceases its prominent role in the Clinton years, before resurfacing in the George W. Bush administration. In particular, we can see that the taxation topic in the economic reports lag the "Economic Growth and Tax Relief Reconciliation Act" of 2001 and is consistently high around the "Jobs and Growth Tax Relief Reconciliation Act" of 2003, both of which cut tax rates significantly. Under Obama the taxation topic again loses in relevance as no big tax changes are proposed. The final reversal occurs with the Trump administration which highlights the "Tax Cuts and Jobs Act" of 2017 in its first two economic reports. For comparison the LDA in panel (d) has a very similar behavior and thus provides strong robustness to our NMF results. Overall, we can thus see that the importance of this topic follows the individual priority of each administration and thus faces frequent and strong reversals. We can also see a clear divide along party lines, with republican administrations stressing taxation more than democratic ones - given the prominence of taxation in Republican party dogma this is not a surprising result.

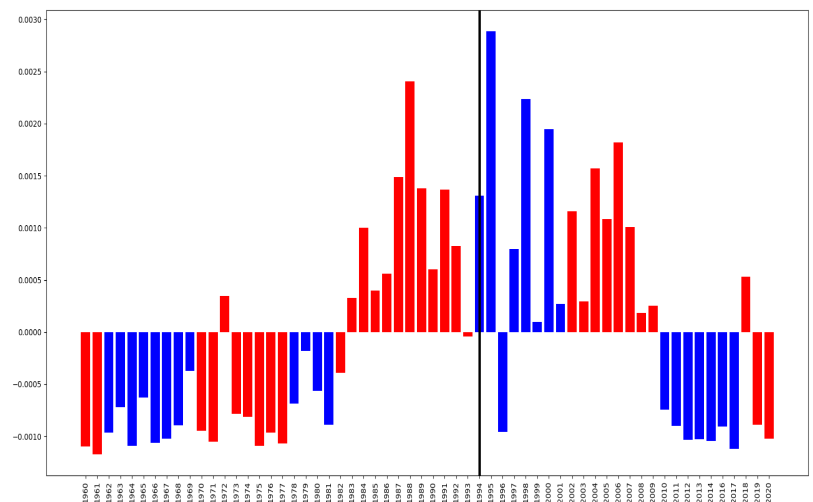
Next, we turn to panel (a) in Figure 7, which depicts the "Health Care" topic introduced above. We observe a clear trend, with health care gaining importance over the years. Until the 1990s the topic is below its mean, afterwards consistently and significantly above. However, this should not distract from the variation in the pre-1990 period. Consider for example the introduction of Medicare and Medicaid in 1965, which leads to an uptick in the health care topic. In the 1990s we can observe that health care became a frequently mentioned topic starting with the 1994 Clinton Health Care Initiative. Most interestingly, after its failure the Clinton administration significantly reduces its emphasis on health care. Nevertheless, health care remained a hot topic with a large number of proposals throughout the Bush years, which is again reflected by the topic. Eventually, this debate culminated in the "Patient Protection and Affordable Care Act" of 2010. Finally, the topic also showcases the continued debate about the extension or repeal of this contentious reform as it remains above its long-term average in the late Obama and early Trump years. Once again, the LDA



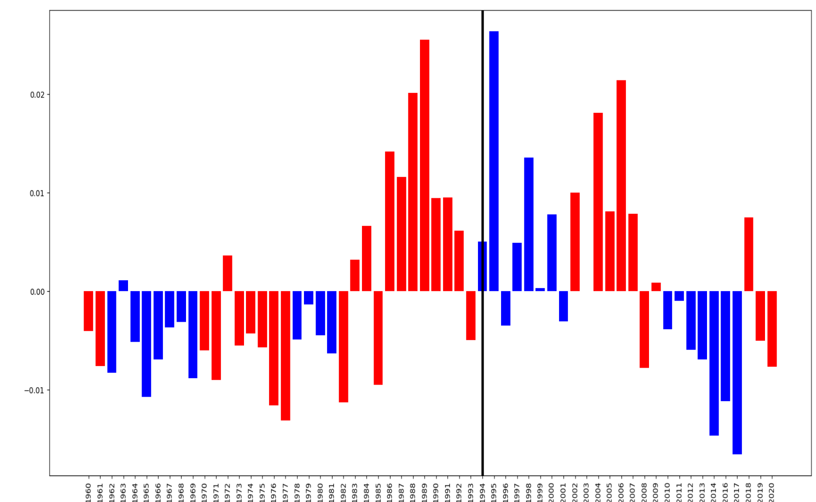
(a) Health Care Topic NMF



(b) Health Care Topic LDA



(c) Trade Topic NMF



(d) Trade Topic LDA

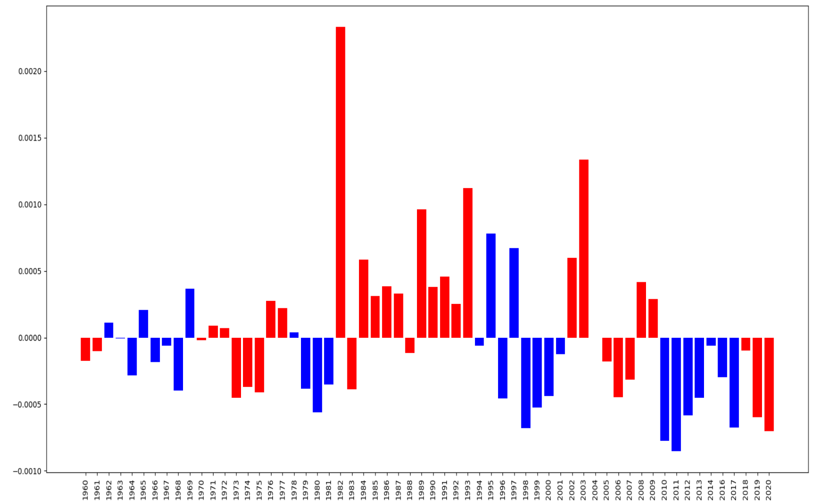
Figure 7: Time series: the left column depicts the time series behavior of two primary topics identified by the NMF, and the right column the time series of the equivalent LDA topics.

in panel (b) confirms the NMF results, placing even more importance on the Obama and Trump years.

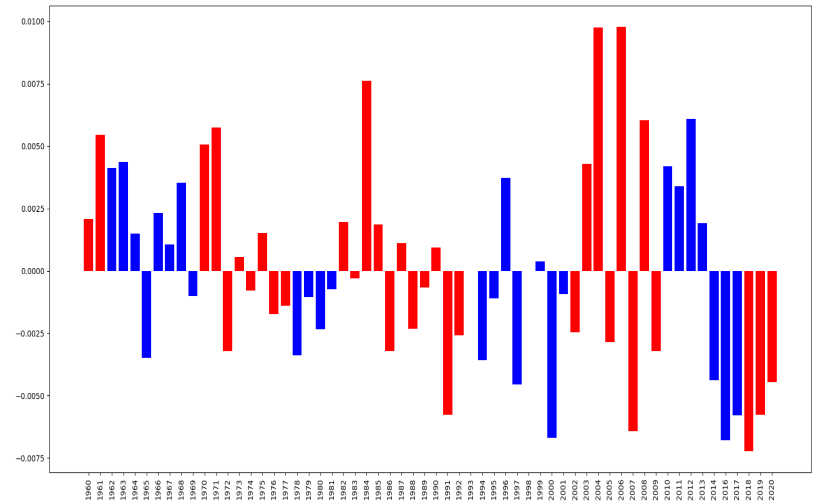
Panel (c) of Figure 7 displays the "Trade" topic. Until the early Reagan years, trade remains below its long-term average. This may at first appear somewhat surprising given the continuous tariff reductions with the various GATT rounds happening in those years. However, since the graph expresses relative size, these low scores are explained by the intense focus trade receives from the early 1980s until the Great Recession, which dominates all other periods. Indeed, in this era trade was on the forefront of the policy agenda for both the Reagan, Bush and Clinton administrations culminating in the North American Free trade Agreement of 1994. As witnessed by the time series plot, trade did not abandon its prominent role on the policy agenda until the great recession, at which point the topic frequency drops sharply. With the transition from the Obama to the Trump administration, the policy focus shifts from the negotiation over new trade agreements to an increased focus on the downsides of trade. This leads to a brief resurgence of the trade topic in 2017 in line with the "trade war" rhetoric of the time. The LDA in panel (d) confirms the results of the NMF almost perfectly, despite the somewhat different focus of the word groups defining the topics.

Moving on, we consider panel (a) in Figure 8 which displays the time series for the "Government Expenditure" topic. Whilst the overall time series lacks a clear trend, the topic is more pronounced during the Reagan and Bush administrations. Given the small-government agenda of these periods, this hardly comes as a surprise. Government expenditure thus is another topic that arises more under republican administrations and their small government focus. Note that again the LDA results in panel (b) agree broadly with the NMF results, in that they also do not exhibit a clear time trend and highlight the same peaks.

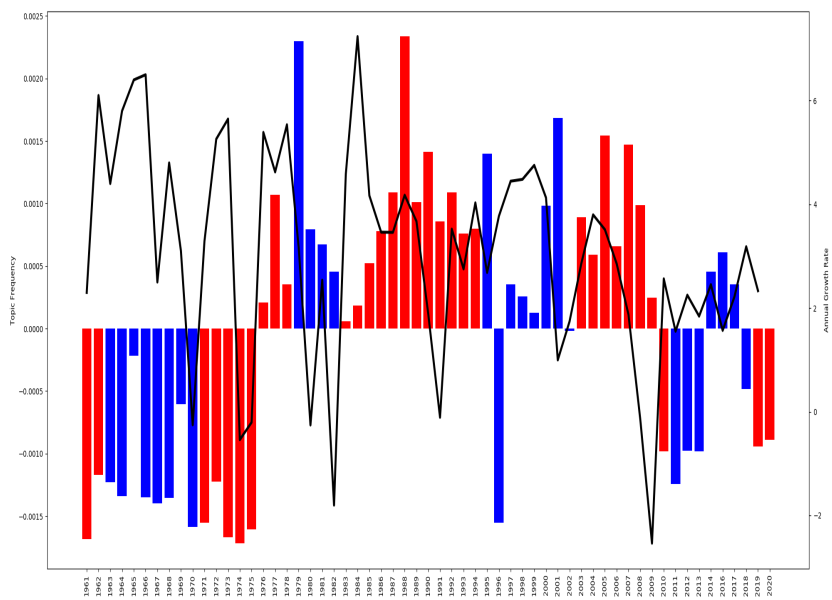
Panel (c) displays the "GDP/Growth Topic" overlaid with the yearly real GDP growth rate. We can see that up to the mid 1970s this topic is significantly below its average. This is explained by the sustained and high growth of these years which made it less of an immediate policy concern. In light of the oil crises in the 1970s and the respective recessions, the topic surfaces for the first time. It then remains prominent until the late



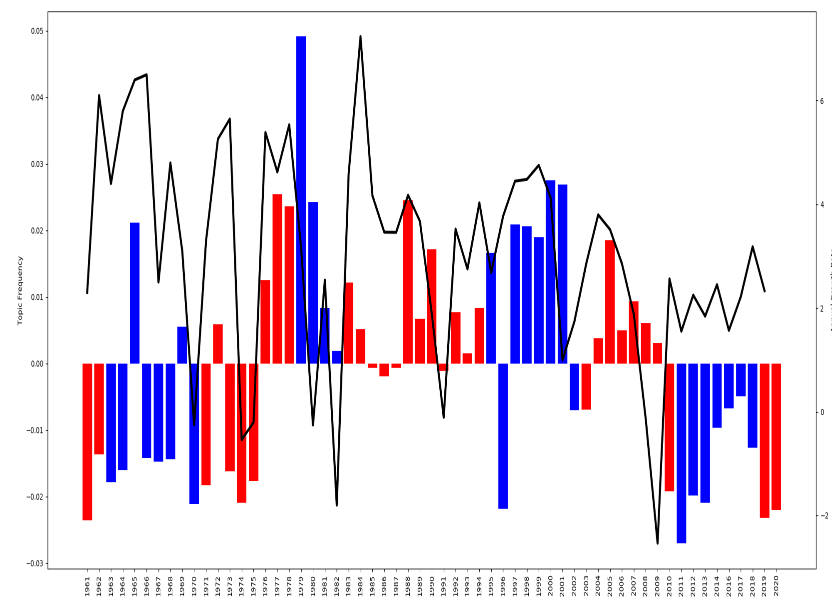
(a) Government Expenditure Topic NMF



(b) Government Expenditure Topic LDA



(c) GDP/Growth Topic NMF



(d) GDP/Growth Topic LDA

Figure 8: Time series: the left column depicts the time series behavior of two secondary topics identified by the NMF, and the right column the time series of the equivalent LDA topics.

Bush years, relatively independent of the growth rate, which roughly fluctuates around 4% in this interval. Somewhat surprisingly, we do not see a strong uptick of the growth topic in the great recession - in fact, we observe a significant drop. Interestingly, the LDA results in panel (d) confirm both the overall pattern and this eventual drop. The drop is thus not an artifact of the NMF but instead a robust feature of the presidential reports. One possible explanation is that in the great recession, the Obama administration focuses not directly on the growth rate but instead on other aspects of the crisis. Examples could be the instability of the financial system or unemployment.

We now consider the Figure **??**, where panel (a) showcases the "Investment" topic. Its behavior is similar to that of the growth topic, being most pronounced from the mid 1970s until the late 2000s. This is in line with the general focus of the administrations in those years on stimulating investment. However, the LDA as displayed in panel (b) is not as similar as in the other topics discussed before. Therefore, we will be careful in drawing conclusions from this topic.

Panel (c) displays the "Inequality" topic, which is only identified by the NMF. The topic does not display a strong time trend, indicating that inequality has been regularly discussed in the presidential reports over the entire time period. In light of the recent increase in both income and wealth inequality this might seem somewhat surprising. As shown by our NMF topic this public discourse has not translated into a higher priority in the presidential report. It is thus an exception to the general pattern that we have observed above, where the most prominent economic policy topics feature heavily in the presidential reports.

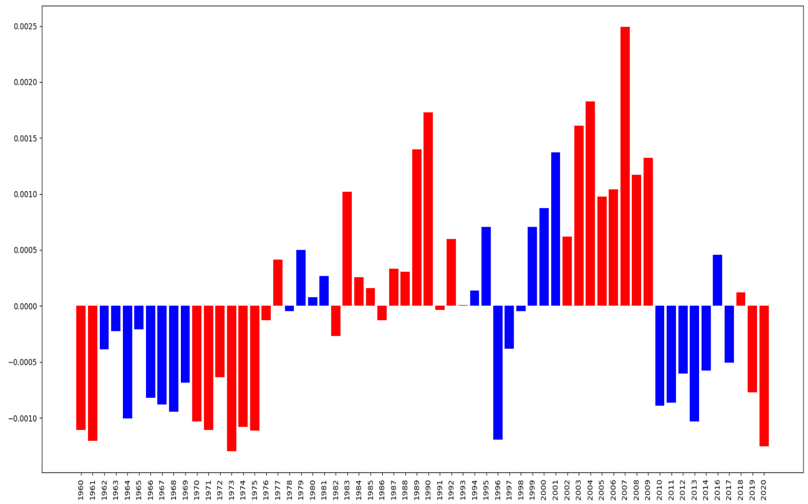
Panel (d) finally shows the "Education" topic. Similar to the inequality topic, it is distributed relatively evenly over the years. However, it does not come as a surprise in this case as education policy has historically not been subject to the same cyclical bouts of policy activity as other policy areas. Despite the absence of an overarching trend, the topic does reflect differences in administration priorities. For example, this is reflected in the low frequency of the topic during the Reagan years, whose administration aimed at abolishing the department of education indicating a lesser focus on federal education policy. In contrast to this, the late Clinton years and early George W. Bush years display a spike in the topic, reflecting the policy discussions in the build up to the "No Child Left Behind Act" of 2002.

Overall, we can also see that with the exception of George W. Bush, this topic is significantly more present under democratic presidents than under republican ones, which is in line with the difference in policy focus between the two parties.

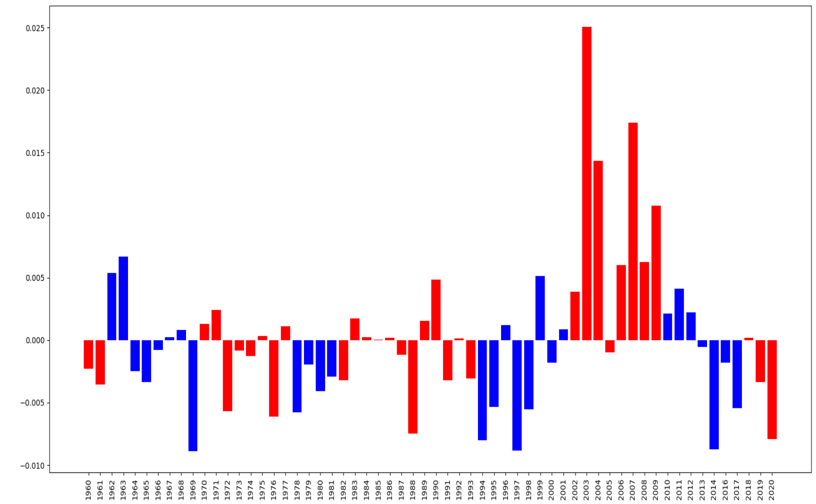
Overall, we can draw several conclusions from this section. First, it is striking how closely the time series behavior of topics is related to the policy focus of the time and the administration. The economic report therefore seems to be a good representation of the economic policy focus of each administration and the major economic issues of the time. We can thus convincingly reject the hypothesis that the economic reports have limited value in determining the policy priorities of a given administration. Second, we note the striking resemblance of the time series of the topics generated by the NMF and the LDA. Only in one topic, investment, do the time series differ markedly. The fact that these two completely different approaches, which are based on different underlying statistical models, yield very similar results provides us with strong confidence that our topics are well-identified both in their definitions and their time-series behavior. The fact that they make intuitive sense at both levels provides even more evidence.

4.3 Comparison of Magnitudes

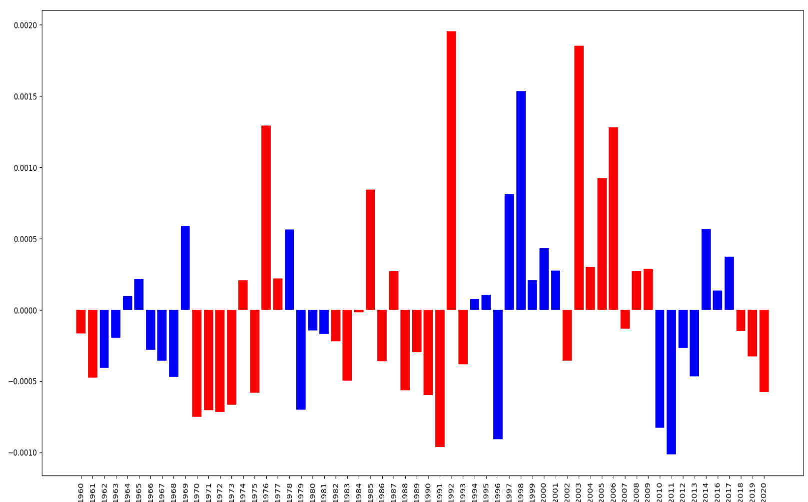
In the above analysis we compare the time-series behavior of each topic in isolation from the other topics. In particular, we focus on the relative frequency of the topic against its long-run average. Whilst we see that each topic waxes and wanes over time, the attentive reader might wonder how the frequency of one topic compares with that of the others. To see why this matters, consider the following case. If one topic is significantly above its average in its time series, in the above analysis we would interpret this as the administration putting strong emphasis on this topic. However, if the magnitude of that topic were significantly lower than that of other topics, this change might not indicate a meaningful change in administrative priorities. This is because in that case the topic is overall only a minor topic, rarely mentioned in the report. On the other hand, if the absolute magnitudes of all topics are at the same level, our above analysis is in fact correct and changes in relative topic frequency directly translate into meaningful priority shifts. Fortunately, for us this is exactly the case as Figure 12 demonstrates. Here, we plot the magnitudes of the first three



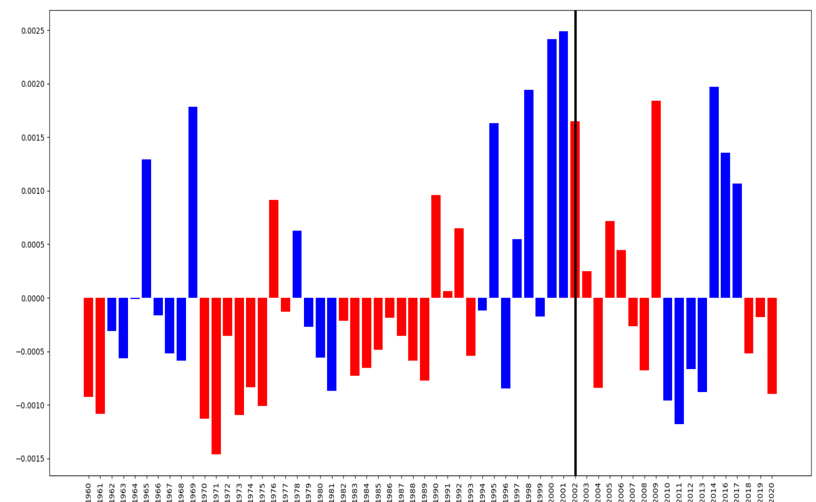
(a) Investment Topic NMF



(b) Investment Topic LDA



(c) Inequality Topic NMF



(d) Education Topic NMF

Figure 9: Time Series: panel (a) and (b) display the time series of the NMF and LDA topic "Investment". Panels (c) and (d) show the time series of the topics only identified by the NMF

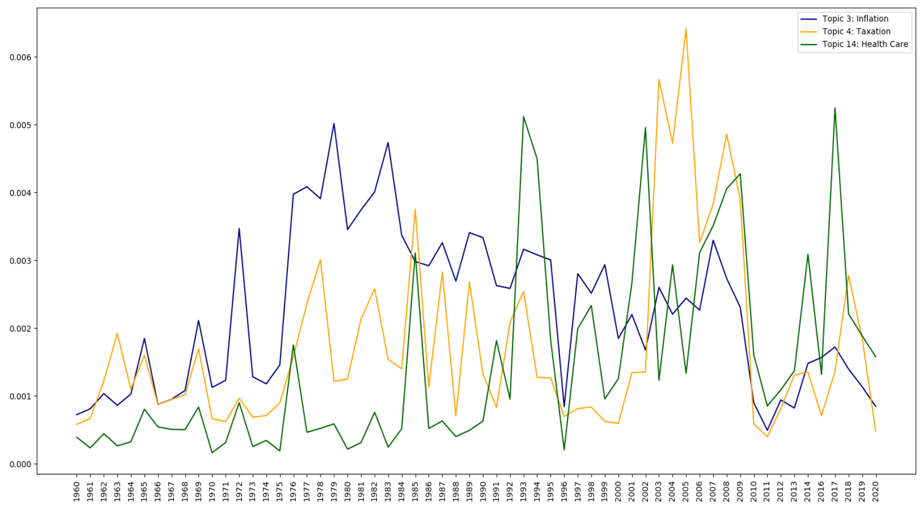


Figure 10: Inter-topic Magnitude Comparison: Time series graphs for inflation, taxation and health care.

NMF topics - inflation, taxation and health care - in one graph. Importantly, this magnitude is the same for all three topics. In the appendix we provide the same graphs for all other remaining topics as well as for the LDA topics. Crucially, all these graphs indicate that the magnitudes of all topics are indeed the same. Therefore, we can rest assured that changes in relative topic frequency as analyzed in the above section, also indicate policy priority shifts, thereby validating our analysis.

4.4 Differences and Similarities between Parties

The analysis of the above time series reveals that on most topics we cannot detect a systematic difference in emphasis between republican and democratic presidents. Indeed, it seems that time trends outweigh party differences. This might be a surprise, since the political dogma of the two parties differ significantly. Moreover, since the LDA and the NMF arrive at similar time series pattern we conclude that this a robust feature of the data. Only in a few topics can we detect a clear difference between the parties. The first of these topics is the "Taxation" topic, which is disproportionately emphasized by republican presidents. Meanwhile, democratic presidents talk more about education than their republican counterparts. Despite these two exceptions it is remarkable that topics, as distinct as "health care" or "inequality", do not exhibit clear party divides in emphasis. Based on these results we conclude that topic analysis of the economic reports of the president reveals issues that are at the top of the political agenda and not just dogmatic party issues. Consequently, it is sensible to study the outputted topics for policy analysis and academic pursuits.

5 Conclusion

In this paper we make several important contributions to the emerging literature on natural language processing in economics. First, we demonstrate the power of modern natural language analysis machinery on a typical economic text corpus. In particular, we apply two machine-learning approaches, Non-negative Matrix Factorization and Latent Dirichlet Allocation, on the Economic Reports of the President. Both of these approaches extract the main topics discussed in each presidential report. It is remarkable how well-defined and consistent these topics are - by inspecting the most frequent words, one can immediately determine what each topic is about and assign a simple English name, encompassing its meaning. Regarding the comparison between the two approaches, we find that in most cases the NMF and the LDA produce topics that are remarkably similar with respect to both their defining word groups and their time-series behavior. However, we also find that the NMF identifies a higher number of policy-relevant topics, so that we recommend the NMF as the text-analysis algorithm of choice for future natural-language applications in economics. With respect to methodology, we thus demonstrate the feasibility and success of two natural-language processing algorithms.

With respect to the contents of the Economic Reports we also find multiple interesting results: First, by analyzing the time-series behavior of each topic, we find that the major topics highlighted in the report are also major policy issues discussed at the time and mostly change in line with big secular changes in policy views. In contrast to this, party dogma and philosophy are not as strongly represented in the Economic Reports. In broad terms, a change between a democratic and a republican administration does not lead to rapid shifts in the topics stressed in the Economic Reports. Instead, these topics change slowly in line with overall economic trends. Finally, whilst the overall trend is clearly important, we can also see that specific events such as important reforms are directly reflected in the reports. These results establish the Economic Reports as a crucial data source for further analysis. Without being overtly obscured by party ideology, they emphasize both the major policy issues of the time and specific policy events.

We believe that this paper paves the way for a plethora of future research in natural language processing in economics. One specific application that we are interested in and will investigate in follow-up research is the classification of taxation changes into endogenous and exogenous changes to investigate the relationship between taxation and macroeconomic outcome variables such as growth. In particular, Romer & Romer (2010) have demonstrated that it is possible to assess the language of the reports to identify whether a change in the tax policy is due to economic conditions, e.g. various stimulus packages, or driven by ideological considerations or other exogenous factors. The significance of being able to tell these apart is that this can give rise to a causal identification strategy. In contrast to Romer and Romer, the natural language processing approach is fully objective and easily generalizable to different data sources.

References

Blei, D. M. (2012). Probabilistic Topic Models. *Communications of the ACM*, *55*(4), 77–84.

Blei, D. M., & Lafferty, J. D. (2006). Dynamic Topic Models.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Boyd-Graber, J., Hu, Y., & Mimno, D. (2017). Applications of Topic Models. *Foundations and Trends in Information Retrieval*, *11*(2-3), 143–296.

Dickey, J. M. (1983, September). Multiple Hypergeometric Functions: Probabilistic Interpretations and Statistical Uses. *Journal of the American Statistical Association*, *78*(383), 628–637.

Hansen, S., & McMahon, M. (2016, March). Shocking language: Understanding the macroeconomic effects of central bank communication. *Journal of International Economics*, *99*, S114–S133.

Hansen, S., McMahon, M., & Prat, A. (2018, May). Transparency and Deliberation Within the FOMC: A Computational Linguistics Approach*. *The Quarterly Journal of Economics*, *133*(2), 801–870.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning.* New York, NY, USA: Springer New York Inc.

Lee, D. D., & Seung, H. S. (1999, October). Learning the parts of objects by non-negative matrix factorization. *Nature*, *401*(6755), 788–791.

Lee, D. D., & Seung, H. S. (2001). Algorithms for non-negative matrix factorization.

O'Callaghan, D., Greene, D., Carthy, J., & Cunningham, P. (2015, August). An analysis of the coherence of descriptors in topic modeling. *Expert Systems with Applications*, *42*(13), 5645–5657.

Porter, M. (1980, March). An algorithm for suffix stripping. *Program*, *14*(3), 130–137.

Romer, C. D., & Romer, D. H. (2010, June). The macroeconomic effects of tax changes: Estimates based on a new measure of fiscal shocks. *American Economic Review*, *100*(3), 763-801.

Salton, G., & Buckley, C. (1988, January). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, *24*(5), 513–523.

Wainwright, M. J., & Jordan, M. I. (2007). Graphical Models, Exponential Families, and Variational Inference. *Foundations and Trends in Machine Learning*, *1*(1–2), 1–305.

6 Appendix

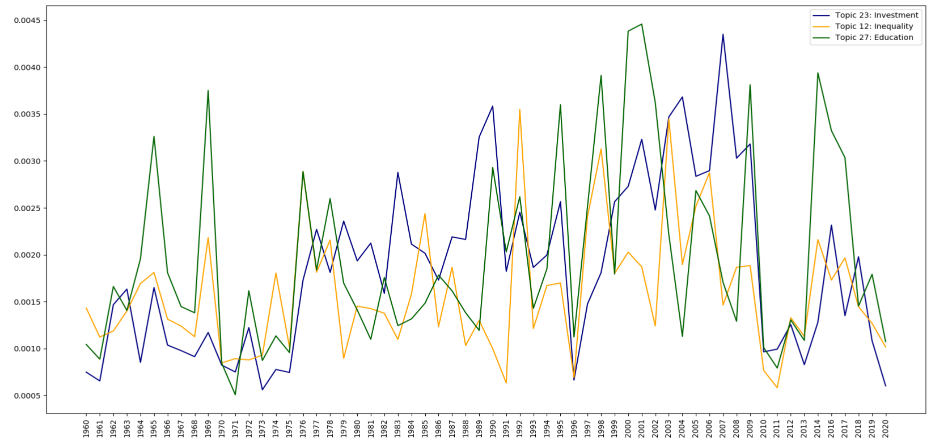


Figure 11: Inter-topic Magnitude Comparison: Time series graphs for investment, inequality and education.

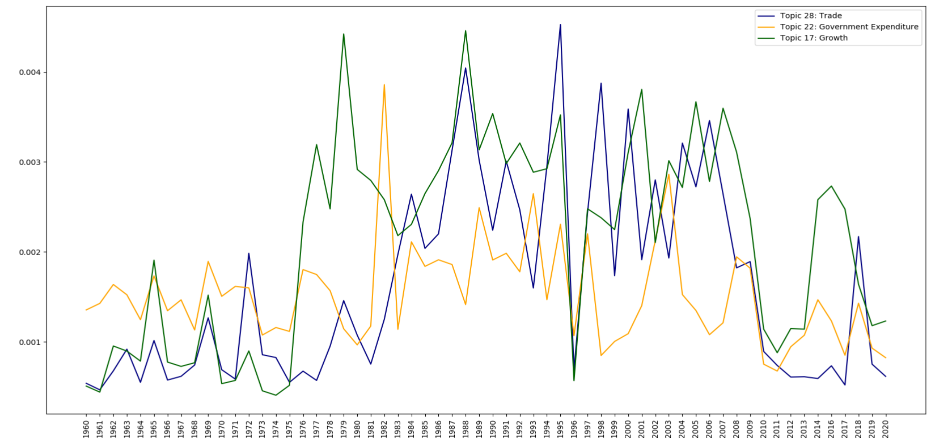


Figure 12: Inter-topic Magnitude Comparison: Time series graphs for trade, government expenditure and growth.

Table 1: Top 20 words by topic

	Topic 0	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9
Rank 1	bank	chang	depart	rate	tax	product	market	data	price	june
Rank 2	reserv	climat	sourc	interest	corpor	gross	financi	includ	index	juli
Rank 3	feder	technolog	commerc	inflat	profit	domest	loan	note	consum	may
Rank 4	deposit	adjust	bureau	annual	cut	farm	credit	begin	oil	august
Rank 5	commerci	measur	analysi	increas	revenu	agricultur	institut	season	increas	januari
Rank 6	loan	effect	censu	exchang	credit	output	mortgag	adjust	inflat	decemb
Rank 7	member	inventori	offic	declin	effect	sector	secur	monthli	energi	novemb
Rank 8	monetari	percentag	statist	expect	reduct	nonfarm	fund	base	rise	octob
Rank 9	gold	structur	econom	rise	would	busi	asset	see	commod	septemb
Rank 10	fund	shift	note	high	reform	produc	debt	estim	wage	april
Rank 11	credit	composit	treasuri	point	margin	per	deposit	quarterli	food	februari
Rank 12	hold	time	busi	lower	individu	type	corpor	except	produc	march
Rank 13	central	base	agricultur	higher	reduc	input	bank	avail	suppli	round
Rank 14	lend	respons	manag	level	deduct	major	borrow	seri	demand	detail
Rank 15	money	reflect	calcul	yield	propos	innov	hous	period	deflat	add
Rank 16	requir	global	budget	percentag	taxpay	improv	commerci	januari	ga	end
Rank 17	target	condit	add	particip	act	gain	stock	exclud	rel	begin
Rank 18	york	affect	cea	low	incom	petroleum	risk	compar	higher	total
Rank 19	currenc	pattern	round	would	burden	oil	interest	preliminari	market	month
Rank 20	agenc	rel	detail	money	capit	food	insur	revis	stabil	see

Table 2: Top 20 words by topic (continued)

	Topic 10	**Topic 11**	**Topic 12**	**Topic 13**	**Topic 14**	**Topic 15**	**Topic 16**	**Topic 17**	**Topic 18**	**Topic 19**
Rank 1	econom	percent	incom	deficit	health	tabl	state	growth	labor	dollar
Rank 2	council	averag	person	budget	care	index	unit	real	forc	billion
Rank 3	advis	increas	famili	fiscal	insur	gross	japan	gdp	statist	adjust
Rank 4	report	total	household	surplu	coverag	select	local	economi	bureau	season
Rank 5	journal	rose	farm	receipt	medic	continu	foreign	output	civilian	annual
Rank 6	presid	gnp	poverti	account	spend	see	germani	slow	particip	quarterli
Rank 7	activ	share	dispos	outlay	medicar	dollar	canada	spend	popul	million
Rank 8	review	gdp	distribut	feder	plan	person	world	econom	depart	except
Rank 9	american	annual	earn	current	cost	group	abroad	gnp	market	currenc
Rank 10	act	per	account	balanc	medicaid	expenditur	posit	per	women	valu
Rank 11	offic	real	inequ	debt	provid	major	europ	project	age	inventori
Rank 12	estim	declin	median	spend	servic	real	million	expans	sourc	monthli
Rank 13	preliminari	fell	consumpt	billion	premium	note	import	period	work	amount
Rank 14	chairman	sinc	save	offic	hospit	civilian	area	contribut	survey	rate
Rank 15	committe	estim	transfer	reduct	afford	relat	gold	demand	suppli	foreign
Rank 16	member	grew	net	save	american	type	agreement	potenti	capit	balanc
Rank 17	analysi	period	individu	project	improv	debt	immigr	strong	men	figur
Rank 18	cea	compar	measur	would	individu	corpor	european	grow	group	averag
Rank 19	develop	nearli	capita	payment	patient	popul	dollar	rapid	trend	payment
Rank 20	univers	less	per	reduc	qualiti	domest	payment	forecast	current	profit

Table 3: Top 20 words by topic (continued)

	Topic 20	**Topic 21**	**Topic 22**	**Topic 23**	**Topic 24**	**Topic 25**	**Topic 26**	**Topic 27**	**Topic 28**
Rank 1	unemploy	quarter	govern	invest	year	countri	cost	program	trade
Rank 2	job	fourth	expenditur	capit	age	develop	regul	educ	intern
Rank 3	insur	third	local	busi	fiscal	intern	would	school	agreement
Rank 4	recess	second	feder	save	last	world	firm	children	barrier
Rank 5	forc	first	privat	privat	past	foreign	use	student	world
Rank 6	rate	real	state	fix	end	currenc	new	assist	tariff
Rank 7	person	inventori	receipt	foreign	increas	exchang	reduc	train	negoti
Rank 8	civilian	declin	sector	net	recent	oecd	may	provid	liber
Rank 9	benefit	half	public	equip	first	major	benefit	famili	open
Rank 10	group	annual	consumpt	consumpt	averag	cooper	regulatori	administr	gatt
Rank 11	week	three	purchas	direct	per	import	competit	work	free
Rank 12	compens	rose	revenu	infrastructur	expect	europ	technolog	colleg	import
Rank 13	age	four	spend	inventori	half	export	energi	support	round
Rank 14	structur	sale	gener	flow	rise	mani	emiss	act	partner
Rank 15	reason	recoveri	transfer	return	million	domest	effici	poverti	global
Rank 16	select	gnp	debt	stock	next	imf	increas	improv	agricultur
Rank 17	economi	peak	role	increas	declin	flow	also	benefit	market
Rank 18	inflat	gdp	includ	new	earlier	less	requir	public	balanc
Rank 19	million	final	financ	depreci	expans	european	innov	fund	rule
Rank 20	high	recess	total	abroad	period	payment	standard	help	region

Table 4: Top 20 words by topic (continued)

	Topic 29	**Topic 30**	**Topic 31**	**Topic 32**	**Topic 33**	**Topic 34**
Rank 1	industri	system	polici	export	employ	nation
Rank 2	manufactur	board	monetari	servic	worker	research
Rank 3	major	governor	economi	good	wage	gross
Rank 4	averag	sourc	fiscal	import	job	paper
Rank 5	profit	reserv	inflat	net	increas	work
Rank 6	select	feder	econom	agricultur	work	expenditur
Rank 7	group	note	stabil	foreign	earn	account
Rank 8	earn	treasuri	administr	purchas	skill	bureau
Rank 9	corpor	except	effect	total	train	econom
Rank 10	privat	organ	goal	demand	salari	receipt
Rank 11	sale	intern	intern	domest	employe	save
Rank 12	sector	exchang	demand	produc	compens	scienc
Rank 13	inventori	standard	chapter	consum	hour	major
Rank 14	competit	calcul	expans	commod	number	relat
Rank 15	capac	cea	respons	includ	occup	defens
Rank 16	index	base	must	provid	nonfarm	cea
Rank 17	util	monetari	continu	account	busi	deflat
Rank 18	hourli	investor	achiev	valu	pay	calcul
Rank 19	construct	poor	macroeconom	abroad	million	develop
Rank 20	ad	reform	need	militari	popul	center