Different Approaches to Topic Modeling with State of the Union Speeches

By Alix Gates, Aleister Montfort, Mario Moreno

Abstract

This paper reveals that there are significant lexical differences in the State of the Union (SOTU henceforth) speeches over time, and finds that 1914 is a clear transitional point in the form and text of SOTU addresses. What's more, in this paper, we seek to emulate a strategy for identifying and grouping topics over long periods of time that was first presented by Rule, Cointet, and Bearman in 2015. While we ultimately fail to replicate all the results, we do manage to identify and group 1000 topics over the 228 years of SOTU addresses using a range of unique pre-processing techniques, similarity scores across co-occurrence matrices, semantic networks, and a combination of community detection algorithms and unsupervised learning methods. Lastly, these results are then compared to a baseline LDA approach.

The Problem and Significance

On January 9, 1790, George Washington went before Congress and delivered the first SOTU. In what's become an annual tradition, every President since has delivered his message to the country. Within those speeches, presidents reflect the challenges and opportunities facing the Nation, outline their policy priorities and, along the way, provide a window into the prevailing national sentiment.

Using the text of every SOTU address, our project attempted to identify United States' priorities given the most prevalent topics covered in State of the Union addresses. In doing so, we hoped to contribute to a better understanding of US priorities over time, group similar topics across time periods, and learn more about United States history. From a computational perspective, we hoped this project would help us better understand how to work with text data, deploy topic modeling algorithms, and explore different approaches to topic modeling described in academia.

Literature Review

There's an extensive amount of research related to the fields of topic modeling, including papers and articles already published that are directly related to our dataset and choice of project. In an academic paper analyzing lexical changes in SOTU addresses over time, Rule, Cointet and Bearman (2015) developed a strategy for identifying categories in texts despite changes in lexicology over time. In doing so, they provide a blueprint for two critical questions in our own analysis: how do we account for changes in the use of language and words in the last 230 years; and how do we identify common themes in these texts in spite of those changes?

To answer those questions, the authors deploy a method known as co-occurrence, which identifies categories based on terms co-appearing over a unit of text. This analysis involves discrete steps. First, the authors find frequently occurring nouns and paired terms such as

‘national security,’ and ‘local government.’ Next, they induce semantic categories by understanding the patterns in co-occurrence of terms in a period of time by identifying how many times joint terms appear together in a given paragraph, compute a proximity score that measures the relatedness of similar terms, and then employ a community detection algorithm to identify cohesive subsets. They first applied these steps to all the SOTU speeches to develop a global semantic network, and then broke it down into local semantic networks by doing the same steps over discrete periods of time. Finally, the relationship between local semantic networks was determined using a river network.

The authors provide three results, two of which are directly applicable to our work. First, they are able to identify a set of clustered points that directly relate to one of nine topics covered by SOTU speeches over time. Second, they are able to track the importance of these nine topics over time-based on how often they are mentioned in each speech.

Other papers use different methods to identify topic clusters in SOTU speeches, though without the significant time component differences explored in the prior paper. Crockett and Lee (2012) use 23 recent SOTU speeches to identify topic clusters. To do so, they employ two different approaches: text mining and topic approach. The difference between these methods is that the former creates an agnostic set of clusters, determined by an algorithm, whereas the latter involves some domain knowledge which defines the “major topics of interests”. These authors only provide an analysis of what are the most important topic clusters for this subset of texts.

Schmidt and Fraas (2015), in an article for *The Atlantic*, developed an interactive tool to analyze the frequency of words mentioned by each president, as well as the evolution of relevant topics or values that are important in United States discourse. For instance, they track the evolution of the words “Liberty” or “War”, and discuss the inclusion of gender topics by tracking the evolution of the word “her”. They also count the mention of foreign countries as a proxy for understanding isolationist tendencies of presidencies.

Our Approach

In this project, we seek to emulate the Rule, Cointet, and Bearman approach in order to compare it to traditional LDA topic modeling approaches. Each approach required the same pre-processing steps, which we’ll cover now.

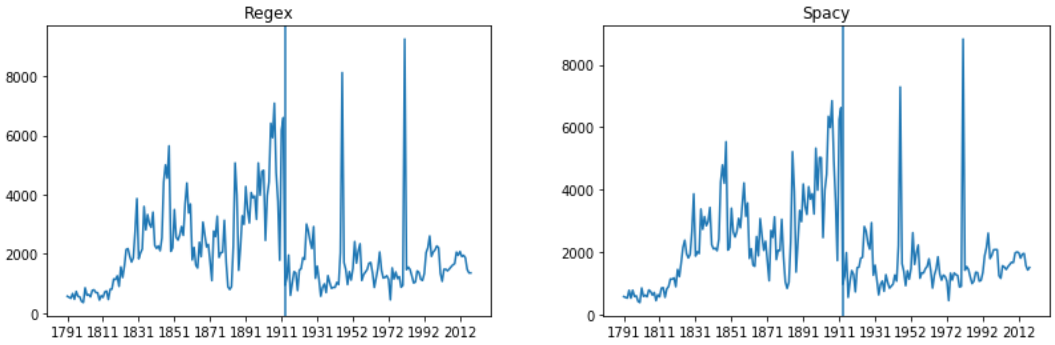
Pre-Processing

One of the most consequential assumptions of the approach pioneered by Rule, Cointet, and Bearman is that the best way to identify topics is to conduct the analysis using noun phrases rather than words, and doing counts at the paragraph level rather than the speech level. Given that there were roughly 400,000 noun phrases in the entire corpus, we also spent a non-trivial amount of time reducing the data to only 1000 noun phrases. Roughly speaking, the steps inherent in pre-processing were as follows: part-of-speech tagging, identifying noun phrases,

implementing similarity algorithms to reduce the number of noun phrases to 1,000, and outputting a co-occurrence matrix. We'll talk about each of these steps below.

Part of Speech Tagging and Identifying Noun Phrases

The first step in pre-processing is to define what we mean by a noun phrase, given that there could be different variations on what constitutes the structure of a noun phrase. We utilize two definitions and run our topic modeling approach on both: 1) a self-defined approach that identifies a noun phrase as a phrase that includes, in order, any number of optional determinants, any number of optional adjectives, and one or two nouns; and 2) a separate approach that's included in the Spacy functionality that has a more expansive definition of a noun phrase. After the next few preprocessing steps, the regex version had more single words in the top 1000 such as person, congress, state, and government, but also had a handful of noun phrases, such as fiscal year and the american people. The spacy version of the top 1000 words included more multi-word phrases such as 'the united state', 'congress', 'us', and 'both the government.' Additionally, both approaches led to similar counts in words for the speeches over time, we as can see in the graphics below.



Inducing Similarity

After part of speech tagging and identification of noun phrases, we lemmatized at both the word and noun phrase level. To lemmatize at the word level, we used Princeton's WordNet. We also applied this lemmatizer to every word within a noun phrase to make lemmatizing at the noun phrase level possible.

Our first thought for lemmatizing at the noun phrase level was compare all 750,000 words and noun phrases to every other word and noun phrase in the corpus. Being an n^n process, this solution ran painfully slow and therefore was not a feasible solution to employ. Our second attempt was broken into a two step process. First, we calculate the jaccard similarity at the character level for all 750,000 words and noun phrases to a random phrase, "america is wonderfully weird", and sort the words and noun phrases in order of their similarity to that random phrase. Our logic for sorting the list based on this calculation was that words that are mostly similar to our random phrase will also be relatively similar to each other. By sorting, we hoped to place similar words near each other to make the next step most effective. We then

split that list into a predefined number of lists. We tested different numbers of lists, including 50, 100, 500, and 1000, to find a balance between accuracy and speed. We found that a higher number of lists yielded a shorter run time, but also a lower accuracy, with mappings missed. Similarly, a lower number of lists increased run time, but also increased accuracy. We ended up using 100 lists which had a good enough balance between the two. Once we had the lists, we compared every word to every other word within each list. This comparison was also a jaccard similarity, but done at the word level. This means that we looked at the intersection of words in the two sets and divided it by the union of the two word sets. At first we tried doing it using a cosine similarity, thinking we could mirror other analysis being done in the paper, but those yielded terrible results.

Once we had the calculation, we mapped noun phrases to their classification. This turned out to be more difficult than expected, but also fairly straightforward once we figured out the logic and all possible corner cases. Since we were iterating over the same list of words twice, we needed to make sure we did not capture classify noun phrases to each other in different parts of the mapping dictionary. Again, while this is simple to understand conceptually, it took longer than anticipated (a couple weekends) to code and debug. Once we mapped the noun phrases to a classifier, we then changed all noun phrases to their mapped classification.

Using the lemmatized words and noun phrases, we then counted how often each word and noun phrase occurred in the corpus and limited the dictionary to the top 1000 most occurring words.

Co-occurrence matrix

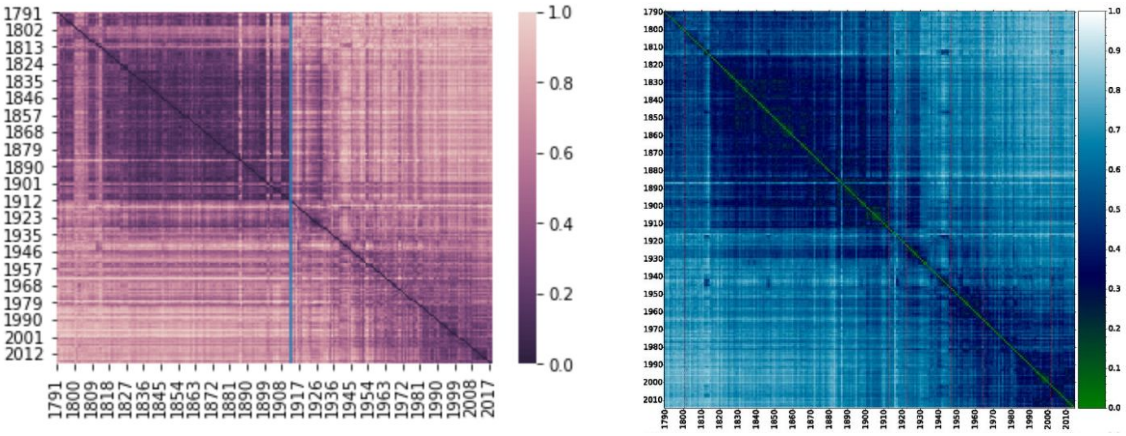
Using the top 1000 words and noun phrases, we constructed a co-occurrence matrix counting the number of times a noun phrase appears in the same paragraph with another noun phrase across the entire corpus. This matrix gives us information about how likely it is that one noun phrase or conceptual category shows up in a political speech in relation to other words. According to Rule, Cointet, and Bearman, this approach makes it easier to identify “relevant and interpretable higher level units of meanings endogenously, and to track their co-evolution through time”. That is, the main point of using this approach is to capture changes in the use of language across political speeches that can be dissimilar due to the passage of time.

Periodization

The next step in the process was to identify the different periods of time to see how topics changed across them. The purpose of this exercise is to capture how language changes over time and see if we can pinpoint changes in the language used to describe languages over time for all analysis following this. We do this using a multistep process. First, we calculate the term frequency - inverse document frequency (TFIDF) for each noun phrase in each speech. We do this by multiplying the frequency of the term in the speech by the log of the number of speeches in the corpus divided by the frequency of the word in the corpus (next page):

$$A^t(w) = f^t(w) \log\left(\frac{225}{f(i)}\right).$$

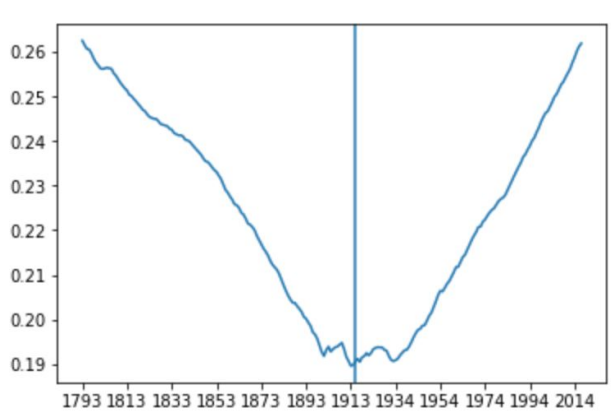
We next calculate a dissimilarity score, which is one minus the cosine similarity for every combination of years across the corpus and make a 227 by 227 matrix. This matrix was then used to create a heat map graphically representing how similar (or dissimilar) each pair of words' language is compared to the others. Even given the different definitions of noun phrases between our analysis and the paper's analysis, the findings are remarkably similar:

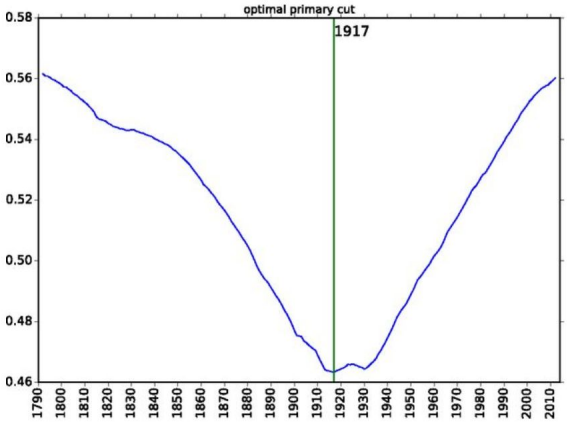


In both of these heat maps, lighter colors indicate less similarity and darker colors indicate higher similarity. The closeness in findings between our analysis and the paper's analysis suggests that there is an underlying relationship in the language of speeches over time and how each of the different presidents speak. Further, both analyses indicate that there is a relationship between major event in American history and the language used in States of the Union. For instance, there is a patch of lesser similarity that overlaps with the years the Second World War was occurring. Further, similar patches occur around the Vietnam War, around the Korean War, and around the occurrence of September 11, 2001. Lastly, as a larger analysis, language for speeches before 1914 were fairly homogenous, with the colors on the heatmap both being fairly dark. Next, between 1914 and the early 1980s, we see that most years do not generally have similar language. Then in the early 1980s, there seems to be a shift in heterogeneity back to language being fairly consistent, and this shift occurs around the beginning of the fall of Communism. Upon additional investigation, 1914 is also the first year States of the Union were given in person, which could explain the shift in homogeneity. These relationships must be explored in more depth, but at first glance, it looks like there is a relationship between the major events in American history and lexical shifts in language said in States of the Union.

Our next step in calculating periodization was finding the exact cut in the data when there was the biggest gap in homogeneity between the two cuts of the data. More specifically, we wanted to find the point in time when the weighted average of the heterogeneity score was at its lowest, indicating that we had found the year when the language in the years before and the years after

were most different. To do this, we looped over every possible cut of the sorted years, calculated the average dissimilarity score in the years before that cut and after that cut, then used those two calculations to find the weighted average using the two average dissimilarity scores and the number of years in each subblock. To calculate the average dissimilarity scores in the years before and years after, we summed the dissimilarity scores for every combination (not permutation) of the years in the before period and in the after period. To compare our findings with the paper, we plotted the weighted averages below. As you can see, the general trends in both graphs are the same, with our cut year defined as 1914 and theirs defined as 1917. Again, both of these years happen to coincide with the start and end of World War 1, respectively. One shortcoming of our graph compared to the paper's graph is that our y axis is almost half of the paper's. We attribute this to our difference in defining noun phrases; however, similar to the heat map above, the fact that the trends still generally hold even with different definitions of noun phrases indicates that these relationships do occur in the data regardless of how words or noun phrases are defined.





Lastly, we attempted to calculate secondary cuts where we repeated the same analysis as above for speeches between 1790 and 1913 and also for speeches between 1914 and 2018; however, while we found similar cut points, the trends we discovered in that work was not as convincing as the trends we discovered in the above graphics, so we did not pursue analysis on the secondary cut level.

Calculating Similarity in Co-occurrence Matrices

Given a co-occurrence matrix, whether it represents the top 1000 noun phrases across the entire corpus or for one of the two time periods determined, the next step of the approach is to understand how similar two noun phrases within each pair are to each other, with the understanding that this score can give a sense of whether or not the pair of noun phrases in question belong to the same topic.

We implemented two approaches to this step: 1) one pioneered by Cointet, Rule, and Bearman (CRB) which proved to be tough to comprehend for Mario's simple mind and performed poorly;

and 2) a pairwise cosine similarity measure that worked relatively well and provided coherent topics on the whole.

The CRB approach is represented by the following equation:

$$S(w_1, w_2) = \frac{\sum_{c \in \mathscr{W} \setminus \{w_1, w_2\}, I(w_1, c) > 0} min(I(w_1, c), I(w_2, c))}{\sum_{c \in \mathscr{W} \setminus \{w_1, w_2\}, I(w_1, c) > 0} I(w_1, c)},$$

Where w1 and w2 represent a pair of noun phrases, and I(w,c) represents a pointwise mutual information between a noun phrase and all noun phrases in the co-occurrence matrix. The strength of this approach is that it manages to factor in context: it first calculates the co-occurrence of the two noun phrases given how many times they occur in the global corpus, and then divides that number by how often that given pair appears relative to every other possible pair. In this paper, we were only able to implement the first part of this approach, to weak results.

The second approach implemented used a pairwise cosine similarity, which is a far simpler approach. Using sklearn's cosine_similarity command, we calculated the cosine similarity for each pair of noun phrases, implicitly assuming that noun phrases that had similar cosine similarity scores were likely to have appeared in a similar context over time. Whether or not that was an appropriate assumption to make is for posterity.

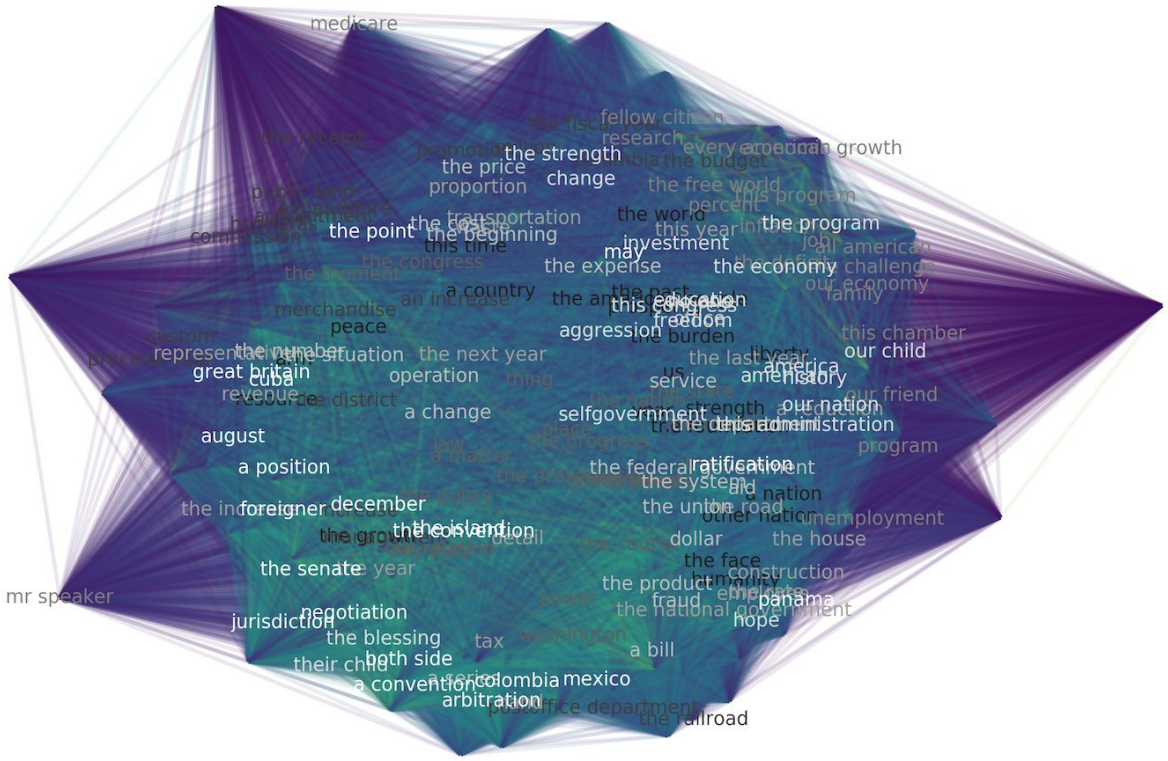
Semantic Networks and Community Detection

Once the co-occurrence matrices have been transformed using either the CBR approach to dissimilarity or pairwise cosine similarity, we draw a series of semantic networks in which the nodes are the 1000 noun phrases and the edges are the weight of the similarity score between the two noun phrases in question. In order to draw the semantic networks, we rely on the networkx library which handles most of the work.

Each node is positioned on the graph using a Fruchterman-Reingold force-directed algorithm. This algorithm moves nodes away from each other stepwise until the weight of the edges pulling nodes apart is equal to the weight of adjacent nodes repulsing any given node. In computational parlance, this is understood as minimizing the energy or force in the network system. Once these nodes are placed, edges are drawn between all nodes. Given that we have one million edges, we proceeded to only draw edges above an arbitrary value of 0.39 (this was a value found by CRB, so not entirely arbitrary but also not calculated directly by us.)

After drawing the semantic networks, we implemented two algorithms to identify clusters within the semantic network: a community detection algorithm and a k-means unsupervised clustering algorithm. The community detection algorithm largely induced either 3 or 4 topics, depending on decisions made earlier in the process as it relates to noun phrase definition and which similarity approach was implemented on the co-occurrence matrix. The kmeans clustering approach by necessity takes a predefined set of topics, which we determined to be eight topics after testing for different values.

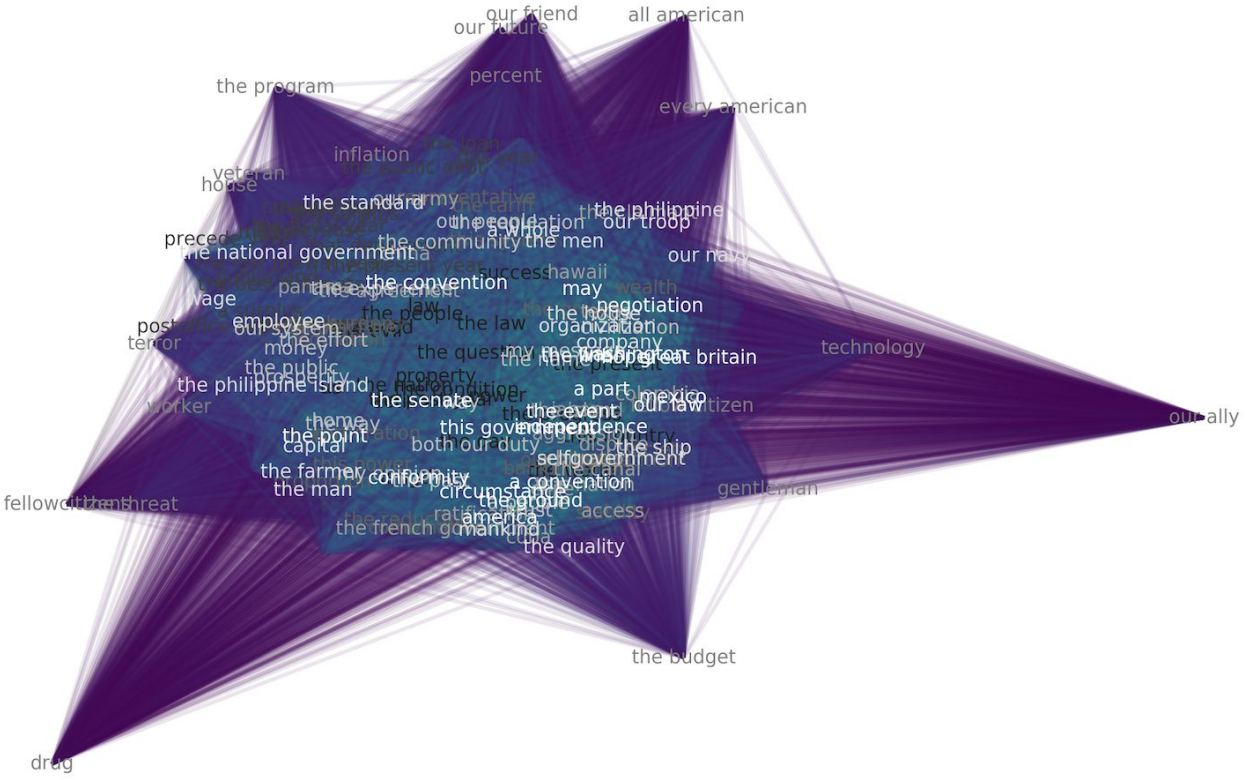
The results for the global corpus are interesting and point to the strategy implemented working relatively well. Of the iteration, the best performing one in terms of coherent topics was implemented using the spacy noun phrase definition, calculating the pairwise cosine similarity between words, and clustering together using a kmeans unsupervised clustering algorithm. It resulted in the following semantic network:

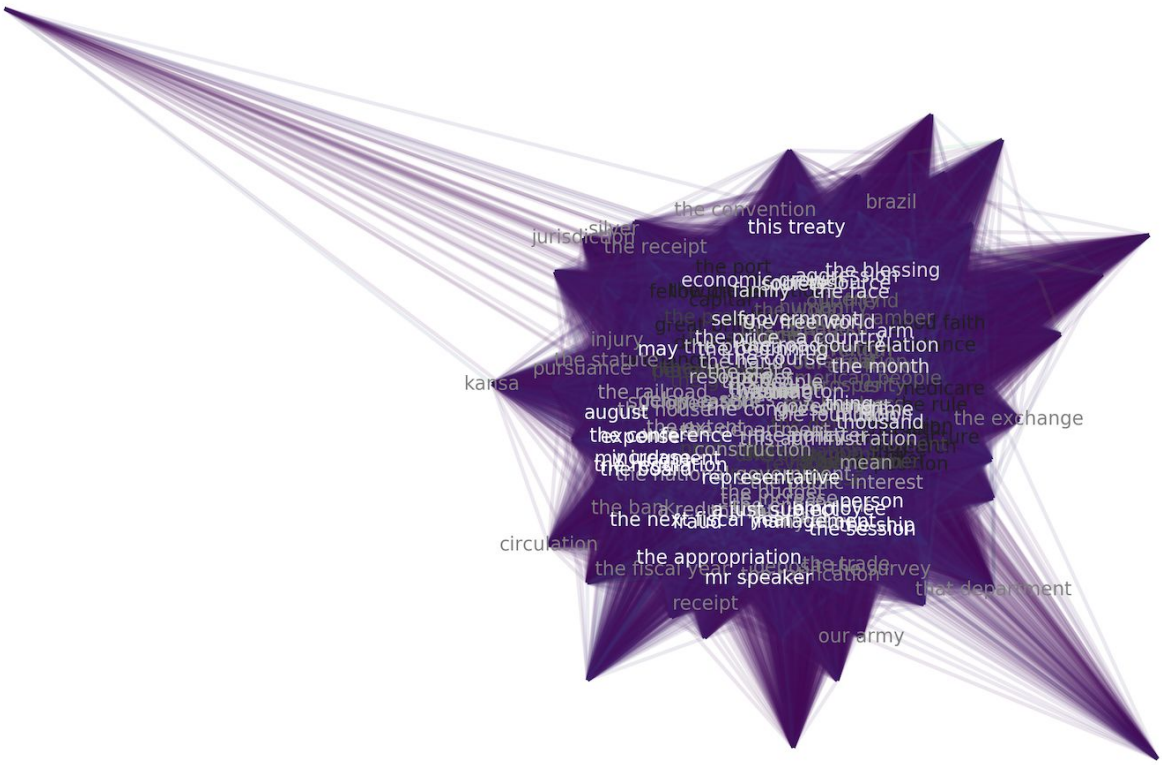


Though hard to visualize in the network graph, each color of node represents a cluster. While there appears to be some overlap in terms of clusters being near each other, we hypothesize this is due to the number of nodes placed in the graph (1000 nodes) and issues of representing multidimensional space in a 2-D representation of a network. A more appropriate method of measuring the coherency of this approach is to look at the words that make up each cluster. Rather than printing out 1000 words in this paper, please refer to the accompanying

presentation for the full list of words. From those lists, eight topics can be somewhat easily inferred: Purpose and Community, Fiscal Matters, Institutions, Domestic Issues, Expansion, The Economy, War, and Foreign Policy. While there are outliers in each, these topics seem to be fairly coherent and could possibly reflect the top 8 topics across SOTU addresses over time.

We also ran different iterations of parameters for the before 1914 and after 1914 time periods, and similarly found coherent topics that corresponded nicely to the time periods in question. The following graphs represent before and after, respectively:





The topics in the before and after periods are surprisingly similar, though are made up of different words (again, in Jupyter notebook or presentation): Foreign Policy, War, Economy, Domestic Issues, Institutions, Opportunity, Expansion, Future, and State-Making.

LDA

To compare the results obtained from modeling topics using a co-occurrence matrix, we also implemented a latent Dirichlet allocation (LDA) topic modeling analysis using the gensim library. Following Tomar from this blog, we explain the LDA approach as follows.

A first thing to understand is what LDA means. “Latent” refers to the fact that we do not know in advance the distribution of words over topics and documents, but we can think of the text (or documents) as generated by a set of topics. Second, Dirichlet refers to how we think about topics and the words that define them in terms of probability distributions. So, in this case, Dirichlet refers to the fact that topics in our speeches are a distribution of distributions: there is a distribution of topics across the documents, and a distribution of words (noun-phrases in our case) in each topic. Lastly, “allocation” implies that we assign topics to documents, and words of that document.

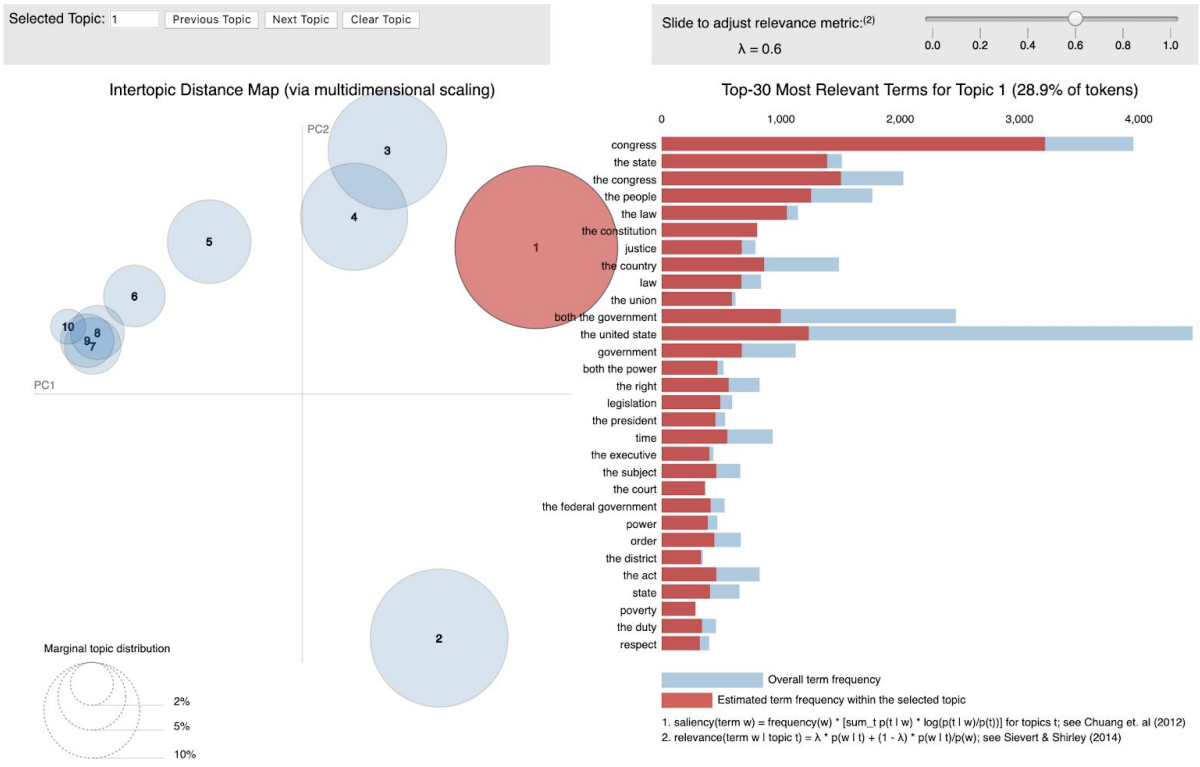
Then, the probability of a picking a word conditional on a document is P(w|d), and this probability can be decomposed into two matrices:

1. Θtd = P(t|d), the probability distribution of topics in documents
2. Φwt = P(w|t), the probability distribution of words in topics

This decomposition comes from the fact that P(w|t,d) = P(w|t), since the model assumes conditional independence of topics and documents. Then the probability of picking a word *w* coming from a document *d* is equal to $\sum_{t \in T}$ p(w|t)p(t|d), which is the dot product of (1) and (2) for each topic. The matrix of probabilities of words across documents can be represented as a singular value decomposition of the matrix **P(w|d)**, where the matrices are the distribution of topics in documents (P(t|d)) and distribution of words in a topic (P(w|t)). Then, the question is how to learn the weights of these matrices such that the likelihood of a given data (document) of belonging to a topic is maximized.

Since the goal of conducting an LDA topic modeling is to have a benchmark of our replication of Rule, Cointet, and Bearman's paper with as much consistency as possible, we used noun phrases to create bags-of-noun-phrases, and considered each paragraph as a document unit. Our first approach to extract topics using LDA was to use the entire corpus, without making any periodization in the corpus, to see what were the most salient topics for the whole history of the State of the Union speeches. The figures below depict the result of this first analysis.

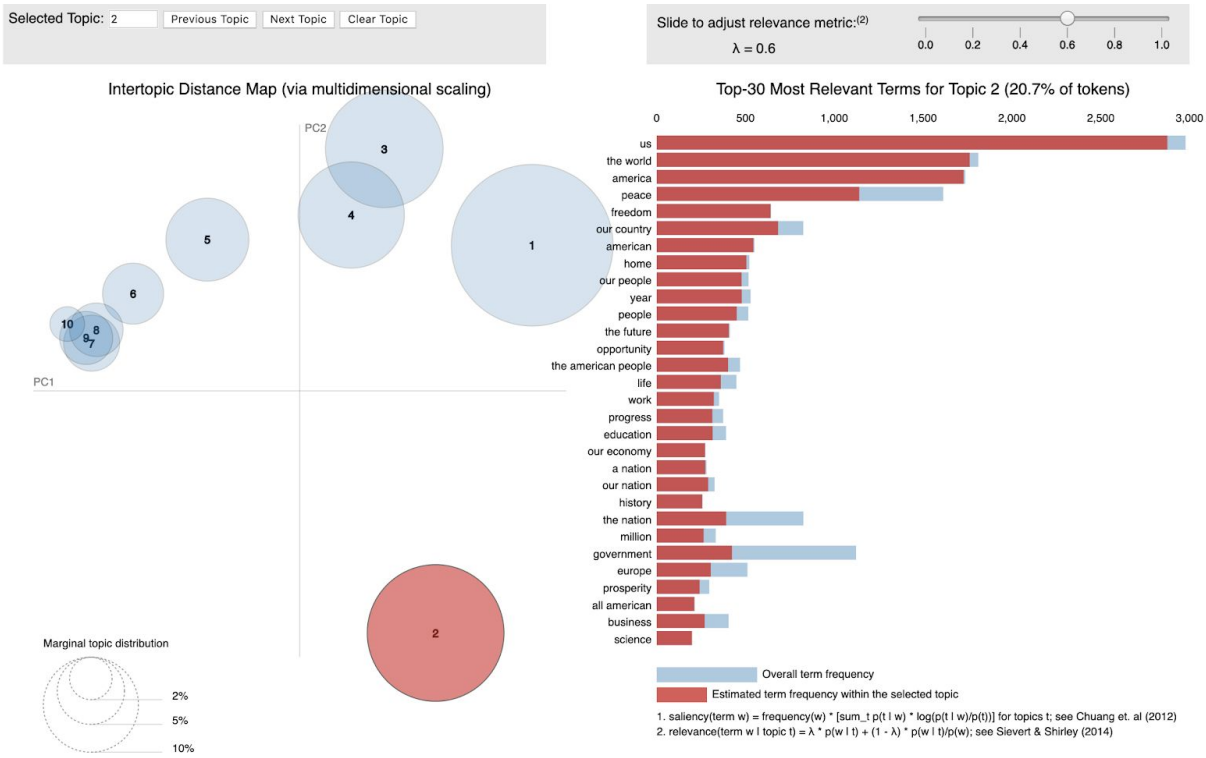
These visualizations were created with the package LDAvis. The left panel of the visualization aims to answer to the questions "How prevalent is each topic?", and "How do the topics relate to each other?" (Sievert and Shirley 2014). Each of the circles on the left represents one topic, and the size of each circle is proportional to how prevalent the topic is across all the documents. To plot the topics in a two dimensional space, the package computes the Jensen–Shannon divergence between topics. This is a measure of divergence or similarity between two probability distributions. Then, across two dimensions (PC1 and PC2), topics that overlap each other are more similar.



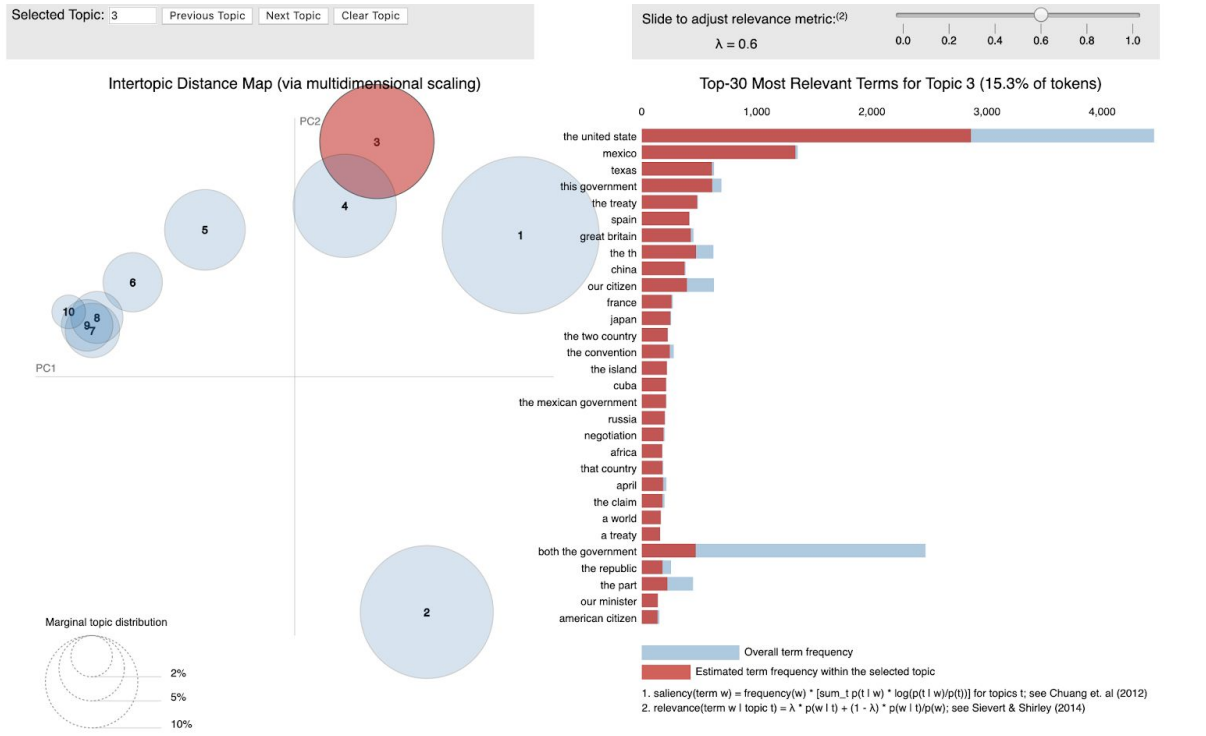
On the right-hand side we see a horizontal bar chart that depicts the ranked words that are more useful to interpret a topic. They aim to answer the question “what is the meaning of each topic”? In blue, it is depicted the frequency of a word across all the corpus, whereas in red we have the topic-specific frequency. Words in which red bars overlap with the blue ones means that that word is particular to a topic. The slide bar that appears on top of the bar charts allows the user to change the ranking of words by weighting how common that word is in the corpus, and how specific that word is given a topic. Values of lambda close to zero highlight exclusive words for a selected topic, whereas values close to one highlight frequent terms, but not necessarily specific to that topic. We adjust the value to 0.6, as suggested by the authors. This way, we can have a more confident interpretation of the importance of words within a topic.

So, for example, we could say that Topic 1, which is the more prevalent when we analyze the whole corpus, is about power, the members of the State or symbols of power: power branches, the Congress, the law, etc.

Topic 2 talks about America, and a sense of community, values of Americans, as well as about the future and the past of the country. Interestingly, it is separated from the other topics along PC2. This is important to underline since this topic is way different from the topics that are above the x-axis (PC1). The ones above PC1 are more related with some kind of policy, whereas the topic on Topic 2 refers more to abstract concepts and values.



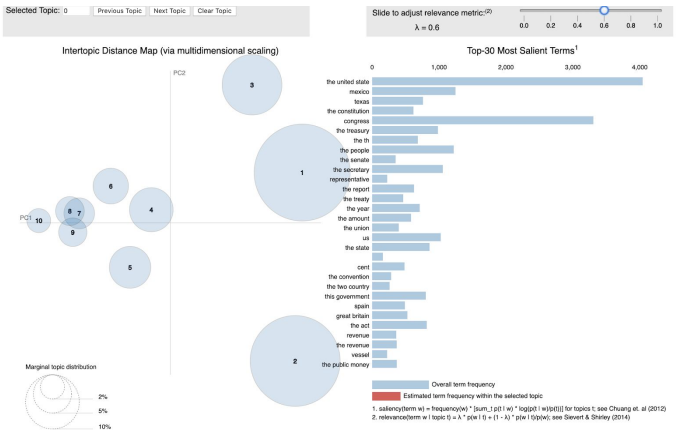
Topic 3 is clearly about foreign policy. Top words include countries (Mexico, Great Britain, China, etc.), as well as concepts related to trade and war processes (convention, treaty, negotiation, etc.).



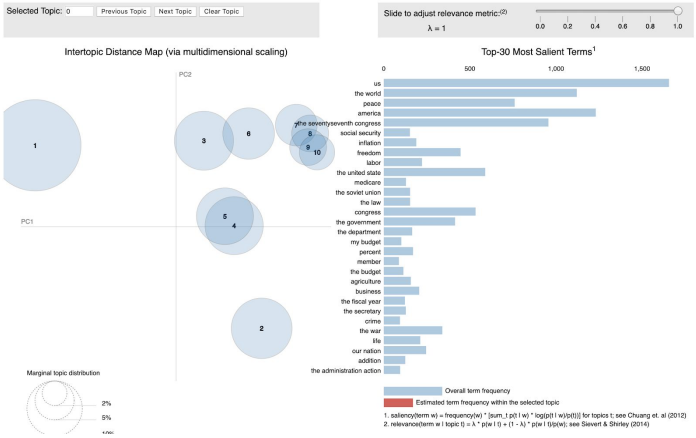
In Topic 4 (not shown), the most important words are navy and army, which are very specific to the topic (red area covers the total frequency in the corpus), and it could be said that this topic is mostly about war and its relation with funding sources. Topic 5 is about economics, and more specifically, about public finance. Topic 6 is about land, and words like mexico, indian, netherlands, or acre are very specific to this topic. In Topic 7, it is difficult to distinguish a clear topic: healthcare is the more salient noun phrase, followed by words like democracy, budget, college, and “this war”. Topic 8 is about US “threats” in the foreign policy: the “Soviet Union” appears in first place, and we can also find noun phrases like “Iran,” “Nuclear weapon,” “tyranny,” “Western Europe”. Topic 9 is interesting: it is also about economics, and it shares some words with topic 5, but it is more about economic growth: among the top noun-phrases we find “inflation” in the first place, followed by “economic growth”, “percent”, “export”, “crime”. For topic 10, it is indistinguishable one main topic among the noun-phrases it represents.

Though this analysis gives some insights about which topics have been present across all State of the Unions, it is hard to believe that the relevant issues in the public arena have remained more or less the same along 200 years. One of the limitations of this analysis is that it does not account for changes that have occurred over time, which is what Rule, Cointet, and Bearman intended to solve with his co-occurrence matrix. However, knowing that there is a point of inflection in how State of the Union speeches are similar to each other in the year 1914 (which is part of the results we showed, too), we split the data into before and after 1914, and conducted the same LDA analysis. We show the topic modeling per period below.

Before 1914



After 1914



Just for the sake of comparison in the distribution of topics, we show the graphs next to each other. As it is evident, the distribution of topics between the two periods is different in terms of its distribution and in how they are similar to each other. For example, along the PC1 axis, Topic 1 in the post-1914 analysis is on the left-hand side, as it is the cluster of topics before 1914. The opposite occurs with the cluster of topics on the right-hand side. Table 1 lists the topics found in each portion of the data.

Before 1914	After 1914
1. Abstract values, State-Nation	1. American values, community
2. Public Expenses	2. Budget assignation
3. Foreign policy (war)	3. Foreign policy (threats)
4. Foreign policy in America (the continent	4. Employment, Growth
5. Indistinguishable	5. Indistinguishable
6. Indistinguishable	6. Medicare
7. Economic Sectors	7. Not clearly distinguishable (weakly related to business creation)
8. American Citizenship	8. Foreign policy
9. Domestic industry	9. Foreign Policy
10. Territorial Expansion (not clearly related to war)	10. Indistinguishable

Interestingly, the first three topics are very similar in terms of the topics themselves (i.e. Topic 1 in before 1914, and Topic 1 after 1914). Both Topic 1 in both subsamples are about abstract political categories (and this topic in both cases is way more different than the rest of them). Both Topic 2 talk about how to assign public resources, and number three is about foreign policy and how the US responds to external threats. Then the rest of the topics and its ranking differ between the samples.

Evaluation of LDA

The way to determine the right number of topics is to calculate the topic coherence score for many LDA model when varying the number of topics. This work did not include such analysis due to the computational time it requires, but future work will include an analysis of topic coherence and the right k topics.

Lessons Learned

Overall, we learned that working with language can be ambiguous, making it difficult to make analysis decisions and determine the “best” solution. For our project, this was further compounded by language changing over time in our states of the union speeches.

We also learned the power of using linear and other non-exponential methods. When re-classifying noun phrases, our first instinct was to compare every word and noun phrase to every other word and noun phrase to ensure the highest degree of accuracy. We discovered that that took forever. To ameliorate this issue, we explored a multi-step process that included a linear pass of the data and then exponential passes over smaller chunks of the data. If we had had time, we could have randomized the chunks or explored other ways of making the function more efficient while achieving better accuracy.

Another lesson learned is that the larger and more distinguishable topics in the pre-1914 and post-1914 are incredibly similar, but the two analyses use different sets of noun phrases to identify those topics. For instance, there are three topics that are common across all periods, all of which talk about the Republican values in America (meaning the 'res publica'). We found this more interesting than talking about the differences between subsamples (eg. the presence of Medicare in post-1914, or the land expansion in pre-1914). For LDA specifically, it was interesting and reassuring to find that LDA distinguished subtopics. For instance, "Economy" was broken down into economic growth and macroeconomic variables (such as public finance).

Lastly, in the topic modeling section of the work, we learned about different approaches to topic modeling. In particular, through our exploration of the bespoke approach to modeling deployed in the paper we tried to emulate, we learned about the composition and importance of co-occurrence matrices in detecting topics, different methods for calculating the similarity of pairwise noun phrases, and experimented with different visualizations that included semantic networks.

Future Work

There is a significant amount of work left to be done. First, which is alluded to in the write up above, the relationship between major wars or events in American history and lexical shifts States of the Union should be explored in more depth, possibly using causal inference. Second, we would want to explore different definitions of noun phrases, or the use of other types of phrasing (such as verb phrases), and the impact those different noun phrases have on the topics identified. It would also be interesting to conduct the analysis with different numbers of top words - in this paper we use the top 1000 words, but we would be curious about the effects of using the top 2000 words, or more. That said, that would be more computationally intensive, so we would have to account for that. For the LDA analysis specifically, we would want to explore and tune the model further to find the appropriate number of topics to include, implement the evaluation, and explore changes in the topics found when we use different definitions of noun phrases.

For the paper's approach to the analysis, we would want to refine the similarity measure deployed by the authors. Our current approach, due to computational and lack of clarity issues, only deployed the first half of their approach -- it doesn't include the contextual comparison in its equation. Any future project would need to capture the complexity of the equation and fully

account for context. What's more, additional work is needed in visualizing the output in a clearer way, as the current network graphs are crowded and don't account for challenges due to dimensionality.

Team Breakdown

Roughly speaking, the work was divided into three steps: literature review, data cleaning and preparation, topic modeling approaches, and visualization of results. The work was divided in the following way.

Data Cleaning and Preparation:

- Read in and prepare data for noun phrase collection: Mario
- Data Exploration: Alix, Aleister, and Mario
- Defining and implementing a definition of noun phrase composition: mainly Mario, some Alix
- Implementing string similarity algorithms to group noun phrases together: Alix
- Finding top 1000 noun phrases across time: Alix
- Periodization: Alix
- Co-occurrence Matrix build: Aleister
- Paper's approach to topic modeling: Mario
- LDA approach to topic modeling: Aleister

In total, we worked nearly 90 hours on this project. The bulk of that time was spent trying to understand the approach the paper took to topic modeling and mapping out what strategies we'd follow to get a workable product. Along the way, we each learned about new libraries or enhanced our understanding of prior libraries. Alix learned a great deal about jaccard similarity and other string matching algorithms, thought critically about computational efficiency in grouping noun phrases together, and deployed an interesting TF-IDF and cosine similarity approach based off the paper to identify periods. Aleister learned about building co-occurrence matrices from scratch, LDA approaches and libraries, and D3 visualization libraries for LDA results. Mario learned about spacy and other methods for part of speech tagging and identifying noun phrases, pairwise proximity scores deployed by the paper as well as cosine similarity, graphing network graphs using the networkx library, and unsupervised clustering methods.