NLP: History of the State of the Union (SOTU) Address

Ali Gowani, Ksenia Luu, and Mitch Thompson

Master of Science in Data Science, Northwestern University

MSDS 453: Natural Language Processing

Dr. Alianna Maren

March 11, 2020

Abstract

A variety of algorithms and methods have blossomed quite recently with regards to natural language processing (NLP). Through this paper, we provide a summation of acquired and pragmatic understanding on selected algorithms within a natural language pipeline as applied to a corpus comprised by historical State of the Union addresses. To do this, the evolution of an NLP pipeline from problem identification, corpus development, and selection of algorithms to applied techniques, and results will be accomplished. We also explain the purpose of the various algorithms in association with the problem and put forward a detailed review and analysis of insight into the corpus provided by the algorithmic results. Each team member contributed on nearly every part of the research project. However, some portions were predominantly accomplished by specific teammates. These portions include Linear Discriminant Analysis accomplished by Ksenia Luu, Decade Analysis accomplished by Ali Gowani, and Sentiment Analysis accomplished by Mitch Thompson.

Keywords: natural language processing, State of the Union, reference term vector, term frequency times inverse document frequency, k-means clustering, linear discriminant analysis, Dirichlet distributions, sentiment analysis, ontology

NLP: History of the State of the Union (SOTU) Address

The State of the Union (SOTU) is a constitutionally mandated message through which the President of the United States "shall from time to time give to the Congress Information of the State of the Union, and recommend to their Consideration such Measures as he shall judge necessary and expedient" (U.S. Const. art. II, § 3). While not articulated to be an annual requirement, history accounts for at least one per year since the first delivery in the year 1790. The modern SOTU, through which the sitting president directly addressed the American population, was born out of the combination technological advances in mass communication and a renewed effort to address the U.S. Congress in-person. Every SOTU since has been an opportunity to lay out the priorities of the administration in addition to the health and wellbeing of the nation.

Employing the technique of Natural Language Processing (NLP) against such a historically rich and instinctively diverse corpus introduced wide ranging choices of algorithms and nearly unbounded paths to explore. As such, the corpus afforded the greatest opportunity to perform investigative and exploratory applications with NLP processes and algorithms. To narrow our scope and provide analytical direction, we posed the following problems:

1. Is an NLP pipeline applicable to a corpus comprised by a history of SOTU addresses?
2. How can we extrapolate key areas from the SOTU speeches by applying different NLP techniques?
3. Are there regular occurring themes or topics across the history of the SOTU?
4. Does the language used in these messages change over time? If so, why or how?
5. Is it possible to implement topic modeling against the history of SOTU messages?
6. What could account for variations in tone or sentiment in each SOTU?

Discussion

Data: SOTU Corpus Development

The complete SOTU corpus consists of 234 documents containing over 1.7 million words, of which 28,195 are unique (Borevitz, 2013). However, for the purpose of this analysis we down-selected to 102 documents between the years 1791 and 2019 which encapsulates 23 unique presidents composed of 8 Democrats and 15 Republicans. These documents were then tokenized and cleansed by traditional means such as removal of punctuation, non-alphabetic characters, and limiting to words with more 4 characters. However, the removal of unwanted words that carries no value to the meaning of the document was not as straightforward. We soon discovered the list of stopwords included as part of the Natural Language Tool Kit (NLTK) was deemed to be woefully inadequate. We found words such as *hereafter*, *according*, *unfortunately*, *toward*, and *welcome* were so common across the history of the SOTU, that their inclusion into a Reference Term Vector (RTV) would severely hinder our modeling. As such, we derived a listing of words commonly found within the corpus and added it to the NLTK listing. This revised list of 558 words was then used to strip the words of no value from the corpus prior to conjoining the documents together to form the corpus. Conversely, we hypothesized terms like *nation*, *citizens*, *economy*, or *prosperity* would rank higher in the RTV due to their meaning and contextual placement within the texts. In the end, the corpus utilized for this project included 25,293 tokens. A word cloud generated from the preprocessed corpus added further weight to this hypothesis (Appendix A).

Challenges Faced and Overcome

The majority of the challenges that our team faced throughout this project were centered around data. We encountered our biggest roadblock while compiling our original corpus. Our

initial approach was to develop a classification model for a teammate's current employer, Riverside Research Corp. The model would map the Requests for Proposals (RFPs) from US Government websites to the core competencies of Riverside Research. Implementation of this model would eliminate hours of manual work resulting in an increased ROI for the company. To create the corpus of appropriate documents, we had to manually pull data from the US Government website that aligned with areas of Riverside Research. After spending many hours manually downloading the data and putting it into an acceptable format our corpus was still not big enough to train an accurate model. We came to the conclusion that we did not have enough resources to pull and aggregate the needed data in time to meet our project deadline. We completely scrapped our project idea and decided to move in a different direction. It is true that getting the data was the most time-consuming portion of the analysis!

We ran into another problem while reformatting the collection of text files for our second project idea into a single corpus file. We kept running into character limits in excel and had difficulty getting the data into the right format to run through our python script. Once we were able to ingest the data, we went through several iterations of cleaning the corpus by removing stop words and common terms used in SOTU addresses that hindered the clustering outputs.

Along with multiple data challenges, some teammates faced a steep learning curve associated with the collaboration tools and software we chose to use for this project. The team dedicated hours to learning Git methodology, GitHub and Jupyter Notebook tools, and various python packages in order to successfully execute this project.

Algorithms and Key Features

During the team’s weekly video conference calls, we agreed that it would be interesting to explore commonalities in topics discussed by different presidents across decades. We decided

to use the Scikit-learn K-Means clustering algorithm to group similar documents and Gensim Linear Discriminant Analysis (LDA) to understand the main topics across the total corpus.

TF-IDF. To prepare the dataset for clustering and topic modeling we used Scikit-learn's `Tfidftransformer` to convert a corpus of raw documents into a matrix of Term Frequency times Inverse Document Frequency (TF-IDF) values. The TF-IDF score is a balanced metric that helps identify terms that are interesting within the context of the corpus. A measure of the TF-IDF is calculated by multiplying the term frequency within the corpus and inverse document frequency resulting in a well-rounded importance score (Lane, Howard & Hapke, 2019, p.90).

K-means Clustering. K-means is an unsupervised learning algorithm that divides the dataset into a pre-defined number (k) of non-overlapping groups using the distances between the specified points. Using the pre-defined number of clusters, the algorithm chooses centroid points and tries to cluster data points with the closest proximity while keeping the other clusters as far as possible. Then new centroids are created by computing the mean of the data points in each cluster. The distance between new and old centroids is calculated and the process is repeated until the delta is less than the defined threshold (Pedregosa, 2011).

Linear Discriminant Analysis. Linear Discriminant Analysis (LDA) is an unsupervised learning algorithm often used to identify topics that best represent a set of documents. LDA reduces the dimension of text data by modeling each document as a distribution of topics and each topic as a distribution of words, mirroring Dirichlet distributions. Working backwards from the prescribed number of topics (k), LDA randomly assigns each word to one of *k* topics to get an initial representation of topics and words. It then adjusts word placement by calculating the probability of each topic in each document and the probability of each word in each topic. LDA

will adjust and move words to different topics until it reaches an optimal state of topic and document representations (Blei, Ng, & Jordan, 2003).

Sentiment Analysis. TextBlob is a Python application programming interface (API) to a collection of algorithms used to perform natural language processing activities such as noun phrase extraction, translation, part of speech (POS) tagging, and sentiment analysis. The library behind the API contains two implementations of sentiment analysis: the NaiveBayes analyzer which comes under NLTK classifier and the pattern analyzer which is dependent on the Pattern libraries. Our use of TextBlob was limited to sentiment analysis which, by default utilizes the Pattern libraries and includes the values of polarity and subjectivity. When called, the function assesses the adjectives contained within the given sequence and returns a (polarity, subjectivity) tuple.

Key Features of Code. Each algorithm is called and applied to the documents and corpus in sequential order within the NLP pipeline. The TF-IDF algorithm is part of the SciKit-learn module and instantiated by assigning a variable to the `TfidfVectorizer()` type. Of the numerous parameters available to this method, we discovered a combination of the ngram range, the maximum threshold, and a cap on the maximum features provided the optimal output. Called by using `ngram_range()`, the range provides lower and upper boundaries for the range of sequenced words as well as the individual words contained within the sequence. The range for this analysis is (1, 3). Next we utilized the `max_df` parameter to ignore terms that are common to over 80% of the documents in the corpus. This helps to set aside those terms that our larger stopword list may have missed. We also relied on the `max_features` parameters to build the vocabulary with consideration to only the top 50,000 features. Finally, the `fit_transform()` method is then applied to the corpus and returns a term-document matrix which can then be

processed further and prepared for clustering. Likewise, the k-means algorithm is also encompassed in the SciKit-learn module and called in similar manners. It also has several parameters but the *k* value, or the number of centroids, is the heart of this algorithm. Unfortunately, initial attempts to discover an optimal value of *k* through the established elbow method came up empty as there was not a clear bend. As a result, the value of *k* for this analysis was arbitrarily set at 6 along with a random state. Identifiable features to the genism LDA algorithm used in this project included the `LdaMulticore()` and `CoherenceModel()` functions. The first function here allows LDA model estimation from a training corpus as well as inference of topic distribution on any new documents that may be introduced after training. However, our use of the multicore function vice the parent `LdaModel()` function. By selecting multicore, we were able to acquire the same training optimization as the parent function at faster rates through the use of parallelized operations. Additionally, calculations of topic coherence were made by using the `c_v` coherence measure. Röder et al. (2015 suggests the meaning of coherence is set forth as "a set of words measures the hanging and fitting together of single words or subsets of them" (para 3). In their introduction of the `c_v` measure, Röder et al. based on a sliding window which derives a one-set segmentation of the top words and an indirect confirmation measure that uses normalized pointwise mutual information (NPMI) and the cosine similarity (Röder, Both, & Hinneburg, 2015). The one-set segmentation of the top words leads to a calculation of similarity between every top word vector and the sum of all top word vectors. Coherence is the arithmetic mean of these similarities. Finally, key features of the TextBlob API used for sentiment analysis included the methods sentiment and polarity. Polarity defines the phase of emotion and falls within the range of -1.0 and +1.0. Similarly, subjectivity helps to determine the personal states of the speaker including emotion, belief, and opinion and is scored

between 0.0 and 1.0 (Centre for Computational Linguistics and Psycholinguistics, 2010). Words without positive or negative leanings are considered neutral, scored appropriately, and not factored into the analysis. Application of TextBlob features was applied to the corpus via a Python lambda function.

Ontology

When addressing problems of any domain with artificial intelligence, understanding of that domain is critical to the application. Moreover, a well-thought out and designed ontology will also help and maintain a proper focus on the task. In the effort to acquire the former and hold the latter, we developed an ontology for the SOTU event to serve as an explicit description of concepts and relations therein (Appendix B). The ontology, together with the set of individual SOTU texts within our corpus, then constitutes a knowledge base from which to work from. We considered this an essential step to understand the domain and process the corpus with the aforementioned algorithms (Estival, Nowak, & Zschorn, 2004).

Algorithmic Results and Analysis

The complexities when dealing with a Bag of Words model is predominately accomplished in an unsupervised and therefore, expectation of results during initial iterations through the pipeline should be nominal at best. This is due to limiting the preprocessing steps prior to applying the common algorithms in order to detect and take advantage of what the dataset offers. This section will review the initial pass results, underlying causes that led to poor algorithmic performance, and techniques employed to achieve improved performance.

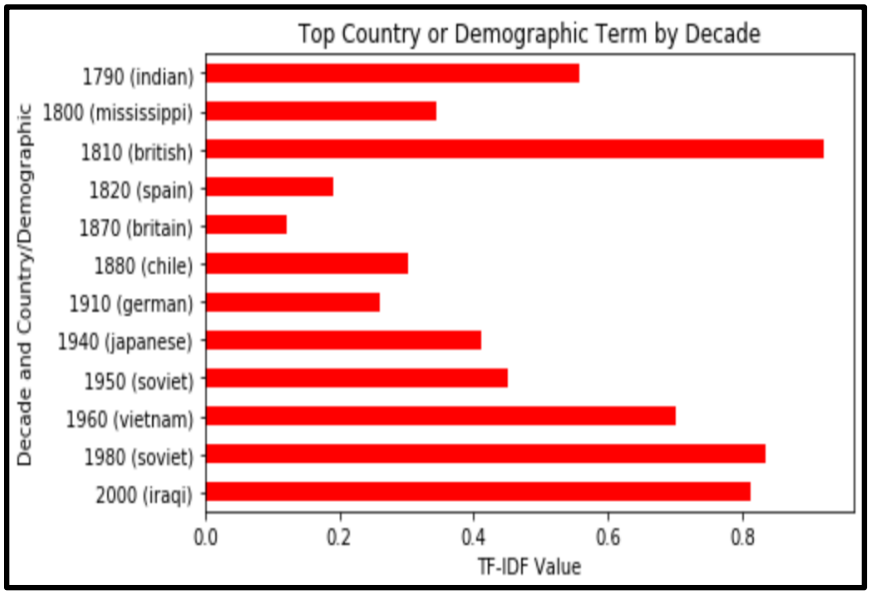
Review of Algorithmic Results

TF-IDF: while we conducted a TF-IDF analysis on our entire corpus, we decided to take a different approach. Usually, each President may have a key area of concern that defines his

presidency. However, there could be areas, especially, countries and people, that transcend political ideologies (e.g.: Republican, Democrat, Federalist, etc.) and even the President's entire tenure. We decided to explore whether terms would stand out based on an entire decade of SOTU speeches rather than by each President's SOTU speech. We found that in 1820, Spain was a country of concern and in 1910 Germany (German) was a key term addressed in SOTU speeches. Many of these terms can be attributed to significant world events (e.g.: World War 1, World War 2, etc.). Japan in 1940 SOTU speeches stood out, as well as, Soviet (Union) in 1950. As we embarked into the new century, Iraq (Iraqi) was a country that defined the focus for SOTU speeches in 2000. What we found quite intriguing is that we can associate timeline to our analysis that can put terms, clusters, etc. in appropriate context as depicted in Figure 1. See Appendix C for Top 10 terms of each decade.

Figure 1

Key country or demographic from each decade of SOTU speeches



K-means Clustering: Using TF-IDF corpus matrix, we ran k-means clustering to analyze the relationships in topics covered by different presidents and to understand if there were

common themes across decades. We evaluated clusters of 4 all the way up to 10. During the first few clustering runs we realized that there are many common terms across most SOTU speeches that resulted in very noisy clusters. We observed terms like *american, tonight, today, freedom, government,* etc. show up in the top ten terms across all clusters. To reduce the noise, we added these terms to the stop-words dictionary to remove them from the corpus. After removing common presidential-speech terms, we saw much cleaner clusters and could better understand the overall topic of each cluster. We concluded that 5 clusters gave us the best results with the highest cross-validation score of 74% as well as the lowest number of documents that could fit into more than one cluster.

Table 1

Clusters and 10 Terms per each Cluster

0	1	2	3	4
program	commerce	treaty	matter	terrorist
federal	british	award	industrial	child
economic	subject	majesty	interstate	family
billion	vessel	chile	labor	budget
problem	spain	relation	tariff	reform
budget	object	convention	action	school
effort	militia	mexico	method	iraqi
soviet	treaty	subject	business	worker
defense	provision	diplomatic	condition	health
increase	indian	republic	price	terror

To understand how well each document fits into its assigned cluster, we created a heatmap that shows the distribution of average TF-IDF values for each cluster of terms calculated for each document. Refer to Appendix D to see the heatmap for the final set of five clusters. You can still

see some overlap of documents between cluster 0 and cluster 4, which could mean that those speeches cover both topics. Looking at top ten terms per cluster in Table 1 we can see that there could be some overlap in topics. Term *defense* in cluster 0 and terms *terrorist* and *terror* in cluster 4 could be correlated with the topic of national security. Seeing this, we decided to perform LDA topic modeling to help us clearly understand the main topics in our corpus.

In the second portion of our cluster analysis we were curious to understand if there was a relationship between the time period of when the speech was written and the cluster assignment. We plotted each SOTU document by year and color-coded according to the assigned cluster.

Figure 2

Scatter Plot of Clusters by Year

As illustrated in Figure 2, documents in the same cluster also tend to aggregate around the same timeframe. This makes sense as each President usually focuses their attention on the highest-

impact issues at the time. Some issues span over multiple presidential terms and are usually discussed in high-profile speeches like SOTU. Cluster 0 stood out as it has *soviet* in the top ten terms and SOTU speeches range from 1940's, beginning stages of the Cold War to the 1990's during the Gulf War.

Linear Discriminant Analysis (LDA): After seeing some overlap in documents across clusters we attempted to run LDA analysis to better understand the main topics within the corpus. Using both the bag-of-word and TF-IDF values we ran LDA for 3 all the way up to 8 topics. We concluded that 4 topics provided the best results and the top terms per topic using the bag-of-words method are shown in Table 2. We can decipher that topic 1 focuses on the future state of the country in terms of peaceful relations and national security, topic 2 focuses on economic stability and American families, topic 3 covers military and current enemies, and topic 4 outlines the theme of treaties and relationships with other nations.

Table 2

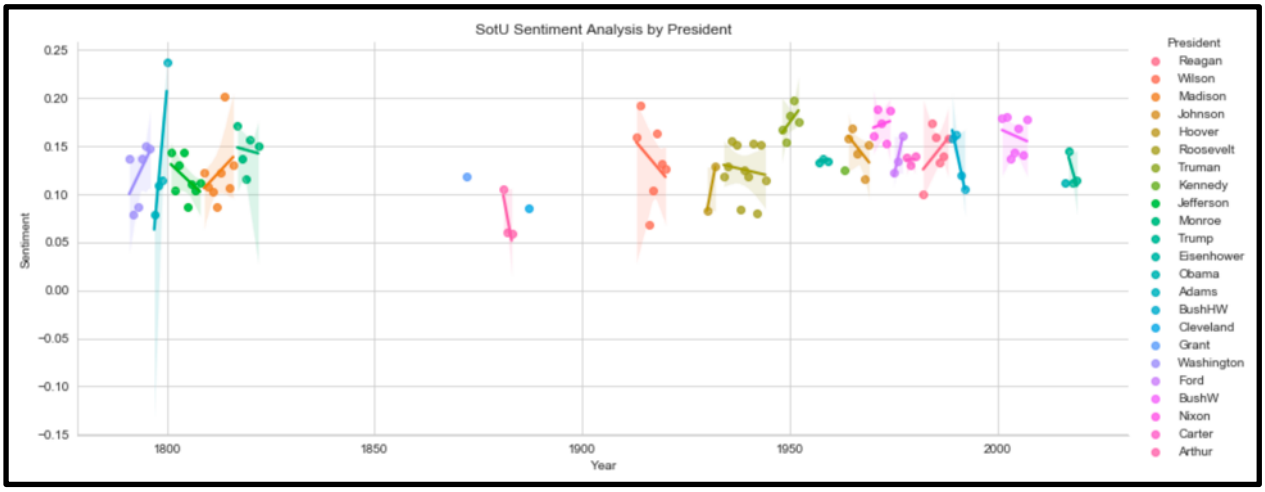
Clusters by Year

Topic 1	Topic 2	Topic 3	Topic 4
peace	program	citizen	treaty
economic	federal	peace	power
power	peace	million	subject
action	security	present	citizen
problem	economy	force	interest
security	child	interest	commerce
force	economic	power	present
federal	budget	british	measure
program	family	enemy	spain
future	future	service	consideration

Sentiment Analysis: Our sentiment analysis analyzed each token within the corpus and assigned a rating based on the positivity or negativity of the connotation. As stated earlier, the TextBlob API relies on a scalar range from -1.0 to 1.0 with the most negative words ranked lower and the words with a very positive meaning at the higher end or near 1.0. The sentiment score of each SOTU within the corpus is depicted in Figure 3. As originally hypothesized, all of the SOTU addresses carried a positive sentiment with scores ranging from 0.05 to over 0.20.

Figure 3

Sentiment Analysis by President



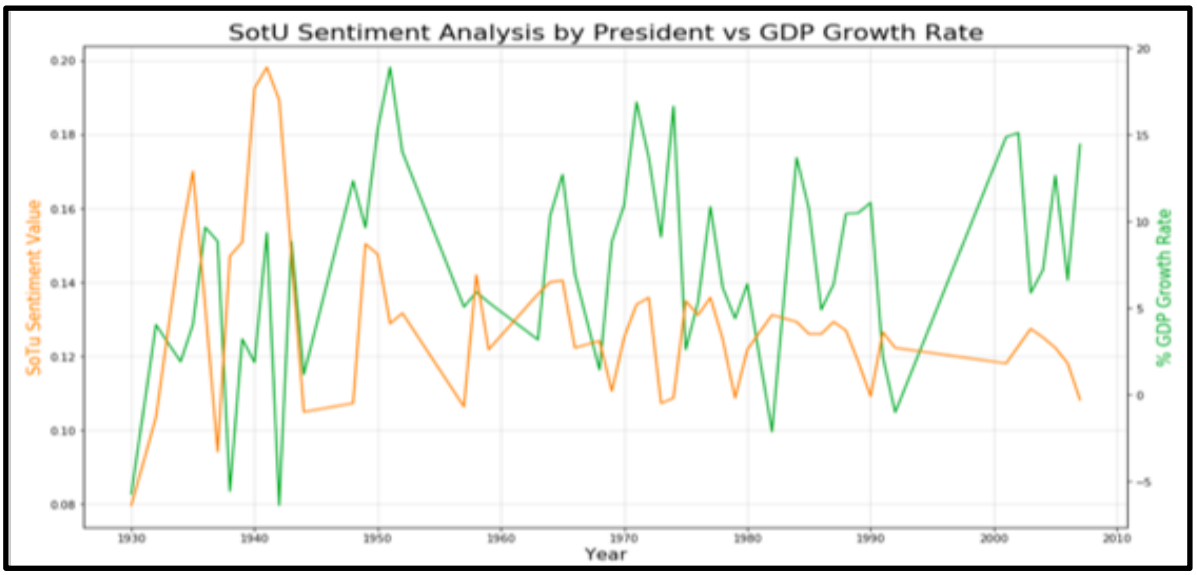
Analysis

Particularly noteworthy in the sentiment analysis of the SOTU is the discovery that only 9 of the 23 presidents represented in the corpus increased the positivity of their sentiment through their time in office. However, it must be noted that the entirety of any president's time in office may not be captured in this reduced corpus. Additionally, the dramatic increase of positive sentiment during the presidency of John Adams from 1797 to 1801. As the second sitting president, his term in office saw rise to the two-party system and the subsequent differences in policy. Through this time as president, he attempted to stay above the partisan squabbling which may have had a

profound impact on the terms he chose to use when addressing the Congress. President Adams was not without his share of scandals and threat of war though which brings the analysis to another question: What could account for variations in tone or sentiment in each SOTU?

Figure 4

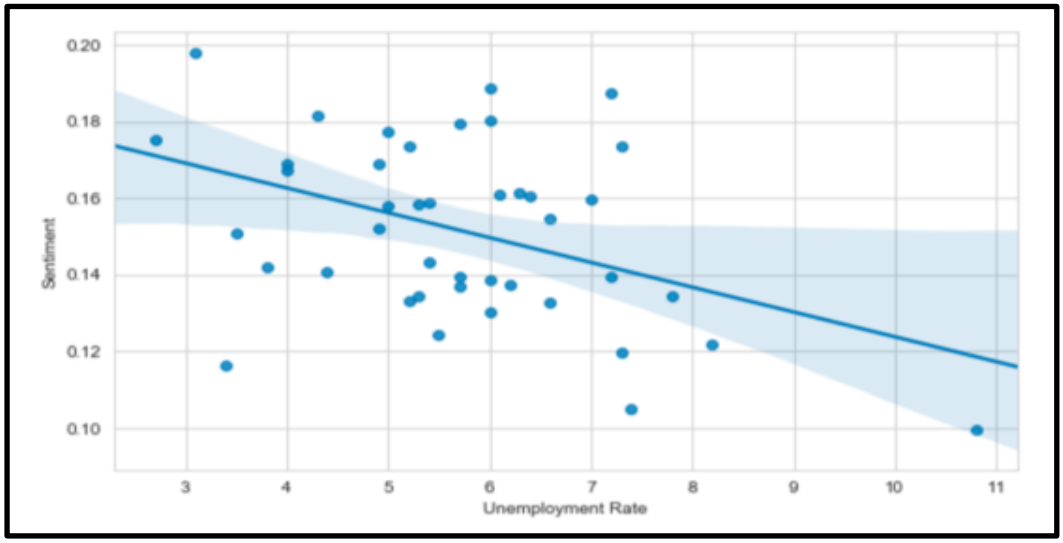
SOTU Sentiment Analysis by President versus GDP Growth Rate



To address this question, we pulled in two additional sources of data that are regularly considered to be indicators of the health of the nation, namely the growth of the gross domestic product (GDP) and employment rations. Surprisingly, we found no correlation between GDP growth and the sentiment of the SOTU as seen in Figure 4. A very good example of this is the GDP growth between 2005 and 2019 compared to the SOTU sentiment of the same time. Conversely, there is an inverse trend between the unemployment rate and sentiment as shown in Figure 5. While this is not evidence for causation it is an expected discovery once contemplated thoroughly. If indeed unemployment of the nation's populace is going up, any demonstration of high positivity in the annual SOTU address may be received as out of touch by unhappy voters.

Figure 5

SOTU Sentiment Analysis by President versus Unemployment Rate



Summary of Key Insights and Findings

This paper described our work in a natural language pipeline as applied to a corpus comprised by historical State of the Union addresses. We have demonstrated a thorough understanding of selected algorithms, proper application, and analysis of algorithmic results. In doing so, we successfully addressed each question listed in the beginning, discovering noteworthy and significant characteristics of the SOTU across history. Our effort emphasized that although language does indeed change over time, similar topics of peace, power, and economic stand the test of time. Additionally, demographical terms by decade were accentuated through the application of natural language processing techniques. This was a very interesting discovery made available by TF-IDF vectorization. Lastly, we showed the sentiment value of each presidential address with the corpus to carry varying degrees of positivity and a correlation to the unemployment rate of the nation's populace. Future work would include the full SOTU history in order to generate stronger analyses and correlations to national or world events.

References

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. JMLR 3, pp. 993-1022. Retrieved from http://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf

Borevitz, B. (2013). State of the Union: Addresses. Retrieved from http://stateoftheunion.onetwothree.net/index.shtml

Centre for Computational Linguistics and Psycholinguistics. (2010, December 26). pattern.en. Retrieved February 23, 2020, from https://www.clips.uantwerpen.be/pages/pattern-en#sentiment

Estival, D., Nowak, C., & Zschorn, A. (2004). Towards Ontology-based Natural Language Processing, Barcelona, Spain. Retrieved from https://www.aclweb.org/anthology/W04-0609

Lane, H., Howard, C., and Hapke, H.M. (2019). *Natural Language Processing in Action* (Shelter Island, NY: Manning Publications). Chapter 3: Math with Words (tf-idf vectors)

Scikit-learn: Machine Learning in Python, Pedregosa et al. (2011). JMLR 12, pp. 2825-2830. Retrieved from https://scikit-learn.org/stable/modules/clustering.html#k-means

United States Constitution. Article II, Sec. 3.

Appendix A

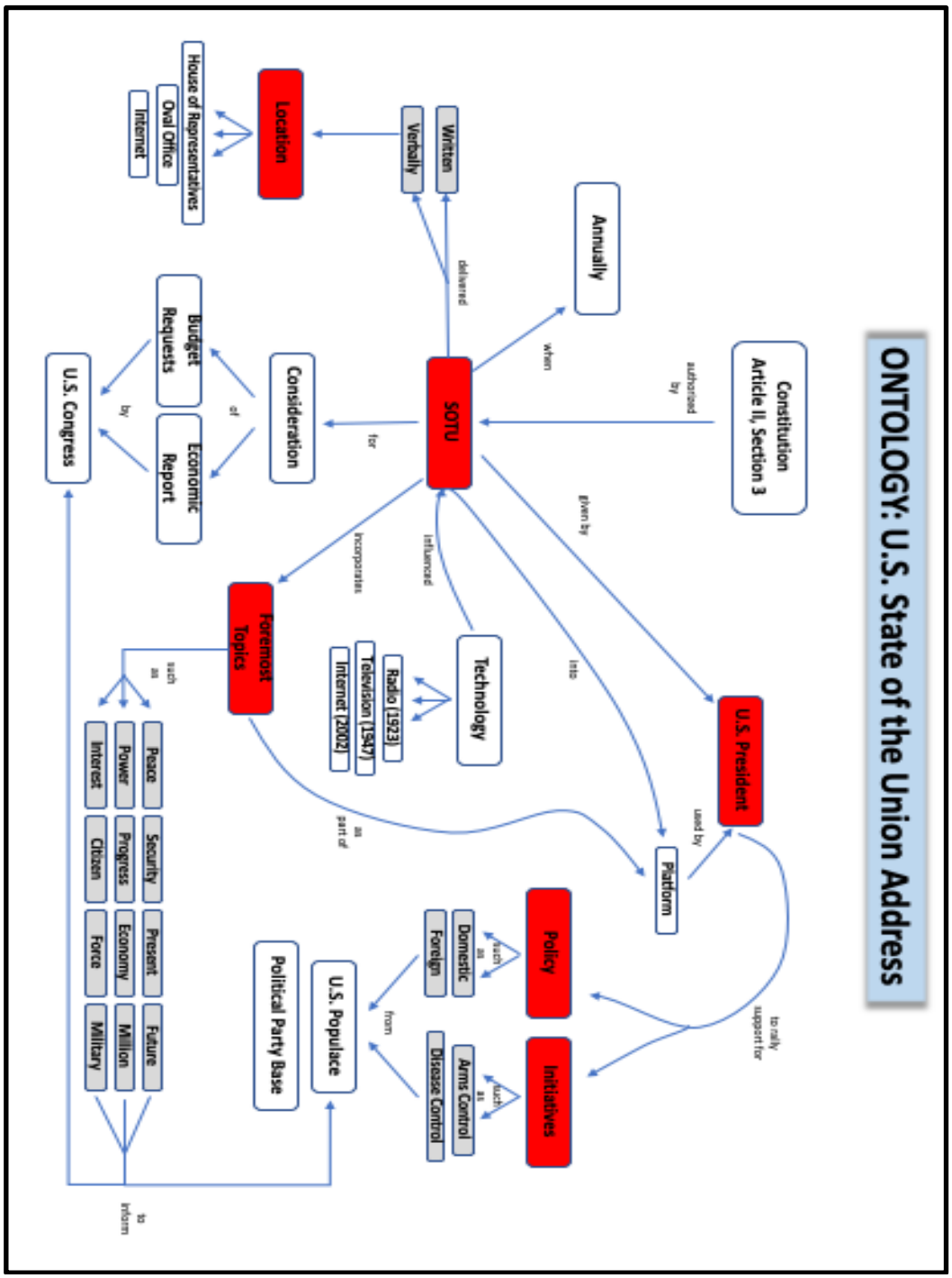
Figure A1

Word Cloud: U.S. State of the Union Address

Appendix B

Figure B1

Ontology: U.S. State of the Union Address

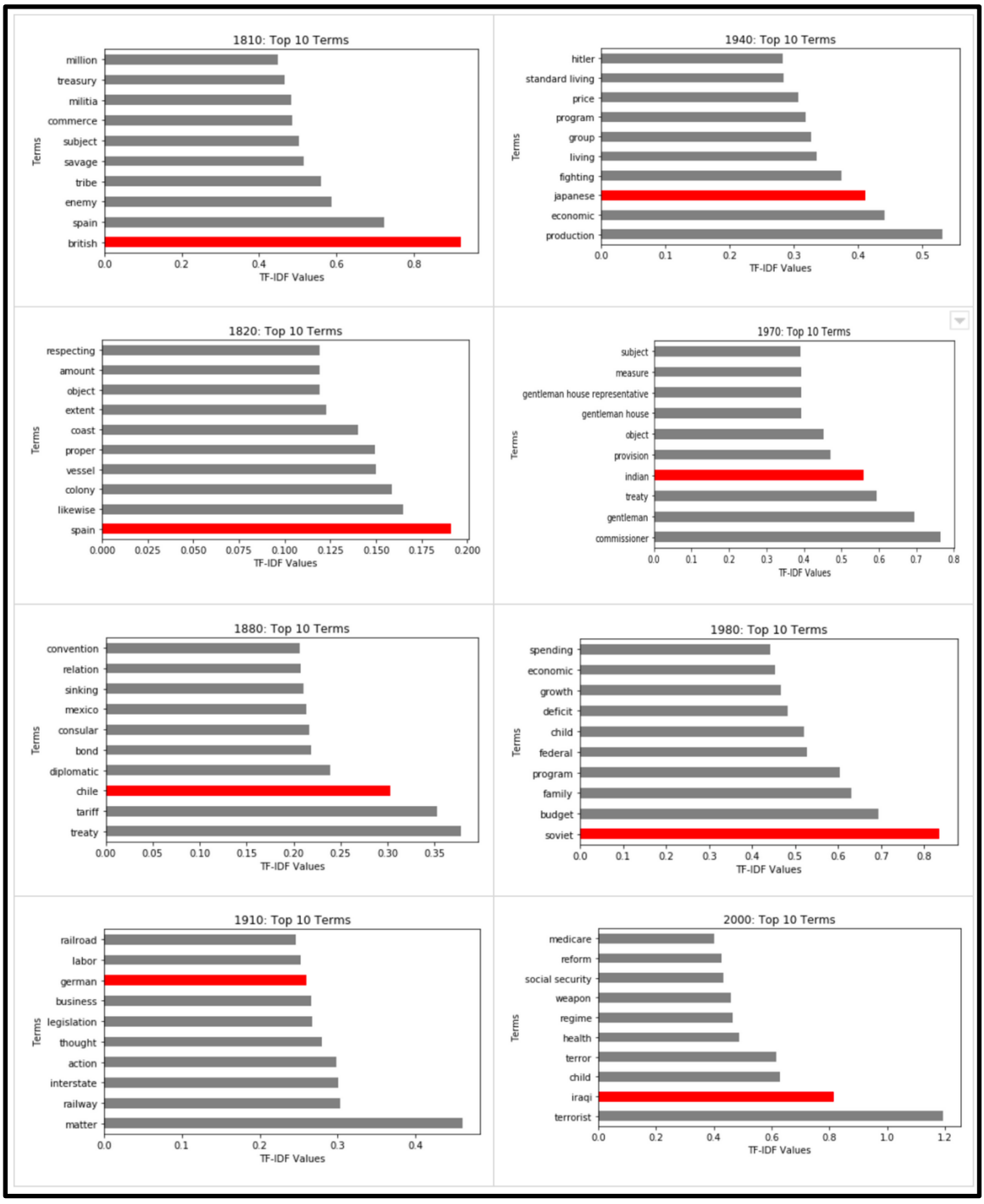


Classes, instances, and relations within the SOTU domain. We used red for classes and grey for instances. Direct links represent slots and internal links such as instance-of and subclass-of.

Appendix C

Figure C1

TF-IDF Values: Top 10 Terms by Decade



Top 10 terms from each decade based on TF-IDF values. The red bar indicates the country or demographic that was a focal point in the SOTU speeches for that decade.

Appendix C

Figure C1

Heatmap: TF-IDF Clustering of SOTU Corpus