Everything Data - CS 216
Project Team - 16: Vincent Lin, Kaitlyn McCarthy, Jacob Baca, Arjun Laud, Anupriya Sivakumar
April 29, 2019

Final Report: Language Analysis of State of the Union Addresses

Motivation:

The State of the Union (SOTU) addresses are speeches that are delivered by the president holding office annually. Furthermore, these annual addresses are mandatory by the United States Constitution. In their addresses, the president covers major topics and issues the country is currently facing and what goals he has for the upcoming year and term. In the past, the SOTU speech has been interpreted much differently than in recent years — as it has become one of the largest media events on the political calendar [1]. Due to this fact, we were interested in exploring how the topic emphasis, word choice, and vocabulary in these addresses have changed recently in comparison to past years. Since all the addresses are available as text files through an online source, we decided to explore the word trends of all SOTUs starting with the first one ever delivered by George Washington in 1790. Our guiding questions are described in the following section and analysis will be conducted using data collected from Kaggle.

Guiding Questions:

1. How have the most common topics differed in recent years compared to the past?
2. Does political affiliation affect topics touched upon in a speech? Does it affect word choice? Does it affect vocabulary?
3. How similar are the speeches when compared to one another? Which presidents have most similar speaking styles in the SOTU addresses?
4. In general, do SOTU addresses follow a similar flow, style, or template?
5. Can we predict what will be said in the next part of a SOTU address using previous parts of the speech as training data?

Related Work:

In order for our project to have a unique analysis of the SOTU addresses, we decided it fitting to explore what research was already out there. Two of the sources we used as inspiration and building blocks for our project are listed below:

The Language of Data: Analyzing the State of the Union [2]

- Berkeley investigated the SOTU speeches in two major ways. First, they looked to assess the different reading levels of the SOTU speeches, seeing who had more and less difficult to understand speeches. Second, they completed a word analysis for various word buckets, such as schools. This word bucket could encompass words such

as “education,” “college,” and “teachers.” They did similar word buckets for the economy, military, policy, people, superlatives, and world.

The Sentiment of the Union: Analyzing Presidential State of the Union Addresses using Sentiment Analysis and Python tools [3]

- This study investigated the SOTU speeches using packages within Python to get more information for sentiment analysis. Once they have retrieved the data necessary, they were able to run some pretty interesting tests. For instance, they investigated if the first and last SOTU speeches by the same president had more, less, or the same amount of positivity and negativity in them. They also calculated the entropy level of speeches to see how much a president was “saying” within their speech. This amount of information given during a SOTU speech determined how much uncertainty they were resolving by giving more information, which was simply using greater amounts of meaningful words.

Data:

State of the Unions (from Kaggle.com)

- This dataset includes every single SOTU speech ranging from the first one delivered by George Washington to the most recent by Donald Trump. Each individual speech is contained in a seperate text file with every word they said during the address transcribed in the file.

The Process:

1. Created a Git repository where we can share codes, visualizations, and other files
2. Cleaned, parsed, and split the data using Python: *pandas*, *re*, *os*, *string*, *csv*.
3. Tokenized each speech and appended them to each corresponding row in the pandas dataframe using Python: *sklearn*.
4. Computed the TFIDF scores and cosine similarities for all documents also using *sklearn*.
5. Used Tableau and WordCloud programming website to do some exploratory data analysis and more easily visualize frequent words.
6. Used Python API, *keras*, to fit a language model and to create a prediction model for the analysis and also used *sklearn* and *stats* to perform cross validation on our prediction model.
7. Created a Markov Chain as a second way to predict text from one starting word.
8. Compare the two predictive text algorithms (language model & Markov Chain) to see how the predicted texts differ from one another and if they make sense.

Visualizations:

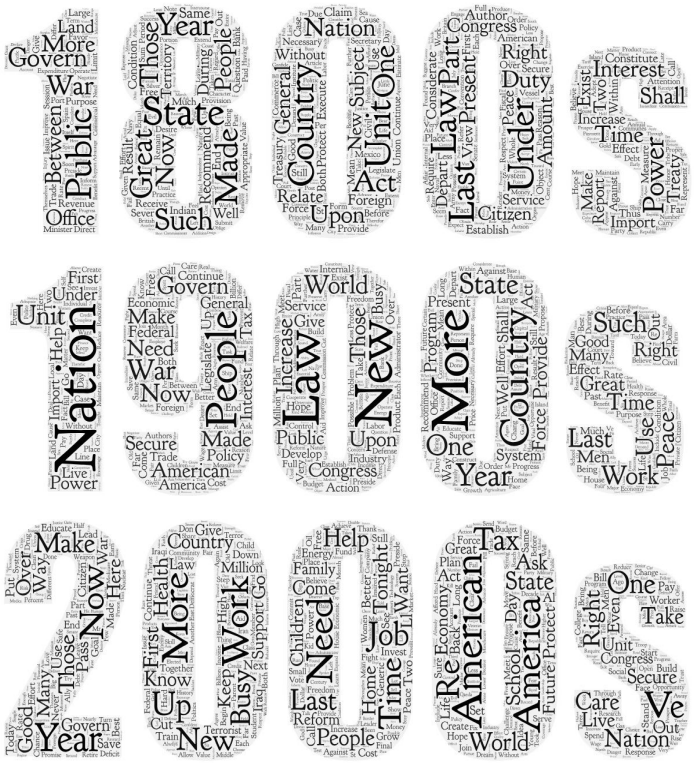

Figure 1

The visualizations of our data reveal some very interesting patterns. For example, the collection of word clouds in Figure 1, which utilizes the TFIDF scores calculated previously to identify the most commonly used words in a certain time period, show the changing emphasis on political issues across time, specifically centuries. The 1800s word cloud shows that the most frequently used words across all parties were unifying words, such as unit, one, and under, that emphasized the unity of America. This makes logical sense as America in the 1800s was focused on promoting its independence as a country from Britain. Similarly, the 1900s word cloud reflects how America moved towards establishing its power in the world, as evidenced by frequent use of words such as “world”, “public”, and “peace.” The 2000s word cloud reflects the economic turmoil that America has been dealing with in the last century, as seen through the use of words such as “work”, “job”, “economy”, and “tax.”

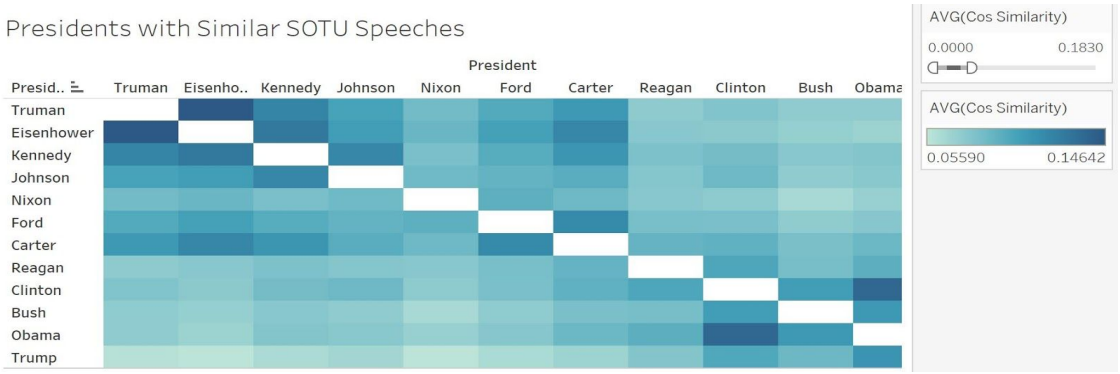

Figure 2

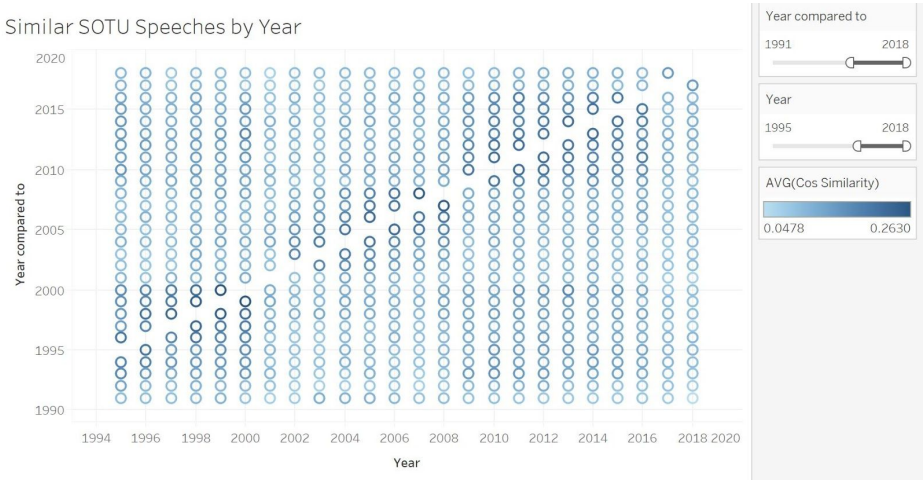

Figure 3

Figures 2 and 3 enable us to see if there are any patterns in the SOTU speeches across time and among presidents. For both of these figures, the darker the color, the higher the cosine similarity level. Additionally, when viewed in Tableau, the visualizations can be filtered for certain time periods. Both of these figures elucidate how the most similar SOTU speeches are clustered together in time — for example, the 2008 SOTU speech is most similar to the 2006 SOTU speech and the 2000 SOTU speech is most similar to the 1999 SOTU address (*see Figure 3).* This may be due to differences in terminology across periods of time. This makes sense; the manner of speaking in the 19th century is markedly different than the common tongue in the 20th century. This pattern is reflected in Figure 2 as the presidents with the most similar speeches are those that are closer together in time served (the presidents are listed chronologically in this graph). For example, Eisenhower's SOTU address is most similar to Truman's, who served directly before him. This may be because the political issues that they addressed are likely similar since they were facing the same national issues during both their presidencies. These figures reveal certain limitations of our analysis. For instance, we were unable to account for solely the semantic style of the speeches. Knowing this weakness, we took the content and overarching common way of speaking into consideration for our results.

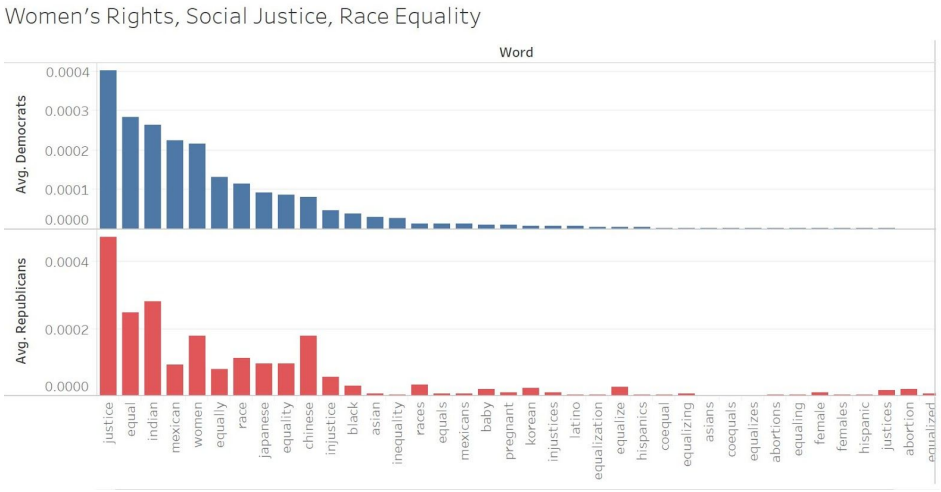

Figure 4

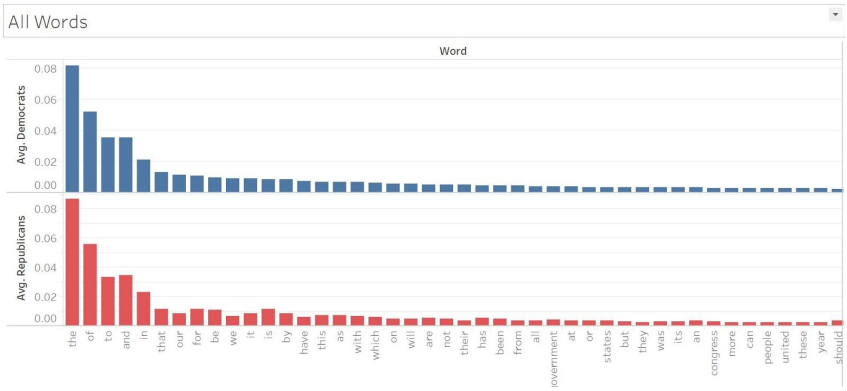

Figure 5

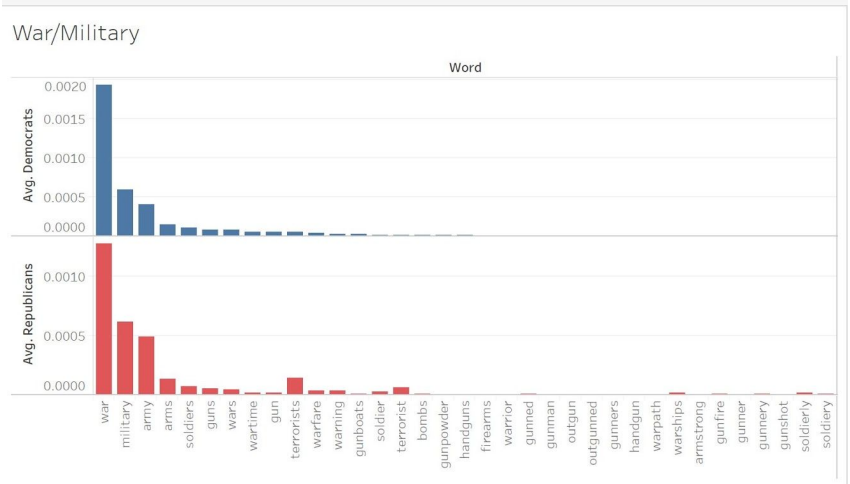

Figure 6

Next, we visualized how the topics addressed in the SOTU differed by party. By visualizing the most frequently used words for each party (as measured by our TFIDF scores in *Figure 5*), we see that the most commonly used words for both parties are common prepositions and articles. Thus, to get to more pertinent subject matter, we grouped words associated with common topics such as Military/War (*see Figure 6*) and Social Justice/Women's Rights (*see Figure 4*). These figures reveal that there is not much, if any, difference in the frequency with which these topics are addressed by Democrats or Republicans. This is rather counterintuitive because in the current political climate, one would expect that Democrats speak more about social justice issues and Republicans speak more about war. Therefore, these figures suggest that the political divide between these parties may not be as drastic as mainstream media makes it seem. However, limitations for this exploratory analysis include that we did not account for the ideological shift between the Democratic and Republican parties that occured around Reagan's presidency, and the words utilized in these groupings are not all-inclusive. For example, not every possible word associated with war is included in *Figure 6*.

Language Model:

In order to create a language model, we decided to only take addresses from a single year to have the code chunks run faster and more efficiently, compared to multiple texts combined together. The speech that we chose was the 2017 SOTU address by Donald Trump. The text was tokenized and was split into sequences with each sequence being a new line in the text file and containing 50 tokens joined together. Using the package *keras* in Python, we fit a language neural network model to learn the sequences of words and to calculate probabilities of specific words following each other.

Prediction:

We were unsure as to how exactly we could incorporate the *keras* package since we were originally thinking about producing the code in R, so we used incite from an online resource to build our code in Python [4,5]. When we were running the model with a batch size of 100 and the number of epochs (trials) set to 100, we ran the accuracy of the model through all 100 trials. The accuracy improves with every few trials. By the time the model reached the last 10 trials, the accuracy averaged 0.95. So, we are fairly certain that the model learns the sequences of words and the ones that should presumably follow. We created a function in order to generate a sequence of words from the model. Even though we had a high predictive accuracy returned from the model, we wanted to check if the model was able to predict sequences of words that made sense from the 50 words that were the initial starting point inputted into the model.

Markov Chain Text Generator:

We wanted to create another algorithm to see how it differed in accuracy from the generated text from the predictive model described in the section above. We decided a good comparison implementation would be creating our own Markov chain function. First, the function would find all the pairs of two words — with each pair being words that are adjacent to one another — within the text and then used these pairs to make a dictionary. This dictionary has single words as keys with corresponding list values. These lists represented all the words that followed the given key. Then, we created a predictive function that take one single would as the start word and generates however many words the user decides to choose.

Results:

After cleaning and organizing the data, we examined the SOTU speeches with a focus on topic emphasis, word choice, and vocabulary, finding a number of interesting results. Our visualizations of the data can offer insight into particular topics across time, presidents and political parties. For example, our word cloud visualizations for each century show the most common words in SOTU speeches during that respective century, indicating what the important topics and themes were during that time. In the 1800s visualization, we notice words such as unit and one, along with related synonyms, suggesting that speeches focused on emphasizing the unity of the United States as a nation. This makes sense given the nation was recently founded in opposition to Great Britain, so unity was important to its success.

We also produced some similarity visualizations in Tableau using TFIDF scores and cosine similarity data. We began by examining which speeches were most similar to ones given by a particular president. We found that the most similar speeches were generally by presidents in close succession, such as Clinton and Obama's similarity or Truman and Eisenhower. This is also logical as certain topics and issues will carry more importance during different periods of time, persisting through multiple presidencies. Finally, our Tableau graphs have the most interesting and surprising result of our visualizations. When comparing the frequency of certain topics addressed by each party, we found that both parties address specific political topics with similar frequency. This is surprising as current voters may expect that the Democratic Party would address more "liberal" topics, such as feminism and racial equality, with a higher frequency than Republicans. Thus, our analysis suggests that there is not as big a divide between parties as we may expect currently, and that the perception of this division may be a result of increasing media portrayal of such conflict and polarization.

After creating two different generative text algorithms, we compared them side by side with an example sequence from Trump's 2017 SOTU speech shown below:

of congress the first lady of the united states and citizens of america tonight as we mark the conclusion of our celebration of black history month we are reminded of our nation's path toward civil rights and the work that still remains to be done recent threats recent threats targeting

We inputted this sequence into both the model and the Markov Chain function and returned the next 100 predicted words to see if either predicted text made any sense.

jewish community centers and vandalism of jewish cemeteries as well as last week's shooting in kansas city remind us that while we may be a nation divided on policies we are a country that stands united in condemning hate and evil in all of its very ugly forms each american generation passes the torch of truth liberty and justice a new love that have slaughtered muslims and conflict we must enrich the chance his symbol of a father then in value since the middle east or in the help of a tax credit and an electric child scalia i

---

Generated Text from Model

of the audience tonight as an earthquake and abroad we will require us pay very brave men and get her master s what would have announced that threaten the defense sequester and republicans and terrorism and to create the departments of american taxpayers many other presidents combined we share of american people for our great milestones in america is to live past and it's time we have announced that stir our neglected inner cities of our neglected inner cities of lawless savages that you to enter the rule straining the time has been so many cases is looking down

---

Generated Text from Markov Chain

As seen, the text from each algorithm are indeed different, but they do make sense to an extent. Even though the text differs, we do not necessarily want them to output the same text or the following sequences because we want these functions to have some sort of predictive power. Unfortunately, it was difficult to validate the Markov Chain algorithm — especially with text — so we had to decide whether the algorithm was sound by looking at many outputs and seeing if the next words made sense given the sequence.

Conclusion:

The results from our research suggest that voters in the 2020 election should keep in mind that both parties likely focus on the same political agenda. They should focus on how the parties' specific strategies differ for these universal topics in determining what party to vote for — rather than focus superficially on what topics these parties address. However, while our research is enlightening in many ways, it does have several limitations. For instance, we did not account for the shifting ideologies between the parties or the significant language differences across time. Future steps for this project could include not only accounting for these limitations, but also determining the overall sentiment regarding a topic. For instance, deciding if a particular text has positive or negative sentiments regarding a specific topic, such as abortion rights. This will help to see if these sentiments vary over time, parties, and presidential terms.

References

[1] BBCNews, “What is the State of the Union speech?” https://www.bbc.com/news/world-us-canada-12221823, 2018

[2] Berkeley School of Information, “The Language of Data: Analyzing the State of the Union” https://datascience.berkeley.edu/blog/trump-state-of-the-union-analysis/, 2019

[3] Daniel Bashir, “The Sentiment of the Union: Analyzing Presidential State of the Union Addresses using Sentiment Analysis and Python tools” https://towardsdatascience.com/sentiment-of-the-union-analyzing-presidential-state-of-the-union-addresses-with-python-2a8667a578b9, 2019

[4] Jason Brownlee, “Making Predictions with Sequences” https://machinelearningmastery.com/sequence-prediction/, 2017

[5] Jason Brownlee, “How to Develop a Word-Level Neural Language Model and Use it to Generate Text” https://machinelearningmastery.com/how-to-develop-a-word-level-neural-language-model-in-keras/, 2017