Introduction

The words that we as humans speak have an incredible amount of weight behind them. As the main form of communication in our society, words are ever more important for how we convey our ideas and our thoughts. Words are crucial in how we communicate to others and no other person's word carries more meaning and power than the president of the United States. The President of the United States is the leader of the free world and what he says shapes the identity of the United States and also much of the Western world. Each President's words can provide great insight into the current state of the country and the world at large. Depending on the tone, context, verbiage, and many other linguistic signals, one can interpret what is being said beyond the words actually being used. While this is a daunting task programmatically, there are interesting patterns that can be identified by just analyzing the words themselves. The analysis of Presidential State of the Union Addresses form the basis for this thesis. The goal of this research is to examine the deeper meaning behind these speeches to gain greater insight into the status of the United States and the world at large at the time the address was delivered. The central question here being: can one use the words a President is speaking to accurately predict his position on the political spectrum, as well as predict how their speech is impacted by the current events at the time.

Sources and Methodologies

This thesis will center around Natural Language Processing (NLP) of textual data. The NLTK, Natural Language Toolkit, which is a useful set of tools that implement many Natural Language Processing processes, will be used to develop the code needed to process the texts. Research for the thesis will be collected from various sources across the web that analyze the

Presidential addresses from a historical perspective rather than a linguistic one. One of the central goals of this thesis will be to determine if Natural Language Processing techniques can be used successfully to learn trends in speech and etymology to make predictions about the people making those speeches.

NLP and NLTK

The Natural Language ToolKit provides a library of NLP methods to process text and extract meaningful trends and patterns from the text of interest. The NLTK is implemented using Python, which is a simple, yet powerful language with excellent functionality for processing linguistic data [NLTK Book].

State of the Union Addresses

The focus of this research centers around the State of the Union (SOTU) addresses, but the methods could be applied to other corpora as well. These annual SOTU addresses allow the President to opine the issues of the time and also shape the narrative for the entire country. The consistent collection of this text data makes it a good corpus to examine and it is very relevant to all United States citizens. Also, because these addresses are heavily examined and discussed, more meaningful observations can be made about them. And finally, because the political figures giving these speeches have such a public persona with many speeches and quotes on the record, it should be possible to categorize their speech patterns and use them as a basis for comparison and to make accurate predictions concerning the President's speech patterns and political leaning.

Historical Research

Historical research will be crucial to this thesis as a standard to compare the findings against. NLTK will be used to predict where the Presidents reside on the political spectrum, then historical research will be used to determine whether the prediction is correct or incorrect, based on where the President actually resides on that spectrum. Besides the main goal of determining position on the political spectrum there are many other historical trends and questions that can be explored when it comes to language and the evolution of the spoken word over time. Many linguistic trends can be seen when examining the addresses and how language has developed as well as how certain historical events (Great Depression, World Wars, etc.) have impacted the language and tone of the speeches over time. A secondary goal of this thesis is to construct a chronological language map that puts the language from each address in different bins based on the speech content and then map the words in each bin to historical events. This should allow the identification of certain events based on the sentiment analysis of the SOTU addresses, which might provide insight into the outlook of the nation or the personal attitude of the President himself.

Methodology

Much can be said about the accuracy of the political spectrum, as many Political Scientists argue that a single left-right axis is insufficient for attempting to express the complicated political beliefs of many individuals, but for this project, it will provide reasonable ground truth. The political spectrum will range from liberal on the left, to moderate in the middle, to conservative on the right. A political lexicon will then be constructed that scores certain words based on their political connotation to create quantitative values on which

the addresses can be compared. Words with more liberal intent will have scores closer to -100, moderate intent 0, conservative 100, and filler words will not be counted. The words and phrases that constitute the domain lexicon try and capture the meaning behind what the Presidents are saying. This lexicon will then be applied to each one of the Presidential Addresses and a score will be output from -100 to 100 that indicates where the President falls on the political spectrum by averaging the scores from each of the words or phrases that were tagged as important in each of their addresses. Once each score is produced, it will be discretized and based on that score each president will be classified as Conservative, Liberal, or Moderate. Then, by cross-referencing historical and political data, each President's ideology according to the literature will be compared to the predicted ideology from the program and the accuracy of the program will be judged based off of the percentage of correct predictions.

Review of Existing Literature

The literature review for this thesis will mainly be centered around NLTK book, as well as NLP research concerning Political ideology, and then the SOTU Addresses and the analyses of each of them, as well as presidential profiles. One of the main sources that will be used when looking at the change in speech over the course of the State of the Union addresses is a book from author Ryan Teten titled "Evolution of the Modern Rhetorical Presidency: Presidential Presentation and Development of the State of the Union Address" [Teten (2003)]. This book analyzes the State of the Union addresses in structure and rhetoric and talks about their change over time. It will provide a good baseline to determine the accuracy of the changes detected by the NLTK to see if meaningful conclusions can be

drawn. Another source is "Addressing the State of the Union: The Evolution And Impact of the Presidents's Big Speech" which gives a more holistic view of the State of the Union addresses and how they played out politically and socially, providing some essential context to the speeches and the words being spoken in them. There have been several studies similar to the topic of this research with some slight differences that will be useful references. A more popular approach to this task is to not classify speakers based on ideology but based on their values, smaller ideals that are easier to classify based on the words they say. A useful resource dealing with NLP and Political Ideologies is a paper entitled "A Scaling Model for Estimating Time-Series Party Positions from Texts" that implements NLP to process political texts and output their political standing by utilizing a statistical model that examines word usage across each political party in Germany and uses it as a basis for comparison [Slapin u. Proksch (2008)]. The standard approach among many of this research is to examine word usage within the confines of the speeches themselves and create a statistical model to make predictions instead of relying on an objective political ideological dictionary that is used in this thesis. [Sim u. a. (2013)] The analysis won't mean all that much if there isn't anything to compare the results to so these sources, along with others to come, will play a key role in contextualizing the results from the Natural Language Processing.

Conclusions and Preliminary Hypothesis

The conclusions for this thesis will most likely be mixed but there should be some rather high degree of accuracy with predicting where a person falls on the political spectrum if the lexicon is constructed in an intelligent manner. There is much information and data to pull from the words a person is speaking and a person's thoughts and feelings are conveyed

in every word, so analyzing them should paint a fairly vivid picture on where they stand overall. The major limitation of Natural Language Processing and raw text data in general is the lack of context and the lack of a deeper meaning which could definitely manifest itself in the results of this thesis. The subtleties of language and the spoken word make for skewed results when analyzing just the text as tone is hard to quantify and sarcasm is impossible to detect. Regardless of this, the results should still have some degree of accuracy based on how political figures generally speak, as they are more likely to be straight-forward, as well as each political party usually has their "talking points" that should be easy to identify and quantify into an intelligent prediction. For the more exploratory aspects of the thesis, the conclusions will be determined by whether or not the processed text accurately reflects the important discussions being held at the national stage at the time.

Bibliography

[Sim u. a. 2013] Sim, Yanchuan ; Acree, Brice D. ; Gross, Justin H. ; Smith, Noah A.: Measuring ideological proportions in political speeches. (2013)

[Slapin u. Proksch 2008] Slapin, Jonathan B. ; Proksch, Sven-Oliver: A Scaling Model for Estimating Time-Series Party Positions from Texts. In: *American Journal of Political Science* 52 (2008), Nr. 3, 705–722. http://dx.doi.org/10.1111/j.1540-5907.2008.00338.x. – DOI 10.1111/j.1540–5907.2008.00338.x. – ISSN 1540–5907

[Teten 2003] Teten, Ryan L.: Evolution of the Modern Rhetorical Presidency: Presidential Presentation and Development of the State of the Union Address. In: *Presidential Studies Quarterly* 33 (2003), Nr. 2, 333–346. http://dx.doi.org/10.1111/j.1741-5705.2003.tb00033.x. – DOI 10.1111/j.1741–5705.2003.tb00033.x. – ISSN 1741–5705