Chapter 1

Motivation

Language is our main way of communication and the words an individual chooses to express their ideas can be very telling about their mood and opinion towards a particular subject or issue. This is extremely pertinent for the President of the United States as their words and messages set the tone for the United States as a whole, and their words can often be analyzed to look for further meaning in the things that they say. This analysis has the potential to reveal underlying themes and patterns of speech in how president's speak and what their word choice indicates about the State of the United States and also how they view certain issues and topics. The driving force behind the research conducted here was to see if there was a way to reliably predict a president's political party purely based on their words they use in their State of the Union Addresses. This research can extend far beyond this central question as well, expanding to include more data sources to increase the accuracy of the predictions. On top of this, there is potential to predict further characteristics beyond political party such as ideology and other personality traits of the speaker.

The other crucial part of this research is more historical in nature in that each sentiment score derived for a president must be contextualized by the important historical events that occurred during their presidency. The biggest challenge here being separating out the relative importance of the president's own outlook on the world versus that of the event itself. Some

presidents may have a more positive tone and outlook on the world and try to use their State of the Union address to encourage the public even if the events of the time may be dire such as the Great Depression, so that will be an interesting challenge to weigh the relative importance of each.

Chapter 2

Corpus

2.1 State of the Union Addresses

The main textual data that was collected to be processed was all of the Presidential State of the Unions from George Washington's first address to Barack Obama's last address. The text source was initially pulled from a Presidential Address Repository. The initial source data of speeches came in a long text file, but python was used to split each address into it's own file and tagged using the identifiers of the President and the speech, i.e. $179001_Washington.txt.Initiallyonlytheyearwasplacedatthebeginningoftheaddressfiles, butsomePreside$

2.2 Topic Classifier Sets

An important part of the latter half of the preprocessing work for this research was the topic classifier sets. At first, the sentiment score was calculated for each presidential address with an overall score from -1 to 1, indicating their tone when delivering that address. After these were calculated, they were analyzed to look for trends in each president's tone to see if there were any interesting observations to be seen. As an additional breakdown to see if there was any more context-specific information that could help determine a president's political party, topic categories were added to diversify the scores of the president's even more.

Four Presidential Addresses were chosen (Washington, Lincoln, Kennedy, Obama) and manually read through to discover what words were being used when talking about certain topics within the United States. The topics that were chosen were: crime, economy, education, energy, environment, family, foreign, government, job, religion, terrorism, and war. Text files were created using the trigger words for each major topic, the trigger words being pulled from the four addresses mentioned above and throughout various other addresses as they were skimmed through, that would add the entire sentence to an array named for what topic it was going to collect information on. This processing was conducted on every address and the sentiment score for each topic was found for each president, which resulted in a vector for each address that had their overall sentiment score and the sentiment score for each topic covered in the address. These scores for each address were then averaged together to create an overall vector for each president that could be used for classification and learning to learn their political party and predict others.

Chapter 3

Preprocessing

3.1 NLP and NLTK

3.2 Normalization

3.3 Intensifiers

Chapter 4

Visualizations

4.1 D3

4.2 Line Plot

Description / explanation of data sources / examples (screenshots)

4.3 Word Clouds

Description / explanation of data sources / examples (screenshots)

Chapter 5

Results

Chapter 6

Discussion / Reflections

Bibliography