Brandon Falcona (bmf855)
Noah Pang (np9692)
LIN 350 “Analyzing Linguistic Data”
Dr. Katrin Erk
8 April 2020

Intermediate Project Report

Research questions

After obtaining a collection of presidential speeches in plaintext form and briefly parsing through them, we realized that focusing specifically on the topic of immigration would not be as doable as we previously imagined. The main issue with this stemmed from the fact that the presidents in question varied drastically in the amount which they talked about immigration during their speeches. Eisenhower, for instance, very rarely talked about immigration at all, whereas Trump has an extensive amount of speech content related to immigration. This would have caused issues not only with the time-intensive task of defining and isolating only immigration-related excerpts, but also with the great degree of variability in the amount of data available for each president.

As such, while we are still interested in examining Republican president’s speech patterns with regards to immigration, we will be broadening the scope. We will now focus on studying how the policy priorities of the party have changed over time. In the context of immgration, we expect to see more recent presidents talk more about immigration-related topics than past presidents. However, we will also examine in general which topics were talked about most by each president, which presidents were the most and least similar with their policy focuses, and

Immigration is but one of many issues on which the Republican Party has drastically changed position over time, so it will be interesting to see how the party’s focus has evolved. Furthermore, this will highlight the historical events that have shaped conversations in U.S. politics. We may even examine general shifts in language and tone, noting such contrasts as shifts to the more brash language of Donald Trump’s populist message.

Methods

Despite the change in scope, the same methods of topic modeling and clustering are still being used to analyze the data. Topic modeling and clustering remain useful tools for determining the relevant topics in the speeches given from each president and can be used to compare similarities and differences in speech topics between different Republican presidents.

Status

We obtained an online corpus of speeches by U.S. presidents from The Grammar Lab. This online archive contains speeches from all U.S. presidents—from George Washington to Donald Trump—but for the purposes of this project, we only obtained speeches from all Republican presidents since Dwight D. Eisenhower. These speeches include Inaugural Addresses, State of the Union Addresses, and press conferences. Data extracted from these texts can be used for topic modeling, which can 2The number of speeches and total word count for each president studied is as follows: Dwight D. Eisenhower, 6 speeches (18,097 words total); Richard Nixon, 23 speeches (66,482 words total); Gerald Ford, 14 speeches (40,446 words total); Ronald Reagan, 59 speeches (196,553 words total); George H. W. Bush, 23 speeches (71,160

words total); George W. Bush, 39 speeches (107,737 words total); and Donald Trump, 62 speeches (approximately 40,000 words total).

We chose to exclude texts that included significant amounts of speech from people other than the presidents, such as debates and interviews, since the effort required to isolate only excerpts of the presidents' words would be too great in order to salvage the text. After running a sample topic modeling analysis on Ronald Reagan's speeches, we also realized we needed to expand the list of words to exclude. Though by default we excluded stopwords, we also realized we needed to exclude common words like "government", "United States", "America", and "nation". Words like these turned out to dominate the top of the topic modeling lists with the highest probabilities, but since we presumed these words were used ubiquitously by all presidents, we figured they would not tell us anything about the presidents' policy focuses.

Though only Reagan's speeches were analyzed, the code can be easily modified to work for any of the other presidents. With all in mind, we should be on track to finish our planned work within the given time.