Brandon Falcona (bmf855)
Noah Pang (np9692)
LIN 350 "Analyzing Linguistic Data"
Dr. Katrin Erk
12 May 2020

Final Report

Introduction

The Republican Party underwent major shifts throughout the twentieth century, both in its policies and constituencies. Immigration has become an especially "hot-button" issue in the United States since Donald Trump's election to the presidency in 2016. Trump's populist and nativist rhetoric has frequently included staunchly anti-immigrant rhetoric. However, this general sentiment on immigration has not always been found within the Party, but has dramatically shifted and developed over time.

With immigration becoming more of a policy priority in recent years, the original research question we wanted to answer was: how well do the thematic shifts in the speeches of the United States presidents from the Republican Party correlate with the Party's shift in policies on immigration since Dwight Eisenhower? Our hypothesis was that immigrant-friendly presidents would use more words related to morality while immigrant-hostile presidents would use more words related to social issues.

However, after obtaining a collection of presidential speeches in plaintext form and parsing through them, we realized that focusing specifically on the topic of immigration would not be as feasible as previously thought. This was because speeches mentioning immigration varied drastically between different Republican presidents; for instance, Eisenhower barely mentioned the issue of immigration while Trump had many speeches dedicated to the issue of immigration. Isolating only the parts of Republican president's speeches related to immigration would be very difficult but there was also the possibility that there were too few mentions of immigration for a proper comparison to be made between the different Republican presidents.

This resulted in the formulation of a more broad research question: How have the Republican Party's policy priorities changed over time? We still examined how each Republican president spoke about the issue of immigration, but shifted the focus to the general policy priorities of the Republican Party for which our obtained data would be more suitable. Our revised hypothesis was that recent Republican presidents would talk more about immigration-related issues than past Republican presidents and that historical events would shape policy focus.

Data

We obtained an online corpus of speeches by U.S. presidents from The Grammar Lab[1]. 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 to determine the topics present and frequency of words in each

[1] Corpus of Presidential Speeches

president's speeches. The number of speeches and total word count for each president analyzed is as follows:

President	Number of Speeches	Total Word Count
Dwight D. Eisenhower	6	18,097
Richard Nixon	23	66,482
Gerald Ford	14	40,446
Ronald Reagan	59	196,553
George H. W. Bush	23	71,160
George W. Bush	39	107,737
Donald Trump	62	approx. 40,000

Some of the obtained text included significant amounts of speech from people other than the presidents. These texts, notably including debates and interviews, were excluded since it would have been too time consuming to isolate only the excerpts spoken by the president. After running topic modeling analysis on the Republican president's speeches, there were certain words that frequently appeared which did not give much information about the policy focuses of the presidents. By default we excluded stopwords, but we also realized we needed to exclude common words such as "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. After the stopwords and common words were excluded, meaningful topics and word frequencies from each Republican president were present in the topic modeling analysis.

Methods

The method of statistical analysis used for the Republican president's speeches was topic modeling. An example of how this method was implemented is as follows, represented arbitrarily by the Python code used for analysis of Richard Nixon's speeches.

First, each of the text files of Nixon's speeches were loaded in and stored in the list **`nixon_files`** via the pathway **`nixon_path`**:

```
nixon_path = Path("presidential_speeches/nixon")

nixon_files = [ ]
for file in nixon_path.iterdir():
    if file.name != ".DS_Store":
        nixon_files.append(file)
```

This list of text files was combined into one text file and stored in the variable **nixon_uncleaned**:

```
with open("output_file", "w") as outfile:
    for fname in nixon_files:
        with open(fname) as infile:
            outfile.write(infile.read())

with open("output_file") as f:
    nixon_uncleaned = f.read()
```

As the first step of preprocessing, the text in **nixon_uncleaned** was *tokenized* (split into a list of words with punctuation separated out) via the **nltk** (Natural Language Toolkit) package and stored in **nixon_tagged**:

```
nixon_tagged = nltk.pos_tag(nltk.word_tokenize(nixon_uncleaned))
```

From **nixon_tagged**, the words were *lemmatized* (transformed into their *lemmas*, or *dictionary forms*) and contained, along with part-of-speech tags for each word, in the list **nixon_lemmatized**:

```
nixon_lemmatized = [ ]
for word, tag in nixon_tagged:
    wntag = penntag_to_wordnettag(tag)
    nixon_lemmatized.append(lemmatizer.lemmatize(word, wntag))
```

The lemmatized text was further processed by making the words lowercase and removing stopwords, common terms (from a manually created list **common_terms**), words with less than four characters, and punctuation. This was stored in the list **nixon_text**:

```
nixon_text = [ ]
for word in nixon_lemmatized:
    if word not in stopwords and word not in common_terms and len(word) >= 4 and
    word.strip(string.punctuation) != "":
        nixon_text.append(word.lower())
```

This cleaned text was then appended to **nixon_corpus**, transformed into the gensim-internal format, and again appended into **nixon_gensim** so that topic modeling could be performed:

```
nixon_corpus = [ ]
nixon_corpus.append(nixon_text)

nixon_dictionary = gensim.corpora.Dictionary(nixon_corpus)
nixon_gensim = [ ]
for document in nixon_corpus:
    nixon_gensim.append(nixon_dictionary.doc2bow(document))
```

Finally, topic modeling (via **nixon_model**) was performed:

```
nixon_model = gensim.models.ldamodel.LdaModel(nixon_gensim,
                                              num_topics = 5,
                                              id2word = nixon_dictionary,
                                              passes = 20, iterations = 500,
                                              alpha = "asymmetric",
                                              random_state = 0)

for topic in nixon_model.print_topics(num_words = 10):
    print(topic)
```

This model created five topics with ten words in each topic, the results of which can be seen in the following section.

The same process was performed for all other studied presidents, with appropriate variations—according each corresponding president's name—of object names.

Findings

The most successful results from performing topic modeling on each of the presidents' speeches are as follows:

Dwight D. Eisenhower (1953–1961):

```
(1, '0.011*"peace" + 0.009*"free" + 0.006*"soviet" + 0.005*"hope"
+ 0.005*"future" + 0.005*"party" + 0.005*"atomic" + 0.005*"power"
+ 0.004*"security" + 0.004*"freedom"')
```

Richard Nixon (1969–1974):

```
(4, '0.014*"peace" + 0.010*"vietnam" + 0.004*"force" +
0.004*"action" + 0.003*"responsibility" + 0.003*"vietnamese" +
0.003*"future" + 0.003*"watergate" + 0.003*"negotiation" +
0.003*"house"')
```

Gerald Ford (1974–1977):

```
(4, '0.008*"energy" + 0.005*"peace" + 0.005*"future" +
0.004*"economic" + 0.004*"foreign" + 0.003*"food" +
0.003*"security" + 0.003*"economy" + 0.003*"price" +
0.003*"union"')
```

Ronald Reagan (1981–1989):

```
(2, '0.007*"peace" + 0.006*"freedom" + 0.005*"soviet" +
0.004*"force" + 0.004*"economic" + 0.003*"free" + 0.003*"hope" +
0.003*"family" + 0.003*"union" + 0.003*"future"')
```

George H. W. Bush (1989–1993):

```
(0, '0.007*"force" + 0.004*"iraq" + 0.004*"peace" +
0.004*"soviet" + 0.004*"child" + 0.004*"future" +
0.004*"security" + 0.004*"freedom" + 0.004*"change" +
0.003*"union"')
```

George W. Bush (2001–2009):

```
(2, '0.008*"iraq" + 0.006*"terrorist" + 0.005*"security" +
0.005*"freedom" + 0.004*"citizen" + 0.004*"iraqi" + 0.004*"child"
+ 0.003*"peace" + 0.003*"health" + 0.003*"force"')
```

Donald Trump (2017–):

```
(0, '0.006*"deal" + 0.005*"build" + 0.005*"money" + 0.005*"trade"
+ 0.005*"wall" + 0.004*"border" + 0.004*"remember" +
0.004*"change" + 0.004*"fight" + 0.004*"mexico"')
```

Each entry lists the estimated *word frequencies* of each word in the topic occurring in the president's speeches as a whole. For example, the topic modeling result for Dwight D.

Eisenhower estimates a probability of 0.011 for the word “peace”, a probability of 0.009 for the word “free”, and so on.

The above entries represent the topics with the highest probabilities in its entries. Interestingly, for each president, each of the other four topics not presented above listed a probability of 0.000 for all ten words. While this in fact means each of the words had a *near* zero probability of occurring, it was surprising that one and only one significant topic appeared for each president. Based on cursory research, this problem may be related to dataset size, number of topics used, or precision of the model. This issue is a topic of potential further study unto itself.

These results give rise to several interesting conclusions. First, by far the most probable indicators of speech topic were words relating to historical events and issues of each president’s tenure. For instance, the Cold War-era presidents Eisenhower, Nixon, Ford, Nixon, and H. W. Bush featured time-appropriate words like “Soviet”, “atomic”, and “Vietnam”. Both of the Bushes’ presidencies took place during military conflicts with Iraq (the Gulf War and later the Iraq War) and so mentioned words such as “Iraq” and “terrorist”, the latter of which also may refer to the September 11 terrorist attacks during W. Bush’s presidency. Another timely word was “Watergate” in the results of Richard Nixon, clearly referring to the Watergate Scandal.

A second conclusion can be made about the tone of some of the modeled words. Earlier presidents (Eisenhower, Nixon, Ford, and Reagan) tend to use more diplomatic terms, whereas more recent presidents (Reagan, H. W. Bush, W. Bush, and Trump) tend to use more aggressive terms (here, Reagan is considered to be a “transition” between the earlier and more recent presidents). This conclusion is based on words like “peace”, “negotiation”, and “hope” for the earlier presidents; and words like “force”, “security”, and “fight” for the more recent presidents. We suspect this coincides with the Republican Party’s shift towards a more aggressive and militaristic foreign policy, starting most notably with Ronald Reagan. This result was not initially anticipated, so it would be particularly interesting to study this potential trend further.

For several presidents, mentions regarding immigration specifically were minimal; for no president other than Donald Trump did immigration-related words appear in the topic lists. This was especially true for the earliest presidents, like Eisenhower and Ford, whose speeches barely mention immigration-related issues at all. This lack of consistency in frequency of mentions made it difficult to specifically compare immigration policies between the presidents, prompting us to shift our focus away from immigration in particular and towards policy focus in general.

Apart from the previously mentioned trend from diplomatic to aggressive term usage, there are several areas in which further study may be warranted. First, we would also be interested in a similar study done for recent presidents from the Democratic Party. These presidents’ speech would be interesting not only for their own right, but also as a means of comparison of word usage and tone with that of their Republican counterparts. Another area of research expansion would be Republican presidential candidates who failed to win their respective elections. Some of those candidates—most notably Barry Goldwater—were, despite losing their elections, nonetheless very influential in the Republican Party’s history, and thus their study would be justified. Finally, more emphasis could be placed on the tone of each president’s word usage, rather than simply just topics discussed. This may warrant discussion of how the *nature* of the ways in which the Republican Party conducts politics has changed; for instance, as an expansion on the “diplomatic vs. aggressive” terminology conclusion we found.