Second Capstone Milestone Report 1

Laura Friese

Understanding the linguistic and word choice differences between political parties can provide a lot of information of interest. One clear example of presidential speech is executive orders, which will be examined here. The tone of executive orders and the words chosen likely vary according to factors such as time period and the political party of the sitting president. I wish to examine the relationship between these factors and the word choice of the executive orders. It would also be easy to examine the topics addressed in these executive orders, and see how they correlate to each political party. This could provide insights about the values of each party, as well as the typical persuasive strategies and rhetoric each party uses. We could also examine the use of these strategies over time, or for each president individually. This would inform the public of strategies used by each president or party and the use of those strategies over time. Knowledge of common rhetoric is a vital tool in the dismantlement of misinformation, which is so important in our current political climate of "fake news." Thinking critically about the tactics used to convey information, such as creating a sense of fear or duty, helps us better understand our political system.

To investigate the relationship between various linguistic elements of executive orders and the year, president, and political party in control, I will conduct some natural language processing on a dataset of executive orders given since 1994. This dataset is from the national

register, via the Office of the Federal Register, and thus is quite reliable. Before analysis with NLP(Natural Language Processing) can begin, however, the data must be cleaned and wrangled, and some initial probing using statistical inference must be completed. The details of this process are outlined below.

The Executive Order dataset contained 928 Executive Orders, represented as rows, and 13 columns: the citation, the document number, the end page, the html url, the pdf url, the type (all were Presidential Documents), the subtype (Executive Order), the publication date, the signed date, the start page, the title, the disposition notes, and the executive order number. Most of these columns won't be used, but we will use the url and the publication date. The other numbers are largely referential.

This dataset didn't contain the text of each order directly, only links to the PDF and HTML versions of the text. Thus, pulling the text from these URLs and attaching it to the DataFrame was the bulk of the data cleaning process. Before extracting the text from the URLs, I wanted to create another column called "President" and one called "Party" so that we have those pre-labeled before we make any transformations. I can see online that orders 13765-13851 were signed by Trump, 13490-13764 were signed by Obama, 13199-13487 were signed by Bush and 12890-13197 were signed by Clinton. This will help us assign each order to the president that gave it. This will be of use to differentiate the orders for later analysis. To create these columns, I made custom functions filling the new columns based on values in pre existing columns.

The next step was to extract the HTML text of the executive orders from the HTML links. This was done using Beautiful Soup and urllib to access the link information and then clean up the HTML results to make them more readable. After the text was added to the dataset, it was still in messy HTML format. To extract just the text of the order, I identified phrases at the end and beginning of the orders and used them to slice the text. First I looked through the documents to try and find a phrase that all the executive orders started with. I used this to split the document so we only get the text of the order, not all the HTML information. I chose "Use the PDF linked in the document sidebar for the official electronic format". It's long, but I has issues with short phrases appearing earlier in the text. Below, I do the same with the end of the order and the phrase "[FR Doc.". This slicing, while not perfect, pared down the data so that only a few words from the HTML formatting remained. After creating the code to extract a single order, I then wrapped that code in a for loop so it would go down the entire dataset pulling the text for each order.

There was an issue with the splitting, however: many of the executive orders were not in HTML format. I discovered this when the for loop didn't work for every row in the dataset. I examined the links for one of the data points that didn't work, and found that if you click on the link in the actual dataset, it takes you to a page that says that the text of the executive order is not available in this format, and links you to another flat text page. To fix this, I removed all pages that did not have an HTML copy of the executive order, removing 223 of our 928 observances. This is a large portion and they were disproportionately taken from Clinton's executive orders, as

it was mostly the older executive orders they weren't in HTML. This could be a source of error later, although we will later see that there ended up being a fairly even amount of Republican and Democratic executive orders after these were removed. The end result of this cleaning was 705 total executive orders to work with. To further clean the textual data, all of the "\n"s, denoting a new line, were removed so they wouldn't interfere with our results.

The next step in the analysis is to separate the data out by President and political party in order to make comparisons. After the data cleaning, the dataset contained 87 executive orders written by Trump, 272 written by Obama, 291 written by Bush, and 55 written by Clinton. This amounts to 327 executive orders written by Democrats and 378 by Republicans.

After the data cleaning and wrangling were complete, I performed some basic statistical analysis on the data. As much of the data is text, not a lot could be done without using NLP. I created another column that contained the length of each document in characters, and compared this number across president and party. I found that on average Republican executive orders were 6545.6 characters long, and Democrat's executive orders were 7998.9 characters long. I then broke it down by President and found that Trump's executive orders were 8724.5 characters long on average, Obama's were 8187.8, Bush's were 5894.1 and Clinton's were 7064.6. The high number of Trump's was likely balanced out by Bush's low order length, especially as Bush has many more executive orders in this dataset.

After recording these basic statistics, some graphs were constructed to compare the executive orders. The box plots showed that executive orders written by Republicans were

significantly shorter, however another plot also showed that Trump, a Republican, had the longest executive orders on average. This observation is explained by the low length of Bush's executive orders. As they make up the bulk of the Republican executive orders, these short observations lowered the Republican executive order length average. If Trump continues this pattern, the difference will likely become negligible. Our initial statistical observations also indicated an increase of executive order length over time, but this increase was so small that it is likely clinically insignificant, even if it is statistically significant. To test this, I did a linear regression of the relationship between time and the length of executive orders. The linear regression produced a model with an R-squared value of 0.447, which is a moderate association. However, the coefficient of of the publication date (our time variable) was only 0.0098, which is quite small. This further reinforces the idea that, although this trend might be statistically significant, the actual difference is not clinically significant, meaning that is doesn't matter in reality. This small difference indicates a change of only about 3 characters per year. We also created a model using only the non-extreme values of this dataset, which we chose based on the initial visualization to be below about 21000. This model produced an R-squared value of 0.667, which is a moderate association and larger than our previous value, indicating a stronger relationship. However, the coefficient of of the publication date (our time variable) was still only 0.0084, which is even smaller than it was previously. We can conclude that this relationship is likely not a useful one for our investigative purposes.

The nest steps in this process will use NLP to analyze the linguistic associations with each party and president. This will hopefully reveal some information about the two parties or the terms of interest.