Midterm - Project Proposal Outline

Name: Michael Roberts

Project Title: The Frequency of Key Words In Presidential Speeches Over Time

I. Introduction / Background

Over the past decade or so, within American politics there has been a trend of increased polarization within our politics. Both sides of the political aisle focus on their differences and what separates them from their opponents. Each election cycle, you see how what the current administration's goals are through how they speak during many of their public appearances, as well as how the focus of the administration shifts throughout each term as the political landscape changes.

II. Audience

The target audience for my visualization might be future candidates who want to run for president, political pundits, or regular people who would just like to be more informed about the trending topics on both sides of the aisle. Basically anybody who follows American politics, or may have a vested interest in what types of topics would be good to focus on in an election may find my proposed visualization informative.

III. Dataset(s)

1. <u>Data Collection:</u>

The dataset I would like to use I would have to curate myself to be usable for my purposes. I would like to use a text dataset that I would generate from the presidential speeches from 2016-present.

I plan to collect that data from: https://millercenter.org/the-presidency/presidential-speeches

I will use the BeautifulSoup web-scraping python library to pull each of the speeches from Donald Trump's term and from the Joe Biden term into either text files, or into a csv file with each of the speeches dated and labled.

From there I will use the spacy library alongside the nltk library to do text lemmatization on the speeches. This will transform the text data from each speech into a list of words, and then turn each word into a vector so that it will be easier to work with when doing my natural language processing.

Next I will use a good stopword list to remove words that hold little meaning from the speeches, such as "a", "the", "is", etc. Once this has been done I can proceed with my visualization. I plan to use the functionality within spacy and within nltk, alongside libraries such as the scattertext library and the genism library to do my visualizations.

2. **Data Breakdown:**

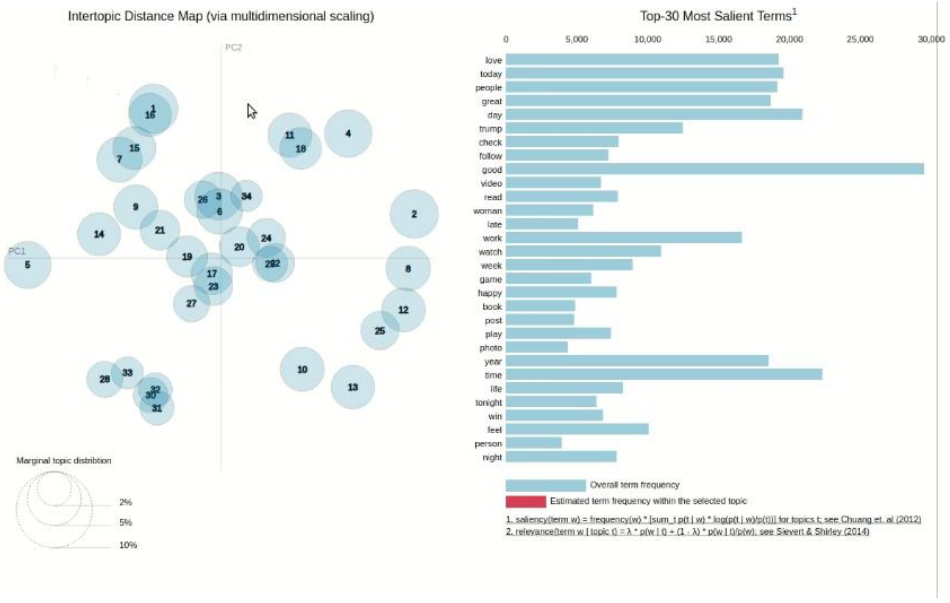
I will likely be using the most recent quarter of the 43 speeches that Donald Trump gave during his term as president, alongside the 11 speeches that the Miller Center has collected from the Joe Biden term as president.

IV. Proposed Visualization

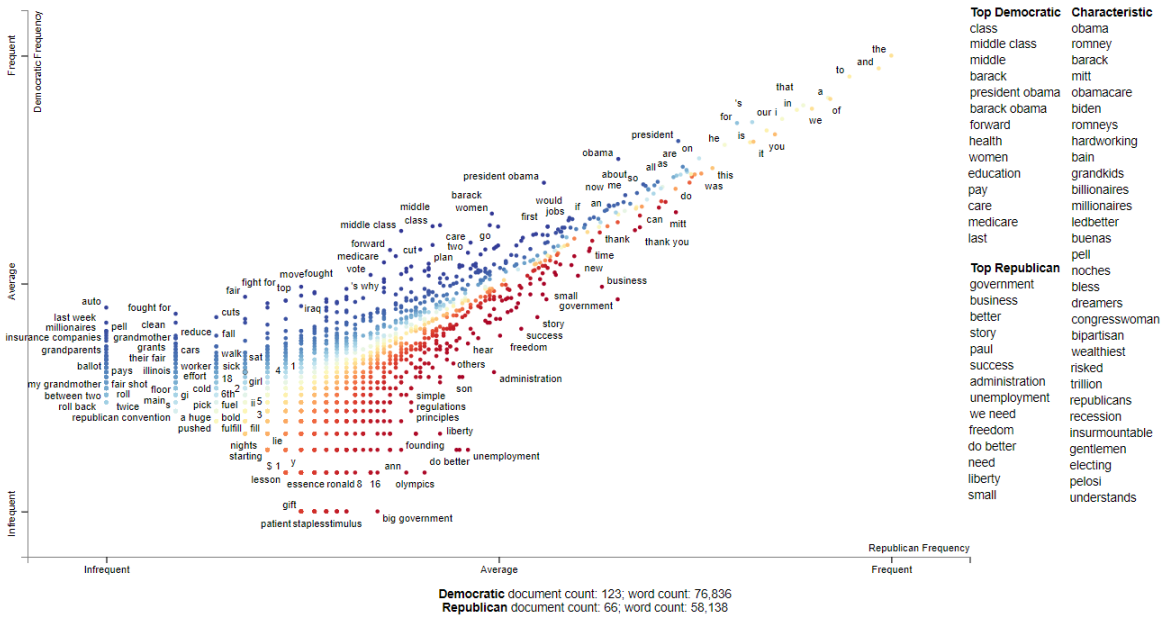
The visualizations I plan to create will be used to showcase how the focus of the American public has shifted over time, since presidential speeches are generally a reflection of the state of the nation. Three types of text visualization in particular stand out to me as potentially useful for my purposes.

1. **A Word Cloud – Using the wordcloud library:**

2. **An LDA Topic Model – Using the LDAVis and Gensim libraries:**



3. **A Word Frequency Scatter Plot – Using the Scattertext library:**



V. Plan

A. Ideal Plan:

Ideally, I plan to take my lemmatized dataset that I collect and categorize the data by years. Then I will create a document corpra for each year containing text from all of the speeches from that year. Then, I will be using LDA topic modeling to group like-text together into likely similar groupings. Then the viewer of my visualization will be able to navigate each year by topics that have been focused on a lot in presidential speeches and be able to see what words were used relating to each topic. In the resulting visualization, the viewer will be able to navigate interactively to see how the breakdown of each group in the more frequently or less frequently referenced groupings shakes out as well as how each term the president may focus differently on similar topics.

B. Backup Plan:

The backup plan is for my scope to change from broadly year-by-year looking at presidential speeches, to instead having the past two presidential terms be how the text data is being categorized. Then for each presidential term, I will create a text scatterplot showing the frequency of individual terms in presidential speeches. If it would be easier to clean up my data and use it with this form of visualization, then I may broaden the scope back out and break it down by year again. If that would be too broad, then I would result with two of the text scatterplots with each presidential term beside the other so that the viewer could compare the two for how the focus was different as measured by word frequency.

Citations:

- Word Cloud Visualization: https://www.geeksforgeeks.org/generating-word-cloud-python/#:~:text=For%20generating%20word%20cloud%20in,from%20UCI%20Machine%20Learning%20Repository.
- LDA Topic Model: https://neptune.ai/blog/pyldavis-topic-modelling-exploration-tool-that-every-nlp-data-scientist-should-know
- Scattertext Plot : https://kanoki.org/2019/03/17/text-data-visualization-in-python/
- Data Source: https://millercenter.org/the-presidency/presidential-speeches
- BeautifulSoup: https://www.crummy.com/software/BeautifulSoup/bs4/doc/