Anne Ku, Software Design Section 2

Project Overview:

I used the Wikipedia pages that outlined Hillary and Bernie's political views. For this project I stored the content into pickle files my repository, and looped through the file to obtain the most common words in both pages. I used the histogram method from

Implementation:

The first major component stores the words into the pickle files, with one for each candidate. The data-obtaining file relies upon the Wikipedia object and the API calls to obtain the results from the specific searches. After returning the search results in a non-string, I pickled the string so that it can be stored in a .pickle file. I chose to approach data collection with Wikipedia because I wanted to compare the similarities between people of the same organization, and I chose to use pickle because converting the results to one string does not seem worth the effort.

The second major script loops through the file to find words that are especially prominent in those pages. In my case, I decided to make an organized histogram and eliminate the common words that don't add meaning, such as “a”, “by”, “who”, and “but”. This function was helpful in finding the words that people used to describe each candidate. I made another function called “similarities” which takes in both of the organized histograms, loops through each histogram to determine what words they have in common, and returns the list of words in both pages.

Results:

This script helped me accomplish finding the most politically relevant words in each wikipedia page. I decided to look at the 20 words that sounded somewhat political and obtained the following results for Bernie and Hillary:

The top 20 relevant (in a political sense) words in Bernie's political wikipedia page are:
['united', 'people', 'state', 'states', 'act', 'u.s', 'american', 'than', 'tax', 'democratic', 'senator', 'post', 'against', 'vermont', 'their', 'socialist', 'bill', 'should', 'senate', 'one']

The top 20 relevant (in a political sense) words in Hillary's political wikipedia page are:
['new', 'our', 'president', 'act', 'york', '25', 'will', 'u.s', 'senate', '2008', 'american', 'more', 'iraq', 'support', 'state', 'all', 'bill', 'about', 'united', 'their']

I was surprised to see that 'president' was not mentioned in Bernie's top 20, despite the traction of his presidential campaign. For Hillary's list, I was not surprised to see a greater variety of words, from 'more' to 'our', to 'united'.

The shared-word list presented a few themes and similarities between the two candidates. Some of the shared words include 'marriage', 'system', and 'campaign', as both of these words are relevant by virtue of running for president.

Reflection:

If I had more time, I would spend more time processing the content to eliminate irrelevant characters such as slashes and numbers. I would also think of ways to separate

conjoined words, the combination of the last word of the title to the first word of the section. With those issues, I could have improved upon making the word count more accurate to show more relevant results. The scope of the project worked out pretty well in my situation, as I was able to practice pickling data and writing functions as well as explore the commonalities between Hillary and Bernie.