Anne Regel

Data Science Career Track

Milestone Report – Capstone Project 1

**Define the Problem**

Twitter has become a prominent tool for political communication. During the 2016 elections twitter was used extensively by politicians to speak directly to their political opponents and the American people. All of the 100 US senators have an official verified twitter account (<https://www.birdsonganalytics.com/blog/2017/02/21/full-list-of-us-senators-on-twitter-2017/>). According to George Washington University all but 7 congresspersons have twitter accounts (<https://gwu-libraries.github.io/sfm-ui/posts/2017-05-23-congress-seed-list>).

Many news companies and websites have been dedicated to analyzing political tweets. The BBC published 7 articles and 5 TV segments involving Donald Trump’s tweets between June and September 2017. USA Today published an opinion article on July 10th, 2017 suggesting President Trump’s twitter account may be against the law. TweetCongress.org is a website that provides information about all the tweets issued by members of congress. They also offer statistics about most followed and new followers for each congressperson. However, even with all of this monitoring and analysis of political tweets, there is no easy way to compare statements made *via* twitter to official governmental action.

**Identify the Client**

My clients would include anyone interested in using twitter statements in order to anticipate future actions made by the President of the United States of America.

**Describe Data Set (Cleaning/Wrangling)**

My data consists of two sets:

1. The tweets from President Trump

The data was obtained using the Trump Twitter Archive (http://www.trumptwitterarchive.com/archive), the data is from 01/20/2017 - 03/02/2018 2:38 PM MST in json format.

I converted the json into a pandas dataframe. I then used regular expressions to the hashtags and @mentions for each tweet. I also used the nltk library to tokenize the text of every tweet.

1. All documents published by the Executive Office of the United States of America

The Federal Register's website hosts the pdfs of all of the actions published by the Executive Office. I used the requests library to scrape all of the pdfs and the pdfminer (pdfminer.six for python 3.6) to convert the pdfs to text, then regular expressions to find the date each document was published and finally the nltk library to tokenize the text of each document.

I then used my data to determine the most used words in each data set, the words used by both data sets, and the percentage of documents each joint word was used for both datasets.

**Other Potential Data Sets to Use**

I focused on the executive office, however many elected officials have twitter accounts. I could have used the tweets from a Senate member and the documents published by the Senate of the United States of America.

**Explain Initial Findings**

Plotting the percent of documents/tweets the joint words occur in, shows that the tweets and federal documents to not share many words. The word great was used in 17.5% of the tweets, but only 2% of federal documents. Hundred, Lord, and virtue were the top joint words used in the federal data appearing in 3.9%, 3.8%, and 3.2% of the documents, respectively. However, those same words only appeared in 0.033% of tweets. In fact, of the words that appeared in the same percentage of tweets and federal document, had very low usage. The top one being, left, which appeared in 0.5% of both the tweets and federal documents.

These initial finds lead me to believe that word choice alone would not be a good metric to compare the two datasets. Instead I will be using cosine similarity, which vectorizes the features (words) of a document and compares the vectors between documents to score similarity.