Data Expedition 2019 Summary: Text as Data (Advanced)

Becky Tang and Graham Tierney

### Project Summary

Text analysis is an increasingly useful skill for data scientists, particularly those who work closely with the social sciences and humanities. It allows for robustly extracting and classifying information from text documents such as books, tweets, and emails. This data expedition introduced students to the concept of text as data and how to perform text analysis on bodies of text using R. Tweets of four Democratic candidates for the 2020 Primary were used as data to demonstrate basic text analysis techniques such as word frequencies, log-odds ratios for word usage, and pairwise word correlations. Students then learned about more advanced techniques such as classification and topic modeling using State of the Union addresses as a working dataset.

*Graduate Students*: Becky Tang & Graham Tierney, Department of Statistical Science

*Faculty*: Fan Li

*Course*: Sta 440: Case Studies in the Practice of Statistics

### Summary

We first taught this data expedition to Sta 199 students earlier in the Fall 2019 semester. However, students in Sta 440 have a stronger statistical background so we incorporated more advanced topics for this expedition. Students were introduced to the idea of text as data, and we provided two datasets to demonstrate several text analysis tools. The first dataset is comprised of tweets from four Democratic 2020 Primary candidates–Joe Biden, Kamala Harris, Bernie Sanders, and Elizabeth Warren. With this dataset, students were introduced to concepts such as sentiment analysis and comparing documents and text by word frequencies, sentiment, and log-odds of word usage.

The expedition then shifted focus to more complicated text analysis techniques of classification and topic modeling, demonstrated using transcripts from State of the Union addresses beginning with President Nixon’s 1970 address. In particular, the classification task of interest was recovering the political party from the paragraphs of a speech. We employed logistic regression using sentiment analysis as a dimension reduction tool to predict political party. Then moving away from sentiment, we introduced a bag-of-words generative model for classification using Dirichlet, Multinomial, and Dirichlet-Multinomial distributions. Students saw that these methods were able to correctly classify Republican presidents from the text, but sometimes misclassified Democratic presidents. However, classification requires labeled data and it is not always the case that labels are provided, nor does classification tell us about the content of a document. Thus, we finished the expedition by introducing topic models using the Latent Dirichlet Allocation (LDA).

This data expedition gave students an opportunity to see data analysis in action, and were encouraged them to perform text analysis for their final project.

### Data

The Twitter data were scraped using the rtweet() package in R. We obtained the 1,200 most recent tweets at the time of the data expedition from Joe Biden, Kamala Harris, Bernie Sanders, and Elizabeth Warren. In addition to the author and text, other variables include the timestamp of each tweet, if/which hashtag was used, and how many retweets and favorites a tweet garnered (as of July 24, 2019, the time of scraping). However, the analysis mainly focuses on the author and text. A benefit to this dataset is that each tweet is a complete body of text, allowing for easier determination or classification of the sentiment in a tweet as compared to analyzing an entire book or speech. This data also demonstrates the challenge of working with short text, as tweets are constrained to 280 characters.

The State of the Union address transcripts were obtained from The American Presidency Project (<https://www.presidency.ucsb.edu/>), and our dataset begins with President Nixon’s 1970 address. Each transcript is provided as a separate .txt file. Also included is an Excel file specifying the year, president, and the president’s political party corresponding to each address.

The following R packages were used for this expedition: tidytext, tidyverse, stringr, scales, textdata, reshape2, lubridate, widyr, tidyr, igraph, ggraph, quanteda, and topicmodels.