Data Expedition Proposal

Becky Tang and Graham Tierney

8/5/2019

1. **Sponsoring faculty members**: Fan Li (STA 440) and Maria Tackett (STA 199)

2. **Title of datasets**: (1) Democratic Candidate Tweets; (2) Democratic Debate Transcripts, June 2019; (3) State of the Union Transcripts, 1970-present

3. **Description**: The following datasets were chosen for their quality of being flexible to different kinds of analyses which explore real and relevant social science questions.

- **Democratic Candidate Tweets**: The dem_cand_tweets.csv file contains the 1200 most recent tweets from each of the following major Democratic presidential candidates: Joe Biden, Kamala Harris, Bernie Sanders, and Elizabeth Warren. 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). This Twitter data can be used to compare and contrast the four candidates based on the language of their tweets. 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 in a book or speech. The data were scraped via the `rtweet` package in R, which also lends itself as an opportunity for the students to learn how to obtain/scrape their own data (provided they have/create a Twitter account). The dem_cand_tweets.csv file contains 4800 observations of 90 variables.
- **Democratic Debate Transcripts, June 2019**: Each row in the dem_debates.csv file represents a candidate's continuous block of speech before being interrupted by another candidate, with variables specifying the night, placement, speaker, and the text of each block. We also provide text files of the raw transcripts from both debates. The transcripts were obtained from the New York Times, copy-and-pasted into two plain text files. Then the transcripts were then combined into to a single csv file. We choose to provide the debate transcripts as a way to compare/contrast to the aforementioned Twitter data. The scale of the debate transcripts data is much larger, and may require a deeper analysis to characterize text. The dem_debates.csv file has 1136 observations of 4 variables.
- **State of the Union Transcripts, 1970-present**: The data associated with this project are transcripts from State of the Union addresses beginning with President Nixon's 1970 address. Also included is an Excel file stateofunion.xlsx specifying the year, president, and the president's political party corresponding to each address. A major theme presented in *The New Jim Crow* is that mass incarceration in America began when President Nixon coined the term "War on Drugs". We can use the State of the Union addresses to investigate how different presidents have championed this issue. In particular, these speeches allow us to analyze how language has changed over time, and lend themselves nicely to classifying/characterizing this change. The data consist of 50 transcripts saved as text files which vary in length/size. The transcripts were obtained by copy-and-pasting each transcript from the website The American Presidency Project into a .txt file, although in the future we could use a web scraper.

4. **Potential classroom exercises**: We can ask questions at the following levels of granularity. Students will learn to create informative visualizations/graphs and employ simple statistical hypothesis tests to answer many of these questions:

- **Document specific**: What are the major themes/topics presented in Document Z? What is the overall sentiment associated with Document Z?
- **Author/speaker specific** How does Person X's language change over time? How does Person X feel about Issue Y? What words/phrases does Person X use the most?
- **Comparisons across documents and authors** How often does Person X speak about Issue Z in comparison to Person Y? How similar is the language used by different authors or in different documents?

Given a line of text, can we determine who the author is?

5. **Techniques**: All analysis will be performed in R, and the installation of several R packages will be necessary. Therefore to maximize classtime, we would like to have the following packages pre-loaded using a virtual environment or RStudio in the Cloud: `tidytext`, `tidyverse`, `rtweet`, `gutenbergr`, `topicmodels`, `igraph`, `ggraph`, `wordcloud`, `reshape2`. While we describe the major avenues of analysis below, each method provides students the freedom to explore and create several kinds of visualizations, including bar graphs, histograms, network graphs, PCA, and word clouds.

- **Sentiment Analysis**: Sentiment analysis provides a simple way to reduce the dimensionality of text data. Certain words are mapped to a given (short) list of sentiments. Each document is then scored by the number or proportion of words of the given sentiment. Some sentiment dictionaries map words to a continuous positive or negative scale, and then a document's sentiment is the average or total of the score. These scores can be easily visualized in histograms and density plots for different categories of documents or across different levels of some document feature (such as time of creation). As EDA, we can apply PCA using the sentiment scores to observe any clustering of documents. Logistic regression can be used to classify documents based on their sentiment scores, with k-fold cross-validation to assess accuracy. In the case of State of the Union addresses, students can try to predict the party of the author of a given paragraph via the sentiment information. They can also evaluate which Presidents are most distinctive, by comparing the accuracy rate of their logistic regression predictions for each President.
- **Multiple testing for word proportions**: A method to identify words that are used "significantly" differently across authors. If documents in category A use word i at frequency p_{iA} and documents in category B use word i at frequency p_{iB}, a simple two-sample t-test can be used to test the null hypothesis that $p_{iA} = p_{iB}$. However, since the number of unique words in the corpus is quite large, multiple testing corrections are necessary. The Bonferoni correction is useful if one is trying to assess whether there is any difference in word use between the categories. The Benjamini and Hochberg (1995) correction is more useful to identify the subset of words that differentiate the two categories. This method was used in Airoldi et al. (2006).
- **Bag-of-words generative models** (Advanced): The general assumption of these approaches is that authors pick words according to some unknown set of frequencies. The likelihood of an observed word is thus proportional to the unknown frequency. The order of words is generally ignored in these models. With a Dirichlet prior on the frequency vector, the posterior can be easily computed via conjugacy. Students in STA 440 with more statistics background can be assigned to compute these posteriors and their properties as homework problems. For instance, the probability that a text with unknown authorship comes from a certain author is proportional to the prior probably times the posterior-predictive likelihood for that author. This probability can be used to measure classification accuracy and quantification of uncertainty. This method generally requires labeled data from two or more authors. Examples are found in Airoldi et al. (2006) and Gentzcow, Shapiro, and Taddy (2019).
- **Topic modeling with LDA**: LDA is an unsupervised generative model that does not require labeled data. A "topic" is modeled by a frequency vector over all possible words. After estimation, one can inspect the topic-word frequencies to see what concepts they relate to and the document-topic frequencies to see which document come mostly from each topic. Documents can be classified into the topic their words most frequently come from. Students in STA 199 can use the the LDA package, whereas students in STA 440 could program Gibbs samplers (provided they have taken STA 360). (Details: If a word W comes from topic t, then $P(W = w_i) = \beta_{ti}$. For a fixed number of topics, a corpus of documents can be generated by the following process. Choose the length of the document (number of words) via $N \sim Pois(\lambda)$. Choose each document's topic frequency by $\theta_d \sim Dir(\alpha)$. For each word w_i in the document d, draw z_{di} Categorical(θ_d), then draw word w_i Categorical(β_{z_i}).)

6. **Sources**

- https://www.nytimes.com/2019/06/26/us/politics/democratic-debate-transcript.html
- https://www.nytimes.com/2019/06/28/us/politics/transcript-debate.html
- The American Presidency Project (https://www.presidency.ucsb.edu/)