Data Expedition Proposal

Becky Tang and Graham Tierney

7/23/2019

1. Sponsoring faculty member:

- Fan Li (STA 440)
- Maria Tackett (STA 199)

2. Title of dataset(s)

-

- State of the Union Transcripts, 1970-present

3. Description:

- - One-two sentence description of data file
  - Source(s)
  - Why this dataset?
  - How dataset was put together
  - Dimensions of dataset

- 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.
  - The data consist of 50 plain text files. The transcripts vary in length/size, with the 1973nixon.txt file being notably larger than the other transcripts. In 1973 President Nixon gave six separate addresses, and we decided to concatenate these addresses into one large file.
  - The transcripts were obtained from the website The American Presidency Project. We simply copy-and-pasted each transcript into a .txt file, although in the future we could use a web scraper.
  - Motivation: Graham and I recently read *The New Jim Crow: Mass Incarceration in the Age of Colorblindness*, and one major theme presented in the book is that mass incarceration in America began when President Nixon coined the term "War on Drugs". We are interested in exploring this topic further, and see the benefits of connecting statistical methods with the social sciences. In terms of the choice of this specific text, we believe that State of the Union addresses lend themselves nicely to various forms of text analysis. These include topic modeling, classification to president or party, examining what policies were considered most important at the time of the speech based on word/topic frequencies (ex- how crime has been treated by the presidents over time), and much more.

4. Potential classroom exercises

- With these datasets, we can ask questions at different levels of granularity:
  - Document specific (ex. What are the major themes/topics presented in Document Z? What is the overall sentiment associated with Document Z?)
  - Author/speaker specific (ex. How does Person X's language change over time? How does Person X feel about Issue Y?)
  - Comparisons across documents and authors (ex. How often does Person X speak about Issue Y in comparison to Person Y? How similar is the language used by different authors or in different documents? Does Person X use more or less positive language than Person Y? Given text from Document Z, can we determine who the author is?)
- These sorts of questions fall into the two broad categories of sentiment analysis and topic modeling. We envision utilizing the tidytext R package to format the data. At the very base level, we can calculate

raw word frequencies/proportions. We can then build up to correlation tests across authors/documents, difference in proportion tests, PCA analysis, and LDA (more details in the following bullet).

5. Techniques

- List of computational techniques (do we need to ask for VM?)

6. Source(s)

- State of the Union Address Transcripts obtained from The American Presidency Project (https://www.presidency.ucsb.edu/)