February 22, 2017
DS-GA 3001-004, Text as Data
Prof Arthur Spirling

Homework 1

This homework must be returned to Patrick Chester's mailbox (2nd floor, 19 West 4th Street) by **5pm, March 6, 2017**. Late work will incur penalties of the equivalent of one third of a letter grade per day late.

It must be your own work, and your own work only—you must not copy anyone's work, or allow them to copy yours. This extends to writing code. You may consult with others, but when you write up, you must do so alone. Also, while the TA is available to help you and you should make use of his help, please do not wait until the last minute to consult him.

You must include an `R` code appendix which is clearly commented, such that it your understanding of the problems can be assessed. You must turn in a paper copy: **no electronic copies will be accepted**.

---

1. First we'll use the data from the United States Inauguration speeches. Let's look only at the inaugurations given by Bush in 2001 and Obama in 2009.

   (a) Calculate the TTR of each of these speeches.

   (b) Now create a document feature matrix of the two speeches, with no pre-processing other than to remove the punctuation–be sure to check the options on "dfm" in R as appropriate.

   (c) Calculate the cosine distance between the two documents with `quanteda`.

   (d) Now, think about several different pre-processing choices you could make. For each of the following options, make a theoretical argument for how it should affect the TTR of each document and the similarity of the two documents. Then, re-do parts a, b and c with each of these different pre-processing options. To be clear, do each pre-processing option separately.

      - Stemming the words?

      - Removing stop words?

      - Converting all words to lowercase?

      - Does tf-idf weighting make sense here? Explain why or why not

(e) Calculate the MLTD of each speech, with the TTR limit set at .72. Since Quanteda does not support the MLTD function, you may use the function provided in the supplementary code or one of your own creation.

2. Take the following two sentences:

   ```
   Whenever you find you are on the side of the majority, it is time to pause and reflect.

   A jury consists of twelve people who determine which client has the better lawyer.
   ```

   Calculate the Euclidean and Manhattan distance between and the cosine similarity of these two sentences by hand–that is, you can use R, but you can't use the built in functions. Use whatever pre-processing of the text you want, but justify your choice.

3. In their 1963 paper, Mosteller and Wallace used function words to predict the authorship of one of the Federalist Papers. You will perform similar analysis on these texts using the methods we've learned so far. You've been given 5 machine-readable texts each written by Charles Dickens and Jane Austen, as well as a mysterious 11th document. Your task is to determine which of these authors wrote it.

   (a) Before you begin, list one methodological assumption upon which Mosteller and Wallace's attribution of the mystery text depended. Explain how the authors' attribution of authorship might be changed if this assumption was incorrect. Feel free to use an assumption described by Mosteller and Wallace (the paper will be available on NYU Classes and on Patrick's github), one described in Lecture 4, or another that you identified on your own.

   (b) Create a document feature matrix from the documents containing counts of the following function words: the, may, it, was, which, not, be, and upon.

   (c) Using the DFM, calculate the average occurrence rate of *each* function word for both authors and the mystery text. Graphically compare the terms' average rates between the mystery text and each author.

   Hint: To calculate the average occurrence rate of ***one*** *feature, you would take the count of the feature of each document of an author, divide those counts by the total number of features in each document. Then you would average the frequencies for that term across all documents written by that author to obtain the average rate for that feature.*

   (d) For each function word, calculate the absolute difference between their average occurrence in the mystery text and their average occurrence in the texts of each author. Based on your findings, which author do you believe wrote the mystery text?

4. Using all of the Dickens and Austen text, make a graph demonstrating Zipf's law. Also, find the value of and b that best fit this data to Heap's law, fixing $k = 44$.

5. Both Dickens' *Tale of Two Cities* and Austen's *Pride and Prejudice* examine the role of class in British society, but in very different ways. Choose a few Key Words in Context and give a brief discussion of how the two novels treat this theme differently.

6. In this exercise, you will apply the bootstrapping techniques that were demonstrated in Recitation 4. For this problem, you will use the 230 State of the Union Speeches (accessible within the quantedaData package) to produce an estimate of the FRE scores of these speeches aggregated at the presidential level ordered by time (for simplicity's sake, exclude all speeches given by Grover Cleveland, as he was the only non-consecutively elected President). One important change: use sentence-level bootstraps instead of the speech-level bootstraps used in the code from Recitation 4. Also, use the bootstraps to calculate the mean FRE for each President and the standard error of those estimated means.

   - Create a graph that displays the average FRE of each President obtained from the bootstrapped results, the 95% confidence interval associated with each of those means, and the FRE means obtained for each President calculated directly from the unsampled corpus; order your plot by the time the Presidents were in office.

   - Discuss the contrast between the FRE means you estimated with bootstrapping and those you estimated directly from the texts. Are any of the differences you observe statistically significant? If you do observe any mean FRE scores that are significantly different from zero, are there more scores that are significantly different than you would expect given the number of Presidents you have analyzed and the size of the estimated confidence intervals?

   - Finally, for all Presidents, calculate both the FRE scores and the Dale-Chall scores. Report the correlation between the two sets of scores.

   Hint 1: After you split up each speech into sentences, some of the sentences will begin with a number, or not be "sentences" at all, but formal language breaking up the text. Run the attached code on the data frame containing your text split up into sentences to solve these problems.
   Hint 2: To graph the average FRE scores of Presidents ordered over time, feel free to adapt the code used to visualize the bootstrap results from Recitation 4.
   Hint 3: Make sure to choose a large enough number of sentences to sample for each bootstrap so that each of the 40 presidents show up in the sample (again, not including G. Cleveland).