DS-GA 3001-004, Text as Data
Prof. Arthur Spirling
Assignment date: February 15, 2018

Homework 1

This homework must be returned to Leslie Huang's mailbox (2nd floor, 19 West 4th Street) by **5pm, February 28, 2018**. 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 anyone to copy yours. This extends to writing code. You may consult with others, but when you write up, you must do so alone.

Your homework submission must be in one of the following formats: (1) A set of answers and a clearly commented `R` code appendix (use comments to identify code relevant to each answer you produced), (2) A report consisting of clearly marked answers, each accompanied by the relevant code (e.g., a report generated using `rmarkdown`, `knitr`, or similar). **In either case, your code must be included in full, such that 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 State Of the Union speeches available in `quanteda.corpora`. Let's first look at the States of the Union given by Franklin D. Roosevelt in 1936 and 1945.

   (a) Calculate the TTR of each of these speeches and report your findings.

   (b) 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. Calculate the cosine distance between the two documents with `quanteda`. Report your findings.

2. Consider different pre-processing choices you could make. For each of the following parts of this question, you have three tasks: (i), make a theoretical argument for how it should affect the TTR of each document and the similarity of the two documents, and (ii), re-do question (1a) with the pre-processing option indicated, and (iii) redo question(1b) with the pre-processing option indicated.

   To be clear, you must repeat tasks (i-iii) for each pre-processing option below. You should remove punctuation in each step.

   (a) Stemming the words?

(b) Removing stop words?

(c) Converting all words to lowercase?

(d) Does tf-idf weighting make sense here? Explain why or why not.

3. Calculate the MLTD of each of the two speeches by FDR, with the TTR limit set at .72. Rather than covering the entire speech, you can find the Mean Lengths starting from 25 different places in each speech, as long is there is no overlap between the snippets.

   Hint: If you get stuck on this problem, examine the documentation for the library `koRpus` *or contact the TA for a code hint.*

4. Take the following two sentences:

   `"Mr. and Mrs. Dursley, of number four, Privet Drive, were proud to say that they were normal, thank you very much."`

   `"The Dursleys had everything they wanted, but they also had a secret, and their greatest fear was that somebody would discover it."`

   (a) Calculate the Euclidean distance between these sentences by hand—that is, you can use base R, but you can't use functions from `quanteda` or similar. Use whatever preprocessing of the text you want, but justify your choice. Report your findings.

   (b) Calculate the Manhattan distance between these sentences by hand. Report your findings.

   (c) Calculate the cosine similarity between these sentences by hand. Report your findings.

5. One of the earliest and most famous applications of statistical textual analysis was to determine the authorship of texts. You now get to do the same! You've been given 10 machine-readable texts 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 getting started, consider the outliers discussed in Figure 3 of Peng & Hengartner (2002). You need to make sure that the texts you analyze are actually what you want to analyze. Summarize your decisions about how to address this.

   (b) Now choose the features of the texts you want to analyze, and explain your choice.

(c) Prepare and divide up the texts in the way that P&H do. Include a brief discussion of why they (you) make this choice. At the end of this step, you should have a document-feature matrix.

(d) Using the DFM, calculate the average occurrence of the features that you chose in part (b) for each author and the mystery text. Generate a graph that compares the average frequency of the terms you chose in the mystery text and for each author. (Your answer should be a graph.)

(e) Explain how the average term frequency of the mystery text compares with each author. Based on your findings, which author do you believe wrote the mystery text?

6. Using the 10 labeled Dickens and Austen texts, make a graph demonstrating Zipf's law. Include this graph and also discuss any pre-processing decisions you made.

7. Find the value of b that best fit the 10 labeled Dickens and Austen texts to Heap's law, fixing $k = 44$. Report the value of b as well as any pre-processing decisions you made.

8. 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 discuss the different context in which those words are used by each author. Give a brief discussion of how the two novels treat this theme differently.

9. Consider the bootstrapping of the texts we used to calculate the standard errors of the Flesch reading scores of Irish budget speeches in Recitation 4.

   (a) Obtain the UK Labour Party's manifestos from `quanteda.corpora`. Generate estimates of the FRE scores of these manifestos over time, using sentence-level bootstraps instead of the speech-level bootstraps used in Recitation 4. Include a graph of these estimates.

   (b) Report the means of the bootstrapped results and the means observed in the data. Discuss the contrast.

   (c) For the empirical values of each text, calculate the FRE score and the Dale-Chall score. Report the FRE and Dale-Chall scores and the correlation between them.

   Hint: After you split up each speech into sentences, some of the sentences will begin with a number, or not be "sentences" at all (e.g. headings). Regular expressions are one way to remove this kind of text. If you get stuck on this problem, contact the TA for a code hint.
   Hint: Make sure to choose a large enough number of sentences to sample for each bootstrap so that each of the years shows up in the sample.