Homework I

PLSC 497 – Text as Data

Professor Kevin Munger

This homework is due electronically by **11:59 p.m. EST on Wednesday, March 10, 2021**. You can submit your homework by **emailing copies both** to Prof. Munger (kmm7999@psu.edu) and Mr. Villegas-Cruz (amv5718@psu.edu). 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.**

---

1. First we'll use the data from the U.S. inaugural addresses available in quanteda.corpora. Let's first look at the inaugural addresses given by Ronald Reagan in 1981 and 1985.

   (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 similarity between the two documents with ***quanteda***. Report your findings

2. Consider different preprocessing 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 (ii) re-do question (1a) with the preprocessing option indicated and (iii) redo question(1b) with the preprocessing 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. Take the following two headlines:

   "Trump Says He's 'Not Happy' With Border Deal, but Doesn't Say if He Will Sign It."
   "Trump 'not happy' with border deal, weighing options for building wall."

   (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.

4. Collect all of the novels written by Charles Dickens from Project Gutenberg.

   (a) Calculate the following statistics:

      i. Which one has the most tokens?

      ii. Which one is the most lexically diverse?

      iii. Using cosine similarity, which two are the most similar to each other?

   (b) Calculate the Flesch Reading Ease score for each book, and plot these scores on the y-axis with the dates of publication for each book on the x-axis. Is Dickens getting more or less complex over time?

5. Using James Joyce's "A Portrait of the Artist as a Young Man" (gutenberg id = 4217) and Mark Twain's "The Adventures of Tom Sawyer" (gutenberg id = 74), make a graph demonstrating Zipf's law. Include this graph and also discuss any pre-processing decisions you made.

6. Both James Joyce's "A Portrait of the Artist as a Young Man" and Mark Twain's "The Adventures of Tom Sawyer" broach the topic of religion, 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.