PLSC 497 Text as Data
Prof. Kevin Munger
Assignment date: September 13, 2019

Practice Homework

This homework must be returned to Kevin Munger's mailbox (Mailroom, 2nd floor, Pond Lab) by **5pm, September 20, 2019**. 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.

Conceptual Questions:

Question 1) What are latent variables?

Question 2) What is stemming? How is it different from lemmatization?

Question 3) What is a document term matrix? Why is it usually sparse?

Question 4) Explain the tf-idf statistic and the advantage of using it

Question 5) Explain Zipf's Law as it applies to text data

Coding Tasks:

Question 1) Use the Quanteda R package and load in the corpus of presidential inaugural addresses, 'data_corpus_inaugural'. Summarize the corpus.

Question 2) Using the docvars function, save the last name of the presidents in a vector

Question 3) Use the tokens function to split Lincoln's first address in the corpus into words. Remove punctuation and convert the entire text into lowercase.

Question 4) Create a document term matrix to create a matrix of counts of the occurrences of each word in each document. Report how sparse this matrix is.

Question 5) Make a figure which depicts the number of word used by year. Here, the x-axis will depict the year and the y-axis, the number of words used in each inaugural address.