DS-GA 1015, Text as Data
Prof. Arthur Spirling
Assignment date: February 20, 2019

Homework 1

This homework must be returned to Pedro Rodriguez' mailbox (2nd floor, 19 West 4th Street) by **5pm, March 5, 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**.

---

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. Calculate the MLTD of each of the two speeches by RR, 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 as there is no overlap between the snippets.

   Hint: If you get stuck on this problem, examine the documentation for the library ***`koRpus`****.*

4. 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.

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 will be using Leslie Huang’s (a PhD student at NYU’s CDS) `stylest` package. To get the texts for this exercise you will need the `gutenbergr` package.

   (a) First you will need to get the data from Project Gutenberg using their `gutenbergr` package. Download the <u>first four</u> novels for each of the following authors: `Austen, Jane` (*Persuasion*, *Northanger Abbey*, *Mansfield Park* and *Emma*), `Twain, Mark` (*What Is Man? and Other Essays*, *The Adventures of Tom Sawyer*, *Adventures of Huckleberry Finn* and *A Connecticut Yankee in King Arthur’s Cour*t), `Joyce, James` (*Dubliners*, *Chamber Music*, *A Portrait of the Artist as a Young Man* and *Ulysses*) and `Dickens, Charles` (*A Christmas Carol in Prose; Being a Ghost Story of Christmas*, *A Tale of Two Cities*, *The Mystery of Edwin Drood* and *The Pickwick Papers*). From each of these novels extract a short excerpt (e.g. 500 lines of text).

(b) Next you will need to organize the data as required by the package. Create a table (i.e. a dataframe) with one column for the text excerpts and one column identifying the author of each excerpt (although not required to fit the model, also create a column for the title of the novel which the excerpt belongs to).

(c) Now use the `stylest_select_vocab` function to select the terms you will include in your model. Note, this function allows you to include some pre-processing options. Justify any pre-processing choices you make. What percentile (of term frequency) has the best prediction rate? Also report the rate of incorrectly predicted speakers of held-out texts.

(d) Use your optimal percentile from above to subset the terms to be included in your model (this requires you use the `stylest_terms` function). Now go ahead and fit the model using `stylest_fit`. The output of this function includes information on the rate at which each author uses each term (the value is labeled `rate`). Report the top 5 terms (in terms of usage rate) for each author. Do these terms make sense?

(e) Choose any two authors, take the ratio of their rate vectors (make sure dimensions are in the same order) and arrange the resulting vector from largest to smallest values. What are the top 5 terms according to this ratio? How would you interpret this ordering?

(f) Load the mystery excerpt provided. According to your fitted model, who is the most likely author?

6. For this question we will use the `sophistication` package discussed in the lab. The corpus for this exercise will be the U.S. inaugural addresses.

   (a) Using the aforementioned corpus make snippets between 150 to 350 characters in length and clean the snippets (print the top 10).

   (b) Randomly sample 1000 snippets and use these to generate pairs for a minimum spanning tree. From these generate 10 gold pairs (print these —only each pair of text— in your HW). Without looking at the automated classification, read each pair and select whichever you think is "easiest" to read. Now compare your classification with those made by the package. What proportion of the ten gold pairs were you in agreement with the automated classification? Any reasons why you may have arrived at a different judgment?

7. 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.

8. Find the value of b that best fit the two novels from the previous question to Heap's law, fixing $k = 44$. Report the value of b as well as any pre-processing decisions you made.

9. 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.

10. 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 Conservative 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.