POL 504, Text as Data
Prof. Arthur Spirling
Assignment date: September 28, 2023

Homework 1

This homework must be turned in on Canvas by **2pm, October 12, 2023**. 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 a PDF or HTML report, containing all written answers and code, generated from `RMarkdown`. **Raw `.R` or `.Rmd` files will not be accepted.**

Please remember the following:

- Each question part should be clearly labeled in your submission.
- Do not include written answers as code comments. We will not grade code comments.
- The code used to obtain the answer for each question part should accompany the written answer.
- **Your code must be included in full, such that your understanding of the problems can be assessed.**

There are plenty of resources to get started with `RMarkdown` online, see for instance here. You can also use the `template.Rmd` template in our GitHub repo to get started. Using this template is not required.

---

1. Let's begin by loading in data on State of the Union speeches using the `sotu` R package. Create a corpus of SOTU speeches for the years 2007-10.

   (a) Calculate the type-token ratio and Guiraud's index of lexical richness for each of these speeches and report your findings.

   (b) Create a document feature matrix of the speeches, with no preprocessing other than to remove the punctuation–be sure to check the options on "dfm" and "tokens" in R as appropriate. Calculate the cosine similarity between the 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 documents (ii) re-do question (1a) with the preprocessing option indicated and (iii) re-do question (1b) with the preprocessing option indicated.

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

(a) Stem the words

(b) Remove stop words

(c) Convert all words to lowercase

(d) Does tf-idf weighting make sense here? Calculate it and explain why or why not.

3. Recreate the tf-idf table from Table 2, Column 4 of Ballandonne and Cersosimo (2023)[1] for Adam Smith's *Wealth of Nations*. You can find the specific table (and a discussion of the authors' process) in the lecture notes from Week 1. See if you can recreate their top tf-idf terms:

   (a) Load the texts as a corpus.

   (b) Tokenize each text *as single documents*, following the preprocessing of the authors

   (c) Generate a tf-idf document-feature matrix, following the approach of the authors. Report the top 10 features for *Wealth of Nations*. Are they in agreement with Ballandonne and Cersosimo?

   ** Don't stress if you can't get each term to align exactly. Just do your best to follow the process of the authors.*

4. Take the following two headlines:

```
"Biden Administration Loses Expensive Aircraft Because Pilot Scared of Bad Weather"

"U.S. Marine Pilot Ejects From F-35 Aircraft Following Mishap in South Carolina"
```

   (a) Create a DFM of the two sentences. Make sure to remove punctuation and convert the sentences to lower case.

[1] Ballandonne, Matthieu, and Igor Cersosimo. "Towards a "Text as Data" Approach in the History of Economics: An Application to Adam Smith's Classics." *Journal of the History of Economic Thought* 45.1 (2023): 27-49.

(b) Calculate the Euclidean distance between these sentences **by hand—that is, you can use base R, but you can't use distance functions from `quanteda` or similar.** Report your findings.

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

(d) Calculate the Jaccard similarity between these sentences by hand. Report your findings.

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

(f) Calculate the Levenshtein distance between *surveillance* and *surveyance* 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 the `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:

- `Fitzgerald, F. Scott (Francis Scott)` (*This Side of Paradise*, *Flappers and Philosophers*, *The Beautiful and Damned* and *The Great Gatsby*)
- `Melville, Herman` (*Typee: A Romance of the South Seas*, *Moby Dick; Or, The Whale*, *I and My Chimney* and *Omoo: Adventures in the South Seas*)
- `Austen, Jane` (*Persuasion*, *Northanger Abbey*, *Mansfield Park* and *Emma*)
- `Dickens, Charles` (*A Christmas Carol in Prose; Being a Ghost Story of Christmas*, *The Mystery of Edwin Drood*, *The Pickwick Papers* and *A Child's History of England*).

From each of these novels extract a short excerpt (e.g. 500 (random) 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). Print the `str()` of your table.

(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 preprocessing options. Justify any preprocessing choices you make. What percentile (of term frequency) has the best prediction rate? Also report the mean 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 UK Political Party Manifestos data from `quanteda` (data corpus `data_corpus_ukmanifestos`).

   (a) First, extract and concatenate the entire text of the corpus, remove punctuation and set all characters to lower case. Use this text to produce a contingency table for the collocation "United Kingdom" *[Hint: Regular expressions with look-aheads and look-behinds are useful here]*. Calculate the expected frequency of "United Kingdom" under independence. Compare the observed and expected frequency. Based on this comparison, is "United Kingdom" a meaningful multi-word expression in this corpus?

   (b) Now use `quanteda`'s `textstat_collocation` to inspect the same 2-gram "United Kingdom". Report the λ and z values. How do these results relate to your conclusions in question 5(a)?

   (c) Finally, use `textstat_collocation` to inspect all 2-grams with `min_count = 5`. Report the 10 collocations with the largest λ value. Report the 10 collocations with the largest count. Discuss which set of n-grams are likely to be multi-word expressions.

7. Using F. Scott Fitzgerald's "The Great Gatsby" (gutenberg_id = 64317) and Herman Melville's "Moby Dick; Or, The Whale" (gutenberg_id = 2489), make a graph demonstrating Zipf's Law. Include this graph and also discuss any preprocessing decisions you made.

8. Find the value of b that best fits the two novels from the previous question to Heaps' Law, fixing $k = 44$. Report the value of b as well as any preprocessing decisions you made.

9. Let's focus on the UK political party manifestos data again. Pick two parties and consider the context in which speakers from these parties talk about *nation* and *industry.* Use `quanteda`'s `kwic` function to discuss the different context in which those words are used by different countries.

10. Consider the bootstrapping of the texts we used to calculate the standard errors of the Flesch reading scores of Irish budget speeches in Precept 3.

    (a) Obtain the SOTU speeches for US Presidents from 1982-present, using the `sotu` data again. Generate estimates of the FRE scores of these speeches over time (i.e. per year), using sentence-level bootstraps instead of the speech-level bootstraps used in lab. 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 Gunning's-Fog Index score. Report the FRE and Gunning's-Fog Index scores and the correlation between them.

    Hint: After you split up each speech into sentences, some of the sentences will not be "sentences" at all (e.g. headings). Regular expressions are one way to remove this kind of text.

11. As part of the Republican National Convention in 2016, Melania Trump gave a speech. Mrs Trump, or her speechwriters, were subsequently accused of plagiarizing an address given by Michelle Obama at the 2008 Democratic National Convention. Your task now is to scientifically assess the accusations, using aligment techniques. The relevant speeches are called `melania` and `michelle` respectively.

    (a) Using `text.alignment::smith_waterman()` and the standard scoring defaults therein, report the local alignment score for the speeches. Perform your analysis at the `words` (not `characters`) level.

    (b) Repeat the analysis, but changing the gap cost to -5. What is the local alignment score now? Does it imply more, or less, of the speech was plagiarized? Why—what is the mechanism for this change?