DS-GA 1015, Text as Data
Marco Morucci
Assignment date: February 14, 2024

Homework 1

This homework must be turned in on NYU Brightspace by **11pm, Feb 28, 2024**. 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 `IPython`. **Raw `.py` or `.ipynb` files will not be accepted.** You are responsible for making sure that your homework is fully readable in .pdf format. Anything that is not readable will not be graded.

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.**
- Please make sure that code lines are not cut (by breaking up any line of code longer than 80 characters).

---

1. (4 pts.) First we'll use the data from the U.S. inaugural addresses. Let's first look at the inaugural addresses given by Ronald Reagan in 1981 and 1985. Load them in as strings from the files 1981-Reagan.txt and 1985-Reagan.txt respectively.

   (a) (2 pts.) Calculate the TTR and Guiraud's index of lexical richness for each of these speeches and report your findings.

   (b) (2 pts.) Create a document feature matrix of the two speeches, with no preprocessing other than to remove the punctuation. Calculate the cosine similarity between the two documents with `nltk`. Report your findings.

2. (8 pts.) 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 preprocessing option below. You should remove punctuation in each step.

   (a) (2 pts.) Stem the words

   (b) (2 pts.) Remove stop words

   (c) (2 pts.) Convert all words to lowercase

   (d) (2 pts.) Does tf-idf weighting make sense here? Calculate it and explain why or why not.

3. (7 pts.) Take the following two headlines:

   `"China Condemns U.S. Decision to Shoot Down Spy Balloon."`

   `"U.S. Shoots Down Suspected Chinese Spy Balloon, Recovery Under Way."`

   (a) (2 pts.) Create a DTM of the two sentences. Make sure to remove punctuation and convert the sentences to lower case.

   (b) (1 pt.) Calculate the Euclidean distance between these sentences **by hand—that is, you can use base Python, but you can't use distance functions from any library.** Report your findings.

   (c) (1 pt.)Calculate the Manhattan distance between these sentences by hand. Report your findings.

   (d) (1 pt.)Calculate the Jaccard similarity between these sentences by hand. Report your findings.

   (e) (1 pt.)Calculate the cosine similarity between these sentences by hand. Report your findings.

   (f) (1 pt.)Calculate the Levenshtein distance between *surveillance* and *surveyance* by hand. Report your findings.

4. (12 pts.) 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!

   (a) (2 pts.) First you will need to get the data from Project Gutenberg using their `gutenberg` library. Download the first four novels for each of the following authors:

   - `Shelley, Mary Wollstonecraft` (*Frankenstein; Or, The Modern Prometheus*, *Notes to the Complete Poetical Works of Percy Bysshe Shelley*, *Proserpine and Midas* and *Mathilda*)
   - `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 Court*)
   - `Joyce, James` (*Dubliners*, *Chamber Music*, *A Portrait of the Artist as a Young Man* and *Ulysses*)
   - `Hume, Fergus` (*The Green Mummy*, *The Mystery of a Hansom Cab*, *The Secret Passage* and *Madame Midas*).

   From each of these novels extract a short excerpt (e.g. 500 (random) lines of text).

   (b) (2 pts.) Now preprocess and tokenize each of these excerpts using any of the techniques you learned. Concatenate each author's token and create a DTM where each row is one author. Print the first few columns of your DTM.

   (c) (4 pts.) Now take the ratio of vectors of each author vs. the rest. To do so fairly, for each author, sum the rows of the remaining three and divide the resulting frequencies by 3. Then take the ratio of frequencies of the held-out author to the resulting average. Sort the ratios so obtained and display the top-5 highest ratio words for each author.

   (d) (4 pts.) Load the mystery excerpt provided. Using each author's highest-ratio word obtainedin the previous part, perform a χ^2 test where the mystery excerpt is considered as not being that author's work. Which author/word leads to the highest p-value? Which to the lowest? Using this evidence who is most likely to be the author of the mystery excerpt?

5. (8pts.) For this question we will use the UN general debate data available in the UN archive provided.

   (a) (4pts.) Load, 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 Nations". Calculate the expected frequency of "United Nations" under independence. Compare the observed and expected frequency. Based on this comparison, is "United Nations" a meaningful multi-word expression in this corpus?

   (b) (4pts.) Finally, use `nltk's BigramCollocationFinder` 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.

6. (6 pts.) For this question download 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) using the `gutenberg` library.

   (a) (4pts.) Make a graph demonstrating Zipf's law. Include this graph and also discuss any preprocessing decisions you made.

   (b) (2pts.) Find the value of b that best fits the two novels to Heap's law, fixing $k = 44$. Report the value of b.

7. (13 pts.) Consider the bootstrapping of the texts we used to calculate the standard errors of the Flesch reading scores of presidential speeches in lab 3.

   (a) (4 pts.) Load the inaugural speeches for US Presidents from Woodrow Wilson onwards from the text files provided. Split each speech into chunks of 100 tokens each. The last chunk should have 100 tokens plus all the excess tokens.

   (b) (4 pts.) Generate estimates of the FRE scores of these speeches over time (i.e. per year), using a chunk-level bootstrap, that is, for each document, apply the bootstrap procedure to the chunks that it is divided into. Compute and store bootstrapped mean FREs obtained in this way.

   (c) (1 pt.) Plot the estimates you obtained in the previous part.

   (d) (2 pts.) Report the means of the bootstrapped results and the means observed in the data. Discuss the contrast.

   (e) (2 pts.) 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.