Text As Data

Problem Set 1

Columbia University
School of International & Public Affairs
Spring 2024

Please submit your answers in a PDF format by **Wednesday, 2/14 at 11:59pm**. *Any figures that you generate should be added to your answer sheet.* Attach your R code to your submission. You can confer with your colleagues, but your answer must be your own. Plagiarism will be penalized.

Access the data sources by following this link.

Text Preprocessing

The CNN and Fox News corpus (`cnn_fox_corpus.rdata`) contains the transcripts, as well as metadata, for news segments that were aired on CNN and Fox News between 1/7/2021–1/7/2022 that included the keywords *January 6* and *Capitol*. Load the dataset into `R` and answer the following questions.

1. **Tokenization.** Generate a tokenized version of the corpus where tokens are *sentences*. (*Hint*: look at the documentation for the `tokens()` function).
   (a) Inspect the tokenized version of article number 4. How many sentences (tokens) does it have?
   (b) What is the second sentence in article number 600?
2. **Document Feature Matrices.** Generate three different DFM versions from the CNN and Fox News corpus, drawing on various preprocessing steps we discussed in class. (*Note*: you do not have to use sentences as tokens, you can use individual words or n-grams).
   (a) What are the dimensions of each DFM?
   (b) Inspect the three DFMs you generated in step 2a using tools learned in class. In your opinion, which preprocessing version is best for this corpus? Why?
   (c) Create a wordcloud from your chosen DFM. Which word/phrase is the largest?

Descriptive Analysis

3. **Co-occurrences.**

(a) The function `dfm_subset()` allows subsetting a DFM based on metadata. Use this function to generate two separate DFMs from your most preferred DFM from section 2b, where one DFM includes transcripts from CNN and the other from Fox News. What are the dimensions of each DFM?

(b) Generate a *document level* Feature Co-occurrence Matrix (FCM) from the Fox News DFM and create a figure that shows the relationship between the features. Which features have the highest co-occurrence? (*Tip:* calculating an FCM can be computationally intensive if the DFM from which it is derived has many dimensions. To speed up the calculation, make sure that your DFM has $< 10,000$ features.)

(c) Replicate your analysis in Question 3b using the CNN DFM.

4. **Zipf Law.** Generate rank frequency plots for the CNN and Fox News transcripts (*Hint:* you can use the FCMs that you generated in Question 3 for this purpose). Do they have a Zipfian distribution? What are the top 10 features for each source?

5. **Web scraping.** Using the tools learned in class, scrape the content in the Wikipedia page describing the January 6th attack on the United States capitol. The URL of the page can be found here.

   (a) How many paragraphs of text does the page have?

   (b) Preprocess the text using the tools learned in class. What the the top 5 features?

   (c) Write code to scrape the images in the article. How many images are there?

6. **Cosine Similarity.** Now we will examine the similarity between the Wikipedia page that you scraped in Question 5 and the CNN and Fox News transcripts.

   (a) Create a single document of text from the Wikipedia page (*hint:* you can use the `paste()` function with the `collapse` argument for this purpose).

   (b) Create a DFM from the Wikipedia text using the same preprocessing steps that you used for your preferred DFM in Question 2b. What are the dimensions of the Wikipedia DFM?

   (c) Use the `textstat_simil()` function to calculate the cosine similarity between the Wikipedia page and the CNN and Fox News transcripts. (*Tip*: in the function, set `x` to be equal to the DFM of the news transcripts and `y` to be equal to the DFM of the Wikipedia page). Which transcript is most similar to the Wikipedia page? (you can simply provide the document number). Can you tell if this is a CNN or Fox News transcript?