CIV 337 Homework 11

due: **11/20/24**

Instructions: Please submit solutions on canvas. Only a knitted pdf of an `Rmarkdown` file will be accepted.

Problem 1: Text analysis gives researchers a powerful set of tools for extracting information from a large body of documents.[1]

We will analyze data from newspapers across the country to see what topics they cover and how those topics are related to their ideological bias. The authors computed a measure of a newspaper's "slant" by comparing its language to speeches made by Democrats and Republicans in the U.S. Congress.

You will use three data sources for this analysis. The first, `dtm`, is a document term matrix with one row per newspaper, containing the 1000 phrases – stemmed and processed – that do the best job of identifying the speaker as a Republican or a Democrat. For example, "living in poverty" is a phrase most frequently spoken by Democrats, while "global war on terror" is a phrase most frequently spoken by Republicans; a phrase like "exchange rate" would not be included in this dataset, as it is used often by members of both parties and is thus a poor indicator of ideology.

The second object, `papers`, contains some data on the newspapers on which `dtm` is based. The row names in `dtm` correspond to the `newsid` variable in `papers`. The variables are:

Variable	*Description*
`newsid`	The newspaper ID
`paper`	The newspaper name
`city`	The city in which the newspaper is based
`state`	The state in which the newspaper is based
`district`	Congressional district where the newspaper is based (data for Texas only)
`nslant`	The "ideological slant" (lower numbers mean more Democratic)

The third object, `cong`, contains data on members of Congress based on their political speech, which we will compare to the ideological slant of newspapers from the areas that these legislators represent. The variables are:

Variable	*Description*
`legname`	Legislator's name
`state`	Legislator's state
`district`	Legislator's Congressional district
`chamber`	Chamber in which legislator serves (House or Senate)
`party`	Legislator's party
`cslant`	Ideological slant based on legislator's speech (lower numbers mean more Democratic)

a. We will first focus on the slant of newspapers, which the authors define as the tendency to use language that would sway readers to the political left or right. Load the data and plot the distribution of `nslant` in the `papers` data frame, with a vertical line at the median. Which newspaper in the country has the largest left-wing slant? What about right?

[1] This exercise is based on Gentzkow, M. and Shapiro, J. M. – What Drives Media Slant? Evidence From U.S. Daily Newspapers. *Econometrica*. 2010. 78(1): 35-71.

b. We will explore the content of these newspapers using the `wordcloud` package.

First load the `wordcloud` package. Make a word cloud of the top words (at most 20) in the `dtm` object. What were the biggest topics in the news in 2005 when these data were collected? Hint: first convert `dtm` into a `matrix`.

Now subset the data to the tenth of newspapers with the leftmost (lowest) political slant and the rightmost (highest) political slant. Make two word clouds showing the words most commonly used by each group of newspapers (again, at most 20 words). How does their language differ? Do they have anything in common? Hint: to use your usual subsetting/indexing tools, convert your `dtm` matrix into a data frame using the `data.frame` function.

Pay close attention to your warnings, as they contain important information. For extra credit, see if you can make them go away.

c. We will now explore the relationship between the political slant of newspapers and the language used by members of Congress.

Using the dataset `cong`, compute average slant by state separately for the House and Senate. Now use `papers` to compute the average newspaper slant by state. Make two plots with Congressional slant on the x-axis and newspaper slant on the y-axis – one for the House, one for the Senate. Include a best-fit line in each plot – a red one for the Senate and a green one for the House. Label your axes, title your plots, and make sure the axes are the same for comparability. Can you conclude that newspapers are influenced by the political language of elected officials? How else can you interpret the results?

d. We will now take a closer look at the relationship between congressional and media slant at the district level, for one particular state – Texas. To do so, subset the two datasets to Texas alone, then merge them by district and state, keeping only the observations that appear in both datasets. Then, produce the same plot as in question 3 above, but at the district level (just for the House). What do you find? Which results do you think are more informative, and why?

e. Identify the most important terms for capturing regional variation in what is considered newsworthy – the terms that appear frequently in some documents, but not across all documents. To do so, compute the **term frequency-inverse document frequency (TF-IDF)** for each phrase and newspaper combination in the dataset (for this, use the `tm` package and the `dtm` object originally provided).

Subset the TF-IDF transformed matrix you created to contain the newspaper closest to Princeton, the "Home News Tribune" of East Brunswick, NJ. Print the terms with the largest TF-IDF in decreasing order. What topics are of interest to our region, but not likely to make the national news?