Portfolio Problem 3: Text Processing

Data Overview

The President of the United States has given a State of the Union speech every year since 1790. For background on these addresses, you can read the introductory essay published by The American Presidency Project. The file stateofunion1790_2022_tibble.rds (on https://www.math.carleton.edu/ckelling/data_science/) contains the text of every State of the Union speech until 2022 formatted as a tibble, and can be imported using the read_rds() function from {readr}. The text of the speeches was obtained from The American Presidency Project.

Here are the first few rows of the tibble to give you an idea about the structure of the data

```
## # A tibble: 232 x 3
##    name              date              text
##    <chr>             <chr>             <chr>
##  1 George Washington January 8, 1790   "I embrace with great satisfaction the o~
##  2 George Washington December 8, 1790  "In meeting you again I feel much satisf~
##  3 George Washington October 25, 1791  "\"In vain may we expect peace with the ~
##  4 George Washington November 6, 1792  "It is some abatement of the satisfactio~
##  5 George Washington December 3, 1793  "Since the commencement of the term for ~
##  6 George Washington November 19, 1794 "When we call to mind the gracious indul~
##  7 George Washington December 8, 1795  "I trust I do not deceive myself when I ~
##  8 George Washington December 7, 1796  "In recurring to the internal situation ~
##  9 John Adams        November 22, 1797 "I was for some time apprehensive that i~
## 10 John Adams        December 8, 1798  "While with reverence and resignation we~
## # i 222 more rows
```

As you work with these speeches, there are a few things to keep in mind:

- Capitalization isn't important, so deal with this before you start.
- There are unspoken reactions in the text. For example, [laughter] points out where the audience laughed, but this shouldn't count as part of the speech.
- I recommend removing characters that aren't letters or white space
- You **may** use data processing functions that we haven't learned in class (such as `stringr` functions). However, you should not use statistical techniques that you have not learned, such as topic modelling. These methods can easily be used incorrectly, so this project is meant to be an exploratory analysis.
- Some functions that you may consider investigating/using (among many others): `str_split_1`, `group_by`, `str_extract`, `str_detect`, `unnest_tokens` (in the `tidytext` package). We will cover more about text data in class later in the term, time allowing, but this is supposed to be a preliminary exploration.
- A place to start: How often do certain words appear over time? Do not make your whole analysis word clouds...

Assignment

Your goal is to explore how the State of the Union speeches have changed over time. What you explore is entirely up to you, as is the time frame you explore. For example, you might decide to track the usage of key words over time, but I'm not providing specific questions to answer. I want you to craft an interesting question (or set of questions) and work to answer them.

In addition to submitting your R code, you will also record a lightning talk video. A lightning talk is five minutes or less where you describe your research questions and outline your main findings. That's it. I'm looking for a short presentation that is engaging, insightful, and well-crafted.

For your lightning talk, you should create a short slide deck describing your findings. You may use any software to prepare your slides (e.g., Google slides, PowerPoint, Keynote). Once you have your slide deck, you should record a screencast where you present the findings. You can do this using Panopto or any other software you're comfortable with. Once you have recorded your video, please upload it to Panopto and share it with me. (You can click share and then enter my email address: ckelling@carleton.edu.)

To submit your work, push your code (as an .Rmd file) to GitHub and link it to Gradescope. You also need to share the Panopto video with me: please share it with me through Panopto and include a document with a link in your Github/Gradescope submission.

Only one person needs to submit per group.

Assessment

Below, I've included some items that I will consider for assessment of your final portfolio problem.

Professionalism and Reproducibility

- All necessary components of the problem were pushed to GitHub and linked to Gradescope
- The .Rmd file contains all necessary code to reproduce the analysis
- The .Rmd file does not contain any unnecessary code
- The code runs without errors
- The .Rmd file does not produce any unnecessary content (package loading messages, warnings, etc.)
- Code is **well commented**, so a randomly selected classmate could easily navigate it

Analysis

- String processing tools—regular expressions and/or `{stringr}` functionality—was used in the analysis
- `{stringr}` functions were correctly used to derive information from the text column of the data set
- Regular expressions used are correctly implemented
- Analysis (e.g., summary statistics, plot creation) is appropriate for the research question(s)
- Analysis is correctly implemented in R
- **Depth of analysis:** pursue something insightful!

Communication

- Time limit: The lightning talk is 5 minutes or less. (6 minute limit for a group of 3)
- Slides: The slides were clear and readable on the video. Content on the slides emphasizes the story you are telling, they don't distract from the story.
- Organization/clarity: The presentation was clear and easy to follow. It had a clear introduction and conclusion.
- Content: The presentation gives an effective overview of your main findings.
- Delivery: The presentation is well rehearsed, and easy to follow and understand. It is clear that you rehearsed what you would say. There is good pacing (not too quick, not too slow).