

## Homework # 6

Attached you will find a transcript of the recent presidential debate. Open up the file and look through it to familiarize yourself with the format.

1. Write a script that converts the raw text file containing the transcript to a processed dataset. To process your dataset, create a `data.frame` containing the columns `linenumber`, `speaker`, `statement`, and `time`. For each line in the dataset file, the `speaker` column should contain the speaker name, e.g. TRUMP, BIDEN, WALLACE, and the `statement` column should contain the entry of the given line. The `time` column should be a `character.vector` giving the time in minutes and seconds. For example, the time 2:30 would be entered as "2min30sec". Your processed dataset should be this `data.frame` saved as a csv file. (Your identification of the speakers does not have to be perfect. Often, there are a few special cases in a dataset and it is not possible to account for every one. Of course, you need to correctly identify the speakers outside of such special cases.)
2. . Implement the following functions
  - (a) Write a function `get_word_counts(d, speaker)`. This function should return a `data.frame` with the columns, `word` and `count`. The `word` column gives each unique word spoken by the speaker and the `count` column gives the number of times the speaker said the word. The variable `speaker` specifies the speaker, e.g. "Trump". `d` is the processed `data.frame`.
  - (b) Write a function `total_word_counts(d, speakers)`. The variable `speakers` is a `character.vector` containing some combination of the three speakers, e.g. `c("Trump", "Biden")`. This function should return a word count `data.frame`, as in (a), except the word and counts reflect any word spoken by any of the specified speakers.
3. The package `wordcloud` creates word clouds given a collection of words and their frequencies. Install the package and read its documentation, the key function you need to know is `wordcloud`. Write a function `prepare_word_cloud(d, speaker)`.

This function should return a `data.frame` with columns `word` and `weight`, reflecting words spoken by the speaker. The `weight` column should measure how important you think the word is in characterizing the speaker. For example, all speakers say the word “the” many times, but this should have low weight because it is not informative and we do not want “the” in our word clouds. Using your `prepare_word_cloud` function, use the `wordcloud` package to create word clouds for President Trump and Vice President Biden separately, and one for them jointly. To generate informative clouds, you will likely have to iterate a few times between making the clouds and rewriting your `prepare_word_cloud` function. You may find the file `word_frequency.csv` useful. It contains the 5000 most commonly used words in the English language and the relative frequency with which they are used.