**OPTION 1**

**List of websites I plan to use:**

* <https://www.wordfrequency.info/samples.asp>
* <https://www.kaggle.com/datasets/rtatman/english-word-frequency>
* <http://norvig.com/google-books-common-words.txt>
* <https://github.com/filiph/english_words/blob/master/data/word-freq-top5000.csv>

All of these websites provide lists of the some of the most common English words used on the Internet as well as their approximate frequencies. The first site allows you to download the top 5000 entries of a 60,000-word table that is based on data from the Corpus of Contemporary American English. The next website is a dataset from Kaggle of the 333,333 most used words, which comes from the Google Web Trillion Word Corpus. I found the last two websites looking through forums of people asking for large datasets of the most commonly used words on the Internet. From what I understand, they each draw data from studies of the most popular English words and/or popular websites on the Internet. For three of these four sources, I should be able to simply download the raw data as a csv file. For the one that I cannot, I will likely copy the data from the website and paste it into Excel where it can be saved as a csv file. Once I have downloaded all of this data and have them each in their own csv file, I will then eventually combine all of the data into one csv file with updated word frequencies. I will most likely store this data locally on my MacBook for easy accessibility, but if this becomes an issue for some reason, I will plan to move it to the cloud.

**OPTION 2**

**List of websites I plan to use:**

* New York Times
* Washington Post
* NPR
* CBS News

If I were to do this option, I would use this website (<https://wordcounter.net/website-word-count>) to acquire a list of all of the different words and their frequencies from the homepages of each of these popular online news outlets. From there, I would copy the list of words from each homepage and paste them into their own csv file in Excel. Similar to the previous option, I would then eventually combine all of the data into one single csv file with the frequency for each word added up between the different websites. I would once again store this data locally on my MacBook. The potential problems with this option compared to the last option is that each source in option 1 has at least around 5000 unique words. On the other hand, from what I could tell, each source in this option averages around 2000 (probably a little less) words, thus providing a much less comprehensive list (and the site providing the data does give some words that are not actually valid words). Additionally, the frequencies/sample set of words are much smaller for the news homepages than they are in the datasets in option 1 (leading me to be less confident in the results). However, this option does have more of a specific domain than the first option does.