Project 3

The objective of this project is to perform word frequency analysis.

This link provides Twitter data of Elon Musk from 2010-2022. For analysis consider the years 2017-2021 (last 5 complete years). Each year has thousands of tweets. Assume each year to be a document (all the tweets in one year will be considered as a document)

1. Compute the term frequencies for each year. They should be normalized (scale of [0, 1]). Exclude stopwords.
2. Show the top 10 words (for each year) by highest value of word frequency.
3. Plot a histogram of word frequencies for each year
4. Demonstrate Zipf's law by plotting log-log plots of word frequencies v. rank for each year
5. Use TF-IDF to calculate and show the 5 most "important" words for each year

Submission Format

1. Submit all the solutions as a Python notebook (.ipynb) or PDF
2. Students can create their own custom functions if necessary
3. This is a group effort; only one member from each group needs to submit the solution
4. Submit the solutions by 12pm PT on December 12