Shubham Mahajan

A quick overview of my proposal is that I am building a web application that will simplify complex English text. I plan to do this by scrapping articles and calculating term frequencies for each word. I assumed that simpler words have a higher term frequency. I will use the nltk library to determine the synonyms of any low-frequency words and replace them with a synonym that has a higher frequency. All of this is done in Python.

So far, I found some New York Times articles from the past 100 years (from 1920 – 2020). I needed to partition the file as Github would only allow a max file size of 100MB. I then needed to scrape and tokenize all these articles using regular expressions. After this, I can calculate the term frequencies. I managed to spin up an AWS EC2 instance to do all of these computationally intensive tasks. I stored all of these term frequencies in a text file.

I need to calculate the inverse document frequency of each term to determine the stop words. This should be a similar process to calculating the term frequencies. Once I do that, I can need to build an endpoint that will determine which words in the text have a low term frequency and not a high inverse document frequency (prevents replacing stopwords with excluding high inverse document frequency). This endpoint will determine the synonyms of the low-term frequency words and try to use words with a higher-term frequency. The endpoint will return the new simplified text. I need to determine an appropriate threshold to figure out how low the term frequency has to be. After this, I will need to build a web application (most likely will build a Flask app) which uses the endpoint mentioned before and display the simplified text.

As of right now, the only challenge I faced was being able to process all the data locally. Thankfully, I could easily spin up an AWS EC2 instance that could process all the data. Also uploading all the data to Github was a pain due to slow upload speeds. Once again, using EC2 helped with this. I am sure I will have some problems attempting to build a flask app since I do not have experience with flask.