Project 1

This partial data set, sourced from Kaggle, contains research articles related to various specializations related to COVID-19. This corpus has around 56000+ files.

In this assignment, you must perform a set of tasks from the JSON-encoded partial COVID-19 dataset as given below.

Tasks

1. Extract the text content from the JSON-encoded data set and create a text corpus. You may use any JSON library to extract the text
2. Develop your preprocessing steps and order of steps
3. Count the frequency of the word in the vocabulary and compute its corresponding rank. Using this table, find the average value of alpha
4. Plot Tokens Vs Vocabulary graph using Heaps' empirical law. Find Vocabulary count for every 10000 tokens. You may use a log scale for plotting

Note

1. Try your assignment with a smaller corpus initially. Once the output of your functions is as expected, you may proceed to extract all the content and perform the tasks to complete the assignment.
2. Use functional-style programming
3. Write at least 1-3 lines of comments for every function.
4. You may use any JSON library to extract the text from the files.
5. You may use NLTK or SpaCy for lemmatization and stemming operations.
6. You may use regex libraries to remove unwanted words from the corpus.
7. Keep the processed corpus safe. It will be of use in the next assignment.

https://drive.google.com/file/d/1wh4roCvvSbDQqWGwj1SkhLU70-5e7YFD/view?usp=sharing