CMSC 676: Project Phase 5 Akarsh Kashamshetty

**PHASE 5 – REPORT**

**Preprocessing:**

In this phase, I made sure to remove the words given in the stopwords list and I have removed the unique words that found in the documents whose frequency is 1. Also, the words that are of length 1. After this preprocessing most of the tokens looked meaningful and better when compared to the phase 2 Tokenizer.

**Approach**:

After creating tokens for each document. I concatenated it after preprocessing is done. and stored in dictionary with key as document it and values as a string of words in that document.

Later to find the similarities of the document I used the package spacy. Using spacy found the similarity of documents. Used two for loops to select one document and compare it with the rest in the corpus and done same for the rest. And maintained two variables for storing the maximum similarity and minimum similarity along with the document names.

As this this approach takes time complexity O(N2) the program took a while to finish.

**Other approaches:**

As the above process takes much time I tried using the sklearn package using the feature extraction and similarity function. However, I was unable to handle few errors thrown.

For the document that is close to the centroid. I tried using the Agglomerative clustering of sklearn package to find the labels of the clusters. And for finding the document that is close to the centroid of the whole corpus thought of using NearestCentroid package from the sklearn. But I could not be able to do that due to some errors.

**Output**

**Text

Description automatically generated**

Here we can see that document 296.htmland 206.html are most similar and documents 233.html and 392.html are most dissimilar.

**Command**

python3 main5.py files/

Improvements: Can make the algorithm for find the similar documents to me more efficient by using other packages in python.