Assignment 2 Basic Text Processing

Approach and Code Development

I created the text processor using the NLTK library, the collections tool and other basic list processing.

The first task was to remove the integers and real numbers. I do two scans over the tokenized list. In the first pass I remove the '.' character if present between two integers in order to take care of cases like 95.27%. This has to be done because we can easily remove all the integers from the file but once the integers are removed such '.' characters need not be in the list. In the second pass the tokens are checked for integers and are accordingly removed.

The next part is about stemming or lemmatization. Here for the purposes of this assignment, I thought that the usage of the Porter Stemmer could give maximum accuracy and hence I used it.

Since we might want to retrieve information from the tokens we have generated, it is necessary to remove stop words. I used a NLTK library function to do the same.

In the end we just had to find the count of each token. Here I have first converted all characters in all tokens to the lower case and then used the collections toolkit to count the frequency.

Datasets and Experiments

I ran a large number of test cases myself to check for the accuracy of my code and to check out for any corner cases such as the removal of the '.' characters in the first task.

Collaborators

Since the assignment was pretty straight forward and I had a clear idea of what all was required to be done in the assignment, I did not have to collaborate with any of my classmates for this assignment.