IR Term Project Report Part 3
Sanskar Patni 19CH10046

The contribution I had in the Part 3 of default term project are:

TASK #3A(Relevance Feedback)

For this task I contributed in:
I did the **preprocessing, using nltk libraries**. Removed stopwords and performed leammatization
I created the **term frequency and document frequency matrix** for the corpus.
The main problem faced here was the size of the tf matrix.Even after reducing it to 20k vocabulary the tf matrix(20k*50k) was too large to be loaded onto RAM. The solution that I came up with was to store only the non-zero values, this way the previous sparces tf matrix was reduced a lot in size.
One more design choice was for **defining the vocabulary**, I chose to take the words which occur in maximum number of documents. Therefore the 20k vocabulary is for the words that occur in most number of documents.

Using the tf and df values, I wrote the code for calculating **tf-idf** index for documents and queries.
I wrote the function for cosine normalization of the tf-idf matrix, to make the code more clean.