**Information Retrieval**

The IR Engine that has been designed for the class project is as per the specifications contains an initial Text parser followed by the Indexer in this current update.

The system has the following implementations:

1. Load the files i.e. TrecData that is provided.
2. Identify the document numbers and its corresponding text tags. Thus, place the document names by assigning a document number in a dedicated hashmap.
3. Also, place by identifying the corresponding individual words in the text, by first:
   1. Tokenize the words, i.e. remove words with numbers, lower case the words, remove special characters.
   2. Check that the word is not present in stop list, if not present continue to stem the word and store it in a uniquely sorted word dictionary.
4. In this current version, the code is updated to incorporate:
   1. Forward index which contains: DocId – WordId and freq. count respectively.
   2. Inverted Index which contains: WordID – DocId and freq. count of the respective word in the document.
5. The system also contains a test class which takes in the path parse from the user and ask for a word to search, then prints out the documents which contain that word.

* The current system takes around **13 seconds** for the total indexing which can be cross verified by executing the current IR system.
* The forward index size is 5368
* The Inverted index size is 32842

In the third assignment, a txt file with queries has been taken and the queries were parsed as suggested in steps 2,3&4 for the TREC data files.

* A weighted index has been calculated for each word in the queries file and for all the matched words in TREC Data file.
* The Cosine similarity is then calculated for the queries and TREC Data.
* Based on the similarity values, ranks have been assigned to each word uniquely.

The program has been executed in two ways by taking just the Title tag and then considering both Title and Description tags in the query file. When only the Title tag was taken, about 2000 and odd documents were retrieved. And when the Description tag was also considered along with the Title tag. About 3000 documents were retrieved. Hence, we are considering the Precision to be 0.0017 approx and recall to be 0.019. In the given main.qrels they have given documents which are not in trec data I am including them also.