**IR ASSIGNMENT 4 README FILE**

**Pre-processing steps:** The pre-processing steps used for the dataset.

* All the text is converted to lower case
* All meta-data is removed from the text
* All email-ids are removed from the text
* Tokens which consist of letters as well as digits like play2 are removed from the text.
* All the punctuation marks are removed from the text
* All stop-words are removed from the text

Lemmatization is applied on the text of the dataset as lemmatization is a better technique to apply in case of search engine type applications.

All the above pre-processing steps except removing the metadata are done for the query taken as input from the user as well.

**Methodology Used:**

**Question 1:**

* First of all, in Question 1 the vectors of length equal to the size of the vocabulary are prepared for all the documents using the tf-idf values of the vocab terms.
* The doc name consists of the containing folder name appended with the doc name.
* All the vectors are stored as lists. Now these lists are the values of the dictionary. The dictionary is pickled to be used further in Rocchio algorithm.

**Question 2:**

* Now the value of k is taken as input from the user.
* The no. of queries the user wishes to enter is taken as input from the user.
* Also for each of the queries the folder whose docs are to be considered as the ground truth is also taken as input from the user.
* All the pre – processing steps are applied on the query and the query is word-tokenized.
* Now vector of size equal to the vocab size are prepared for each of the queries.
* The cosine similarity score of the query vector with each of the document vectors in the corpus is calculated.
* The user is asked for the value of p. These p% of k documents which are also present in the ground truth folder also are displayed with a \*.
* Now for the initial query the top k documents are displayed according to the descending cosine similarity scores.
* The PR curve is displayed.
* The updated query is found using the equation of Rocchio algorithm.
* Further the top k documents are found using updated query.
* Further the Mean Average Precision values are calculated.
* Also the TSNE curve is plotted.