Assignment 3

Q1.

Preprocessing: I used os liberary to walk in entire directories and make a all directory documents list and then one by one i took all the documents and open the documents one by one. By using nltk word_tokenizer i tokenize the entire document content and then i made inverted index on the bases of tf and make high and low list for each term according to their static scores. In the last i stored all the dictionaries in pickle files.

Methodology: After preprocessing I took the dictionary and take a query preprocess this query similar to the above documents and make a query vector now I take all the high list of each present term if they are less than k than also consider the low lists of the terms in query now I computed the cosine similarity between the query and the documents. After all these steps I ranked all the resultant documents with g(d)+cos(query, doc) score.

Q2.

methodology, preprocessing steps and assumptions

Preprocessing: After open the file i used reader as line by line and store only those lines where i have qid:4 and thake these urls for the questions.

Methodology: I sorted these lines on the basis of the first column as a relevance score for max dcg and store in a file.

For the second part I calculate the dcg for 50 documents for URL with qid:4 and the same calculate idcg for 50 documents after the sort of the lines by its relevance score. Now i calculated ndcg=dcg/idcg.

Same for all the dataset with qid:4

In the third part, I calculated precision and recall and store these values in a list and by using matplotlib library I plot the Precision-Recall curve.