ASSIGNMENT-3

SWAGATAM CHAKRABORTI(MT18146)

1. Tf-idf based document retrieval and relevance feedback

PREPROCESSING:

1. Each line of the file is pre processed individually
2. Word tokenization is done using nltk library using the regextokenization which handles the formation of the tokens and also the removal of the punctuations
3. Stopwords have been removed from the tokens formed.
4. Lemmatization have been performed over the tokens using the nltk library
5. If the line contains any numbers, it is converted into words using inject library and is stored in the vocab along with the number itself

METHODOLOGY:

1. Traverse through all the documents, preprocess the data and maintain a vocab dictionary for each documents having word as the key and corresponding term frequency of the word in the document.
2. Create a list of dictionaries, having document name as the key for each dictionary and vocab dictionary of the corresponding calculated in the previous step as the value.
3. For the query entered by the user is pre-processed and a processed query is obtained.
4. For each words in the query, calculate the inverse document frequency, fot each documents, if the word matches in the vocab of the document, the tf-idf score is multiplied and appended in the dictionary with the document name as the key and tf-idf score as the value
5. Finally the dictionary is sorted based on the if-idf score.
6. For displaying the top k relevant documents, firstly the documents present in the match with the title displayed and remaining documents among the k documents is displayed from the sorted dictionary of document tf-idf score.

7. Based on the top K documents user provide the feedback for the K documents.
8. From the relevance documents we find the relevance centroid.
9. From the non relevance documents we find the non relevance centroid.
10. Taking alpha value as 1 and beta value as 0.7 we obtain the modify query of length of vocab based on the rochio algorithm. Repeat from step 5.

RESULT AND ANALYSIS:

Query: Navy engineering software systems

Top 10 documents:

ASSIGNMENT_1Q2_dataset\comp.graphics\38609 0.23634924248300748

ASSIGNMENT_1Q2_dataset\rec.motorcycles\104529 0.22736486159739

ASSIGNMENT_1Q2_dataset\comp.graphics\37261 0.21992594302269824

ASSIGNMENT_1Q2_dataset\comp.graphics\38947 0.1891285382086621

ASSIGNMENT_1Q2_dataset\comp.graphics\38625 0.1761849536586591

ASSIGNMENT_1Q2_dataset\comp.graphics\39738 0.1620484101023273

ASSIGNMENT_1Q2_dataset\comp.graphics\39655 0.15846910912665435

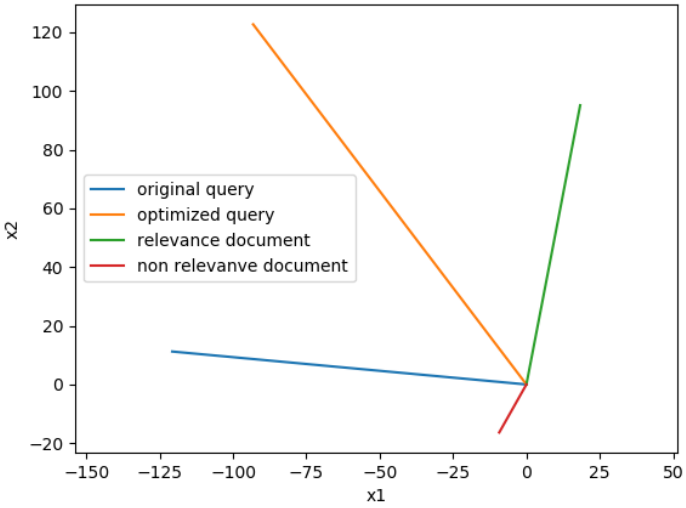
ASSIGNMENT_1Q2_dataset\comp.graphics\38976 0.1562327075167356

ASSIGNMENT_1Q2_dataset\comp.graphics\38241 0.1555031122478027

ASSIGNMENT_1Q2_dataset\comp.graphics\38855 0.15298591881092605

Relevance documents:

comp.graphics\38609, rec.motorcycles\104529, comp.graphics\37261, comp.graphics\38947, comp.graphics\38625



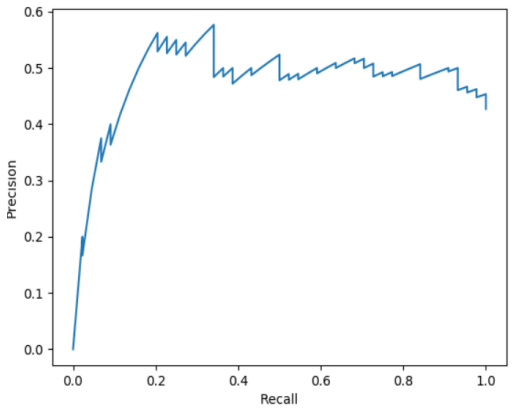
INFERENCE: From the 2D TSME plot it is evident that the modified query vector gets shifted away from the non-relevance document centroid vector.

2. PRECISION RECALL CURVE

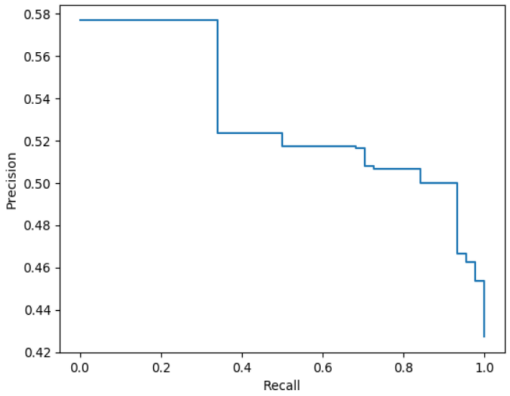
METHODOLOGY:

1. From the dataset obtain the queries having quid:4
2. From the quid:4 data obtain the feature 75 and the 0^{th} feature i.e the relevance score.
3. Sort the data based on the 75^{th} feature in descending order.
4. Calculate the precision and recall for each document retrieval.
5. Plot the precision and recall curve i.e Interpolated curve and 11 point precision and recall curve.

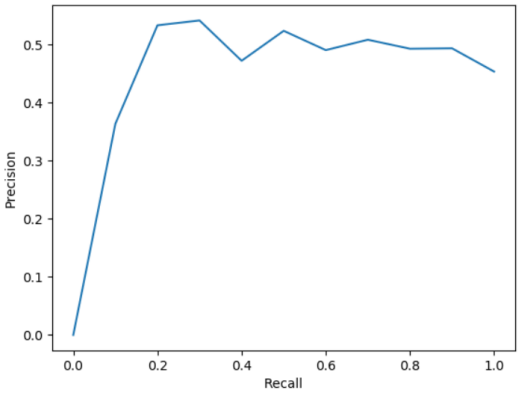
RESULT ANALYSIS:



Normal precision-recall curve



Interpolated precision-recall curve



11 point precision recall curve