**IR Assignment 4**

Rocchio Algorithm

Shalini Bhardwaj( MT19045 )

**1.Preprocessing**

Dataset: 20Newsgroup’s 5 folders

1.com.graphics 2.rec.sport.hockey 3.sci.med 4.sci.space 5.talk.politcs.misc

Attached Folder Named: 20\_newsgroups\_assi4 has all these folders.

**1.Preprocessing**

Preprocessing steps: Nltk library is used for preprocessing

1. To Lowercase: query is transformed to lower case

2. Punctuation is removed.

3. Stop words are removed

4. Tokenization : converted to tokens

5. Lemmatization : converting to root dictionary word

Both query and dataset are preprocessed.

Files from all folders are kept in one place and then read.

**2.Methodology**

1.A dictionary of dictionary is built for storing tfs of all the vocab terms corresponding to documents in which they are present. It is named invIndex. While calculatinf tf it is normalized by documents length.

2.In other dictionary ids are stored for voacb terms and stored.

3.Finally these are mapped to calculate tf-idf values of vocab term.So that this can be used to calculate cosine similarity score of terms in documents.It is named dict\_tfidf.It is dictionary of dictionary having term and docs as keys and tfidf as final value.

4.All 5000 documents of all folders are represented as vectors with dimensions as all vocab terms.Vectors has value tfidf.

5. Similarly query is also transformed to vector with value as tfidfs with respect to query vector tf and idf are calculated.

6.Cosine similarity between query and document is calculated then top K(entered by user )

are retrieved and shown to user.

7.To improvise the query to have better results feedback is taken from the user.User will enter the folder from which query is taken (so that we can analyse our algorithm).

8.Using Rochio Algorithm and given alpha beta and gamma values various weightage to relevant and non relevant docs is given and hence docs are evaluated.

10.Four iterations are performed in which user gives feedback and hence gets better values.

9.TO do analysis of our algorithm once user mark the doc are relevant \* is appended to it and its position in next iteration can be as comparison to improvement in query done.

10.Once relevant docs are marked they are not counted when query is updated in next .Hence each time new p%of k docs are kept in relevant docs of rocchio algorithm. As previous relevant docs are accommodated in the new query vector.

11.Performance of algorithm is measured using precision recall curve,MAP and TSNE.

Assumptions:

1.Original Document name are mapped to document number starting from 1 to 5000. Its mapping is shown in code.

DOC Name DOC Number

1.com.graphics 1-1000

2.rec.sport.hockey 1001-2000

3.sci.med 2001-3000

4.sci.space 3001-4000

5.talk.politcs.misc 4001-5000

Hence we can get the document to which query is most similar.

2.P is chosen as 10% of total documents user wants to retrieved.

3.Automated relevant document selection from top of these all retrieved documents is being done as per ground truth folder entered by user.