ASSIGNMENT 4
NAME:-Anand Sharma (MT19059)
READ ME FILE

DATASET USED :- 20_newsgroups:- comp.graphics , sci.med, talk.politics.misc , rec.sport.hockey , sci.space.

• Question 1

PREPROCESSING:-

- converted query and tokens of document to lower case
- used RegexTokenizer to tokenize both query and tokens of document which will convert it to tokens and remove extra things from it
- Used lemmatization and converted similiar words to same tokens in both query and document tokens
- Removed stopwords from the given text
- Used stemming for both data and query
- Converted numbers to words for data and query

Assumptions :-

- as given in question all things are assumed

Methodology:-

- Dataset is opened and various preprocessing steps are applied and then documents and words are used to create a proper dictionary to calculate TF ,IDF and DF .
- Using these values 2 numpy matrix is created for query and dataset.
- These matrix are used to calculate cosine scores

Question 2:-

PREPROCESSING:-

1.converted query and tokens of document to lower case
2.used RegexTokenizer to tokenize both query and
tokens of document which will convert it to tokens
and remove extra things from it
3.Used lemmatization and converted similiar words to
same tokens in both query and document tokens
4.Removed stopwords from the given text
5.Used stemming for both data and query

6. Converted numbers to words for data and query

Assumptions :-

- as given in question all things are assumed
- ground truth values are assumed as specified in question
- Used $\alpha= 1$, $\beta= 0.75$, and $\gamma=0.25$ as parameters for the Rocchio’s algorithm.
- Assumption of docid are:-
- (0-999):- 20_newsgroups/sci.med
  (1000-1999):-20_newsgroups/comp.graphics
  (2000-2999):-20_newsgroups/rec.sport.hockey
  (3000-3999):-20_newsgroups/talk.politics.misc
  (4000-4999):-20_newsgroups/sci.space

Methodology:-

- Dataset is opened and various preprocessing steps are applied and then documents and words are used to create a proper dictionary to calculate TF ,IDF and DF .
- Using these values 2 numpy matrix is created for query and dataset.
- These matrix are used to calculate cosine scores
- after calculating cosine score results are shown to user and then feedback is taken from the user regarding relevant docs. User mark p% of total k retrieved docs to be relevant
- Applying Rocchio algorithm to docs . 3 vectors are created. For relevant docs ,non relevant docs and modified query vector
- applying rocchio alforithm formula Q_m=alpha*old_Q+(beta*meanof(relevant)-gamma*meanof(nonrelevant))
- then new cosine are calculated based on modified query vector
- Precision Recall curve is plotted and TSNE is ploted to visualize the changes per iteration