ASSIGNMENT 5
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 tokens of document to lower case
- used RegexTokenizer to tokenize wordsof document which will convert it to tokens and remove extra things from it
- Used lemmatization and converted similiar words to same tokens for document words
- Removed stopwords from the given text of dataset
- Used stemming for data
- Converted numbers to words for data and query

Assumptions :-

- as given in question all things are assumed
- For TF-IDF feature selection classwise top k% features are selected which are common for each class
- For Mutual Information based feature selection top K% features for each class are selected.
- Means that each class has their own top k% features

Methodology:-

- Two types of Feature Selection Technique are applied top k% features are selected based on TF-IDF score and Mutual Information
- Two types of Machine Learning Algorithms are applied ie Naive Bayes classififcation and KNN classifiers
- First class Frequency is calculated for each word in documents and then Inverse class frequency .Then Term frequency is calculated for each term in class . Based on this we calculate TF-IDF score with word . Then sorted and taken top k % features for classifiers
- For Mutual Information first N11,N01,N10,N00 are calculated which has their usual meaning for each term and each class. Based on this formula of Mutual Information is applied and MI is calculated for each term and ecah class. Top k % words are selected for each class by sorting them reverse order
- Naive Bayes Classification is applied based on features selection of both techniques and results are compared . Frequency of a term in a document is calculated and total length for each class .

Using normalized naive bayes formula a class is predicted . Which is checked with actual class for accuracy and confusion matrix

- In KNN classifiers 2 matrix are created for training and testing of size no of document*selected features . These matrix are used to calculate cosine similarity between training and testing document . And according to value of K final class is taken as most occuring predicted class .
- These actual and predicted class is used to calculate accuracy and confusion matrix .

How to run .py file:

- **Open terminal and run question1.py**
- **ipynb file is also submitted for running**
- **Dataset file 20newsgroup should be in same folder while running program**