Readme(Assignment 5)

(MT19111) Akshaj Patil

1:PreProcessing Steps:

- Case folding (all terms are converted to lower case).
- Stop words removal.
- Removal of punctuations.
- Numbers are converted to Words.
- Lemmatization.

Same pre-processing is done for train data and test data.

2:Assumptions:

- Meta data of the documents is considered as part of document.

3:Methodology:

First user will be asked amount of data he want to keep as training set and remaining will be considered as testing set.

While data pre-processing we have to calculate tf-idf in two ways. 1] Class wise ti-idf (used in Naïve Bayes) here each class is considered as one document and tf-idf is calculated accordingly. 2]Document wise tf-idf (used in KNN) here each document is considered separately and tf-idf is calculated accordingly.

- MI(Mutual Information) is calculated as class-document-term basis.

Below formula is used for calculating MI .

$$I(U;C) = \frac{N_{11}}{N}\log_2\frac{NN_{11}}{N_{1.}N_{.1}} + \frac{N_{01}}{N}\log_2\frac{NN_{01}}{N_{0.}N_{.1}} + \frac{N_{10}}{N}\log_2\frac{NN_{10}}{N_{1.}N_{.0}} + \frac{N_{00}}{N}\log_2\frac{NN_{00}}{N_{0.}N_{.0}}$$

Now user will be given a choice to select algorithm for feature selection. 1:]TF-IDF 2:]MI.

And according to users choice features will be selected accordingly.

For Tf-idf for each class terms will be arranged in decreasing order of tf-idf values. And first 30% terms are selected as features.

For MI each doc terms will be arranged in decreasing order of MI values. And first 60% terms are selected as Features.

After feature selection Naïve Bayes algorithm is used for classification. And confusion matrix along with accuracy is printed.

After feature selection KNN algorithm is used for classification with k=1,3,5. And coufusion matrix and accuracy is printed along with accuracy graph at k=1,3,5.

Functions:

def calculateIdf(df): To calculate idf according to class basis.

def calculateIdfdoc(df): To calculate idf according to document basis.

def calculate_tfIdf(dictIdf): To calculate tf-idf according to class basis.

def calculate_cond_prob():To calculate conditional probability of each term.

def calculate_prior_prob(): To calculate prior probability of each class.

def calculate_MI(): To calculate MI of each term.

def select_MI_features(): Select features according to MI.

def feature_selection_tfidf(): Select features according to tf-idf.

def classify_naive(test_doc): To classify document according to naïve bayes.

def calculateCosineScore(queryVectorTemp,docVectorTemp): To calculate cosine score between two vectors.

def knn(test_doc,k): Classify documents according to KNN.

How to Run Program.

There are total 13 cells

Run 1st cell which contains header file.

Run 2nd cell which contaion functions.

Run 3rd cell and enter percentage of train data you want.

Run 4th cell which is for pre processing data.

Run 5th cell which is for calculating tfidf.

Run 6th cell which calculates prior probability.

Run 7th cell which contains MI function.

Run 8th cell which will ask you to choose algorithm for feature selection.

Run 9th cell which will classify your test docs.

Run 10th cell which will print confusion matrix with accuracy of Naïve bayes.

Run 11th cell which will classify test docs according to KNN.

Run 12th cell which will plot graph of KNN for k=1,2,3.

Run 13th cell which will plot graph of Naïve bayes for different splits. (Note:- here acc is manually entered and pic of those accuracies are attached in analysis file).

Now if you want to run program on different split then run cells from 3 to 13 again.