Assumptions :

- While removing the punctuations the letters are separated. For example u.s.a. will become u s a
- The distance measure is Euclidean Distance(RSS)

KNN for Text Classification :

a).Pre-Processing :

- Tokenization : Longer strings are splits into tokens. In this the text is tokenized into sentences and then sentences are tokenized into words. During tokenization the punctuations are also removed.
- After tokenizing the documents the number is converted to words(like 100 is converted to one hundred).If size is greater than each digit of number is converted to word.(like 100 is converted to one zero zero)
- Case-Folding(Lower Case) : Converting all words to lower case to make searching easy.
- Stop 54Words Removal : Stop words are removed.
- Stemming : It is process of eliminating affixes from the word to get the stem word. Generally it makes word running to run.

b).Methodology :

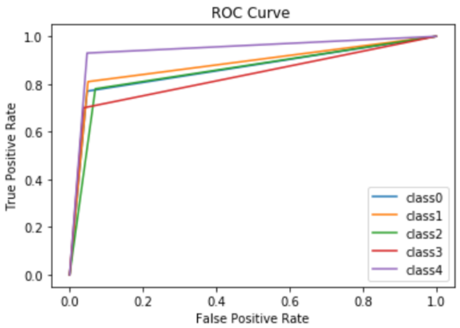
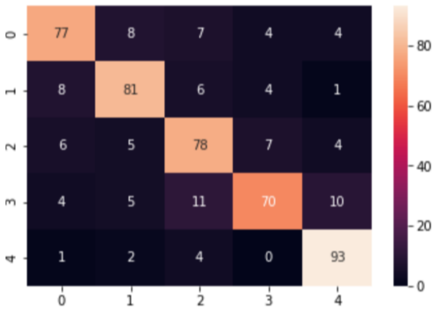
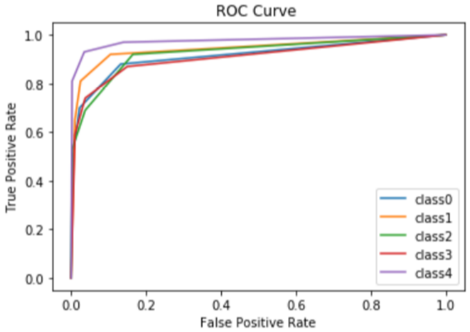
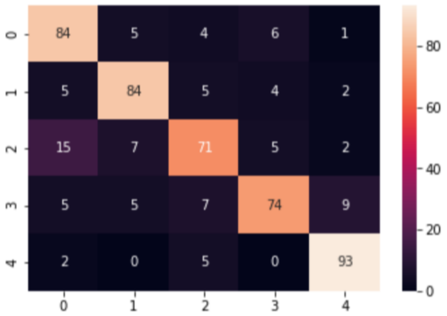
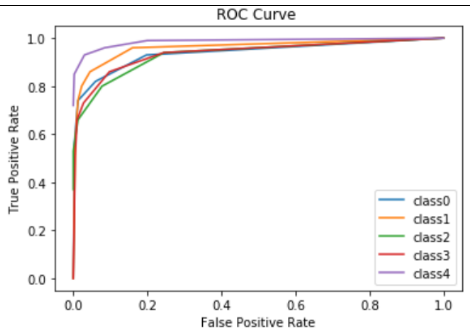
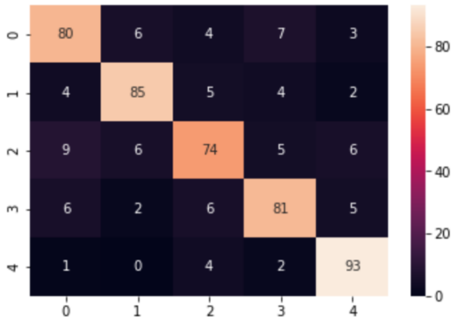
- First step is to load the files from directory and reading the files using csopen() function.
- Then we split the train and test data into x:y ratio where x is training data and y is test data.
- Once the file is read the all pre-processing is done like tokenization, conversion of number to words and stemming etc in train data.
- For each document in train and test data we will create Document Vector and the document vector will store normalized term frequency.
- Now for applying k-nearest neighbour, for each document we will find the cosine similarity and on the basis of value of k we will perform majority voting and assign label to each test documents.
- Then we have predicted labels and actual labels, we will find accuracy , Roc curve and confusion matrix.

  Accuracy= Correctly Labeled Documents/Total Documents

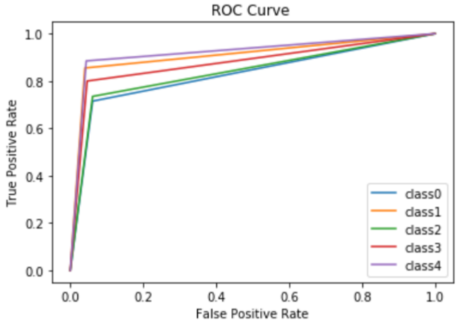
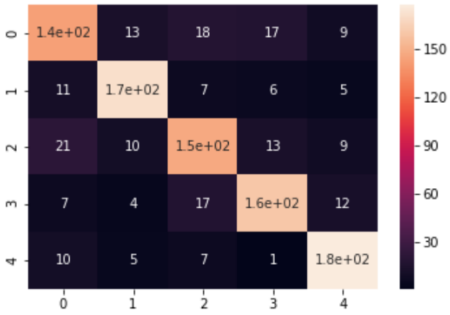
c). Test Cases:

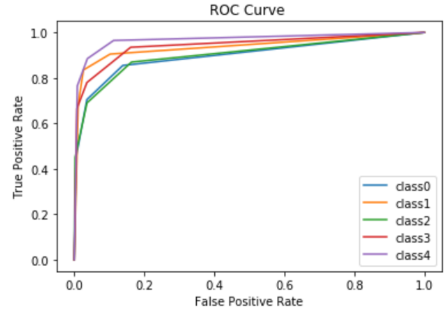
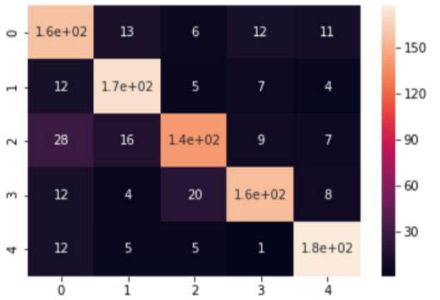
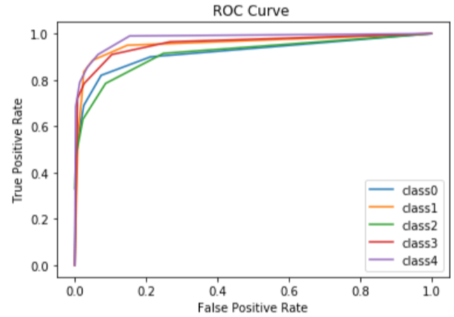
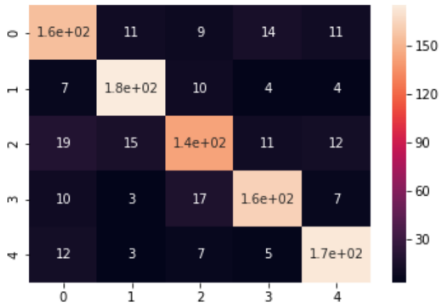
- Training data=90, Test Data=10

Number of Nearest Neighbour	Accuracy	ROC	Confusion Matrix

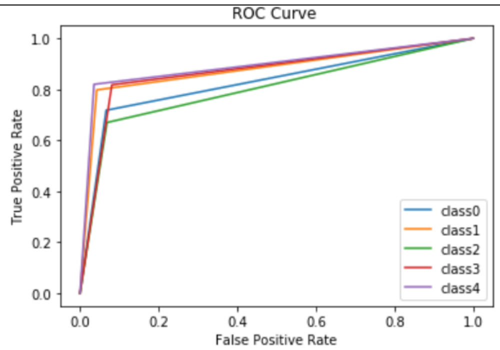
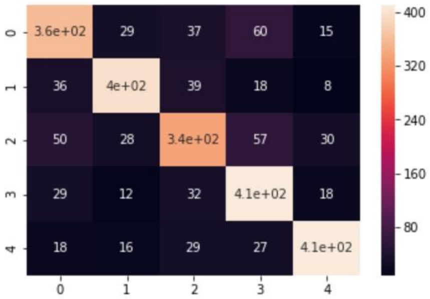
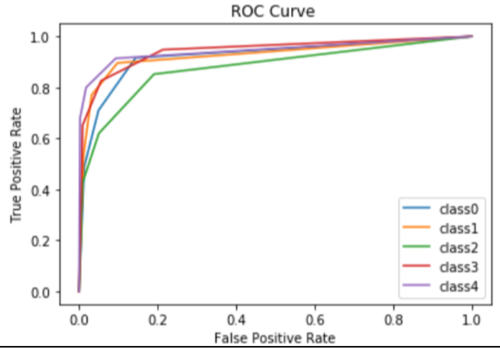
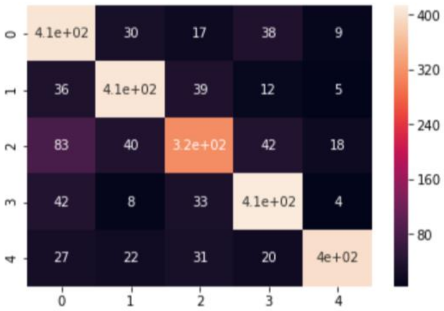
K=1	83.7%		
K=3	85.2		
K=5	86.4		

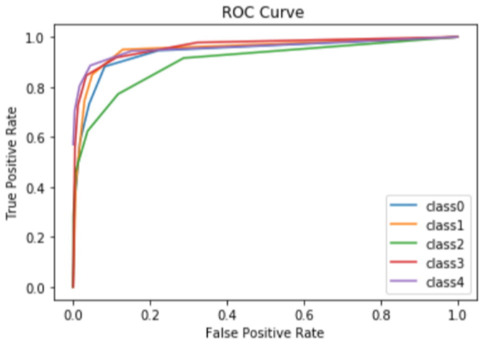
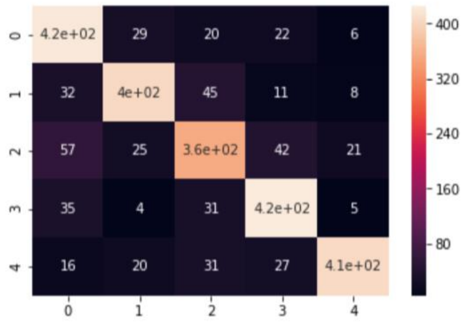
- Training data=80, Test Data=20

Number of Nearest Neighbour	Accuracy	ROC	Confusion Matrix
K=1	81.2		

K=3	82.67		
K=5	81.9		

- Training data=50, Test Data=50

Number of Nearest Neighbour	Accuracy	ROC	Confusion Matrix
K=1	76.48		
K=3	77.76		

K=5	80.52	ROC Curve True Positive Rate False Positive Rate 0.0 0.2 0.4 0.6 0.8 1.0 class0 class1 class2 class3 class4	4.2e+02 29 20 22 6 32 4e+02 45 11 8 57 25 3.6e+02 42 21 35 4 31 4.2e+02 5 16 20 31 27 4.1e+02 0 1 2 3 4 400 320 240 160 80