Assignmen4- Report

Q1.

Assumption:

Ignoring all the unknown words if present in test during the creation of word count feature matrice of test

- Loading and Preprocessing of data is done:
  - Preprocessing of data includes:
    - Converting into lower case
    - Removing URLs
    - Removing punctuations
    - Removing stopwords
- 1. vocabulary of unique words from the training set :
  - length of vocabulary list is 1698

2. Training word count feature matrices created of size: 800 rows × 1698 columns Testing word count feature matrices created of size: 200 rows × 1698 columns Both the feature matrices are different

- Multinomial Naïve Bayes applied with add-1 smoothing
-

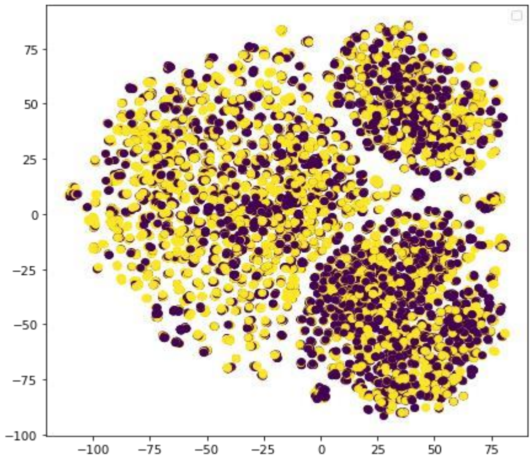
Metric Measure	Training Set	Test Set
Accuracy	0.956	0.74
F1-Score	0.961	0.633

Q2.

Assumption: One hot feature encoding is applied on the data set before splitting

- Data is splitted into 75:25 ratio

  tsne plot to visualise the dataset



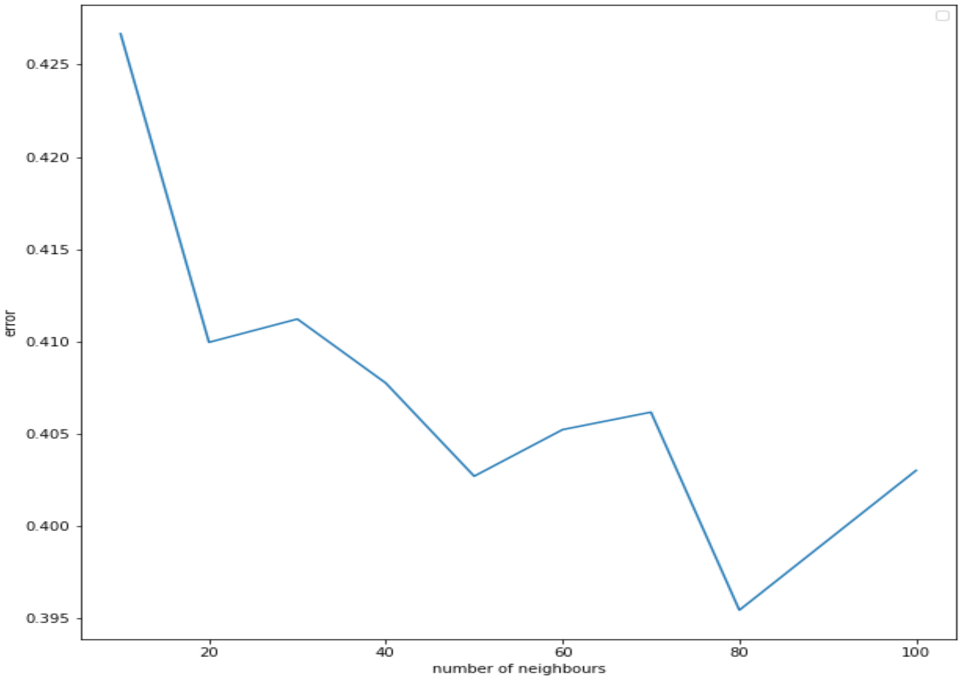
- Knn implemented from scratch

Grid Search on K values and error is as follows:

```
10: 0.4266792809839167,
20: 0.4099653106275623,
30: 0.4112267423525702,
40: 0.4077578051087985,
50: 0.402712078208767,
60: 0.4052349416587827,
70: 0.40618101545253865,
80: 0.3954588457899716,
100: 0.4030274361400189

Optimal K value is 80
```

```
Error Vs Number of neigbours graph:
```



- `Sklearn Train accuracy for optimal K: 0.70 Sklearn Test Accuracy for optimal K: 0.631`

- Comparison b/w my implementation and the inbuilt sklearn function accuracy for optimal k value

Knn Implementation	Training Accuracy	Test Accuracy
Scratch(my implementation)	0.72	0.59
Sklearn Function	0.70	0.63

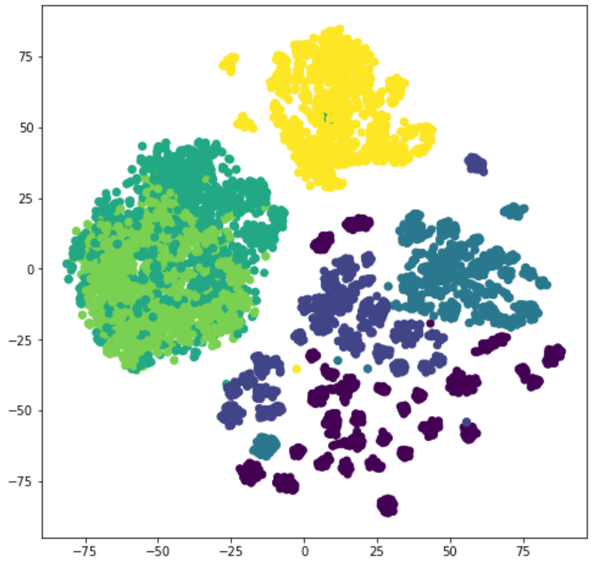
There is a deviation between accuracies. Sklearn function is giving more accuracy for test and train set for optimal k i.e80 value compared to my implementation.

Q3.

Approach:

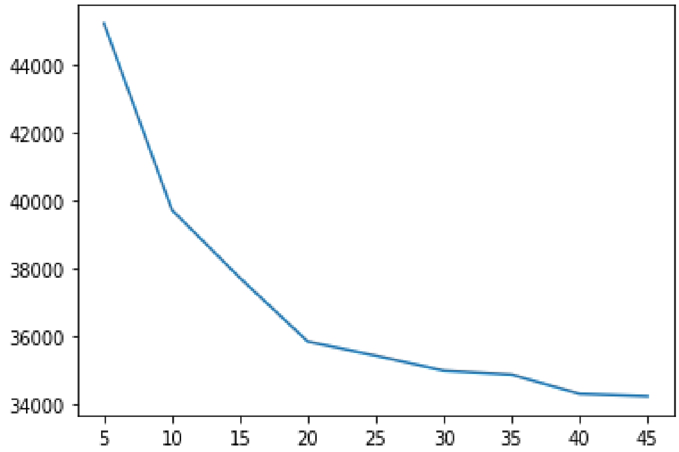
In Kmeans after providing cluster numbers to every cluster formed. Within cluster voting is done whichever data point has maximum occurrence that label is given to the cluster and whenever a new data point is closed to any cluster that cluserts label is given to the data point in this way prediction is done in K means implementation from scratch

- Visualize data through tsne plot different colors are labels of the data

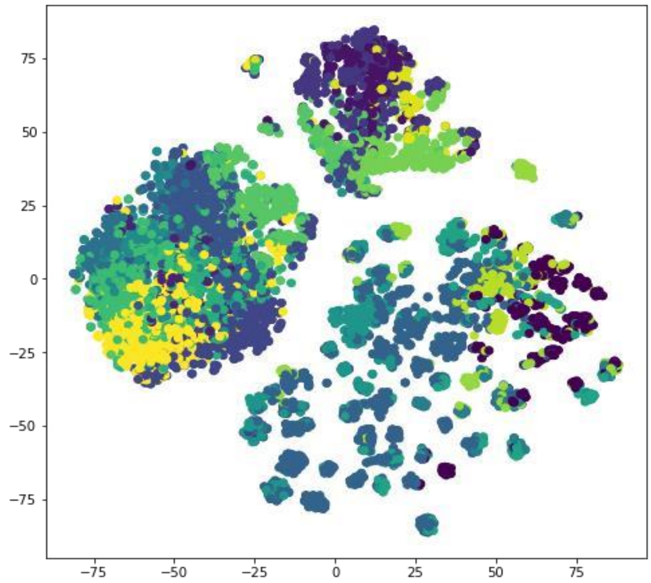


- Optimal Value of k is found using elbow method i.e. 20(clusters)

  Plot of error vs number of clusters graph



- Scatter plot to visualize the dataset to depict the clusters formed(optimal)



- Training Accuracy for optimal K for Kmeans Scratch: 0.645
  Testing Accuracy for optimal K for Kmeans Scratch: 0.70

  Comparison b/w my implementation and the inbuilt sklearn function accuracy for optimal k value for Kmeans

K-Means Implementation	Training Accuracy	Test Accuracy
Scratch Implementation	0.64	0.70
Sklearn Kmeans	0.75	0.77

Sklearn is giving better accuracy for training and test set compared to implementation of K means from scratch