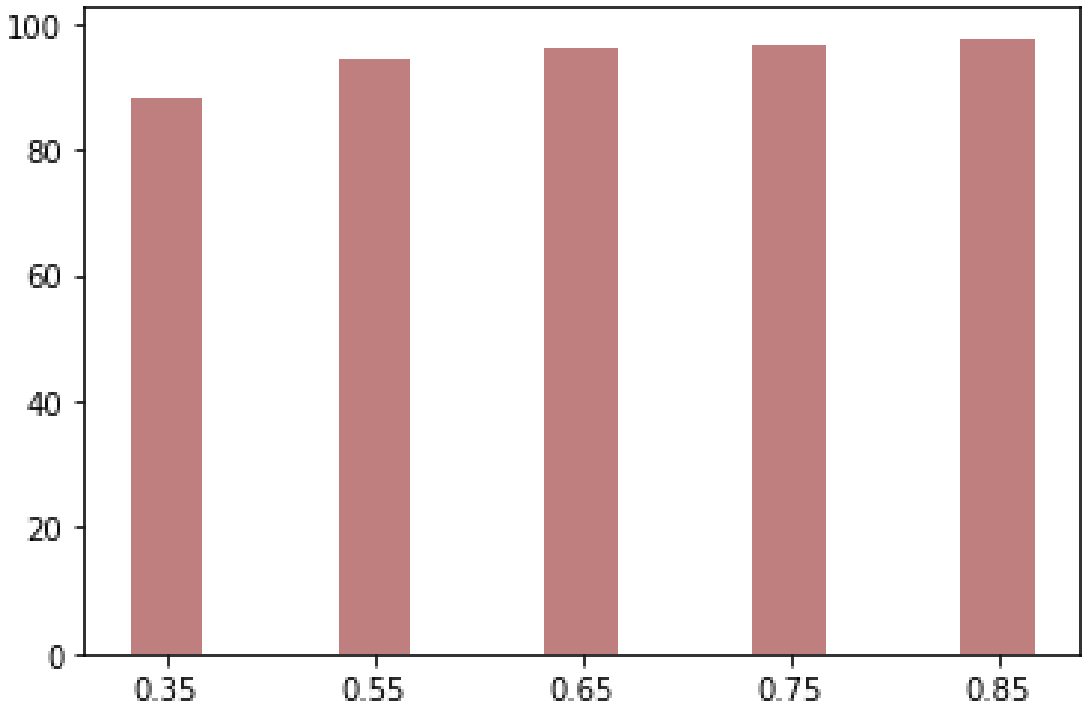
Kshitija Shelar | Text Analytics COMP4760 | Date: 08/11/20

20200155 | Practical 6

1. i.
    - K nearest neighbour is a supervised algorithm which compares the query point with k of its nearest neighbours.
    - The distance is calculated using Euclidean distance or Manhattan distance. Depending upon the majority number of nearest neighbours, classifier classifies the query point under the respective class.
    - Split is one of the parameters given and it is used to split the data set into train and test set. Ideally, the test size should be such that majority of the data set should be train data. It is because the more data we provide for training, the better the model will learn.
    - K value here, is the number of nearest neighbours to be considered. I have chosen values like 5, 10,15, 20,25, 30,35, 40,45,50.
    - Split values used in the code are 0.35, .55, .65, .75, .9.
    - However, in the given code, there is a line of code random.random() which generates random number between 0.0 and 1.0. Because of this, every time the model gives different accuracy. It is because every time different data goes into train and test set.



K = 25

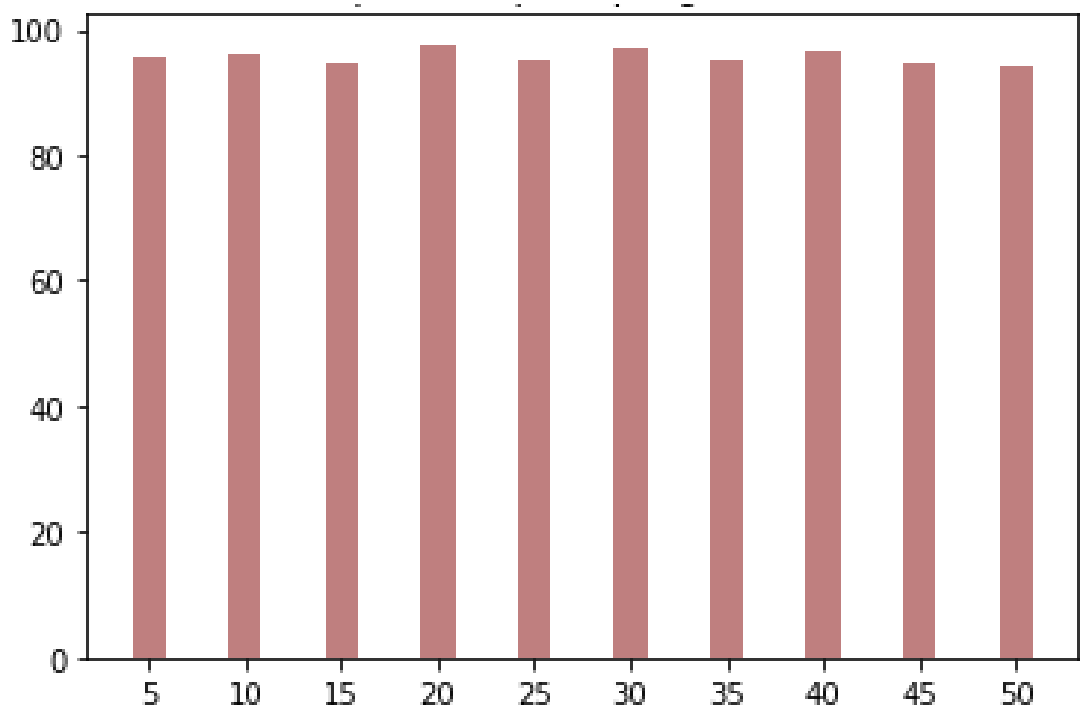
Accuracy: [88.23, 94.03, 96.05, 96.55, 97.69]%

ii. For the following figure:

- On X axis are the split value for k=25. We can see that as the split value increases, the accuracy increases. That is evident because the higher the split value, the algorithm will be better trained on larger data set.
- It gives the highest accuracy for split value of 0.85. That means 85% of the data is used for training and 15% for testing. It is performing the worst for split value of 0.35 where only 35% of data is used for training.

X axis denotes the different value of K for a fixed split size of 0.85

Since .85 split value gave the best accuracy in part 1 of question. I have kept that split value here constant and changing value of K.



Accuracy: [95.45, 96.0, 94.66, 97.6, 95.19, 97.00, 95.20, 96.66, 94.76, 94.19] %

Here we can observe that at k = 20, accuracy is the highest. We can also observe that after k=30, the accuracy keeps decreasing.

2.

- Naïve Bayes Classifier: It is a supervised learning algorithm. It is a classifier which uses probability of features occurring in samples with respect to a category to predict a query object. It assumes attributes have independent distribution. It is fast and space efficient.
- Running the code with feature as last letter (given code) gives the following results:
  Most Informative Features

```
last_letter = 'a'          female : male   =     35.6 : 1.0
last_letter = 'k'            male : female =     31.5 : 1.0
last_letter = 'v'            male : female =     16.5 : 1.0
last_letter = 'f'            male : female =     16.1 : 1.0
last_letter = 'p'            male : female =     12.7 : 1.0
```

  Mark is: male

  Precilla is: female

  Accuracy :0.768

- Here we can see that one of the 5 most informative features is K and it has a ratio of 31.5. That means, the number of names ending in K are 31.5 times more male names than female names. Hence, whenever a new name is tested on this data, any name ending with K will be classified as male even if it's a female name.

- This feature selection is also giving a good result.
  New feature that I have chosen for training is the first 2 letters of the name.

  {'First_2letter': word[0:2]}

- The following is the output for this change in feature.
  Mark is: female

  Precilla is: female

  Accuracy: 0.698

- We can see that Mark is classified as Female. This is happening because off all the names in dataset, 1.8 times the names starting with ‘Ma’ belong to females than males. We can see this from the following piece of information.

  First_2letter = 'Ma' female : male = 1.8 : 1.0

- This is the output of most_informative_feature when I checked for top 100 values.
  We can also see that the accuracy is very low compared to the previous results.

- It is so because of the same reason. There is an imbalance in data for male and female names for the selected features. This can be fixed by providing equal number of data set for both male and female names. This will reduce this bias and give better accuracy and results.