Homework 2

Name: Rohit Garla Net ID: rg3365

Part 2: Programming Exercise

1(a)

For K=1

i) The predicted class is 1

ii) Confusion Matrix when K=1

	1	0
1	TP = 209	FN = 64
0	FP= 134	TN = 93

iii) Accuracy, True positive Rate, False Positive rate for k=1

Accuracy = 60.4

TPR = 0.765

FPR = 0.590

For k=5

iv) Predicted class is 1

v) Confusion Matrix when K=5

	1	0
1	TP = 212	FN = 61
0	FP= 136	TN = 91

vi) Accuracy, True positive Rate, False Positive rate for k=1

Accuracy = 60.6

TPR = 0.776

FPR = 0.599

vii) Accuracy = 60.6

viii) ZeroR Confusion Matrix

	1	0
1	TP = 273	FN = 0
0	FP= 227	TN = 0

1(b)

The following reasons show why such a function cannot be used for a larger dataset

i) Knn has high time complexity and when the data set increases it takes a lot of time to computer predictions

ii) Increasing the dataset increases the frequency of redundant words like, ’the’, ’he’, ‘,’ , ‘all’ , ‘then’ all this effect the accuracy even though they shouldn’t be contributing to the prediction

1(c)

i) Cross validation Accuracies are:

K=3 Accuracy= 66.066
K=7 Accuracy= 65.966
K=99 Accuracy=64.37777

ii) K=3 had the highest accuracy

Accuracy = 59.0

	1	0
1	TP = 212	FN = 61
0	FP= 144	TN = 83

1(d)

i) Distance Function after stemming and removing stop words

This distance function manipulates the string text by removing redundant texts which can be considered for comparison and then calculates the distance using our same formula. For this method, we will get different distance scores for some of the text which refines our accuracy. The following function performs below manipulation before taking distance:

- Removed special characters
- Removed stop words
- Stemming
- Removing short words

ii) On removing the small words, stop words, punctuations we have removed the redundant texts which gives us only relevant tokens for comparing and calculating distance.

iii) K=1 Confusion Matrix

	1	0
1	TP = 211	FN = 62
0	FP= 98	TN = 129

iv) K=1
Accuracy=0.68
TPR = 0.7728937728937729

FPR = 0.43171806167400884

v) K=5

	1	0
1	TP = 211	FN = 62
0	FP= 91	TN = 136

vi) K=5
Accuracy=0.694
TPR = 0.7728937728937729

FPR = 0.4008810572687225

vii) Yeah, the accuracy increased in both the cases because of the reason

- On removing the small words, stop words, punctuations we have removed the redundant texts which gives us only relevant tokens for comparing and calculating distance.