Assignment 3

Answer 1

The 679 words from 'keywords679.txt' were adopted as dictionary and used to train the model using the 3 algorithms. The one-vs-rest and one-vs-one logistic regressions were given a max iteration of 1000, while the SVM was tested with values of 1000,100 and 1.

Finally, the proportion of correctly classified test samples vs the total test data samples was calculated as the 'accuracy'. The accuracy of the algorithm will be used to find the superior data algorithm. The output of the code is given as follows: -

```
--- Experiments with dictionary: keywords769.txt (length=769) ---
Logistic Regression One-vs-Rest accuracy: 0.4959
Logistic Regression One-vs-One accuracy: 0.4745
SVM (linear) C=1000 accuracy: 0.3971
SVM (linear) C=100 accuracy: 0.3966
SVM (linear) C=1 accuracy: 0.4275
```

Figure 1: The accuracy scores for 679 words dictionary

From the output we can see make the accuracy percent table below

Table 1: Algorithms and their respective accuracy for 679 keywords.

Method	Accuracy (%)
Logistic Regression one-vs-rest	49.59
Logistic Regression one-vs-one	47.45
SVM C = 1000	39.71
SVM C = 100	39.66
SVM C = 1	42.75

Logistic regression one-vs-rest had the highest accuracy followed by the one-vs-one logistic regression which implies that these two methods were able to classify the data the best. Meanwhile SVM remained between the low 39-42% range. Overall, all the methods stayed in the below 50% accuracy rates showing moderate predictive ability.

For the SVM method the C value influences the strictness of classification, a higher C value will try harder to classify the data correctly at the risk of overfitting. Here the accuracy rose very slightly for lower C values showing that the SVM struggles to categorize the data regardless of the C strength.

Answer 2

The same procedure from part 1 is followed with a different dictionary 'keywords35.txt' which only has 35 entries. The aim is the see how the smaller dictionary size affects the accuracy of the results. After running the program, we get the following results: -

```
--- Experiments with dictionary: keywords35.txt (length=35) ---
Logistic Regression One-vs-Rest accuracy: 0.1293
Logistic Regression One-vs-One accuracy: 0.1281
SVM (linear) C=1000 accuracy: 0.1200
SVM (linear) C=100 accuracy: 0.1188
SVM (linear) C=1 accuracy: 0.1210
```

Figure 2: The accuracy scores for 35 words dictionary

Making the percentages table we get

Table 2: Algorithms and their respective accuracy for 35 keywords.

Method	Accuracy (%)
Logistic Regression one-vs-rest	12.93
Logistic Regression one-vs-one	12.81
SVM C = 1000	12
SVM C = 100	11.88
SVM C = 1	12.10

The accuracy percentage has dropped significantly for the 35 words dictionary with the percentage of logistic regression and SVM being in the same ballpark range of 12-13%. Given that there were only 20 classes in the dataset this accuracy suggests that the model resorted to almost random guessing rather than any meaningful classification. Based on these results the dictionary size appears to be too small to meaningfully classify the data.

Overall, the larger dictionary still generates only a moderate accuracy while the smaller size dictionary has almost zero accuracy. Logistic regression appears to be better suited to classifying these types of datasets as compared to linear SVM.

Finally, to get higher accuracy rates even larger dictionary files with more data might be required.