1. Q1

2. Q2

```
1. Command to be entered on terminal:

python lr_svm.py trainingSet.csv testSet.csv 1

Output: Training Accuracy LR: 0.65

Testing Accuracy LR: 0.65
```

2. Command to be entered on terminal:

python lr_svm.py trainingSet.csv testSet.csv 2

Output: Training Accuracy SVM: 0.56

Testing Accuracy SVM: 0.55

3. Q3

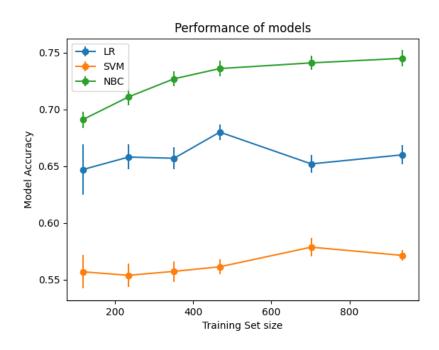
1. Preprocessing dataset for NBC

```
Command to be entered on terminal: python preprocessNBC.py Output: Training and testing files are generated
```

2. Comparing models accuracy

```
Command to be entered on terminal: python cv.py
Output: Plot showing the accuracies for all the three models
```

3. Plot obtained after comparing the accuracies of different models



From the observed plot, we can infer that:

- 1. Accuracy of Naive Bayes classifier improves with increase in the size of the training set.
- 2. Logistic Regression and Linear SVM have similar performance trends but LR performs better than SVM.
- 4. Hypothesis formulated (H_1) :

LR is better than SVM

Null Hypothesis (H_0) : LR is no better than SVM

5. As we are not given any alpha level, we assume it to be 5% i.e 0.05.

The calculated p-value is: 5.914271183792437e-06

As p-value <<<0.05, we can reject the null hypothesis with high confidence and accept alternate hypothesis.

Thus, we conclude that LR is better than SVM