ASSIGNMENT 3 REPORT

Naïve Bayes:

With stop-words being considered:

```
Do you wont to consider stop words - yes/no: yes

Enter Spam training folder path:C:\Users\kxc200005\Desktop\KC\CS 6375 - Machine Learning
\Assignment\Assignment 3\train\spam

Enter Ham training folder path:C:\Users\kxc200005\Desktop\KC\CS 6375 - Machine Learning
\Assignment\Assignment 3\train\ham

Enter Spam testing folder path:C:\Users\kxc200005\Desktop\KC\CS 6375 - Machine Learning
\Assignment\Assignment 3\test\spam

Enter Ham testing folder path:C:\Users\kxc200005\Desktop\KC\CS 6375 - Machine Learning
\Assignment\Assignment 3\test\ham
Accuracy is:1.0
```

Without stop-words being considered:

```
Do you wont to consider stop words - yes/no: no

Enter Spam training folder path:C:\Users\kxc200005\Desktop\KC\CS 6375 - Machine Learning
\Assignment\Assignment 3\train\spam

Enter Ham training folder path:C:\Users\kxc200005\Desktop\KC\CS 6375 - Machine Learning
\Assignment\Assignment 3\train\ham

Enter Spam testing folder path:C:\Users\kxc200005\Desktop\KC\CS 6375 - Machine Learning
\Assignment\Assignment 3\test\spam

Enter Ham testing folder path:C:\Users\kxc200005\Desktop\KC\CS 6375 - Machine Learning
\Assignment\Assignment 3\test\ham
Accuracy is:1.0
```

Observation:

There is no difference to accuracy even if we don't consider stop-words.

Logistic Regression:

Eta =0.1 lambda = 0.1 iterations = 500

With stop words:

```
In [30]: run LR.py train\spam train\ham test\spam test\ham
stopWords.txt yes
The Accuracy of Logistic Regression is:  94.76987447698745
```

Without stop words:

```
In [31]: run LR.py train\spam train\ham test\spam test\ham
stopWords.txt no
The Accuracy of Logistic Regression is:  95.39748953974896
```

Eta =0.1 lambda = 0.01 iterations = 500

With stop words:

```
In [34]: run LR.py train\spam train\ham test\spam test\ham
stopWords.txt yes
The Accuracy of Logistic Regression is:  95.60669456066945
```

Without stop words:

```
In [35]: run LR.py train\spam train\ham test\spam test\ham
stopWords.txt no
The Accuracy of Logistic Regression is:  94.56066945606695
```

λ (Regularization Factor)	η	No of iterations	Accuracy with stop words	Accuracy without stop words
0.1	0.1	500	94.76987447698745	95.39748953974896
0.05	0.1	500	94.14225941422593	94.76987447698745
0.01	0.1	500	95.60669456066945	94.56066945606695
0.1	0.1	1000	94.76987447698745	95.39748953974896
0.05	0.1	1000	93.93305439330544	95.39748953974896
0.01	0.1	1000	94.35146443514645	94.76987447698745

Observations:

After removing the stop words, the accuracy of LR decreases in most of the cases but increases for high values of λ because λ is the penalty on higher values to avoid overfitting.

Before removing the stop words, the accuracy of LR increases, but decreases for high values of λ because some stop words can't be used for classification of a mail as spam or ham but they still interfere with the calculations.