Assignment 3

1) Naïve Bayes

With stop words:

```
Library Loaded.
Start training...
Accuracy of Naive Bayes:
Ham: 0.97989
Spam: 0.84615
Total: 0.94351

Process finished with exit code 0
```

Without stop words:

```
Library Loaded.
Start training...
Accuracy of Naive Bayes (Ignore Stop Words):
Ham: 0.96552
Spam: 0.90769
Total: 0.94979

Process finished with exit code 0
```

In total, the accuracy for ham is quite high, but the accuracy for spam is not that good. This means that the possibility of a ham is considered as a spam than a spam is considered as a ham. This is good because this algorithm would nearly never junk a ham. The total accuracy is over 90% which means the algorithm has potential for practical application.

About stop words, the accuracy for ham decreases a little but the accuracy for spam improves a lot. This means that some significant words might occur in only in spams. Maybe some other stop words should be added to the default stop words list.

2) Logistic Regression

Hard limit: 200 times

With stop words:

```
Library Loaded.
Start training...
Progressing ##############################################
Accuracy of Logistic Regression with lambda = 0.01:
Ham: 0.95690
Spam: 0.73077
Total: 0.89540
Progressing ##############################################
Accuracy of Logistic Regression with lambda = 0.02:
Ham: 0.95690
Spam: 0.73077
Total: 0.89540
Progressing ##############################################
Accuracy of Logistic Regression with lambda = 0.05:
Ham: 0.95690
Spam: 0.73077
Total: 0.89540
Progressing ##############################################
Accuracy of Logistic Regression with lambda = 0.10:
Ham: 0.95977
Spam: 0.73077
Total: 0.89749

Process finished with exit code 0
```

Without stop words:

```
Library Loaded.
Start training...
Progressing ##############################################################
Accuracy of Logistic Regression with lambda = 0.01 (Ignore Stop Words):
Ham: 0.93391
Spam: 0.66154
Total: 0.85983
Progressing ##############################################################
Accuracy of Logistic Regression with lambda = 0.02 (Ignore Stop Words):
Ham: 0.93391
Spam: 0.66154
Total: 0.85983
Progressing ##############################################################
Accuracy of Logistic Regression with lambda = 0.05 (Ignore Stop Words):
Ham: 0.93391
Spam: 0.66154
Total: 0.85983
Progressing ##############################################################
Accuracy of Logistic Regression with lambda = 0.10 (Ignore Stop Words):
Ham: 0.93678
Spam: 0.66154
Total: 0.86192

Process finished with exit code 0
```

In total, the accuracy for ham is still larger than that for spam and total. Hence, the conclusion of Naive Bayes can also be used in Logistic Regression. The algorithm would prefer to let all hams in.

About stop words, the accuracy for ham increases a little but the accuracy for spam stays. This is conflict with the conclusion we got in Naïve Bayes. I would prefer to try until converging to see whether the accuracies would be affected. But for now, we can get that stop words would not affect the algorithm of Logistic Regression.

Besides, the accuracies of Logistic Regression are lower than Naïve Bayes. This might be caused by using hard limit instead of converging.

About lambda, for the given range of lambda [0.01, 0.1], the result nearly never be affected by lambda. This means that we could choose lambda in suitable range without concerning about lambda affects the result. I also try some bigger and smaller lambda only with stop words, and the result has an insignificant change until lambda = 1.0.

Smaller lambda: [0.0001, 0.001, 0.005]

```
Library Loaded.
Start training...
Progressing ##################################################
Accuracy of Logistic Regression with lambda = 0.0001:
Ham: 0.95402
Spam: 0.69231
Total: 0.88285
Progressing ##################################################
Accuracy of Logistic Regression with lambda = 0.0010:
Ham: 0.95402
Spam: 0.69231
Total: 0.88285
Progressing ##################################################
Accuracy of Logistic Regression with lambda = 0.0050:
Ham: 0.95402
Spam: 0.69231
Total: 0.88285
```

Bigger lambda: [0.2, 0.5, 1]

```
Progressing ##################################################
Accuracy of Logistic Regression with lambda = 0.2000:
Ham: 0.95402
Spam: 0.70000
Total: 0.88494
Progressing ##################################################
Accuracy of Logistic Regression with lambda = 0.5000:
Ham: 0.95690
Spam: 0.70000
Total: 0.88703
Progressing ##################################################
Accuracy of Logistic Regression with lambda = 1.0000:
Ham: 0.95402
Spam: 0.73077
Total: 0.89331

Process finished with exit code 0
```

3) Others

a) In train set, 2248.2004-09-23.GP.spam.txt will cause UnicodeDecodeError. Hence, I ignore it when open it.

b) When do exp in Logistic Regression, the sum often above 700 which might cause overflow. In this, we will let the probability be 1 for:

$$\lim_{n\to\infty} \frac{n}{1+n} = 1$$