Conclusions:

1. Number of iterations required drop by a large degree with proper tokenization of the text file. In fact, if I use a regex([A-Za-z'], the vocabulary falls by 1500 words and the number of iterations needed for convergence falls down to less than 20 iterations. Typically:
   1. Case 1:

Split Regex = [a-z0-9':\"\\/,.-]+

Match Regex = [a-z0-9']+

Stop Words Removed: True

eenta = 0.1

size of vocabulary = 9937

noOfIterations needed for convergence = ~10,000 iterations

* 1. Case 2:

Split regex = [ ]

Match regex = [a-z']+

Stop Words Removed: True

eenta = 0.1

sze of vocabulary = 8732

noOfIterations needed for convergence = ~200 iterations

1. As the value of eenta increases, convergence happens a lot faster. The weight vector stop updating sooner for larger values of eenta. Example:

Covergence for learning rate of 0.01 happened in approx 300 iterations.

Iterations:140 Eenta:0.01 Stop Words removed:true

SPAM ACCURACY:83.07692307692308

HAM ACCURACY:89.94252873563218

**Iterations:300 Eenta:0.01 Stop Words removed:true**

**SPAM ACCURACY:86.92307692307692**

**HAM ACCURACY:89.08045977011494**

Iterations:620 Eenta:0.01 Stop Words removed:true

SPAM ACCURACY:86.92307692307692

HAM ACCURACY:89.08045977011494

Convergence for learning rate of 0.02 happened in less than 140 iterations

SPAM ACCURACY:80.0

HAM ACCURACY:89.94252873563218

**Iterations:140 Eenta:0.02 Stop Words removed:true**

**SPAM ACCURACY:83.84615384615384**

**HAM ACCURACY:90.51724137931035**

Iterations:300 Eenta:0.02 Stop Words removed:true

SPAM ACCURACY:83.84615384615384

HAM ACCURACY:90.51724137931035

Convergence for learning rate of 0.08 happened in less than 60 iterations.

Iterations:20 Eenta:0.08 Stop Words removed:true

SPAM ACCURACY:79.23076923076923

HAM ACCURACY:95.97701149425288

**Iterations:60 Eenta:0.08 Stop Words removed:true**

**SPAM ACCURACY:82.3076923076923**

**HAM ACCURACY:95.40229885057471**

Iterations:140 Eenta:0.08 Stop Words removed:true

SPAM ACCURACY:82.3076923076923

HAM ACCURACY:95.40229885057471

We just have to ensure that the exact same value of randomized vector is being used for different values of learning rate. **The number of iterations needed to converge to correct values depends on the combination of initial value of the weight vector and value of learning rate.**

1. When the stop words are removed from the list, the accuracy of the Ham records drops by a small amount. This happens because ham records have more stop words than spam records. However, the accuracy for spam and ham for test set coverges to better values. Example-
   1. With the stop words, removed the following accuracies are obtained for the test set at convergence

Iterations:400 Eenta:0.005 **Stop Words removed:true**

SPAM ACCURACY:86.15384615384616

HAM ACCURACY:88.79310344827586

* 1. With the stop words present, the accuracies converges to more disparate values for spam and ham. Notice that the ham accuracy is higher because of the stop words being more present in ham records than spam.

Iterations:400 Eenta:0.005 **Stop Words removed:false**

SPAM ACCURACY:73.84615384615384

HAM ACCURACY:90.80459770114942

1. Appropriate values of learning rate and number of Iterations and accuracy on test set.
   1. Enron4 test set:
      1. Iterations:8 Eenta:0.04 Stop Words removed:true

SPAM ACCURACY:93.09462915601023

HAM ACCURACY:90.13157894736842

* + 1. Iterations:8 Eenta:0.08 Stop Words removed:true

SPAM ACCURACY:95.9079283887468

HAM ACCURACY:94.73684210526316

* + 1. Iterations:8 Eenta:0.1 Stop Words removed:true

SPAM ACCURACY:96.41943734015345

HAM ACCURACY:94.07894736842105

* + 1. Iterations:7 Eenta:0.2 Stop Words removed:true

SPAM ACCURACY:97.44245524296676

HAM ACCURACY:94.73684210526316

* + 1. Iterations:7 Eenta:0.3 Stop Words removed:true

SPAM ACCURACY:97.44245524296676

HAM ACCURACY:93.42105263157895

* + 1. Iterations:9 Eenta:0.4 Stop Words removed:true

SPAM ACCURACY:97.69820971867007

HAM ACCURACY:96.71052631578948

* + 1. Iterations:8 Eenta:0.4 Stop Words removed:true

SPAM ACCURACY:97.18670076726343

HAM ACCURACY:96.05263157894737

* + 1. Iterations:8 Eenta:0.5 Stop Words removed:true

SPAM ACCURACY:98.20971867007673

HAM ACCURACY:94.73684210526316

* + 1. Iterations:8 Eenta:0.6 Stop Words removed:true

SPAM ACCURACY:98.46547314578005

HAM ACCURACY:96.71052631578948

* + 1. Iterations:8 Eenta:0.8 Stop Words removed:true

SPAM ACCURACY:97.44245524296676

HAM ACCURACY:95.39473684210526

* 1. Enron1 test set:
     1. Iterations:40 Eenta:0.06 Stop Words removed:true

SPAM ACCURACY:91.2751677852349

HAM ACCURACY:89.25081433224756

* + 1. Iterations:30 Eenta:0.07 Stop Words removed:true

SPAM ACCURACY:90.60402684563758

HAM ACCURACY:90.22801302931596

* + 1. Iterations:20 Eenta:0.08 Stop Words removed:true

SPAM ACCURACY:90.60402684563758

HAM ACCURACY:93.15960912052117

* + 1. Iterations:20 Eenta:0.1 Stop Words removed:true

SPAM ACCURACY:87.91946308724832

HAM ACCURACY:93.48534201954398

* + 1. Iterations:19 Eenta:0.11 Stop Words removed:true

SPAM ACCURACY:87.24832214765101

HAM ACCURACY:92.83387622149837

* + 1. Iterations:15 Eenta:0.12 Stop Words removed:true

SPAM ACCURACY:87.24832214765101

HAM ACCURACY:93.15960912052117

* + 1. Iterations:15 Eenta:0.2 Stop Words removed:true

SPAM ACCURACY:83.89261744966443

HAM ACCURACY:95.11400651465799

* + 1. Iterations:15 Eenta:0.26 Stop Words removed:true

SPAM ACCURACY:86.57718120805369

HAM ACCURACY:96.09120521172639

* + 1. Iterations:15 Eenta:0.29000000000000004 Stop Words removed:true

SPAM ACCURACY:86.57718120805369

HAM ACCURACY:96.74267100977198

* + 1. Iterations:15 Eenta:0.32999999999999996 Stop Words removed:true

SPAM ACCURACY:85.90604026845638

HAM ACCURACY:95.43973941368078

1. Enron4 dataset:
   1. NAIVE BAYES:

Accuracy over Spam: 79.79539641943734

Accuracy over Ham: 100.0

* 1. Logistic Regression:

Lemda = 1.3

THE ACCURACY OF LOGISTIC REGRESSION OVER SPAM RECORDS IS: 100.0

THE ACCURACY OF LOGISTIC REGRESSIOn OVER HAM RECORDS IS: 98.49624060150376

* 1. Perceptron:

Iterations:9 Eenta:0.4

SPAM ACCURACY: 97.69820971867007

HAM ACCURACY: 96.71052631578948

1. Enron1 dataset:
   1. Naïve Bayes:

Accuracy over Spam: 94.63087248322148

Accuracy over Ham: 100.0

* 1. Logistic Regression:

lemda = 0.9

SPAM: 97.98657718120805

HAM: 100.0

* 1. Perceptron

Iterations:20 Eenta:0.08

SPAM ACCURACY:90.60402684563758

HAM ACCURACY:93.15960912052117

1. HW2 Dataset
   1. Naïve Bayes:

Accuracy over Spam: 98.46153846153847

Accuracy over Ham: 99.70588235294117

* 1. Logistic Regression:

lemda = 0.9

SPAM: 89.23076923076923

HAM: 100.0

* 1. Perceptron

Iterations:300 Eenta:0.01

SPAM ACCURACY:86.92307692307692

HAM ACCURACY:89.08045977011494

Logistic Regression outperforms Perceptron algorithm for all the three datasets. This is because a sigmoid function will always work as a better classifier as compared to a linear classifier.

Accuracy on test set as it changes with number of iterations on training data:  
X-axis: Number of iterations

Y-axis: Accuracy of Spam and Ham

