**The Naïve Bayes Report:**

For the implementation of Naïve Bayes Classifier, the Below Constraints were used:

1) In the given document only words with digit or words or numbers or ‘are considered. The Regular Expression used is [A-Za-z0-9\']+' .

2) The encoding of the file Is assumed to be of ‘iso-8859-1’

3) The Stop words were used from the list mentioned in the website.

Below are the results of Naïve Bayes Classification:

|  |  |  |
| --- | --- | --- |
|  | Naive Bayes Classifier not removing Stop words | Naive Bayes Classifier removing Stop words |
| Ham Set Accuracy | 97.70114942528735 | 84.61538461538461 |
| Spam Set Accuracy | 97.1264367816092 | 85.38461538461539 |

|  |
| --- |
| Output from program:  C:\Anaconda3\python.exe "C:/Users/Akhilesh/Desktop/Akhilesh Graduate studies/Sem 2/ML/ML assignments/Assignment 2/NB\_LR/NaiveBayes.py" ./hw2\_train/train/ ./hw2\_test/test/  Training path. /hw2\_train/train/  Testing path./hw2\_test/test/  The Naive Bayes Classifier with Stopwords  (hamCount:spamcount=totalcount),filelistlen = (340:8=348),348  Accuracy of the ham 97.70114942528735  (hamCount:spamcount=totalcount),filelistlen = (20:110=130),130  Accuracy of the spam 84.61538461538461  No Stopwords removed from ham 117  No Stopwords removed from spam 113  The Naive Bayes Classifier without Stopwords  (hamCount:spamcount=totalcount),filelistlen = (338:10=348),348  Accuracy of the ham 97.1264367816092  (hamCount:spamcount=totalcount),filelistlen = (19:111=130),130  Accuracy of the spam 85.38461538461539 |

The Accuracy of the Ham is greater than accuracy of spam, the reason is that ham prior is higher than spam prior, because there are more 3 time more ham files than spam files, so there is more chance that file be considered ham over spam. So, most files will be classified as Ham. Only 8 files were not classified as ham, where there were 20 files which were classified as ham, even though they are spam.

With elimination of stop words, the accuracy of the spam is increased , one of the documents in spam is been correctly over common words in ham. However, the Ham accuracy is decreased a little.

**The Logistic Regression:**

Note: It takes lots of time to train around 80 minutes for logistic regression to run, trained with 10,000 words and their weight vector and word vector for both examples.

Learning rate of 0.001, chosen as low as possible to not miss the Gradient ascent point convergence point. The experiment was tried with greater learning rate 0.01, the accuracy was low.

Lambda parameter among the variety of possible values chosen 3 to maintain the weight vector within certain range and not to overfit the data with high values of weight vector.

The number of iterations chosen ranged from 100 to 500 as we increase the range to more than 300, the increase in the weight is very less for each iteration(but takes more time to converge). so for reducing time I have taken 100 iterations.

Here are the results of the Logistic Regression:

|  |  |  |
| --- | --- | --- |
|  | Naive Bayes Classifier not removing Stop words | Naive Bayes Classifier removing Stop words |
| Ham Set Accuracy | 0.9482758620689655 | 0.9454022988505747 |
| Spam Set Accuracy | 0.8076923076923077 | 0. 82307692307692 |

|  |
| --- |
| Output: The Logistic Regression Classifier including Stop words  init Matrix is built  ---Time Taken: 5.652035474777222 seconds ---  ---Time Taken: 5.653039932250977 seconds ---  classification starts  Accuracy of ham file:0.9482758620689655  Accuracy of spam file:0.8076923076923077  ---Time Taken: 2711.1571829319 seconds ---  The Logistic Regression Classifier removing Stop words  init Matrix is built  ---Time Taken: 2719.0055632591248 seconds ---  ---Time Taken: 2719.0065672397614 seconds ---  classification starts  Accuracy of ham file: 0.9454022988505747  Accuracy of spam file: 0. 82307692307692  ---Time Taken: 4851.835983037949 seconds --- |