**Machine Learning, ASSIGNMENT-2**

**For Q1 and Q3, Naive Bayes method can be run as follows:**

python3 NaiveBayes.py <Path of spam training files> <Path of ham training files> <Path of test spam files> <Path of ham test files> <stopWords file path> <yes/no to remove stop-words>

**Below are my results:**

python3 NaiveBayes.py ./hw2\_train/train/spam ./hw2\_train/train/ham ./hw2\_test/test/spam ./hw2\_test/test/ham ./stopWords.txt no

Accuracy of Naive bayes = 0.9476987447698745

python3 NaiveBayes.py ./hw2\_train/train/spam ./hw2\_train/train/ham ./hw2\_test/test/spam ./hw2\_test/test/ham ./stopWords.txt yes

Accuracy of Naive bayes = 0.9456066945606695

So, the accuracy decreases after removing the stop-words.

**For Q2 and Q3, Logistic Regression method can be run as follows:**

python3 LogReg.py <Path of spam training files> <Path of ham training files> <Path of test spam files> <Path of ham test files> <stopWords file path> <yes/no to remove stop-words>

**Below are my results:**

**ƛ**(regularization factor) **= 0.001,**

**η**(learning rate) **= 0.1,**

**Number of iterations** (to update weights) **= 70**

(The values are hardcoded as lambda, eeta, num\_iteration in code)

python3 LogReg.py ./hw2\_train/train/spam ./hw2\_train/train/ham ./hw2\_test/test/spam ./hw2\_test/test/ham ./stopWords.txt no

Accuracy of Logistic Regression is: 0.9518828451882845

python3 LogReg.py ./hw2\_train/train/spam ./hw2\_train/train/ham ./hw2\_test/test/spam ./hw2\_test/test/ham ./stopWords.txt yes

Accuracy of Logistic Regression is: 0.9539748953974896

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **ƛ** | **η** | **No of iterations** | **Accuracy with stop-words** | **Accuracy without stop-words** |
| 0.001 | 0.1 | 100 | 0.9497907949790795 | 0.9539748953974896 |
| 0.005 | 0.1 | 100 | 0.9476987447698745 | 0.9560669456066946 |
| 0.01 | 0.1 | 100 | 0.9476987447698745 | 0.9560669456066946 |
| 0.1 | 0.1 | 100 | 0.9435146443514645 | 0.9560669456066946 |
|  |  |  |  |  |
| 0.001 | 0.1 | 300 | 0.9456066945606695 | 0.9581589958158996 |
| 0.005 | 0.1 | 300 | 0.9476987447698745 | 0.9581589958158996 |
| 0.01 | 0.1 | 300 | 0.9476987447698745 | 0.9560669456066946 |
| 0.1 | 0.1 | 300 | 0.9456066945606695 | 0.9456066945606695 |
| 0.5 | 0.1 | 300 | 0.9372384937238494 | 0.9539748953974896 |
|  |  |  |  |  |
| 0.001 | 0.01 | 100 | 0.9602510460251046 | 0.9518828451882845 |
| 0.005 | 0.01 | 100 | 0.9602510460251046 | 0.9518828451882845 |
| 0.01 | 0.01 | 100 | 0.9602510460251046 | 0.9518828451882845 |

As we can see from the table above, if value of **η** is very small (0.01), then change in value of lambda results in negligible/no change in accuracy.

Otherwise, after removing the stop-words, the accuracy of Logistic regression mostly decreases but then increases for high values of lambda(0.5). This is because lambda is penalty on high values of weight to avoid overfitting.

Before removing the stop words, the accuracy of Logistic Regression increases, but it decreases for high values for lambda(0.5). It can be because of the fact that some stop-words can not be useful for classification of a mail as spam or ham but they still interfere with the calculations.