**Assignment 2**

**CS 6375.001: MACHINE LEARNING**

**Spring 2017**

**Submitted by: Kritika Srivastava (kxs161730)**

**Naïve Bayes Classification**

|  |  |  |  |
| --- | --- | --- | --- |
| **Stopwords Used** | **SPAM Accuracy** | **HAM Accuracy** | **Total Accuracy** |
| No | 85.38 | 97.98 | 94.56 |
| Yes | 86.15 | 97.12 | 94.12 |

As we can see there is no big change after removal of stop words. This is because Naive Bayes algorithm is based on count of tokens and conditional probability corresponding to that. There are not many stop words in the training set so there is not much impact of stop words in our number of features.

**Logistic Regression Classification**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Iterations** | **Learning Rate** | **Lambda** | **Stopwords Used** | **HAM Accuracy** | **SPAM Accuracy** | **Total Accuracy** |
| 1000 | 0.01 | 0 | No | 95.68 | 86.92 | 93.30 |
| 1000 | 0.01 | 0 | Yes | 96.26 | 88.46 | 94.14 |
| 1000 | 0.01 | 0.3 | No | 94.54 | 88.46 | 92.88 |
| 1000 | 0.01 | 0.3 | Yes | 96.55 | 90.76 | 94.97 |
| 1000 | 0.01 | 0.5 | No | 94.54 | 88.46 | 92.88 |
| 1000 | 0.01 | 0.5 | Yes | 96.83 | 90.76 | 95.18 |
| 1000 | 0.01 | 0.8 | No | 94.82 | 89.23 | 93.30 |
| 1000 | 0.01 | 0.8 | Yes | 96.83 | 89.23 | 94.76 |
| 1000 | 0.01 | 2 | No | 95.40 | 89.23 | 93.72 |
| 1000 | 0.01 | 2 | Yes | 97.70 | 85.38 | 94.35 |
| 1000 | 0.01 | 5 | No | 95.40 | 88.46 | 93.51 |
| 1000 | 0.01 | 5 | Yes | 97.40 | 80.0 | 92.88 |

There is an increase in the Ham and Spam Accuracy after removal of Stop words. The reason is reduction in features has removed noise from data. Stop words frequency count was big in numbers. The best practice in Text classification is to always remove stop words from training Data. Removal of Stop words always help in improving accuracy.