**Spam or Ham Detection Tanvi Patil**

1. Naïve Bayes Algorithm

Accuracy for spam detection (Without filtering stop words) : 96.15

Accuracy for ham detection (Without filtering stop words) : 96.83

Accuracy for spam detection (After filtering stop words) : 96.92

Accuracy for ham detection (After filtering stop words) : 96.26

2. Logistic Regression

Below are some statistics of the logistic regression algorithm , which convey the accuracy reports for different values of eta, lambda and hard-limit.

Value of eta = 0.01

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Lambda  Hard limit | | 0.01 | | 0.1 | | 0.5 | |
| With stop words | Without stop words | With stop words | Without stop words | With stop words | Without stop words |
| 50 | Spam | 83.07 | 82.30 | 83.07 | 81.53 | 83.84 | 82.30 |
| Ham | 95.40 | 96.83 | 95.40 | 97.12 | 95.68 | 96.55 |
| 100 | Spam | 83.84 | 83.07 | 83.84 | 82.30 | 83.84 | 83.84 |
| Ham | 95.40 | 96.83 | 95.40 | 96.83 | 95.40 | 97.12 |
| 500 | Spam | 85.38 | 86.15 | 85.38 | 86.92 | 89.23 | 86.92 |
| Ham | 95.40 | 96.26 | 95.68 | 96.26 | 95.40 | 94.82 |

Value of eta = 0.1

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Lambda  Hard limit | | 0.01 | | 0.1 | | 0.5 | |
| With stop words | Without stop words | With stop words | Without stop words | With stop words | Without stop words |
| 50 | Spam | 83.84 | 83.84 | 82.30 | 87.69 | 88.46 | 85.38 |
| Ham | 94.25 | 97.12 | 94.54 | 96.26 | 93.96 | 96.55 |
| 100 | Spam | 83.07 | 85.38 | 83.07 | 86.15 | 79.23 | 82.30 |
| Ham | 94.54 | 97.12 | 94.25 | 96.26 | 95.40 | 96.55 |
| 500 | Spam | 83.07 | 84.61 | 83.07 | 88.46 | 87.69 | 85.38 |
| Ham | 94.54 | 97.12 | 95.40 | 95.97 | 93.67 | 97.41 |

Value of eta = 1

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Lambda  Hard limit | | 0.01 | | 0.1 | | 0.5 | |
| With stop words | Without stop words | With stop words | Without stop words | With stop words | Without stop words |
| 50 | Spam | 83.07 | 88.46 | 87.69 | 100.0 | 0.0 | 0.0 |
| Ham | 94.54 | 95.68 | 94.54 | 68.10 | 100.0 | 100.0 |
| 100 | Spam | 83.07 | 87.69 | 6.15 | 83.84 | 7.69 | 13.84 |
| Ham | 94.54 | 95.68 | 100.0 | 97.12 | 100.0 | 100.0 |
| 500 | Spam | 85.38 | 88.46 | 0.76 | 83.07 | 0.0 | 100.0 |
| Ham | 95.97 | 94.82 | 100.0 | 96.83 | 100.0 | 33.62 |

Does the accuracy improve after removing the stop words? Explain why the accuracy improves or why it does not?

In most of the runs, the accuracy increases after filtering out the stop words. This might be because stop words do not contribute much towards identifying whether a document is spam or not, thus providing more weightage to other words, in deciding whether a mail is spam or not.