Naïve Bayes:

|  |  |  |
| --- | --- | --- |
|  | With StopWords | Without StopWords |
| Vocabulary size: | 11175 | 11135 |
| Hamset | 6401 | 6335 |
| Spamset | 6320 | 6221 |
| Ham Accuracy | 94.8275% | 95.6896% |
| Spam Accuracy | 89.2307% | 90.7692% |

Inference: There is a slight increase in the accuracy on removing the stop words.

Logistic Regression:

With Stop Words:

Iterations: 50

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Lambda | | 0 | 0.001 | 0.005 | 0.01 | 0.05 | 0.1 |
| Learning Rate | |
| 0.001 | Ham Accuracy | 68.10% | 72.12% | 72.70% | 76.14% | 80.17% | 84.48% |
| Spam  Accuracy | 80% | 95.38% | 86.92% | 80% | 90% | 76.15% |
| 0.005 | Ham Accuracy | 72.41% | 75% | 73.27% | 67.24% | 39.08% | 24.42% |
| Spam  Accuracy | 96.92% | 99.23% | 99.23% | 100% | 100% | 100% |
| 0.01 | Ham Accuracy | 71.55% | 73.56% | 65.80% | 66.09% | 32.47% | 18.39% |
| Spam  Accuracy | 100% | 100% | 100% | 95.38% | 90% | 96.15% |
| 0.05 | Ham Accuracy | 67.81% | 64.08% | 38.50% | 22.70% | 0% | 0% |
| Spam  Accuracy | 100% | 98.46% | 100% | 97.69% | 100% | 100% |
| 0.1 | Ham Accuracy | 67.52% | 60.05% | 18.96% | 13.21% | 0% | 0% |
| Spam  Accuracy | 99.23% | 98.46% | 100% | 97.69% | 100% | 100% |

Iterations: 100

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Lambda | | 0 | 0.001 | 0.005 | 0.01 | 0.05 | 0.1 |
| Learning Rate | |
| 0.001 | Ham Accuracy | 73.85% | 78.73% | 76.14% | 77.29% | 71.55% | 57.18% |
| Spam  Accuracy | 96.15% | 90% | 97.69% | 88.46% | 99.23% | 100% |
| 0.005 | Ham Accuracy | 74.13% | 72.70% | 70.97% | 65.80% | 37.93% | 42.81% |
| Spam  Accuracy | 99.23% | 98.46% | 99.23% | 100% | 100% | 86.15% |
| 0.01 | Ham Accuracy | 70.97% | 71.55% | 66.67% | 63.50% | 31.03% | 19.54% |
| Spam  Accuracy | 98.46% | 98.46% | 98.46% | 97.69% | 90.76% | 93.07% |
| 0.05 | Ham Accuracy | 70.40% | 65.51% | 39.94% | 20.97% | 0% | 0% |
| Spam  Accuracy | 100% | 98.46% | 98.46% | 100% | 100% | 100% |
| 0.1 | Ham Accuracy | 69.25% | 60.34% | 22.41% | 6.03% | 0.86% | 0.57% |
| Spam  Accuracy | 100% | 100% | 97.69% | 100% | 96.15% | 98.46% |

Iterations: 150

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Lambda | | 0 | 0.001 | 0.005 | 0.01 | 0.05 | 0.1 |
| Learning Rate | |
| 0.001 | Ham Accuracy | 75% | 75% | 82.18% | 75.57% | 79.02% | 79.59% |
| Spam  Accuracy | 97.69% | 98.46% | 96.92% | 93.84% | 85.38% | 83.07% |
| 0.005 | Ham Accuracy | 71.83% | 73.56% | 68.39% | 69.54% | 40.51% | 26.14% |
| Spam  Accuracy | 100% | 97.69% | 100% | 96.92% | 100% | 100% |
| 0.01 | Ham Accuracy | 73.85% | 73.27% | 64.08% | 64.66% | 20.97% | 25.86% |
| Spam  Accuracy | 99.23% | 100% | 100% | 95.38% | 100% | 80.76% |
| 0.05 | Ham Accuracy | 69.82% | 66.37% | 41.09% | 18.96% | 1.14% | 1.14% |
| Spam  Accuracy | 100% | 98.46% | 97.69% | 100% | 96.15% | 95.38% |
| 0.1 | Ham Accuracy | 70.40% | 60.91% | 19.25% | 9.7% | 1.14% | 0% |
| Spam  Accuracy | 100% | 98.46% | 100% | 100% | 93.84% | 100% |

Iterations: 200

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Lambda | | 0 | 0.001 | 0.005 | 0.01 | 0.05 | 0.1 |
| Learning Rate | |
| 0.001 | Ham Accuracy | 79.59% | 76.14% | 81.03% | 70.68% | 69.82% | 55.45% |
| Spam  Accuracy | 93.84% | 96.92% | 95.38% | 100% | 100% | 100% |
| 0.005 | Ham Accuracy | 71.26% | 71.55% | 72.41% | 66.09% | 39.08% | 45.40% |
| Spam  Accuracy | 100% | 100% | 96.92% | 100% | 100% | 82.30% |
| 0.01 | Ham Accuracy | 73.85% | 72.41% | 67.81% | 59.19% | 33.33% | 21.26% |
| Spam  Accuracy | 99.23% | 98.46% | 97.69% | 100% | 89.23% | 87.69% |
| 0.05 | Ham Accuracy | 70.97% | 65.22% | 38.79% | 19.25% | 7.18% | 0% |
| Spam  Accuracy | 100% | 100% | 100% | 100% | 93.84% | 100% |
| 0.1 | Ham Accuracy | 70.97% | 61.20% | 21.55% | 10.63% | 0% | 0% |
| Spam  Accuracy | 99.23% | 98.46% | 97.69% | 100% | 100% | 100% |

Iterations: 250

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Lambda | | 0 | 0.001 | 0.005 | 0.01 | 0.05 | 0.1 |
| Learning Rate | |
| 0.001 | Ham Accuracy | 75.86% | 75.57% | 74.13% | 78.44% | 67.81% | 81.32% |
| Spam  Accuracy | 96.92% | 97.69% | 99.23% | 94.61% | 100% | 73.07% |
| 0.005 | Ham Accuracy | 72.70% | 72.98% | 70.40% | 63.50% | 42.24% | 25% |
| Spam  Accuracy | 100% | 99.23% | 98.46% | 100% | 100% | 100% |
| 0.01 | Ham Accuracy | 72.70% | 70.97% | 68.67% | 63.50% | 35.05% | 22.70% |
| Spam  Accuracy | 99.23% | 98.46% | 97.69% | 97.69% | 86.92% | 85.38% |
| 0.05 | Ham Accuracy | 70.68% | 65.80% | 41.09% | 19.54% | 7.18% | 1.14% |
| Spam  Accuracy | 100% | 100% | 97.69% | 100% | 93.84% | 96.15% |
| 0.1 | Ham Accuracy | 70.97% | 60.34% | 21.26% | 10.63% | 1.14% | 1.43% |
| Spam  Accuracy | 100% | 100% | 100% | 100% | 95.38% | 93.84% |

Without stop Words:

Iterations: 50

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Lambda | | 0 | 0.001 | 0.005 | 0.01 | 0.05 | 0.1 |
| Learning Rate | |
| 0.001 | Ham Accuracy | 53.44% | 61.49% | 52.87% | 60.91% | 66.95% | 85.05% |
| Spam  Accuracy | 91.53% | 76.15% | 94.61% | 85.38% | 99.23% | 76.92% |
| 0.005 | Ham Accuracy | 74.13% | 65.22% | 69.54% | 65.22 % | 66.09% | 41.95% |
| Spam  Accuracy | 96.92% | 100% | 100% | 100% | 100% | 100% |
| 0.01 | Ham Accuracy | 64.36% | 63.79% | 60.05% | 60.05% | 40.80% | 75% |
| Spam  Accuracy | 100% | 100% | 100% | 100% | 100% | 79.23% |
| 0.05 | Ham Accuracy | 57.47% | 54.02% | 41.09% | 26.43% | 0% | 0% |
| Spam  Accuracy | 100% | 100% | 100% | 100% | 100% | 100% |
| 0.1 | Ham Accuracy | 56.32% | 50.86% | 24.71% | 6.89% | 0% | 0% |
| Spam  Accuracy | 100% | 100% | 100% | 100% | 100% | 100% |

Iterations: 100

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Lambda | | 0 | 0.001 | 0.005 | 0.01 | 0.05 | 0.1 |
| Learning Rate | |
| 0.001 | Ham Accuracy | 77.29% | 63.79% | 69.82% | 72.98% | 79.59% | 64.08% |
| Spam  Accuracy | 85.38% | 95.38% | 98.46% | 95.38% | 100% | 100% |
| 0.005 | Ham Accuracy | 68.67% | 63.79% | 67.52% | 62.06% | 57.18% | 85.34% |
| Spam  Accuracy | 100% | 100% | 100% | 100% | 100% | 73.84% |
| 0.01 | Ham Accuracy | 64.65% | 64.65% | 62.64% | 59.77% | 49.42% | 20.97% |
| Spam  Accuracy | 100% | 100% | 100% | 100% | 100% | 100% |
| 0.05 | Ham Accuracy | 58.62% | 54.88% | 39.08% | 26.43% | 0% | 49.42% |
| Spam  Accuracy | 100% | 100% | 100% | 100% | 100% | 73.07% |
| 0.1 | Ham Accuracy | 57.18% | 50.86% | 25% | 8.04% | 0% | 0% |
| Spam  Accuracy | 100% | 100% | 100% | 100% | 100% | 100% |

Iterations: 150

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Lambda | | 0 | 0.001 | 0.005 | 0.01 | 0.05 | 0.1 |
| Learning Rate | |
| 0.001 | Ham Accuracy | 79.88% | 75% | 70.40% | 79.31% | 69.82% | 89.36% |
| Spam  Accuracy | 88.46% | 96.92% | 100% | 97.69% | 100% | 70.76% |
| 0.005 | Ham Accuracy | 60.34% | 60.63% | 64.36% | 61.49% | 53.44% | 81.03% |
| Spam  Accuracy | 100% | 100% | 100% | 100% | 100% | 93.84% |
| 0.01 | Ham Accuracy | 61.20% | 58.62% | 61.20% | 60.63% | 56.03% | 49.71% |
| Spam  Accuracy | 100% | 100% | 100% | 100% | 100% | 100% |
| 0.05 | Ham Accuracy | 58.33% | 55.74% | 40.22% | 26.43% | 0% | 0% |
| Spam  Accuracy | 100% | 100% | 100% | 100% | 100% | 100% |
| 0.1 | Ham Accuracy | 57.18% | 50.57% | 25.28% | 7.47% | 0% | 0% |
| Spam  Accuracy | 100% | 100% | 100% | 100% | 100% | 100% |

Iterations: 200

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Lambda | | 0 | 0.001 | 0.005 | 0.01 | 0.05 | 0.1 |
| Learning Rate | |
| 0.001 | Ham Accuracy | 74.71% | 79.88% | 66.37% | 71.26% | 74.13% | 61.78% |
| Spam  Accuracy | 98.46% | 93.84% | 100% | 100% | 100% | 100% |
| 0.005 | Ham Accuracy | 63.5% | 66.09% | 63.21% | 64.08% | 75.28% | 52.29% |
| Spam  Accuracy | 100% | 100% | 100% | 100% | 100% | 100% |
| 0.01 | Ham Accuracy | 58.33% | 60.63% | 61.78% | 59.19% | 48.85% | 24.13% |
| Spam  Accuracy | 100% | 100% | 100% | 100% | 100% | 100% |
| 0.05 | Ham Accuracy | 58.90% | 55.17% | 41.37% | 26.43% | 0% | 0% |
| Spam  Accuracy | 100% | 100% | 100% | 100% | 100% | 100% |
| 0.1 | Ham Accuracy | 58.04% | 50.86% | 24.42% | 7.18% | 0% | 0% |
| Spam  Accuracy | 100% | 100% | 100% | 100% | 100% | 100% |

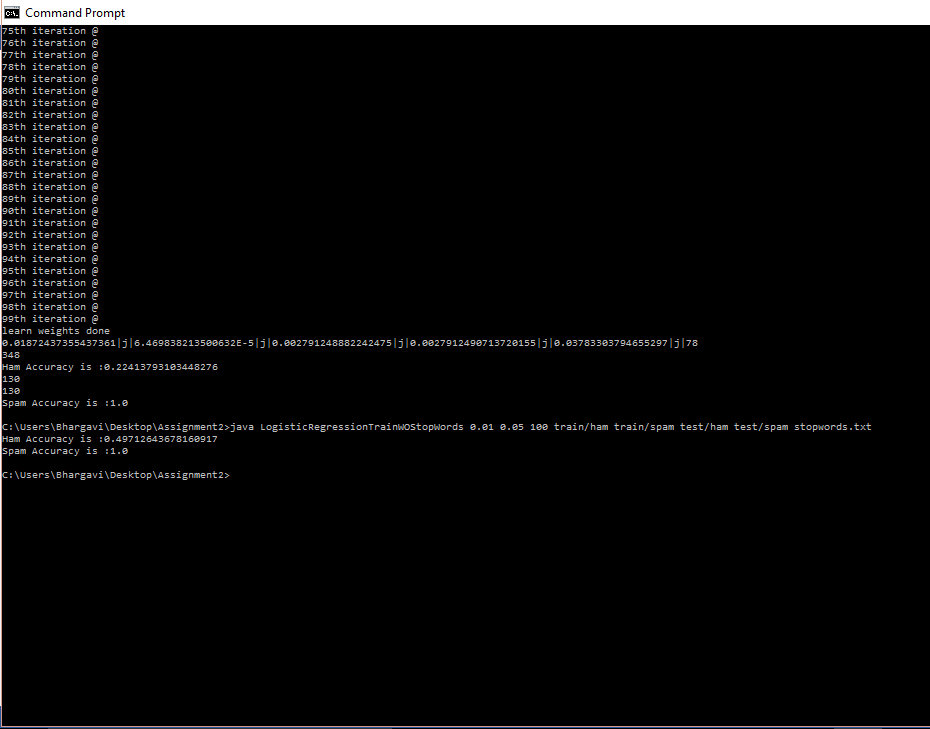
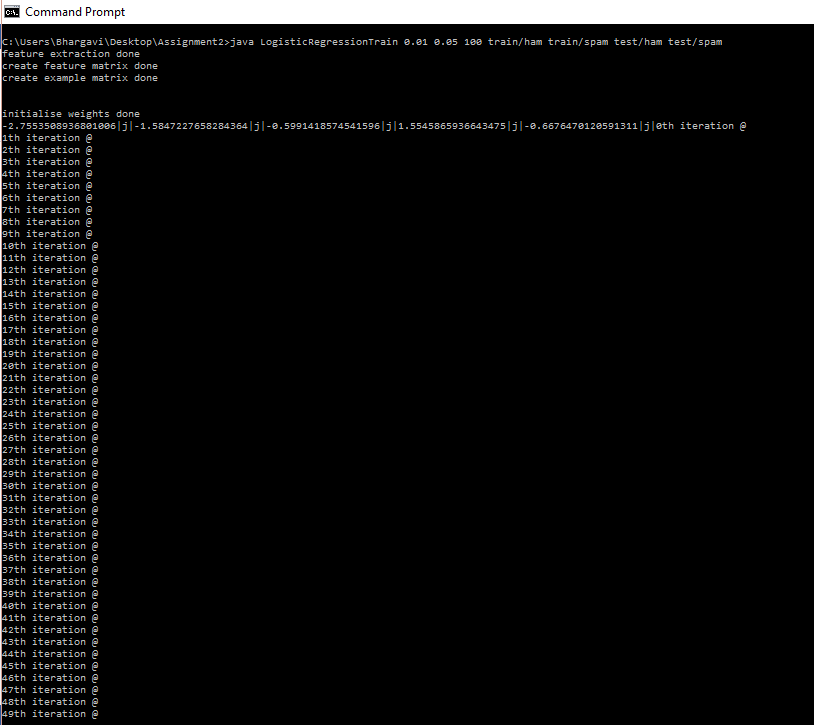
Iterations: 250

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Lambda | | 0 | 0.001 | 0.005 | 0.01 | 0.05 | 0.1 |
| Learning Rate | |
| 0.001 | Ham Accuracy | 74.13% | 76.72% | 70.97% | 72.98% | 71.55% | 89.65% |
| Spam  Accuracy | 100% | 99.23% | 100% | 100% | 100% | 70% |
| 0.005 | Ham Accuracy | 65.22% | 61.78% | 62.64% | 62.93% | 54.31% | 85.91% |
| Spam  Accuracy | 100% | 100% | 100% | 100% | 100% | 71.53% |
| 0.01 | Ham Accuracy | 64.08% | 62.93% | 60.91% | 60.05% | 60.34% | 55.74% |
| Spam  Accuracy | 100% | 100% | 100% | 100% | 100% | 100% |
| 0.05 | Ham Accuracy | 58.04% | 55.17% | 41.09% | 27.58% | 0% | 0% |
| Spam  Accuracy | 100% | 100% | 100% | 100% | 100% | 100% |
| 0.1 | Ham Accuracy | 58.04% | 50.86% | 23.85% | 8.04% | 0% | 0% |
| Spam  Accuracy | 100% | 100% | 100% | 100% | 100% | 100% |

INFERENCE:

When we remove the stop words we are removing the commonly occurring redundant words and thus removing the inconsistences in the data. This will help in improving the accuracies.

Below is shown an example how the accuracies in Logistic Regression are effected:



\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*THANK YOU\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*