Assignment 3: Report

Venkata Bapanapalli

Part 1: Naïve Bayes Algorithm for Text Classification

|  |  |
| --- | --- |
| Classification with stop words: | Classification without stop words: |
| Ham Accuracy: **97.701** | Ham Accuracy: **98.563** |
| Spam Accuracy: **84.615** | Spam Accuracy: **83.076** |
| Total Accuracy: **94.142** | Total Accuracy: **94.351** |

Results:

Performing the Algorithm without stop words increased accuracy on the Ham and Total Accuracy but decreased slightly on the Spam Accuracy. Since the increase in the accuracy is rather small it can be inferred that stop words do not have much impact on classification.

Part 2: Logistic Regression

|  |  |  |
| --- | --- | --- |
|  | With stop words | Without stop words |
| 0.001 | **Ham Accuracy:**  94.82758620689656  **Spam Accuracy:**  88.46153846153845  **Accuracy:**  93.09623430962343 | **Ham Accuracy:** 97.12643678160919  **Spam Accuracy:** 89.23076923076924  **Accuracy:**  94.97907949790795 |
| 0.01 | **Ham Accuracy:**  94.82758620689656  **Spam Accuracy:**  88.46153846153845  **Accuracy:**  93.09623430962343 | **Ham Accuracy:** 97.12643678160919  **Spam Accuracy:** 89.23076923076924  **Accuracy:**  94.97907949790795 |
| 0.1 | **Ham Accuracy:**  94.82758620689656  **Spam Accuracy:**  88.46153846153845  **Accuracy:**  93.09623430962343 | **Ham Accuracy:** 97.12643678160919  **Spam Accuracy:** 89.23076923076924  **Accuracy:**  94.97907949790795 |
| 1 | **Ham Accuracy:**  94.82758620689656  **Spam Accuracy:**  88.46153846153845  **Accuracy:**  93.09623430962343 | **Ham Accuracy:** 97.12643678160919  **Spam Accuracy:** 89.23076923076924  **Accuracy:**  94.97907949790795 |
| 10 | **Ham Accuracy:**  93.96551724137932  **Spam Accuracy:**  89.23076923076924  **Accuracy:**  92.67782426778243 | **Ham Accuracy:** 96.55172413793103  **Spam Accuracy:** 89.23076923076924  **Accuracy:**  94.56066945606695 |
| 100 | **Ham Accuracy:**  83.62068965517241  **Spam Accuracy:**  100.0  **Accuracy:**  88.07531380753139 | **Ham** **Accuracy**: 88.79310344827587  **Spam** **Accuracy**: 93.84615384615384  **Accuracy**:  90.1673640167364 |

Above shows the results of the accuracy on test data with and without stop words and using a lambda value ranging from 0.001 to 100 and 25 iterations. A lambda value greater than 10 has shows to decrease accuracy on the test data. So ideally a lambda value of 10 or less will provide the most accurate results. Iterations of 25 and above have provided comparable results and so 25 seemed ideal since it was least time consuming with best results. We can also see that with stop words the accuracy drops quite considerably. With 100 as the lambda value and with stop words there is overfitting.