# **Machine Learning – Assignment 2 – Report**

**Name:** Satwant Singh

**NetID:** sxs149531

**Accuracy of the Naïve Bayes Classifier on the test data set:**

|  |  |
| --- | --- |
| Vocabulary including stop words | Vocabulary not including stop words |
| 89.7489539749% | 89.539748954% |

**Accuracy of the Logistic Regression Classifier on the test data set:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Eta** | **Lambda** | **Iterations (With Stop Words)** | **Accuracy (With Stop Words)** | **Iterations (Without Stop Words)** | **Accuracy (Without Stop Words)** |
| **0.001** | **0.01** | 5 | 73.0125523013 | 10 | 73.8493723849 |
|  |  | 10 | 73.8493723849 | 12 | 74.0585774059 |
|  |  | 12 | 74.8953974895 | 14 | 74.4769874477 |
|  |  | 13 | 75.3138075314 | 20 | 58.9958158996 |
|  |  | 14 | 58.5774058577 | 21 | 34.30962343096235 |
| **0.001** | **0.0125** | 5 | 73.0125523013 | 10 | 73.8493723849 |
|  |  | 10 | 73.8493723849 | 12 | 74.0585774059 |
|  |  | 12 | 74.8953974895 | 14 | 74.4769874477 |
|  |  | 14 | 58.5774058577 | 16 | 57.949790795 |
| **0.001** | **0.0150** | 5 | 73.0125523013 | 10 | 73.8493723849 |
|  |  | 10 | 73.8493723849 | 13 | 74.0585774059 |
|  |  | 12 | 74.8953974895 | 14 | 74.4769874477 |
|  |  | 13 | 75.3138075314 | 15 | 75.1046025105 |
|  |  | 15 | 77.8242677824 | 17 | 75.7322175732 |
| **0.001** | **0.0175** | 5 | 73.0125523013 | 10 | 73.8493723849 |
|  |  | 10 | 73.8493723849 | 12 | 74.0585774059 |
|  |  | 12 | 74.8953974895 | 13 | 74.0585774059 |
|  |  | 13 | 75.3138075314 | 25 | 75.1046025105 |
| **0.001** | **0.02** | 5 | 73.0125523013 | 10 | 73.8493723849 |
|  |  | 10 | 73.8493723849 | 13 | 74.0585774059 |
|  |  | 12 | 74.8953974895 | 14 | 74.4769874477 |
|  |  | 13 | 75.3138075314 | 15 | 75.3138075314 |

**Graphical comparison of the accuracies with and without stop words:**

**(Orange = Accuracy without stop words, Blue = Accuracy with stop words)**

**Result:**

When the stop words are removed from the vocabulary, the accuracy of the classifier will increase. As seen in the case of Naïve Bayes classifier, the accuracy of the algorithm remains almost the same but in the case of the Logistic Regression classifier, the accuracy increases in most of the cases.

The reason that the accuracy increases is that the stop words are present in both the classes of emails namely ham and spam. But these words are not helpful in any way in the process of classifying an email to either of the classes. Stop words like ‘if’, ‘the’, ‘here’, etc. do not provide us with proof whether the email is ham or spam. Therefore, when we eliminate the stop words like that from our vocabulary, we are left with the words that will help us classify the emails into ham or spam. That is why the accuracy of the classifier increased.