Assignment 2  
CS 6375: Machine Learning

**Report for Naive Bayes and Logistic Regression for Text Classification**

Name: Saraswathi Shanmugamoorthy

Course: CS 6375.003

**Accuracy tables:**

|  |  |  |
| --- | --- | --- |
| Accuracy | Unfiltered data set (before removing stop words) | Filtered Data Set (After removing stop words) |
|  | 94.7699% | 94.3515% |

**Accuracy for Naïve Bayes:**

**Accuracy for Logistic Regression:**

|  |  |  |
| --- | --- | --- |
| **λ values** | **Accuracy on the unfiltered data set (before removing the stop words)** | **Accuracy on the filtered Data Set (After removing the stop words)** |
| λ = 0.1 | 92.8870% | 95.6067% |
| λ = 0.2 | 92.8870% | 95.6067% |
| λ = 1.0 | 93.3054% | 95.8159% |
| λ = 2.0 | 93.3054% | 95.6067% |
| λ = 3.0 | 93.5146% | 95.3975% |
| λ = 4.0 | 93.9331% | 95.1883% |
| λ = 5.0 | 93.5146% | 94.9791% |

The hard limit on the numbers of iterations is **50**.

**Explanation for the above results:**

In the case of Multinomial Naïve Bayes algorithm removing the stop words decreases the accuracy. The accuracy of Logistic Regression is found to increase after removing stop words, the accuracy percentage after filtering stop words is found to lie between 95% and 96%

Stop words are, by definition, words that do not contain information for our classification task. So removing stop words should increase the accuracy. But there is no universal set of stop words that will improve all text classification tasks. Since we are just using an off-the-shelf stop words dictionary, we could be throwing away potentially valuable information. This could be the reason why the accuracy decreases in the case of Naïve Bayes Algorithm.

For example lets consider the word ‘me’, the conditional probability of me is higher than average in ham, which proves that it’s a good word to classify. But since ‘me’ is included in the list of stop words we remove it, this might be one of the reasons for the accuracy to decrease.

Similarly lets consider Logistic regression, if the weights assigned to a particular word is big, then this means that the word is important for the classification task. But if words with higher weights are found in the stop words list and we remove them, then the accuracy will decrease after removing the stop words. But in our case the accuracy increases which proves that words in the stop words list are assigned small weights by our regression model, and hence the stop words do help us in improving the accuracy.

**Some outputs:**

**For Multinomial Naive bayes:**

|  |
| --- |
| python NaiveBayes.py train/spam train/ham test/spam test/ham  \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*    Number of correct predictions without filtering the stop words: 453/478  Accuracy without filtering the stop words: 94.7699%  \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*    Number of correct predictions after filtering the stop words: 451/478  Accuracy after filtering the stop words: 94.3515% |

**For Logistic regression:**

|  |
| --- |
| python LogisticRegression.py train/spam train/ham test/spam test/ham 3.0  \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*    Please wait for unfiltered training set iterations  0  1  2  3  4  5  6  7  8  9  10  11  12  13  14  15  16  17  18  19  20  21  22  23  24  25  26  27  28  29  30  31  32  33  34  35  36  37  38  39  40  41  42  43  44  45  46  47  48  49  \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*    Please wait for filtered training set iterations  0  1  2  3  4  5  6  7  8  9  10  11  12  13  14  15  16  17  18  19  20  21  22  23  24  25  26  27  28  29  30  31  32  33  34  35  36  37  38  39  40  41  42  43  44  45  46  47  48  49  \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*    Number of correct predictions without filtering the stop words: 447/478  Accuracy without filtering the stop words: 93.5146%  \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*    Number of correct predictions after filtering the stop words: 456/478  Accuracy after filtering the stop words: 95.3975% |