ASSIGNMENT - 2: REPORT

This assignment deals with text classification using Naïve Bayes and Logistic Regression.

Part 1:

Logistic Regression

The L2 Regularization algorithm has been applied. The value of learning rate, lambda and number of iterations has been fed as input and the accuracy is reported for test data. The algorithm is run till convergence is obtained through gradient descent. In the same java file I have used a command line argument to either accept or reject stop words. If the stop words are considered and eliminated we find that the accuracy has improved greatly this is mainly because stop words in general help in structuring of sentence but they are of no use to define context of text thus if they are considered they do not give a good accuracy. Also these stop words add negative result and sometimes it may not be more negative enough such that it may affect the classification due to coincidence. Thus I have eliminated the stop words both in the case of Naiive Bayes as well as logistic regression. Some of the stop words I have replaced with null in the algorithm are:

Digits

Special characters like @, #,\* etc.

SGML tags

Apostrophe

Hyphen

A dot separator between words

The accuracy obtained for Logistic Regression on test at set without eliminating stop words for

Alpha=0.5 lambda=0.5 number of iterations= 15 is **76.35983263598327**

After eliminating stop words accuracy is **86.82008368200836**

For couple of learning rate and lambda values the accuracy obtained eliminating stop words is shown in table below:

|  |  |  |  |
| --- | --- | --- | --- |
| Alpha(Learning rate) | Lambda | Iterations | Accuracy (%) |
| 0.5 | 0.75 | 15 | 87.02928870292888 |
| 1 | 1 | 15 | 86.82008368200836 |
| 0.8 | 0.9 | 15 | 86.61087866108787 |

Similarly without eliminating stop words the results are;

|  |  |  |  |
| --- | --- | --- | --- |
| Alpha(Learning rate) | Lambda | Iterations | Accuracy ( %) |
| 0.5 | 0.75 | 15 | 75.73221757322176 |
| 1 | 1 | 15 | 77.61506276150628 |
| 0.8 | 0.9 | 15 | 74.89539748953975 |

Part 2:

NAIIVE BAYES

In the case of Naiive Bayes for text classification Laplace smoothing has been applied to probabilities. The calculation is made in logarithmic scale and to avoid overflow I have restricted it to -100 in the algorithm. Even here the accuracy without eliminating the stop words and after making stop words equal to null has been computed.

The results obtained is as shown below:

Accuracy (in percentage) of Naive Bayes without filtering Stop Words:

**95.68034557235421**

Accuracy (in percentage) of Naive Bayes after filtering Stop Words:

Filtering stop words

Naive Bayes Accuracy is**: 94.38444924406048**

Here we observe there is slight difference in accuracy this is because we are not picking words dynamically. Thus there is a possibility that for restricted or fewer or fixed number of words thus it might not have increased.