**ACCURACY COMPARISON USING VARIOUS MACHINE LEARNING CLASSIFIERS AND FILTERS**

**NO FILTERING OF ATTRIBUTES**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Classifier** | **Attribute Count** | **Stop Words** | **Accuracy** | **Error** |
| Naïve bayes multinomial classification | 64187 | Removed | 80% |  |
| Naïve bayes multinomial classification | 63693 | Not removed | 78.15% |  |
| Logistic regression | 63693 | Removed/not removed | NOT REPORTED | Java heap space error. Even on the cs1.utdallas.edu server. |

**FILTERING ATTRIBUTES USING ATTRIBUTE SELECTION**

GAIN RATIO EVALUATION: Parameter evaluation is done using gain ratio of the attributes with respect to their class.  
RANKER SELECTION: Number of attributes to be selected 10,000 + class attribute  
RUNNING TIME : 20 minutes approximately

NAÏVE BAYES MULTINOMIAL CLASSIFIER: (Running Time: less than 1 minute)  
Default Parameters

|  |  |  |
| --- | --- | --- |
| **Attribute Count** | **Stop Words** | **Accuracy** |
| 10001 | Removed | 79.3% |
| 10001 | Not removed | 79.15% |

LOGISTIC REGRESSION CLASSIFIER: (Running Time: 60 mins approx.)  
Use Conjugate Gradient Descent: True(Faster for large number of parameters)  
Number of iterations(epochs) = 100  
Ridge Parameter = 0.000001

|  |  |  |  |
| --- | --- | --- | --- |
| **Attribute Count** | **Stop Words** | **DATASET** | **Accuracy** |
| 10001 | Removed | Train | 99.79 |
| 10001 | Removed | Test | 95.94 % |
| 10001 | Not removed | Train | 99.77 |
| 10001 | Not removed | Test | 96.47% |

**Explanation:**

Naïve Bayes does not consider correlation between attributes while training the statistical model. Thus, the naïve assumption of all attributes being independent to each other does not work in its favor. Consequently, even with 68,000 attributes the Naïve Bayes classification accuracy is much lower than Logistic Regression with a filtered attribute list of 10,000 attributes. Also, filtering the attribute list to a much smaller size does not affect the classification accuracy significantly. However, the running time is reduced.Logistic Regression is a rather slow learning algorithm compared to Naïve Bayes. With our large vocabulary, it keeps running into Java Heap Space problem during training. Therefore, I decided to filter the attribute list to a smaller number of features. With a smaller feature size, the Logistic Regression converged to acceptable values within 100 training epochs. Since, Logistic Regression takes into account correlation between attributes; its accuracy is much higher than Naïve Bayes even with a smaller number of attributes. Also, filtering the attributes took care of words in the text that do not significantly factor into classification of a document.