COM 527 – Assignment

B)

Using naïve Bayes to classify text documents is a simple approach that assumes that all the conditions are independent meaning that the words used within each of the documents are not considered to be in any specific order or follow a sentence structure, the algorithm uses the probability that each word shows up in data sets for each classification.

An advantage of the naïve Bayes algorithm is that if the conditional independence assumption is correct the classifier can converge very fast requiring much less training data than other models.

The algorithm is incredibly scalable across thousands of words which is typical of the average vocabulary of a human, due the simplicity of the algorithm It scales linearly.

When fitting and training the model, you need to find the best way to represent the text document using feature vectors, one of the models used in language processing is the “bag of words model”. This model creates a collection where any word used in the training sets are stored, no redundant/repeating words are stored and every time a word re-occurs the value associated with the word in key:value pairs.

So, the 2 training data sets of

T1: “Every day hides the terrors of the night”

and

T2: “The night is dark and full of terrors”

will have a vocabulary of

V= {every: 1, day: 1, hides: 1, the: 3, terrors: 2, of: 2, night: 2, is: 1, dark: 1, and: 1, full: 1}

This vocabulary collection would be held for each class, and the new data would use these values to determine how often or likely it is it fit into that category.

C)

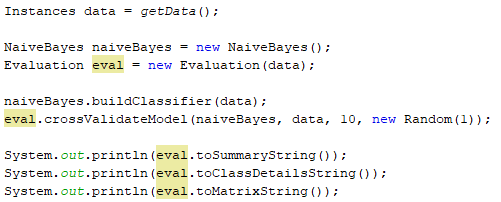
The F-Measure, also known as the F-score, is the weighted harmonic mean of Recall and Precision, also R & P. The measure allows for differential weighting of both the Recall or Precision values but commonly they will be given equal weighting. The Recall value used in F-measure is derived from the amount of responses back on a query, while the precision value is derived from the relevance of those responses.

This means that both the values must trade off against each other as the relevance of the returned data will always decrease when the number of responses is increased.

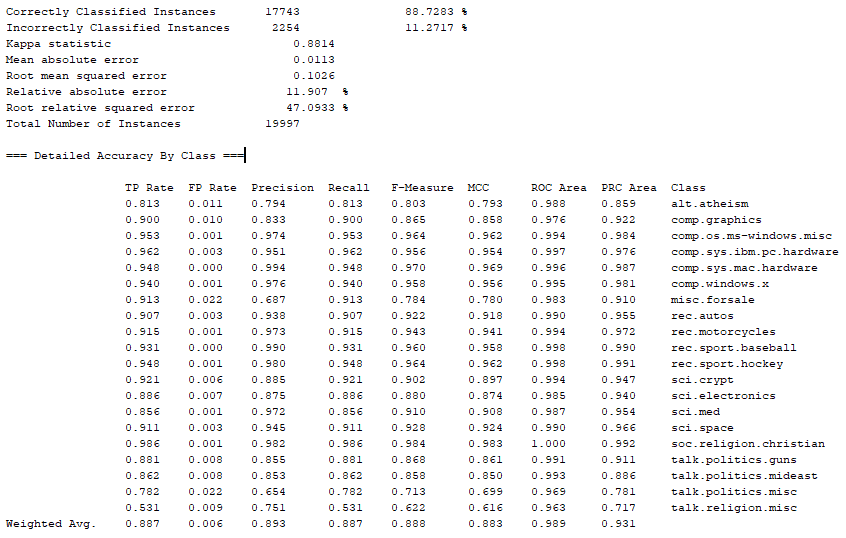
Both these values have radically different importance levels depending on the scenario, while a standard search engine user would want high precision, a para-legal for example would commonly want high recall on a case query.

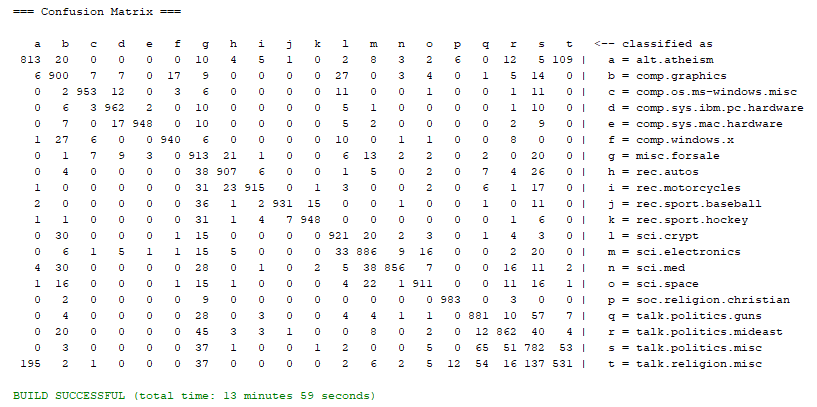
D)

10-fold cross evaluation.



The code above shows how this cross-validation was completed, the classifier was built and then the Evaluation method crossValidateModel() was called with the parameter to create 10 subsets.





E)

Two Partitions at 60%-40%