***Machine Learning Assignment 1***

Topic: Naïve Bayes Text Classifier

|  |  |
| --- | --- |
| Name: | Abu Ayub Ansari Syed |
| UTA ID: | 1001297330 |
| Course ID: | CSE 6363 |
| Email: | abuayubansari.syed@mavs.uta.edu |

**Implementation**:

Splitting the data:

I have manually split the data into training data and testing data. I have kept the first 500 files in the respective classes as training data in the folder named “Train”. The remaining data mostly 500 files in each class is kept in their respective folder classes as the testing data in “Test”. Both there is placed in the “Data” folder

Training the classifier:

In this step I have preprocessed each of the class. In preprocessing, I extract the content of the file, split them by regular expression into words. Make a list of all the words in the file which are not stop words. There is a list of stop words that is available. Next I have stemmed each word to the root words using potter stemmer. For each word list of a file, I have stored them in class data and unique words of all the training data. Have a count of the respective words.

Testing the classifier:

In this case, I use each file individually to test as to which class it belongs. The test file is preprocessed. In preprocessing, read the contents, remove stop words, stem the remaining words to root words, and add them to class data if they are not already present in class data. Send the All class data, present class data with new words, unique words of all data and count to find probability of the test file with respect to the all the classes.

Classifier:

With the data of all classes, I calculate the probability of the current files’ list of words by using Laplace correction for the Naïve Bayes classifier. It returns the max probability of the class it can belong.

Accuracy:

I check if the predicted class is same as the current checking class file, then increment the counter. Calculate accuracy by dividing counter by all the files checked by the classifier.

**Execution:**

* Program code is in **NaiveBayesClassifier.py**
* Keep the data in the same folder as the python file.
* The data locations are hard coded.
* Coded in python 3.4.3, using ntlk, math, regular expression (re) imported libraries.
* Straight forward execution of the python file is sufficient.

**Results**:

The accuracy achieved is 85.08552565769732%.

The files in the same folder have few experimental results. The files are as follows:

Classes Data: Contains the word and their count of all the classes.

Predicted Classes: Contains the predicted classes of the test files.

Result.html: Contains the Anaconda API result console of the program execution.

Program Data: Contains variables data of the program.