**IR ASSIGNMENT 3**

**Haad Fida**

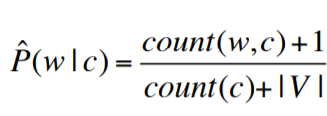
**L16-4032**

**Step 1.** The dataset is considered for analysis which consists of labelled ham and spam folders, the dataset is further divided into training and test data.

**Step 2.** The dataset already has words in lower case, so no need to preprocess that. The only preprocessing is done is to check if a letter is an alphabet or not, this is checked on a letter basis and not on a word basis so e.g. “Subject:” is a word, it will be changed to “Subject” instead of being removed entirely.

**Step 3.** After Step2, it is time to extract features. We would start by finding out the terms in the training folder. Then we will find out the term frequency in each class. We also need to find the total number of tokens in each class. We need the prior probability of a class which is found out by dividing the number of documents in a class by the total number of documents in the training folder.

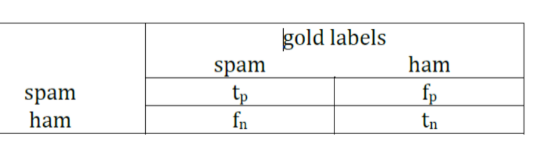
**Step 4.** From the features extracted, we train a Naïve Bayes model. The equation used is as follows:



Where count(w,c) represents the the term frequency in a class. Count(c) represents the total number of tokens in a class and |V| represents the total number of tokens in the training data.

**Note:** Decimal is used as float rounds off very small values to 0.0.

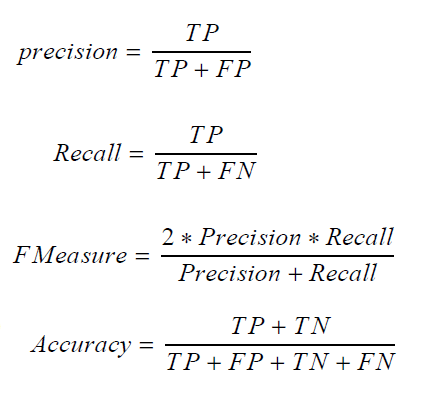
**Step 5.** After training the model, we find out the confusion matrix in this format using the testing data.



Spam in spam represents true positive, spam in ham represents false positive, ham in spam represents false negative and ham in ham represents true negative.

|  |  |  |
| --- | --- | --- |
| **Confusion Matrix** | **spam** | **ham** |
| **spam** | 1455 | 256 |
| **ham** | 45 | 3416 |

**Step 6.** The equations used to find accuracy, precision, recall and F1score are as follows:



**Note:** F Measure/F1 Score is also equivalent to 2TP / (2TP + FP + FN)

Precision: 0.8503798947983635

Recall: 0.97

F1 Score: 0.9062597321706634

Accuracy: 0.9418020108275329