

National University of Computer and Emerging Sciences



IR Assignment 4

Minahil Ahmad 16L-4150

First there is a function in which I use the os library, we open the file directory and return all the filenames in that directory.

Then there’s another function getwordlist which returns all the words from a particular file, preprocessing is done here which removes all the non-alphabet characters and turn it into lower case.

For the training of the Naïve Bayes, we open the directory and first look into the ham folder. We find out the term frequency in the entire ham folder. Similarly we find the term frequency in the entire spam folder. Then we find out the terms in the entire training folder. Probability of each term is calculated using the formula (term frequency in class +1)/(total tokens in the data set + total terms in the data set), this is done for both classes and for all terms. The probability of each term against both classes is returned from this function.

For the testing of Naïve Bayes, the testing data is used. We open up the directory of ham first. For each document, we find out in which class the document belongs to. For this the formula is prior probability multiplied with the probability of all the terms in a class. If a term doesn’t exist in the model, smoothing is applied so it is given some probability. If the cumulative probability of ham is greater, we have classified it as ham and if the cumulative probability of spam is greater, we have classified it as spam. This process is also done for the spam folder in the same manner.

Spam in spam is true positive, ham in spam is false negative, ham in ham is true negative and spam in ham is false positive. Now using the equations

Precision = TP / (TP + FP)

Accuracy = (TP + TN) / (P + N)

Recall = TP / (TP + FN)

F1 = 2TP / (2TP + FP + FN)

We find out the precision, accuracy, recall and f1score of the trained model.