**Information Retrieval Systems**

**Lab Practical and date** – Practical 7, Wednesday 6th October 2020

**Name and Roll Number**- Het Shah, 17BIT103

**Practical Objective**- Email Classification

Given a set of labeled email documents, classify them as spam or non-spam using Naive Bayesian classifier.

**Steps Involved**

* We use the enron1 data set which consists of 5975 emails out of which 1500 are spam
* .We split the dataset in the ratio of 20 percent for testing and the rest 80 percent for the training
* We lemmatize the emails and use the inbuild NaiveBayes classifier for nltk to train the model
* We test the testing data on the model and get the accuracy

**Python Package Used**

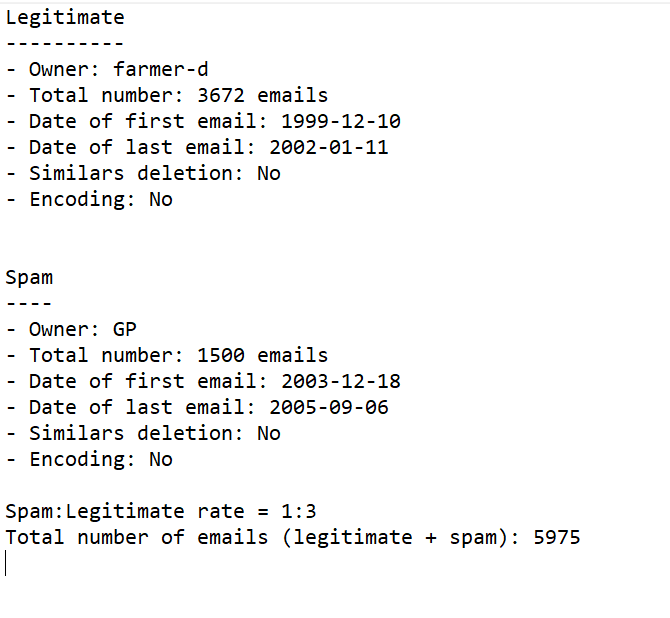
* We used inbuild python data structures such as Dictionary and arrays and to take the input we used the I/O operations from the files
* We used NTLK library to perform the pre-processing steps and to create the naïve bayes model

**Sample Input/Output**

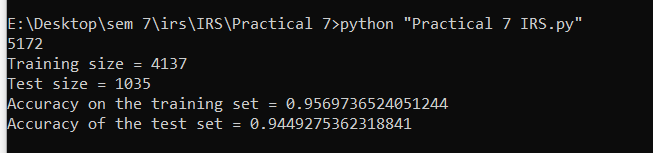
The input is in the form of corpus consisting of 5675 text files located in 2 different directory either spam or ham

The output is presented in the form of the accuracy of the training and the testing data set.

As we can we see we get accuracy of around 95 percent.



Data set description



Output of the program

**Conclusion**

In this practical, did the text preprocessing of an email dataset and perform naïve bayes classification