Problem-1

To solve this problem I have created two data sets as per the requirement. One is Training dataset and other is Testing dataset. In both Training and Testing dataset, I added 10 spam and 10 non spam emails from spam folder and inbox respectively.

Problem-2

To solve this Spam classification using Naive Bayes classifier I have gone through class note and few other articles to understand how it will classify email whether spam or not. This classifier aggregates information using conditional probability. I have modified docclass.py code and test.py to get the result.

In docclass.py I added one more function checkSpamOrNot to train my dataset. In this function I am looping all the 10 spam and 10 non spam emails from train dataset. Read each text file and pass that text to train function for training.

To find spam or not spam email I have used test.py and modified as per the requirement. In order to calculate confusion matrix I used one variable “output” which will store all the True positives, True negatives, False positives and False negatives. Looping all the 10 spam and 10 non spam emails from test dataset to classify each email.

True positives: “not spam email” classifies as “not spam email”

True negatives: “spam email” classifies as “spam email”

False positives: “spam email” classifies as “not spam email”

False negatives: “not spam email” classifies as “spam email”

The test dataset results as given below: