Mahir Efe KAYA
2016400195

Assignment 4

a)

Firstly I have used two seperate vocabularies to make it easier to debug. One is vocabulary for spam mails and the other one is for legitimate mails. When I use all the words as features, they sum up to 15190 distinct words.

b)

['!', '#', '$', '%', '/', '0', '10', '100', '20', '24', '95', 'address', 'advertise', 'anywhere', 'best', 'bulk', 'business', 'buy', 'call', 'card', 'cash', 'check', 'click', 'com', 'company', 'cost', 'credit', 'customer', 'day', 'department', 'dollar', 'earn', 'easy', 'english', 'every', 'free', 'grammar', 'guarantee', 'home', 'hour', 'http', 'hundred', 'income', 'instruction', 'internet', 'language', 'linguist', 'linguistic', 'linguistics', 'list', 'mail', 'mailing', 'market', 'million', 'modern', 'money', 'month', 'name', 'need', 'off', 'offer', 'online', 'order', 'our', 'over', 'pay', 'per', 'personal', 'phone', 'price', 'product', 'profit', 'purchase', 'query', 'receive', 'reference', 'remove', 'return', 'save', 'science', 'sell', 'service', 'site', 'speaker', 'step', 'success', 'syntax', 'theory', 'thousand', 'today', 'university', 'us', 'visit', 'want', 'web', 'win', 'www', 'yours', 'yourself', 'zip']

However, if we want to increase the accuracy, we may change numerical characters to one specific one, so that we wouldn't miss the effect of the numerical words. For instance, what I am saying does not depend on any statistical evidence, spams may have more numerical tokens than legitimate ones, however because of the changing values of these tokens, their effect may be lower than it shall rather be. If we change every number to 0, we might spot the changing prices they show on spams.

c)

System with the whole vocabulary has:

Precision : 0.971357063403782

Recall : 0.9708333333333334

F score : 0.9708252292303418

Hit Rates:

Legitimate hits: 229 / 240

Spam hits: 237 / 240

System with the mutual info of 100 tokens has:

Precision : 0.9739122114163434

Recall : 0.9729166666666667

F score : 0.9729024357409926

Hit Rates:

Legitimate hits: 228 / 240

Spam hits: 239 / 240

Mahir Efe KAYA
2016400195

Macro Averaged Results:

Precision : 0.9726346374100627

Recall : 0.971875

F score : 0.9718638324856672

d)

pValue = 0.99000999000999. Since it is bigger than 0.05, it is approved.

e)


```
Windows PowerShell                                                    —    □    ×

PS C:\Users\mhrfk\dev\493\HW4-SpamEmailFiltering> python filter2.py
For the system  0
Legitimate hits:  229 / 240
Spam hits:  237 / 240


For the system  1
Legitimate hits:  228 / 240
Spam hits:  239 / 240


System with the whole vocabulary has:
 Precision : 0.971357063403782
 Recall : 0.9708333333333334
 F score : 0.9708252292303418


System with the mutual info of 100 tokens has:
 Precision : 0.9739122114163434
 Recall : 0.9729166666666667
 F score : 0.9729024357409926


Macro Averaged Results:
 Precision : 0.9726346374100627
 Recall : 0.971875
 F score : 0.9718638324856672


Testing with R value as 1000
Approved with the P value as 0.98001998001998
PS C:\Users\mhrfk\dev\493\HW4-SpamEmailFiltering>
```