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 : 0.9870689655172413

Recall : 0.9541666666666667

F score : 0.9703389830508474

Hit Rates:

Legitimate hits: 229 / 240

Spam hits: 237 / 240

System with the mutual info of 100 tokens has:

Precision : 0.9956331877729258

Recall : 0.95

F score : 0.9722814498933903

Hit Rates:

Legitimate hits: 228 / 240

Spam hits: 239 / 240

Mahir Efe KAYA
2016400195

Macro Averaged Results:

Precision : 0.9913510766450835

Recall : 0.9520833333333334

F score : 0.9713102164721188

d)

pValue = 0.98001998001998. Since it is bigger than 0.05, it is approved.

e)