Niyazi Ülke – 2017400114
Feb 2021

CMPE 493 Assignment 4 Report

Spam Filtering Using Multinomial Naïve Bayes

Programming Language: Python 3.9
Tested on: Windows 10

The assignment statement does not mention punctuation removal, so it is not done in my system. In addition, I think use of punctuation in an email might be a clue for spam detection. For example, most spam emails I have received contain “!”. On the other hand, legitimate emails usually do not contain exclamation marks.

1- Size of vocabulary is 14670 when all words are used as features.

2- Since there are two classes, their Mutual Information tables are almost the same. So, for two classes only a single set of 100 terms are selected.
Some of these terms might not exist in a class (spam/legitimate) but Laplace smoothing handles this case well.

Selected k = 100 terms are:
['language', '!', '$', 'free', 'remove', 'linguistic', 'http', 'com', 'money', 'linguist', 'check', 'mail', 'linguistics', 'university', 'market', 'best', 'site', 'cost', 'click', 'business', 'our', '0', '100', 'product', 'internet', 'service', 'company', 'day', '#', 'english', 'today', 'advertise', 'million', 'www', 'cash', 'sell', 'hour', 'win', 'dollar', 'pay', 'home', 'bulk', '%', 'web', 'call', 'card', 'query', 'save', 'income', 'credit', 'mailing', 'success', 'offer', 'guarantee', 'thousand', 'purchase', 'hundred', 'yours', 'earn', 'over', 'us', 'department', 'customer', 'instruction', 'easy', 'edu', 'yourself', 'speaker', 'profit', 'reference', 'anywhere', 'online', 'grammar', 'name', 'want', '/', 'visit', 'address', '24', 'order', 'every', 'receive', 'theory', 'zip', 'phone', 'need', 'buy', 'personal', 'price', '20', 'syntax', '10', 'here', 'return', '95', 'off', 'step', 'science', '1998', 'list']

3-

Model 1 (No Feature Selection)
Score Table

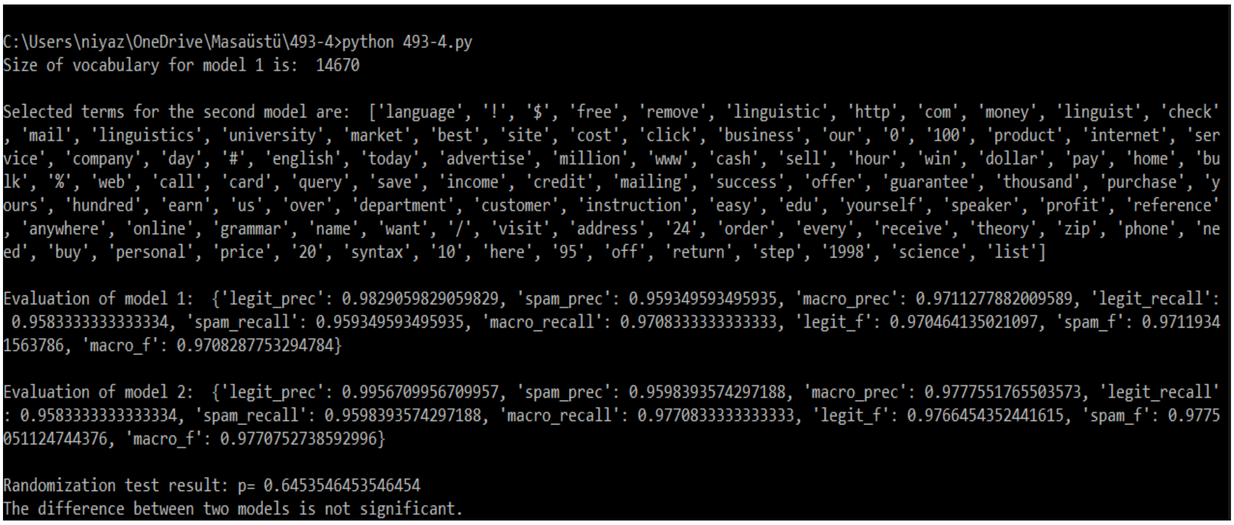
	Precision	Recall	F-Score
Legitimate	0.9829059829059829	0.9583333333333334	0.970464135021097
Spam	0.959349593495935	0.959349593495935	0.97119341563786
Macro-averaged	0.9711277882009589	0.9708333333333333	0.9708287753294784

Model 2 (100 Features Selected with MI)
Score Table

	Precision	Recall	F-Score
Legitimate	0.9956709956709957	0.9583333333333334	0.9766454352441615
Spam	0.9598393574297188	0.9598393574297188	0.9775051124744376
Macro-averaged	0.9777551765503573	0.9770833333333333	0.9770752738592996

Both models perform very well. Model 2 seems better, but it should be checked if the difference is significant or not.

4- The difference between macro-averaged F-scores of models is 0.006. Randomization test parameters: R = 1000, rejection level: 0.05 . The test gives values around p = 0.64 .
This means the models are not significantly different in terms of prediction quality.



```
C:\Users\niyaz\OneDrive\Masaüstü\493-4>python 493-4.py
Size of vocabulary for model 1 is:  14670

Selected terms for the second model are:  ['language', '!', '$', 'free', 'remove', 'linguistic', 'http', 'com', 'money', 'linguist', 'check'
, 'mail', 'linguistics', 'university', 'market', 'best', 'site', 'cost', 'click', 'business', 'our', '0', '100', 'product', 'internet', 'ser
vice', 'company', 'day', '#', 'english', 'today', 'advertise', 'million', 'www', 'cash', 'sell', 'hour', 'win', 'dollar', 'pay', 'home', 'bu
lk', '%', 'web', 'call', 'card', 'query', 'save', 'income', 'credit', 'mailing', 'success', 'offer', 'guarantee', 'thousand', 'purchase', 'y
ours', 'hundred', 'earn', 'us', 'over', 'department', 'customer', 'instruction', 'easy', 'edu', 'yourself', 'speaker', 'profit', 'reference'
, 'anywhere', 'online', 'grammar', 'name', 'want', '/', 'visit', 'address', '24', 'order', 'every', 'receive', 'theory', 'zip', 'phone', 'ne
ed', 'buy', 'personal', 'price', '20', 'syntax', '10', 'here', '95', 'off', 'return', 'step', '1998', 'science', 'list']

Evaluation of model 1:  {'legit_prec': 0.9829059829059829, 'spam_prec': 0.959349593495935, 'macro_prec': 0.9711277882009589, 'legit_recall':
 0.9583333333333334, 'spam_recall': 0.959349593495935, 'macro_recall': 0.9708333333333333, 'legit_f': 0.970464135021097, 'spam_f': 0.9711934
1563786, 'macro_f': 0.9708287753294784}

Evaluation of model 2:  {'legit_prec': 0.9956709956709957, 'spam_prec': 0.9598393574297188, 'macro_prec': 0.9777551765503573, 'legit_recall'
: 0.9583333333333334, 'spam_recall': 0.9598393574297188, 'macro_recall': 0.9770833333333333, 'legit_f': 0.9766454352441615, 'spam_f': 0.9775
051124744376, 'macro_f': 0.9770752738592996}

Randomization test result: p= 0.6453546453546454
The difference between two models is not significant.
```