BOĞAZİÇİ UNIVERSITY

Cmpe 493

Assignment 4 - Report

Arzucan ÖZGÜR

Baran Deniz KORKMAZ

FALL 2020

1. **(a) What is the size of your vocabulary when you use all words as features?**

   17956

2. **(b) Report the k most discriminating words (where k = 100) for each class based on Mutual Information.**

   The most discriminating words sorted in descending order in terms of importance based on Mutual Information are listed below:

   [’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’, ’sell’, ’cash’, ’hour’, ’win’, ’dollar’, ’pay’, ’home’, ’bulk’, ’%’, ’web’, ’call’, ’card’, ’query’, ’save’, ’income’, ’credit’, ’mailing’, ’success’, ’offer’, ’guarantee’, ’thousand’, ’purchase’, ’yours’, ’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’, ’need’, ’buy’, ’personal’, ’price’, ’20’, ’syntax’, ’10’, ’here’, ’95’, ’off’, ’return’, ’step’, ’science’, ’1998’, ’list’]

3. **(c) Report the macro-averaged precision, recall, and F-measure values obtained by the two versions of your classifier on the test set, as well as the performance values obtained for each class separately by using Laplace smoothing with $\alpha = 1$.**

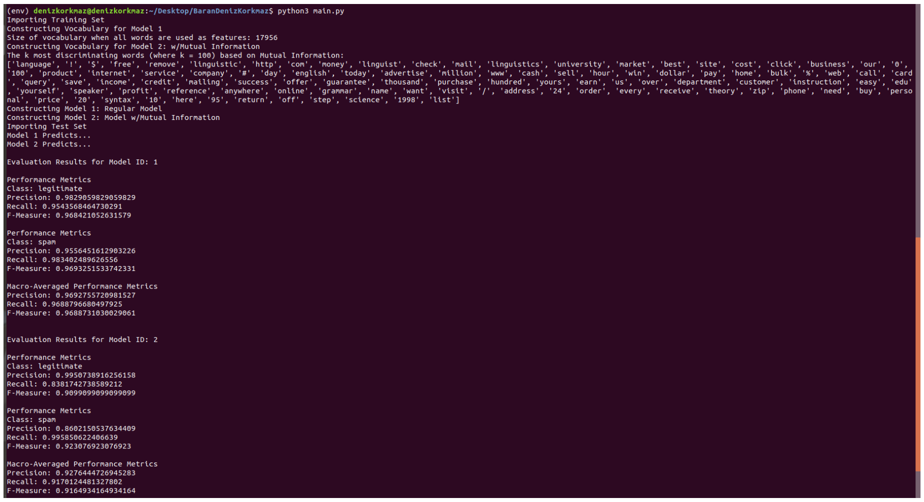
Class	Model	Precision	Recall	F-Measure
legitimate	1	0.9829059829059829	0.9543568464730291	0.968421052631579
spam	1	0.9556451612903226	0.983402489626556	0.9693251533742331
Macro-Avg	1	0.9692755720981527	0.9688796680497925	0.9688731030029061
legitimate	2	0.9950738916256158	0.8381742738589212	0.9099099099099099
spam	2	0.8602150537634409	0.995850622406639	0.923076923076923
Macro-Avg	2	0.9276444726945283	0.9170124481327802	0.9164934164934164

   In the table above, the model with the id of 1 represents the model that uses all the words in documents in the training set as features whereas the model with the id of 2 represents the model that uses the selected features via Mutual Information.

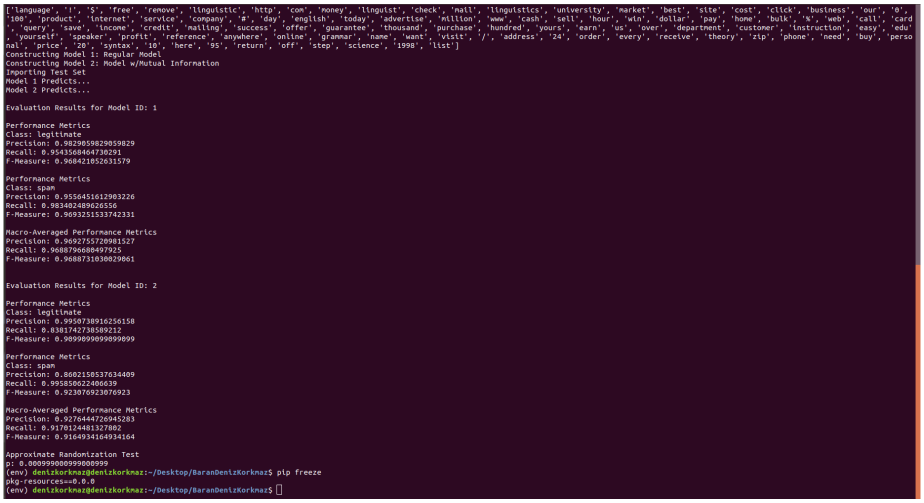
4. **(d) Perform randomization test to measure the significance of the difference between the macro-averaged F-scores of the two versions of your classifier (i.e., without feature selection and with feature selection).**

   The **Approximate Randomization Test** with **R=1000** results in the p-value of **0.000999000999000999**. The smaller the p-value, the more likely it is that the null hypothesis is wrong whereby the null hypothesis is the two systems are not different. The randomization test suggests that the two systems are different. It is sensible given that the two systems provide significantly different **F-Measure** values for both the separate classes and the macro-average of them.

5. **(e) Include a screenshot showing a sample run of your program.**



```
(env) denizkorkmaz@denizkorkmaz:~/Desktop/BaranDenizKorkmaz$ python3 main.py
Importing Training Set
Constructing Vocabulary for Model 1
Size of vocabulary when all words are used as features: 17956
Constructing Vocabulary for Model 2: w/Mutual Information
The k most discriminating words (where k = 100) based on Mutual Information:
['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', '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', 'need', 'buy', 'perso
nal', 'price', '20', 'syntax', '10', 'here', '95', 'return', 'off', 'step', 'science', '1998', 'list']
Constructing Model 1: Regular Model
Constructing Model 2: Model w/Mutual Information
Importing Test Set
Model 1 Predicts...
Model 2 Predicts...

Evaluation Results for Model ID: 1

Performance Metrics
Class: legitimate
Precision: 0.9829059829059829
Recall: 0.9543568464730291
F-Measure: 0.968421052631579

Performance Metrics
Class: spam
Precision: 0.9556451612903226
Recall: 0.983402489626556
F-Measure: 0.9693251533742331

Macro-Averaged Performance Metrics
Precision: 0.9692755720981527
Recall: 0.9688796680497925
F-Measure: 0.9688731030029061


Evaluation Results for Model ID: 2

Performance Metrics
Class: legitimate
Precision: 0.9950738916256158
Recall: 0.8381742738589212
F-Measure: 0.9099099099099099

Performance Metrics
Class: spam
Precision: 0.8602150537634409
Recall: 0.995850622406639
F-Measure: 0.923076923076923

Macro-Averaged Performance Metrics
Precision: 0.9276444726945283
Recall: 0.9170124481327802
F-Measure: 0.9164934164934164
```



```
['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', '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', 'need', 'buy', 'perso
nal', 'price', '20', 'syntax', '10', 'here', '95', 'return', 'off', 'step', 'science', '1998', 'list']
Constructing Model 1: Regular Model
Constructing Model 2: Model w/Mutual Information
Importing Test Set
Model 1 Predicts...
Model 2 Predicts...

Evaluation Results for Model ID: 1

Performance Metrics
Class: legitimate
Precision: 0.9829059829059829
Recall: 0.9543568464730291
F-Measure: 0.968421052631579

Performance Metrics
Class: spam
Precision: 0.9556451612903226
Recall: 0.983402489626556
F-Measure: 0.9693251533742331

Macro-Averaged Performance Metrics
Precision: 0.9692755720981527
Recall: 0.9688796680497925
F-Measure: 0.9688731030029061


Evaluation Results for Model ID: 2

Performance Metrics
Class: legitimate
Precision: 0.9950738916256158
Recall: 0.8381742738589212
F-Measure: 0.9099099099099099

Performance Metrics
Class: spam
Precision: 0.8602150537634409
Recall: 0.995850622406639
F-Measure: 0.923076923076923

Macro-Averaged Performance Metrics
Precision: 0.9276444726945283
Recall: 0.9170124481327802
F-Measure: 0.9164934164934164

Approximate Randomization Test
p: 0.000999000999000999
(env) denizkorkmaz@denizkorkmaz:~/Desktop/BaranDenizKorkmaz$ pip freeze
pkg-resources==0.0.0
(env) denizkorkmaz@denizkorkmaz:~/Desktop/BaranDenizKorkmaz$ 
```

Some prints are added with the purpose of notifying the user about the current stage the program is executing.