CmpE493

Assignment 4

Emre Girgin

2016400099

Report

A. Size of the vocabulary:

a. Including Multiple Occurrences: 112730

b. Excluding Multiple Occurrences: 12848

Applied preprocessing steps:

1. Punctuation removal
2. Integer removal
3. Case folding
4. Remove non-Alphanumerics
5. Remove single letters

B. The most 100 discriminating words:

['language', 'free', 'remove', 'linguistic', 'http', 'com', 'check', 'money', 'linguist', 'linguistics', 'university', 'market', 'site', 'cost', 'best', 'click', 'business', 'our', 'internet', 'product', 'service', 'mail', 'company', 'english', 'today', 'home', 'advertise', 'million', 'day', 'www', 'cash', 'sell', 'hour', 'win', 'dollar', 'pay', 'web', 'bulk', 'call', 'card', 'query', 'save', 'income', 'credit', 'mailing', 'success', 'us', 'offer', 'guarantee', 'easy', 'thousand', 'purchase', 'hundred', 'yours', 'earn', 'department', 'customer', 'instruction', 'edu', 'name', 'over', 'yourself', 'speaker', 'reference', 'anywhere', 'address', 'online', 'grammar', 'want', 'visit', 'zip', 'order', 'theory', 'phone', 'receive', 'need', 'profit', 'buy', 'line', 'every', 'personal', 'price', 'syntax', 'here', 'return', 'step', 'science', 'list', 'per', 'email', 'website', 'modern', 'top', 'watch', 'package', 'month', 'live', 'financial', 'city', 'huge']

C. Output metrics when 1-add Laplace smoothing is applied:

a. Model 1 (All words are used)

- Macro Averaged Precision: 0.99
- Macro Averaged Recall: 0.99
- Macro Averaged F1: 0.99

- Spam Precision: 0.99
- Spam Recall: 0.99
- Spam F1: 0.99

- Legitimate Precision: 0.99
- Legitimate Recall: 0.99
- Legitimate F1: 0.99

b. Model 2 (Only top 100 discriminating words are used)

- Macro Averaged Precision: 0.98
- Macro Averaged Recall: 0.98
- Macro Averaged F1: 0.98

- Spam Precision: 0.97
- Spam Recall: 0.99
- Spam F1: 0.98

- Legitimate Precision: 0.99
- Legitimate Recall: 0.97
- Legitimate F1: 0.98

D. P-Value

P-Value : 0.2827172827172827

E. Screenshots

```
(base) emre@emre-monster:~/Documents/cmpe493/asn4/src$ python preprocess.py
Files from dataset.zip extracted to dataset folder!
Processing Train-Spam...
Processing Train-Legitimate...
Processing Test-Spam...
Processing Test-Legitimate...
Done!
```

```
(base) emre@emre-monster:~/Documents/cmpe493/asn4/src$ python model.py
Selecting features...
Done!
Training model...
Training accomplished! Took 27.57secs!
Training Mutual Information model...
Training accomplished! Took 0.16secs!


 --> Model 1 : W/O Feature Selection
 - Macro Averaged Precision: 0.99
 - Macro Averaged Recall: 0.99
 - Macro Averaged F1: 0.99


 - Spam Precision: 0.99
 - Spam Recall: 0.99
 - Spam F1: 0.99


 - Legitimate Precision: 0.99
 - Legitimate Recall: 0.99
 - Legitimate F1: 0.99


 --> Model 2 : W/ Feature Selection
 - Macro Averaged Precision: 0.98
 - Macro Averaged Recall: 0.98
 - Macro Averaged F1: 0.98


 - Spam Precision: 0.97
 - Spam Recall: 0.99
 - Spam F1: 0.98


 - Legitimate Precision: 0.99
 - Legitimate Recall: 0.97
 - Legitimate F1: 0.98


The models are the same (P-Value : 0.2827172827172827)
```