cmpe 493 Assignment #4

Kayacan Vesek 2016400114

February 12, 2021

Problem description: "Spam/non-spam filter using the Multinomial Naive Bayes (NB) Algorithm."

1 Model

1.1 Size of your vocabulary when all words are used as features:

Size of the vocabulary when all words are used as features 1457

1.2 K most discrimanting words

Following 100 words are the most discrimanting words based on Mutual information. It is same for each class: because it is binary classification

```
['language', 'linguistic', 'de', 'report', '0', 'order', 'email', '20', 'mail', 'university',
'address', 'money', 'free', 'linguistics', 'business', 'linguist', 'english', 'remove', 'internet', 'check',
'com', 'name', 'our', 'department', 'theory', 'science', 'product', 'market', '100', 'grammar',
'service', 'receive', 'edu', 'send', 'site', 'http', 'sell', 'million',
'snow', 'list', 'study', 'en', 'chomsky', 'student', 'credit', 'nbsp', 'analysis', 'advertise', 'la',
'issue', 'bulk', 'y', 'word', 'card', 'click', 'day', 'speaker', 'cash', 'program', '3d',
'income', 'company', 'datum', 'syntax', 'harri', 'query',
'speak', 'conference', '1992', 'reference', 'best', 'pay', 'cost', 'home', 'teacher',
'accent', 'win', 'french', 'term', 'cd', 'over', 'translation', '95', 'dr', 'dialect',
'success', 'capitalfm', 'eskimo', 'dollar', '5', 'financial', 'seem', 'save', 'modern', 'citation', 'papers',
'ac', 'speech', 'web', 'earn']
```

1.3 Results & Sample Run of program

```
kayacan@Kayacans-MBP Spam Email Filtering % python3 main.py
First version:

Results for Class Spam:
Precision:  0.9955947136563876
recall:  0.9416666666666667
F-Measure:  0.9678800856531049

Results for Class Legismate:
Precision:  0.9446640316205533
recall:  0.9958333333333333
F-Measure:  0.9695740365111561

Macro-averaged:
Precision:  0.9701293726384705
recall:  0.96875
F-Measure:  0.9687270610821306


---------

Second version:

Results for Class Spam:
Precision:  1.0
recall:  0.9333333333333333
F-Measure:  0.9655172413793104

Results for Class Legismate:
Precision:  0.9375
recall:  1.0
F-Measure:  0.967741935483871
Precision:  0.96875
recall:  0.9666666666666667
F-Measure:  0.9666295884315907


Randomization Test result:  0.46534653465346537
```

note: last 3 line(Precision, Recall, F) are Macro average results for second version.