CMPE 493, Introduction to Information Retrieval, Fall 2015
Assignment 2 - Spam E-mail Filtering

Utku SARIDEDE
2010400189

In this project, we are supposed to implement a spam e-mail filtering program. For that purpose, the kNN and Rocchio Algorithms are our way to find.

(a). I have used the tokenisation library from python (nltk.tokenize.RegexpTokenizer). On the other hand, I have used the one of the string library (string.punctuation) to filter all punctuation marks from the data.

(b). The size of my vocabulary is 14.578

(c). Top 20 words with highest total tfidf value in spam training data:

100 => 78.5985464689
site => 79.2800104192
product => 79.3001252285
day => 80.2255296248
name => 81.5277067129
internet => 81.6025529049
http => 81.964291669
com => 81.9902497448
20 => 83.3940620941
check => 84.9485301855
address => 86.4093713709
email => 87.59365529
business => 87.7608317282
remove => 87.9701407417
our => 90.0780093024
money => 92.9812047825
order => 93.0188895848
mail => 93.17032383
free => 95.9774011536
0 => 99.9675183841

(d). Top 20 words with highest total tfidf value in legitimate training data:

seem => 47.159192023
query => 47.6205348852
de => 48.090517333
grammar => 48.6946520664
interest => 49.026623644
speak => 49.0963262603
student => 49.4948788683

theory => 50.0316894384
word => 50.1927457429
reference => 50.3032670164
issue => 50.4181023797
study => 53.189604201
edu => 57.6915069971
department => 58.0249182945
english => 71.2174053101
linguistics => 71.8350574505
linguist => 72.4565391356
university => 80.0480609778
linguistic => 85.096880921
language => 102.389021711

(e). kNN = 1:
Precision = 0.940476190476
Recall = 0.9875
F-Measure = 0.963414634146

kNN = 3:
Precision = 0.93359375
Recall = 0.995833333333
F-Measure = 0.963709677419

kNN = 5:
Precision = 0.926070038911
Recall = 0.991666666667
F-Measure = 0.957746478873

kNN = 7:
Precision = 0.919230769231
Recall = 0.995833333333
F-Measure = 0.956

kNN = 9:
Precision = 0.919230769231
Recall = 0.995833333333
F-Measure = 0.956

Rocchio:
Precision = 0.98347107438
Recall = 0.991666666667
F-Measure = 0.98755186722

(f). I have realized that Rocchio algorithm is a way that more faster and reliable than kNN algorithm. We can decide whether kNN or Rocchio is more better than the other with F-Measure value. Also we should take run time into account. But it depends on the spec of computer and as well as current processes. Therefore, it is better to compare F-Measure values.