CMPE493 - Information Retrieval - Assignment 4

Berk Atıl - 2016400102

Introduction

In this assignment, we are asked to implement Multinomial Naive Bayes algorithm for spam/non-spam task. In addition, we are also asked to select most discriminative 100 features via Mutual Information

Preprocessing

The data is read from the files and Subject line is discarded. Since stop word removal and lemmatization processes have been already done, only punctuation removal and case folding are applied. Afterwards, tokenization process is applied

Results

It should be noted that, I used natural logarithm to calculate probabilities and base 2 logarithm for Mutual Information. When all features are used, our vocabulary size is **14192**.
When we select the top 100 discriminative words via Mutual Information, we get the following words:

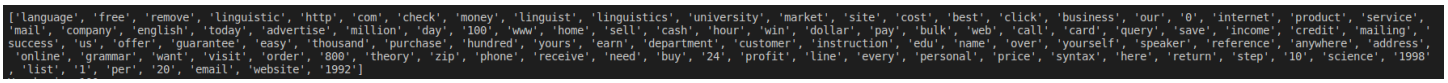

Figure 1: 100 Discriminating Words

Discriminative words are the same for both class because we have 2 classes so all Mutual Information values for each word are the same for both classes.
Precision, Recall, and F Scores are shown below.

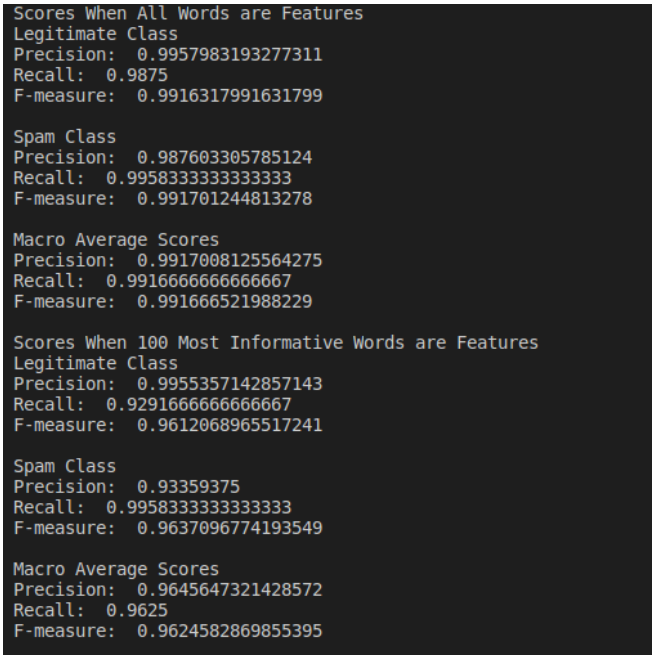

Figure 2: Precision, Recall and F-Measure

I performed Randomization test with R=1000 and R=2000 and the p values are 0.001 for both of them. Hence, it can be

concluded that the the macro-averaged F-measures of models are not equal and when we use all features the model is better.

Sample Run of the Program

```
berk@berk:~/Desktop/Course Books and Slides/CMPE493/InformationRetrieval/Assignment 4$ python3 spam_classifier.py dataset/training/ dataset/test/
Scores When All Words are Features
Legitimate Class
Precision:  0.9957983193277311
Recall:  0.9875
F-measure:  0.9916317991631799

Spam Class
Precision:  0.987603305785124
Recall:  0.9958333333333333
F-measure:  0.991701244813278

Macro Average Scores
Precision:  0.9917008125564275
Recall:  0.9916666666666667
F-measure:  0.991666521988229

Scores When 100 Most Informative Words are Features
Legitimate Class
Precision:  0.9955357142857143
Recall:  0.9291666666666667
F-measure:  0.9612068965517241

Spam Class
Precision:  0.93359375
Recall:  0.9958333333333333
F-measure:  0.9637096774193549

Macro Average Scores
Precision:  0.9645647321428572
Recall:  0.9625
F-measure:  0.9624582869855395

P value is  0.000999000999000999
```