Assignment 4 Report

a. Vocabulary Size without Feature Selection

The size of the size of your vocabulary when you use all words as features:

- **14671**

b. Vocabulary Size with Feature Selection

For both classes k = 100, so there is:

- Number of class "Spam" features: 100
- Total number of tokens in Spam Mega Document: 58268
- Number of class "Legitimate" features: 100
- Total number of tokens in Legitimate Mega Document: 9927

c. Classifier Evaluation Metrics

Spam Classifier without Feature Selection: (α = 1)

- True Positive spam: 236/240
- True Positive legitimate: 232/240

- Macro Precision: 0.9751319811058627
- Recall: 0.975
- F-measure: 0.9750659860868203

Spam Classifier with MI Feature Selection: (α = 1)

- True Positive spam   235/240
- True Positive legitimate 230/240

- Macro Precision: 0.9689535388623534
- Recall: 0.96875
- F-measure: 0.9688517587411852

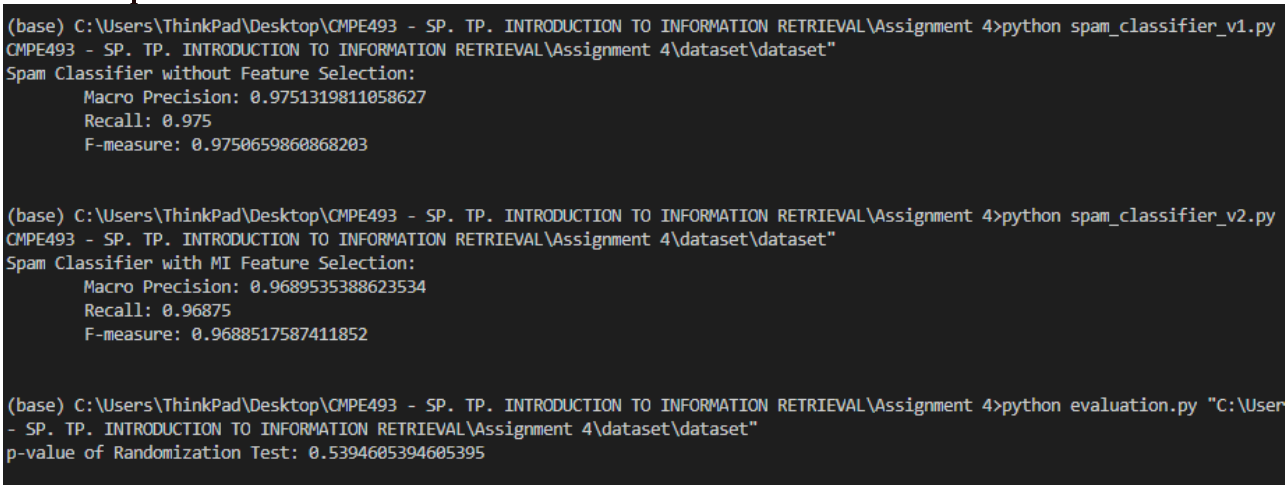
Randomization Test:

- R = 1000
- **p-value of Randomization Test:** 0.5254745254745254

Since the randomization test runs with the null hypothesis which two versions of the classifier namely "spam_classifier_v1" and "spam_classifier_v2" are the same, the p-value above 0.05 concludes that difference between F-measure scores on the individual run does not reflect the .

And **the different between two F-measure score of the classifiers is True**.

d. Sample Runs



```
(base) C:\Users\ThinkPad\Desktop\CMPE493 - SP. TP. INTRODUCTION TO INFORMATION RETRIEVAL\Assignment 4>python spam_classifier_v1.py
CMPE493 - SP. TP. INTRODUCTION TO INFORMATION RETRIEVAL\Assignment 4\dataset\dataset"
Spam Classifier without Feature Selection:
        Macro Precision: 0.9751319811058627
        Recall: 0.975
        F-measure: 0.9750659860868203


(base) C:\Users\ThinkPad\Desktop\CMPE493 - SP. TP. INTRODUCTION TO INFORMATION RETRIEVAL\Assignment 4>python spam_classifier_v2.py
CMPE493 - SP. TP. INTRODUCTION TO INFORMATION RETRIEVAL\Assignment 4\dataset\dataset"
Spam Classifier with MI Feature Selection:
        Macro Precision: 0.9689535388623534
        Recall: 0.96875
        F-measure: 0.9688517587411852


(base) C:\Users\ThinkPad\Desktop\CMPE493 - SP. TP. INTRODUCTION TO INFORMATION RETRIEVAL\Assignment 4>python evaluation.py "C:\User
- SP. TP. INTRODUCTION TO INFORMATION RETRIEVAL\Assignment 4\dataset\dataset"
p-value of Randomization Test: 0.5394605394605395
```

- p-value of Randomization Test: 0.5574425574425574
- p-value of Randomization Test: 0.5404595404595405