Bharatiya Vidya Bhavan’s

Sardar Patel Institute of Technology

Bhavan’s Campus, Munshi Nagar, Andheri (West), Mumbai-400058-India
(Autonomous College Affiliated to University of Mumbai)

Academic SEM: VII **Year: 2022-23**

Experiment: Classification- Text Analysis

Name:	Tripathi Vinayak Ramprakash
UID:	2019110067
Class:	BE ETRX
Batch:	A
Subject:	Data Analytics Lab

Objective: Separating Spam From Ham

System Requirements: CoLab, Python 3.7, SKLearn, SciKit

DataSet:

The dataset contains just two fields:
• text: The text of the email.
• spam: A binary variable indicating if the email was spam.

Code: Please consider the ipynb file in the folder for code and output

Interpretation:

1. **How many emails are in the dataset?**
   → There are total of 5728 emails

2. **How many of the emails are spam?**
   → There are **1368 spams** and **4360 hams**

3. **Which word appears at the beginning of every email in the dataset? Respond as a lower-case word with punctuation removed**
   → The dataset shows the word 'Subject' appears in the beginning of every email.

4. **Could a spam classifier potentially benefit from including the frequency of the word that appears in every email?**
   → If the term "subject" appears more frequently in spam emails than in ham emails, for example, this may potentially help us distinguish between the two types of emails. If the word appears only once in each mail, it won't help us distinguish.

5. **How many terms are in dtm?**
   → There are 878314 terms in dtm

6. **How many of the word stems "enron", "hou", "vinc", and "kaminski" appear in the CART tree?**
   → In the most frequently used Ham words, the words 'Vinc' and 'enron' appear and hence will be used for Ham classification in case of CART algorithm.

7. **What is the training set AUC of spamCART?**
8. **What is the training set AUC of spamRF?**
   → Since we are using training data itself for testing accuracy, It will always be 100%

9. **What is the testing set accuracy of spamRF?**
   → Accuracy is **98.18%**

10. **What is the testing set AUC of spamRF?**
    → The AUC for spamRF is **100%**

11. **What is the testing set accuracy of spamCART?**
    → Accuracy is **93.85%**

12. **What is the testing set AUC of spamCART?**
    → The AUC for spamCART is **96%**

13. **Which model had the best testing set performance, in terms of accuracy and AUC?**
    → From the above accuracy results and AUC for two classifiers we can say that Random Forest is better than CART

Conclusion:

- Data preprocessing is very important for training the model as any noise and irrelevant data can hamper the model performance
- AUC gives a measure of separability for classes
- A good AUC score indicates good separability
- We can see that the training accuracy is 100% in both the models but the testing accuracy drops as this is the case where moel predict on the unseen data