**IT 340 Fall 2024**

**Midterm 2 (100 points)**

A Naïve Bayesian Classifier for Spam (Phishing) Detection

**General Description**

In this assignment, you will be implementing a Naïve Bayesian classifier for phishing detection. The model that you will implement is the **Multi-variate Bernoulli Event Model**. The assignment is due **November 12 at 11:59 PM**

**Dataset Description**

You are provided with 1200 emails, in which 600 of them are non-spam (ham) and the other 600 are phishing emails (spam). The non-spams are collected from the Enron email dataset (<https://www.cs.cmu.edu/~./enron/>), which is a well-known public dataset for email data. The phishing emails are collected from [www.scamdex.com](http://www.scamdex.com), by one of my colleagues at the Dakota State University in 2010.

The punctuation in all the emails have been removed. In addition, each email contains no repeating words.

**Requirements**

* Select 500 non-spam (ham) and 500 phishing emails (spam) as your training set.
* Use the rest of 200 emails as your testing set.
* Implement the Multi-variate Bernoulli Event Model by strictly following the handout.
* You are required to use the provided code for the data manipulation part.
* You cannot make any change to the code provided.
* Your program should display precision, recall, and F1-score in addition to the accuracy.
* You cannot use any Python packages except for the ones coming together with my code.
* You cannot use any online reference for this assignment.
* Once you have completed the implementation, you will find out the accuracy of the model is not very high (approximately 70%). This is to be expected. You must explain why this can happen.
* Your explanation must include two parts:
  1. Show me what caused such low accuracy.
  2. Show me why the cause can happen.
* If your accuracy is somehow higher than 80%, you then need to meet me in person to go over your program in order to avoid receiving a failing grade.
* Grade distribution:
  1. Implementation: 60 points
  2. Explanation: 40 points