**IT 340 Fall 2022**

**Midterm 2 (50 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 10 at 11:55 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. The dataset was originally used for my research project 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.
* You are required to use Python for this assignment.
* Your program should randomly shuffle the emails that are used for training and testing.
* Your program should display precision, recall, and F1-score in addition to the accuracy.
* You cannot use any existing Python machine learning packages or online reference for this assignment.
* This project can be a group project with no more than three members in a group.
* If you worked in a group, only one submission is required, with all the members’ names properly documented.
* If you would like to ask me to check your code, you must come to my office.
* I do not answer assignment-related questions on due date.