Nathan He, Samujjwaal Dey, Yushen Li 11/22/19

CS418 Data Science Professor Bello

**Final Project Proposal**

**Introduction**

Nowadays, spam emails are extremely prevalent in today’s society. Companies and scammers often flood people’s email inbox with a lot of irrelevant information, sometimes even with viruses, to get our attention and give them our sensitive information. Although almost all major companies have some sort of spam detection to filter out spam emails, we want to be able to dive deeper into the inner workings of the spam filtering system. To do so, we will use University of California at Irvine’s dataset on spam and non-spam emails to train our model to predict spam and non-spam emails. Also, we want to perform basic text analysis on those spam and non-spam email dataset. The dataset is used for Hewlett-Packard Internal-only Technical Report; therefore, it only contains sensitive keywords that Hewlett-Packard wants.

**Problem Statement** - What problem are you trying to solve?

We want to be able to classify if emails are spam or not spam by training the current dataset. In addition to solving the first problem, we want to be able to find out which specific words inside the email contribute more to finding out if emails are spam-related or not. Also, we are trying to figure out what is the accuracy of our training model.

**Brief Description of Data Science Solution** - How are you going to solve the problem?

We are going to use classification as a way to identify if emails in the given dataset are spam or not. The next step is to use clustering for text analysis on the dataset of emails.

**Dataset:** We use the dataset called Spambase Dataset is provided by UCI’s Machine Learning repository.

**Dataset Link:** <https://archive.ics.uci.edu/ml/datasets/spambase>

**Methods of Analysis**

1. Classification

We will use different spam keywords, such as money, edu, and others, to train the model and make predictions about whether the email is spam or non-spam when they contain those keywords. 0 is non-spam and 1 is spam.

2. Clustering

We cluster emails on the basis of the frequency of particular words in each email into spam and non-spam. We then determine which words contribute more towards an email being spam based on how true the clusters are.