SPAM EMAIL CLASSIFIER

Project Proposal

UTA_ID:1001861777

VIJAY GANESH PANCHAPAKESAN

Subject: Data Mining

Spam Email Classifier

The Dataset I have used for this classifier is accessible through the following link:

https://www.kaggle.com/veleon/ham-and-spam-dataset

Features:

The main feature of this classifier is to predict whether a given mail is spam or not. The dataset consists of two Folders Spam and Ham. Each folder consist of text file, in each text file there would be a mail, and there were 501 Spam mails and 2551 non-Spam mails. Therefore, in total there were 3052 mails. The goal of this App is to analyse these mails and with the help of these mails, the App needs to predict whether the given mail is spam mail or not .For the prediction I have not used the entire data I have split the data into train and test. The train data consists of 80% of the total data and rest 20% would be the test data. By training the train data, we can apply the results, which was memorized by the training to test the testing data and produce a result that is the accuracy .By applying above-mentioned rules I was able to produce an accuracy of 98%.To recap the main aim of the classifier is to predict whether given mail is spam or not.

Similar Spam Email Classifier Apps:

To effectively handle the threat posed by email spams, leading email providers such as **Gmail, Yahoo mail and Outlook** have employed the combination of different machine learning (ML) techniques such as Neural Networks in its spam filters. These ML techniques have the capacity to learn and identify spam mails and phishing messages by analysing loads of such messages throughout a vast collection of computers. Since machine learning have the capacity to adapt to varying conditions, Gmail and Yahoo mail spam filters do more than just checking junk emails using pre-existing rules

Why Compelling Spam Email Classifier is?

In recent times, unwanted commercial bulk emails called spam has become a huge problem on the internet. The person sending the spam messages is referred to as the spammer. Such a person gathers email addresses from different websites, chatrooms, and viruses. Spam prevents the user from making full and good use of time, storage capacity and network bandwidth. The huge volume of spam mails flowing through the computer networks have destructive effects on the memory space of email servers, communication bandwidth, and CPU power and user time. The menace of spam email is on the increase on yearly basis and is responsible for over 77% of the completely global email traffic. Users who receive spam emails that they did not request find it very irritating. It is also resulted to untold financial loss to many users who have fallen victim of internet frauds and other fraudulent practices of spammers who send emails pretending to be from reputable companies with the intention to persuade individuals to disclose sensitive personal information like passwords, Bank Verification Number (BVN) and credit card numbers.

Thus in order for the customers to have full bandwidth of network and the customers not falling into the trap of the spammers .Thus there must be a good spam mail filtering must be there to prevent the customers form the above mentioned issues.

I have incorporated ML techniques like EDA visualisation Feature engineering and model selection for creating an emulator of a spam mail classifier. I have used Naïve Bayes as my classifier and CountVectorizer as my Vectorizer for my spam mail classifier