Designing a Spam/Ham Classifier

By Ben Chu

1. Introduction

1.1 For this capstone project, I will be designing a classifier to classify spam and ham emails. This project is to practice designing a classifier that will model and predict whether or not an email has keywords belonging to a spam email. We will be extracting data from our email dataset and observe features within the emails.

1.2 Business Problem

For this assignment, we will be addressing the problem of spam emails that cause issues to all email owners. Spam emails cause a huge issue creating false information and clickbait that can scam users of their cash. From creating a classifier for spam and ham emails we want to be able to remove the risk from email usage and eliminate a huge proportion of spam emails.

1.3 Data

We will be utilizing an online dataset of training and testing data for spam/ham emails. The emails will have a variety of features of spam or ham emails. The source of the data encompasses a wide variety of emails from generic sources.

2.1 Methodology

For our methodology, we first cleaned up the data and observed any details that might signify relation to spam or ham categories. For our exploratory data analysis we observed all of the features for our data and identified some keywords that would be perfect for predicting if an email should fall under the spam category. The data featured a great deal of information but many emails that had keywords such as money, free, special, offer, or followed a JSON formatting, had a high rate of identification for spam/ham.

3.1 Results

For the results of our test, we find that there was an unproportional level of emails that had keywords such as 'money', '<html>', 'sold separately', 'opt-in service', 'new and free', 'vip membership', 'are unbelievable', 'need product', '</table>\n \n', 'align="center"', 'we accept visa', 'state, zip', 'can cancel', 'opportunity', 'income', 'potential', 'free', 'dollars', 'sex', 'career', 'increase', '</font></blockquote>', 'grants', 'free money', '<td bordercolor=', 'all mailing lists', 'click here'. We found that there was a significant amount of information in the emails that could lead someone to classifying it through generic keywords. Although we didn’t find a classifier that could predict 99 or 90% of correct emails (spam/ham) we found enough information to calculate and predict around a 85% range for a correct classification.

4.1 Discussion

Based on our experiments and testing, we know that spam emails follow a specific pattern of common keywords. We know that spam emails oftentimes look for people that are seeking out free offers, opportunities, and money. Keywords that generally incite attention for readers or promote usage of credit cards offer a high probability of being within spam emails. The hardest portion of the keyword search is finding words that do not exist in ham emails. Isolating words that only exist in spam emails is difficult, but can appear in specific phrases. Just from adding html, sold separately, money, or vip membership, we drastically increased the probability of predicting the right values. For others that are replicating this classifier, I would search for the most common words in either dataset and not appearing in the other dataset. From identifying the best non-overlapping words in the strings, we are able to identify what works best for either classification category.

5.1 Conclusion.

In conclusion, we require a careful analysis of each individual email keyword system and need to conduct exploratory data analysis to devise our best classifier. From utilizing enough tools within our data science toolkit we are able to navigate through the different spam and ham emails. Overall, our classifier predicted within a 85% probability whether an email is spam or ham.

References :

List of neighborhoods in Toronto:
https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
Foursquare Developer Documentation: https://developer.foursquare.com/docs