# Spam Detection Using Natural Language Processing and Python

Angel Argueta | Creston Altiere | Sahar Sheikholeslami | Wessley Dennis

## Introduction

The ability for a computer system to manipulate, comprehend and interpret human language has long been a focus of the wider computer science community. Natural Language Processing, or NLP for short has long been one of the core focuses of the Machine Learning community. NLP technology is crucial to fully understanding and extracting useful data from articles, and even speech. NLP algorithms are becoming increasingly complex with the availability of large corpuses of voice and text data from various internet mediums like emails, social media, online video and voice assistants.

## Problem Definition

The Goal of this paper is to leverage contemporary Natural Language Processing techniques to construct and train a Spam Email detector. The spam detector will be able to examine components of the message including but not limited to it’s subject and content to determine whether or not an item is likely to be spam.

Ideally when it comes to unwanted emails, whether or not it’s spam could depend largely on what the owner of the inbox is interested in at that time, or whether or not it’s a work account. Even today spam detectors tend to struggle with striking a balance in determining what is considered spam or not.

Some classical approaches to determining this include allowing the user to tag unwanted items as spam to train the spam detector in determining what to look for as they arrive in the inbox. This experiment seeks to examine the possibility of taking a corpus of pre sorted emails, to construct a system that can identify spam emails.

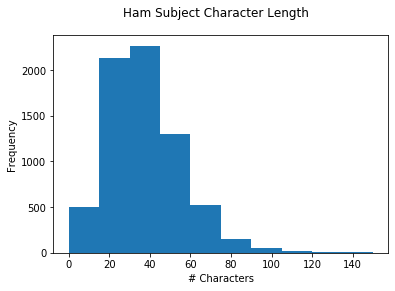
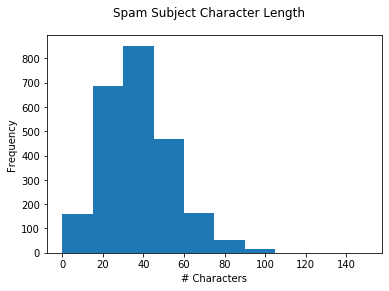
## Data Collection

For this implementation we will be utilizing the SpamAssassin public mail corpus from 2004. An added advantage of using this dataset is we will be able to determine not only how well it will be able to detect spam emails from 2004, but with testing on modern spam emails we can possibly explore if spam emails have improved in their methods of solicitation, by testing it on a few spam emails from current day.

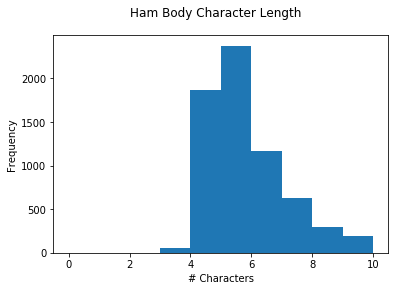
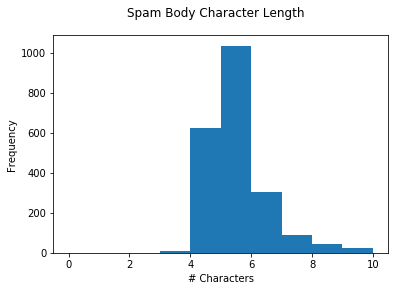
The Spam Assassin public main corpus is widely used in training spam detection systems and should serve as a good dataset for which to train a baseline detector. The hope is to get a system that’s capable to some degree of determining and distinguishing some features of spam emails vs legitimate emails.

In order to see if there were some areas we could examine at a surface statistics level to help determine possible features to focus on we’ve performed a small Exploratory Data analysis to get a better understanding of what the data consists of.

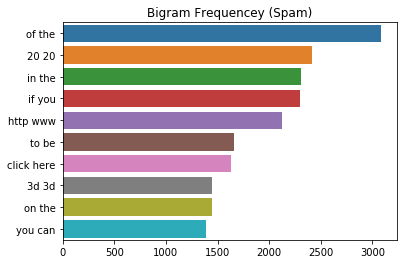
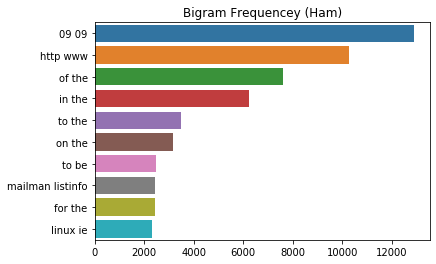
The corpus consists of 9351 emails, 6952 of which are legitimate (Ham) and 2399 of which are actually Spam. Due to this asymmetry in the data we will have to structure our training in such a way that the asymmetry does not effect the likelihood of each estimate. Below are some of the results of the analysis and a small discussion of each:

The above two histograms illustrate the frequency of Subjects with specific character lengths (lower Axis). While one can infer that a slightly larger percentage of spam emails include subjects over 40 characters, we can infer, that on the surface level, this may not be a really good indicator to determine whether or not an items is likely to be spam.

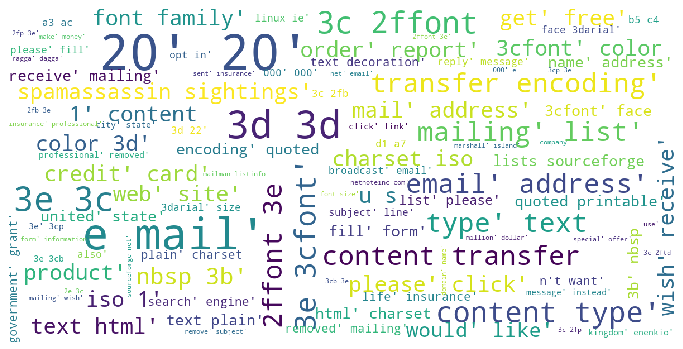
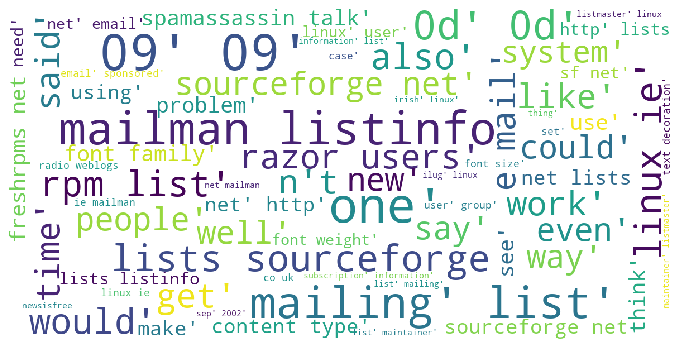
 

Unfortunately other than Ham emails seemingly having much a much more varied word length for body, this may not be a really useful feature for the learning process. Usually most spam is automated, and as such the focus is towards using words that are optimized towards getting the best results from a marketing standpoint, however people tend to use very industry specific terminology that can often be longer. That can be examined as possible explanation here.



Here we have two Bigrams wherein on the X axis is the frequency and on the Y axis the Bigram. When we remove the HTML from the emails we’re able to notice interesting that one of the most frequently appearing words in the Spam dataset body is the phrase ‘click here’ when compared to others. This might be an interesting point of exploration.

We took the liberty of generating two word clouds for our different datasets, to get a better high level look at the data. The Ham wordle is illustrated in blue (on the left) and the Spam wordle is illustrated in red (on the right)



Due to the inability to fully cleanse the email dataset of HTML fragments, the clouds are namely dominated by HTML tags, and formatting symbols. However it’s important to note however, that a closer examination of the spam wordcloud, a lost of marketing terms like product, credit card, please, click, dominate more of the mid section implying that there’s definitely a formula that spam and marketing emails follow, even if loosely.

The above findings might help give insight into understanding our data, and what areas we might be able to effectively employ it in training.

## Data Preprocessing

## Feature Engineering

## Model Selection and Architecture

## Training and Evaluation

## Results

## Discussion