Spam Email Classification

Bryce Wall

9/10

Description

I want to see if it is possible to train a machine-learning model to identify and properly classify spam emails. I have many emails that come in unfiltered that I am able to clearly identify as spam mail. There have also been several important emails incorrectly categorized as spam mail when they should have gone to my inbox. Most spam emails today follow a similar format and are easily detectable, but just as spam is more complex than 20 years ago, it continues to develop past modern detection software. I am curious about what markers in the emails would classify them as a higher possibility of spam mail or a normal email from a coworker or friend.

Data

When thinking of this problem and the digital markers that would contribute to the classification as spam or not, I was able to think of a few right off the bat. Analyzing specific phrases used in the emails, email domain names, punctuation or spelling errors, links with uncommon endings, and reference to specific topics would all contribute to proper spam email classification.

Sender Domain	Is bcc	Subject Punctuation Count	Link Count	Link Domains	Flagged Topic	Spelling Error Count	Is Trusted Sender	Target: **Spam**
@paypall.com	Yes	5	3	.grz	free	13	No	Spam

Gathering Data

There are many different databases that have spam email features, but the one challenge with gathering the data would be combining useful data into one dataset. Some datasets contain entire email bodies as the feature to train on, and some only have information on the sender information. If I were placed on this project, I would also be able to go into my personal email and gather many correctly identified spam emails which we would then be able to extract spam flags from.

There are several datasets of email body data on Kaggle that contain over 4,000 data entries. A simple parser could be written to go through the data and count up specific values like punctuation, spelling mistakes compared to a dictionary library, and counting the lengths of subject lines. UC Irvine has one dataset that includes several thousand emails with the frequency of key words being contained which would be helpful to analyze.