CS 410 Project Progress Report

Topic: Data Set Creation

Team: kms

Kim Li -- kimli2

Kevin Tzeng -- ktzeng2

Shreyas Chandrashekaran -- svc3

1) **Which tasks have been completed?**
    a) We have built a basic web crawler that is able to parse through pages and find the respective sender, recipients, subject line, and email content. We are also able to format the raw data into a CSV file, where further sorting can be applied.
2) **Which tasks are pending?**
    a) Our current web crawler is too generic and is creating a data set that contains a lot of extraneous information, especially because there is no selection criteria other than the content being an email. There are many duplicate emails or emails with overlapping content, which we would ideally like to minimize in our actual data set. Our first goal is to organize all of our results into a smaller collection that still accurately represents all of the data that was initially scraped. From there, we can actually rank the emails based on importance for our actual data set. We also need to solidify how the dataset is being ranked, since there is currently no strict guideline for what makes an email important compared to spam. This specification will also help with developing algorithms in order to sort important emails from spam.
3) **Are you facing any challenges?**
    a) No major challenges, just working on limiting our data set, since the web crawler is returning a few thousand emails, which is too large and tedious for manual ranking.