Emails

Business Understanding

1. There are many scam and spam emails (collectively "junk" emails), they need to be filtered.
2. How can we automatically determine whether an email is "junk" based on its content?

Analytic Approach

This project aims to identify the status of an email (normal or junk). This involves a yes/no answer, hence it requires a classification approach.

Data requirements

We need to identify the necessary data content, formats and sources for initial data collection. We need to define what data we need in addition to the individual emails themselves. The additional data that we need most likely includes personal information of the users.

Data Collection

We need help from the major tech companies that provide email services (e.g. Google, Microsoft, Yahoo etc.). We will compile the data provided by the tech companies. Techniques such as descriptive statistics and visualization can be applied to the data set, to assess the content, quality, and initial insights about the data. Gaps in data will be identified and plans to either fill or make substitutions will have to be made.

Data Understanding and Preparation

Firstly, In order to understand the data related to junk email, statistics will be needed to be run against the data columns that would become variables in the model. Second, pairwise correlations will be used, to see how closely certain variables were related, and find the most relevant ones for modeling. Third, histograms of the variables will be examined to understand their distributions. After understanding the data, we need to prepare the data by doing the necessary transformations into suitable formats. We also have to ensure the missing data and duplication are accounted for.

Modeling and Evaluation

Since this project requires yes/no outcomes, predictive modelling is the most suitable form of modelling. In this case, whether the email is "junk" or "not junk". A decision tree can be used to evaluate if the answer the model can output, is aligned to the initial design. The ROC curve will be used to determine the optimal classification model.