Spam mail classifier

Team 3 - Kunal Chugh & Vighnesh Venkatakrishnan

Data Source

- Following datasets will be used:
- Trec spam dataset (https://trec.nist.gov/data/spam.html)
- Lingspam dataset (http://www.aueb.gr/users/ion/data/lingspam_public.tar.gz)
- PU Dataset (http://www2.aueb.gr/users/ion/data/PU123ACorpora.tar.gz)
- The combined size of the 3 datasets is approximately 600MB
- The data is unstructured, but can be grouped into columns From, Date, Return-Path, Received, Message-ID, Reply-To, To, Subject, Date, X-Mailer, MIME-Version, Content-Type, Subject and Body.

Data Cleaning & Transformation

- Since the data is unstructured, it will be converted to a structured format.
- As part of data cleaning phase, removal of punctuation & additional white spaces will be done. Missing values in categorical columns will be filled with NA, to make another categorical value.
- Stop-word removal, Stemming & Lemmatization, Tokenization will be done to prepare data for feature engineering.

Feature Engineering

- From the tokenized text, a count vectorizer will be used to create a feature.
- TF-IDF will be applied to all words in the corpus, and top 10 words will be selected to create another feature.

Use Case & Acceptance Criteria

- After the model has been trained on a corpus of spam and ham mails, a user can use it to predict whether mails in the test.csv are classified as spam or not.
- We aim to improve the model up to 75% accuracy for classifying spam correctly.