Capstone Project 1 Proposal
Spring Board Data Science Career Track
~ By Zach Nguyen

Problem: As technology improves at exponential rates, everything else improves with it. We seem to enjoy better lives in almost every aspects, except for our privacy and ability to concentrate. Spam sms is one of the major culprit in such area. The need to filter spam accurately is enormous since it afffects just about anyone with a cellular devices. It is a low-hanging fruit problem that can be solved with text classification. Our goal for this project is to build a classification model to accurately predict good sms from the spams.

Client: In addition to ordinary consumers, telecom companies would wants to improve the quality of their products to customers by identifying spam sms and prevent them from reaching customers. This way, they will be able to improve their service quality significantly and win over customers.

Data: A compilation of many different sms datasets with ham/spam labels provided by UCI dataset directory at http://archive.ics.uci.edu/ml/datasets/SMS+Spam+Collection#

Methodology:

- Use text analysis techniques to explore the data (with packages such as Pandas, Seaborn, nltk, Spacy). Analysis include EDA (word count, word length, character count, lexical diversity),TF-IDF analysis, feature engineering.
- Build a Naive Bayes model to predict the label of any sms (sklearn).

Deliverables: Slide Deck, Report, Code