Data Analytics Report

Consumer Complaints Free Text Classification

1. **Executive Summary**

This analytics is a machine-learning project. It uses Python scikit-learn, an open source machine learning library, to classify “Consumer Complaints” data (publicly-available open dataset). The result can be used to automatically classify any new consumer complaints or similar free text.

1. **Introduction**

(Note to Coursera reviewer: This project does not use Foursquare data, due to the fact that in my current job, we do not deal with location data. Rather, we deal with great amount of free texts. Therefore, I would like to use this opportunity to develop an analytical tool that can be leveraged in my daily job. Thank you for your consideration.)

It is estimated that the majority (80% ~ 90%) of current data is un-structured data; that is, there is no pre-defined hierarchy of the data structure and data format. Such un-structured data include emails, text messages, reviews and comments, and other free texts. These data are of great value, since they can be used to provide deep insights and to reveal covered facts that are critical to the business. However, they are highly challenging to be processed for further analysis.

This project aims to develop a preliminary tool to interpret and classify free text. It reads “Consumer Complaints” data, a publicly-available open dataset, undergoes data cleaning process (including removal of stop words, tokenizing and lemmatizing), and uses part of the datasets as the training dataset. Then it will go through the training process to find typical keywords for each Complaint Category (termed as “Product” in the original data source). In the testing process, the tool will evaluate the keywords in the new data entry and classify it into an appropriate category.

1. Data Source

The data is from data.world (<https://data.world/data-society/consumer-complaint-data> ). See snapshot below.

Figure 1. Data.world snapshot of Consumer Complaints data

Figure 2. Data.world snapshot of Consumer Complaints data, Plots

The original data can be found at Consumer Financial Protection Bureau’s website (<https://data.consumerfinance.gov/dataset/Consumer-Complaints/s6ew-h6mp> ). See snapshot below. It also contains a brief description of the data column.

Figure 3. Original data source: Consumer Complaint Database, from Consumer Financial Protection Bureau

This project will only use two data fields (columns): “Product” (the type of products or services); and “”. All unique values in “Product” is below: