**Capstone Project 2: Amazon Reviews**

**Problem**

Many companies have the need to analyze reviews and items purchased by customers. For the purpose of this proposal, ‘items’ refers to goods, services, or digital media such as movies, TV shows, video clips, songs, or albums that are provided to customers. This type of data can be used for many purposes including product recommendations and targeted advertising. Analysis of ratings can help determine which products are likely to be more highly rated in the future and which are not, leading to choices for which products should be recommended by taking an item-based approach. Similar purchases by different customers can be used to recommend items to customers that have similar purchasing history, taking a user-based approach. A better approach would be to combine these into a collaborative-based approach that also takes into account the different categories the items fall into. It would be very beneficial for any company that provides these ‘items’ to provide useful recommendations to customers, so there are many potential applications for these techniques.

In addition to recommendation systems, many companies and businesses have the need to analyze text information to assess customer sentiment. Models or programs that are capable of quickly analyzing text have many applications from assessing overall sentiment or satisfaction to providing meaningful responses or directing customers using chatbots. Focusing on the former, sentiment analysis can help companies quickly identify positive vs. negative text allowing a response either way. This can also assist with identifying products that tend to be associated with positive text vs. negative text for the purpose of marketing or making other decisions for products.

**The Dataset**

The dataset for this project is publicly available from the link at the end of this section. There are 233.1 million total reviews available from this site. The data span from May 1996 to October 2018. The primary data consists of ratings, text review, and helpfulness votes from other users. The data also includes product metadata (tem descriptions, category information, price, brand, and image features) and links to other items viewed/also bought graphs. The metadata has also recently been supplemented with technical details and a similar products table.

<https://nijianmo.github.io/amazon/index.html>

**Approach**

Given the large size of this dataset and the many options it provides, a very specific approach will be taken to keep the project manageable, though additional analyses could either improve the recommendation system or supplement it using different data or techniques.

The first analysis will be to use a collaborative-based approach to provide recommendations to customers. A matrix will be created to compare users and items purchased, and the Pearson correlation coefficient will be used to assign similarity to users, taking a user-based approach. A minimum number of purchases in common will likely need to be set to avoid high correlations between users with few purchases, and this value has yet to be determined. Mean item ratings will also be used to help recommend the more highly rated items over items with lower ratings. A content-based approach will also be incorporated with these recommendations by using item descriptions, categories, and brand information to assist with item recommendations.

The second analysis will use clustering to identify patterns in purchasing that may not have been detected with the first approach. Recommendations will be created for customers in the same cluster that have not yet been purchased.

The third analysis will utilize natural language processing (NLP) to predict ratings from the text reviews. Supervised machine learning algorithms will be used to train models based on a subset of the data, and the models will be validated against a hold-out validation set. Different techniques may be applied to process the text data including the use of different vectorizers or combinations of n-grams and stop words to produce the best predictive model.

**Audience**

Any company providing goods or services should be analyzing data from users to help improve recommendations and analyze text responses from users/customers regarding products or company sentiment. The techniques to be applied in this project can be applied to many different types of products including retail, streaming services, health care, service industry, and many others. The tools and methods for this analysis should therefore be relatively ubiquitous.