Name of Team members: Gagan Shantha Kumar, Jatin Taneja, Raghavendra Venkatesh

Dataset:

We will be working on the Amazon Customer Reviews (a.k.a. Product Reviews) dataset. In a period of over two decades since the first review in 1995, millions of Amazon customers have contributed over a hundred million reviews to express opinions and describe their experiences regarding products on the Amazon.com website. Over 130+ million customer reviews are available for us to apply our algorithms to generate useful information and analysis using this dataset. The data is available in TSV files in the amazon-reviews-pds S3 bucket in AWS US East Region. Each TSV file will contain all the reviews for the respective product category. Each line in the data files corresponds to an individual review (tab delimited, with no quote and escape characters). And each review will contain fields such as country of sale, customer id, product id, product title, product category, star rating, verified customer or not, review content and the date the review was made.

The link to the dataset can be found here : https://registry.opendata.aws/amazon-reviews/

Data Analysis Challenges:

Overreach Goal: Amazon product reviews and ratings are very important to business. Customers on Amazon often make purchasing decisions based on those reviews, and a single bad review can cause a potential purchaser to reconsider. We are interested in analyzing two aspects from the product review dataset. In the first part, we wanted to check the possibility of classifying a set of customers who might have bought similar products and left reviews which are similar as well. This will help us in answering questions such as does the users who bought similar product more likely to leave similar ratings? What is the likelihood that if customer A has reviewed the product P1 and P2 and customer B has reviewed the product P1, is she most likely to leave a comment on product P2 as well? We will try to answer this question as part of this project. In the second part, we will check the prospect of classifying the customers based on the reviews for a set of products. This will help the sellers to determine if a customer can be classified to a group so that the sellers can make product suggestion to the other customers in the same group.

Main Task 1(primary): We are planning to build a model which is used for **Classification and prediction using existing data mining libraries**. On the given product review dataset, our model should be able to differentiate the customers for different clusters based on the reviews they have written for the product and their categories. We can consider the ratings they have left and as well as the product category to classify the users to groups.

Main Task 2(primary): We are also planning to suggest any given user a set of product which she might be most likely to buy by making use of the reviews provided by other peers of hers in the category in which she is classified. We are planning to make use of the verified customer

field in the dataset to confirm if a person has bought a product. It is like the “Users who bought this also bought” feature in Amazon. We will be using a parallel version of Apriori **frequent itemset mining** algorithm.

Main Task 3(primary): The dataset has 130 Million reviews. For initial testing we will need smaller samples since we have to check the sanity of the algorithm we are going to build. A simple random sample of choosing the records in random order will not work well since we will not consider any undiscovered relation between the records and generating a set of sample reviews just for the purpose of checking the sanity of our algorithm will also not make much sense since we might miss the importance of randomness, hence we need to design a special **graph sampling algorithm** which will provide us the sampling records from the actual input dataset and also maintaining any undiscovered dependencies with the records if any.

Main Task 4(secondary): Gagan Shantha Kumar has not completed Homework 2 and 3 completely which were a part of assignments on the Twitter Dataset. As part of this project, he is planning to provide the solution for Homeworks 2 and 3 completely under the task listed as **Revisiting Homework 2 and 3 Twitter Problems**.