**Recommender Engine for Amazon products based on user’s ratings**

**Data Description**

The data we used for our analysis was retrieved from the University of California San Diego’s Repository of recommender systems datasets. We choose 2 datasets from the repository for Amazon’s clothing, shoes & jewelry products. This data contained more than 32 million user ratings for almost 200 k products, the datasets we chose are as follows:

* One was the ratings data for the products
* Second was the metadata containing the description of the products sold on Amazon

**Ratings Data Meta Data**

**A screenshot of a cell phone

Description automatically generated**A screenshot of a cell phone

Description automatically generated

**Techniques**

**Data Pre-Processing:**

* We dropped unnecessary columns from the dataset which were not necessary for our analysis, we kept only user\_id, item\_id & ratings from the ratings data and categories, user\_id & title of the products from the metadata
* We encoded the alphanumeric user\_id and item\_id values to integers with the help of **LabelEncoder** from **sklearn**, for this data to be further used for Alternative Least Sqaures method.

**Analysis:**

* We used collaborative filtering approach to build a recommender system on the Amazon products data.
* The data pre-processing and model building was done on **google collaboratory.**
* Filtering is done based on about interests of a user by collecting likeability information from many users (collaborating).
* PySpark’s **MLlib** library was used for the Collaborative filtering implementations using **Alternating Least Squares**. The implementation in **MLlib** has the following parameters.
* We also used cross-validation to tune the parameters of the model.

**Results**

We build bubble graph using **Tableau** as the visualization tool to show the highly likeable products among the Amazon products data. Based on the popularity rankings, which was decided by analyzing the historical data of user’s likeability top recommendations were given using the ALS technique of MLlib. The following graph shows that **“Sports and Outdoors”** was the most popular category with the highest user ratings, followed by **“Toys & Games”.** The size of the bubbles represents the sum of ratings and color represents different categories of Amazon products.

The Data frame below represent top 10 recommendation for users by **ALS**.

**Top Recommendations for users**

**A close up of text on a white background

Description automatically generated**A picture containing screenshot

Description automatically generated