CS5344 Lab 2

AY2018/2019 Semester 1

This lab requires you to consider the different approaches to build a recommender system. You will need to understand Spark SQL, DataFrame and Spark MLlib. **Please do this assignment in pairs.**

Online retailers often increase their sales revenue by recommending additional products to existing customers who are already making a purchase (cross-selling). **Write Spark programs that implement the following methods of product recommendation.**

1. Frequently browsed together by the customers
2. Collaborative filtering

Which of the above would you consider to be a better recommender system? Justify your answer with empirical results obtained by running experiments to compare the Conversion Rate (CR) of the recommendations. A user has obtained at least one good recommendation if s/he purchased at least one product from the recommended list of top K items. If L is the list of recommended products and L' is the list of products actually purchased by the user, then the conversion rate is given by:

$$\mathbf{ConversionRate@K} = \begin{cases} 1 & if\,|\mathrm{L} \cap \mathrm{L}'| > 0 \\ 0 & otherwise \end{cases}$$

Dataset: Amazon product data http://jmcauley.ucsd.edu/data/amazon/links.html

Choose any ONE product category. The product metadata captures the user browsing behavior (“also_viewed”) and the actual purchase (“also_bought”).

Use Spark DataFrame to load your dataset.

Run at least three SQL aggregate queries to learn the basic features of the dataset.

Methods:

- You can use the Apriori algorithm to find products which are frequently browsed together.
- You can use Spark MLlib to implement collaborative filtering recommendations. You may need to pre-process the dataset to retain users who have bought some minimum number of products.

Deliverables:

Upload the following to the Lab2 folder in IVLE. All the deliverables should be zipped into one file and named as your group number (e.g., Group_XX).

(a) Spark programs (with documentation within the code).

(b) Report that includes:

- Visualization of the results of the SQL aggregate queries.
- Data processing carried out on the downloaded dataset.
- Experiment results comparing the ConversionRate@K for various K values for each recommender methods. You can vary K from 1 to 5.
- Analysis of the results.