

Assignment

Q1. Consider the two data files (users.csv, transactions.csv). Users file has the following fields:

- a) UserID
- b) EmailID
- c) NativeLanguage
- d) Location

Transactions file has the following fields:

- a) Transaction_ID
- b) Product_ID
- c) UserID
- d) Price
- e) Product_Description

By making use of Spark Core (i.e. without using Spark SQL) find out:

- a) Count of unique locations where each product is sold.
- b) Find out products bought by each user.
- c) Total spending done by each user on each product.

Remember, you have to make use of Spark Core for this question. (15 Marks)

Q2. Consider the dataset file Olympics.csv. This file contains information about the Olympic games, players participating in the games, and details of medals won by them. Using Spark core and the data file, compute the following:

- a) Total medals that each country won in a particular sport (such as Gymnastics).
- b) In each Olympic games, how many medals has India won?
- c) Compute top 3 countries in terms of total medals by each Olympic games year.

Remember, you have to make use of Spark Core for this question. (10 Marks).

Q3. Consider the Movie Recommendation code and problem that was discussed during the class (Session 4 and 5).

- a) Please provide a brief write-up on the problem, steps needed to arrive at the solution (recommendation system), and how exactly those steps are implemented in the code. You can make use of the PPT file that discusses the broad solution. While you are doing so, please also mention what each line of code does (It is not sufficient to mention what each block of code does, you would have to provide explanation for each line). (10 Marks)
- b) Create a variation of the code by recommending only good movies (i.e. remove those movies that have bad ratings). (10 Marks)
- c) Rather than using the cosine similarity score, make use of another correlation method (for example, Pearson) to compute similarity. Comment on whether your results are similar or very different. (5 Marks)