ITC 6107: Storing and Retrieving Data

(Winter 2021)

Homework Assignment #2

Due Date: Friday Feb 19, 2021, 23:59.

Exercise 1 (40 points)

Use the data in file BX-Book-Rating.csv to write a Python code in Spark to find

1. The ten most popular books (have been rated most times)
2. The ten books with the highest average rating
3. The average rating of all books
4. The ten users who have rated most books

Each line of the file has the form <user id>;<ISBN>;< rating>.

Exercise #2 (60 points)

Use the file url_pages.txt file that contains pages with links to other pages, with the following format:

<url_number> <list of neighbor_number>

Write Python/Spark code to implement the PageRank algorithm. Use as number of iterations a value n > 5 and to produce a file url_ranks.txt with the following format

<url_number> <rank_value0> <rank_value1> … <rank_vaue(n-1)>

Each line should contain successive ranks the algorithm computes for the respective page. Each rank should be formatted with two decimal digits.

Write a Python code that reads the file url_ranks.txt and plots the successive ranks for a given page.