Big Data 2019 17CS313

Assignment 2

Ranking of Players using PageRank on Spark

1. Assignment Objectives and Outcomes

a. The objective of this assignment is for the students to install Spark and to become familiar with the PySpark libraries and the Spark programming paradigm.

   For this assignment, students will have to implement PageRank on Spark using PySpark and understand the iterative computation of PageRank. The objective of the analysis is to rank IPL players according to their batting and bowling prowess.

b. At the end of this assignment, the student will be able to write and debug PySpark code.

2. Ethical practices

Please submit original code only. You can discuss your approach with your friends but you must write original code. All solutions must be submitted through the portal. We will perform a plagiarism check on the code.

3. Dataset:

a. Use the datasets given at

   i. Bowler Rank: Dataset

   ii. Batsman Rank: Dataset

b. The Bowler rank dataset has the following columns

   i. Batsman,

   ii. Bowler,

   iii. Number of wickets the batsman has lost to bowler,

   iv. Number of deliveries of the bowler, the batsman has faced

c. The Batsman rank dataset has the following columns

   i. Bowler,

   ii. Batsman,

   iii. Number of runs the batsman has scored against this bowler,

   iv. Number of deliveries of the bowler, the batsman has faced

4. Software/Languages to be used:

a. Python (Version **3.5.2**) with PySpark

b. Spark (Version 2.4.4)

c. Please make sure to use Hadoop Version **3.2.0** with Java version 1.8.0 only.

5. Marks:

a. The assignment is for 5 marks.

b. Each task is for 2 marks.

6. Submission Date:

a. 11th October, 2019

7. Tasks:

a. Set up Spark environment by referring to Spark Install

b. These two tasks are to be completed, each of which carry 2 marks :

   i. Ranking bowlers by their bowling prowess.

   ii. Ranking batsmen by their batting prowess

c. Take note of the number of iterations pagerank runs for in both cases but the final code you submit should not print it.

d. Your code should accept 2 command line parameters (both will only be integer inputs):

   i. Number of iterations PageRank should run for. If 0 then your code should run till convergence.

ii. Weight of ranks - the weight in percentage that each iteration of PageRank will attribute to the calculated ranks while 100 - x percentage will be attributed to the default rank - 1. If this is 0 then give 80% weightage to the calculated rank and 20% to default rank (1).

e. Submit one-page report based on the template and answer the questions in the report.

8. Task Specifications:

a. Task 1:

i. Problem Statement:

Use page rank to rank players according to their bowling prowess

ii. Description:

Write python code using PySpark library to rank players as bowlers using PageRank on Spark. Use the bowling rank dataset given and print out the players and their ranks in descending order of rank. For PageRank consider each player to be a page and initial rank of each player as sum of their bowling average against all players and the links should be from the batsman to the bowler for every pair.

iii. Comments:

1. Sort the output in **descending** order of rank of each player.
2. In the case of 2 players having the same rank sort the ones having the same rank in lexicographical order of their name.
3. Initial ranks of the players should be equal to sum of their bowling average against all players or 1, whichever is greater.
    a. Bowling average = number of wickets taken / number of deliveries, for each bowler-batsman pair
4. Page rank should run till convergence with precision of 4 decimal places ie. your algorithm should converge when the difference between ranks in consecutive iterations for each player is less than 0.0001
5. Page rank should be printed till 12 decimal places.
6. Make sure the output includes all players contained in the input file.
7. Output should be printed in the following format:
    a. Player Name, Rank

    Eg. SL Malinga, 3.555903992693

b. Task 2:

i. Problem Statement:

Use page rank to rank players according to their batting prowess

ii. Description:

Write python code using PySpark library to rank players as batsmen using PageRank on Spark. Use the batting rank dataset given and print out the players and their ranks in descending order of rank. For PageRank consider each player to be a page and initial rank of each player as sum of their batting average against all players and the links should be from the bowler to the batsman for every pair.

iii. Comments:

1. Sort the output in **descending** order of rank of each player.
2. In the case of 2 players having the same rank, sort the ones having the same rank in lexicographical order of their name.
3. Initial ranks of the players should be equal to sum on their batting average against all players or 1, whichever is greater.
    a. Batting average = number of runs scored / number of deliveries, for each batsman-bowler pair
4. Page rank should run till convergence with precision of 4 decimal places ie. your algorithm should converge when the difference between ranks in consecutive iterations for each player is less than 0.0001
5. Page rank should be printed till 12 decimal places.
6. Make sure the output includes all players contained in the input file.
7. Output should be printed in the following format:
    a. Player Name, Rank
    Eg. SK Raina, 1.377116600446

9. Submission Link:

Will be shared with you on Piazza at a later date.