CSCE 4783/5783 Cloud Computing and Security
Programming Assignment 8

1 Assignments

- PageRank: develop a Spark program for calculating the page rank.
  - Use `sc.textFile("input.txt")` to read the input.
    - Each line of the `input.txt` file describes a link in the graph, i.e., source_node space(s) destination_node. For example, "1 2" represents a link from node_1 to node_2. Use the provided `input.txt` as a reference.
  - Use `rdd.saveAsTextFile("output")` to save the output.
    - You may save the result of page rank to multiple files.
    - In each file, each line records the page rank mass of a node. The records in a file need to be in an ascending order of the nodes.
- Requirements:
  - The Spark program needs to take one argument to specify how many iterations the algorithm needs to go through. The default number of iterations is 10, which is specified inside the program.
  - The total pagerank mass is 1.
  - The program needs to deal with the dangling nodes, which have no outgoing links, and the random jump as shown in Equation (1).
    - α: random jump factor, use 0.1 in this programming assignment.
    - $|G|$: total number of nodes in the graph.
    - m: the missing page rank mass due to dangling nodes.
  - Initialize the page rank mass of each node to $\frac{1}{|G|}$.

$$p' = \alpha\left(\frac{1}{|G|}\right) + (1-\alpha)\left(\frac{m}{|G|} + p\right) \tag{1}$$

2 Submission

- Due date: December 11, 2020 @ 11:59 pm.
- Submission
  - Name the program as pagerank.scala and upload it to blackboard before the deadline.