Big Data and Cloud Computing

2020-2021

Assignment 1

The input data file graph.txt (from now on, the dataset) is a part of the social graph of a social network. It is organized in two columns representing user relationships (namely, follower – followee relationships). More specifically:

Column 1 represents a user ID (the follower).

Column 2 represents another user ID (the followee).

Example: The 50th line in the input file contains the pair (2, 534). This means that the user with ID=2 is a follower of the user with ID 534.

Task 1: Implement a MapReduce job that creates a list of followers for each user in the dataset.

Example: the list of followers of user 534 is: [2, 16, 37, 73, 156, 210, 308, 347, 446, 455, 487, 519].

Task 2: Implement a MapReduce job that creates a list of followees for each user in the dataset.

Example: the list of followees of user 534 derives by reading the value of the second column in the lines 97097 – 97187.

Task 3: Implement a MapReduce job that identifies the 100 most followed users in the dataset.

Hint: We are not interested in creating lists of followers here. We just need to count the followers of each user in the Reduce phase. This is called the in-degree of a user. Moreover, a temporary data structure D of fixed 100 positions is required. This structure will be initially filled with the first 100 users that are processed in the Reduce phase. Then, the next users (101, 102, 103…) will replace a user in D, only if their in-degree is greater than the in-degree of the least-followed user in D. Notice that the ideal data structure for D is a min-heap (https://docs.python.org/3/library/heapq.html). However, it is totally acceptable to appropriately use a data structure such as a dictionary, or a list.