CES Data Scientist - PageRank Exam

Youcef KACER (youcef.kacer@telecom-paristech.fr)

28 septembre 2016

Exercise 1 : We expose here after the system of linear equations that solves PageRank scores associated to graph G (r_i being PageRank score for page i) :

$$\begin{aligned} r_1 &= r_4 \\ r_2 &= 1/2 * r_1 \\ r_3 &= 1/2 * r_1 + r_2 \\ r_4 &= r_3 \\ 1 &= r_1 + r_2 + r_3 + r_4 \end{aligned}$$

We can solve manually the system :

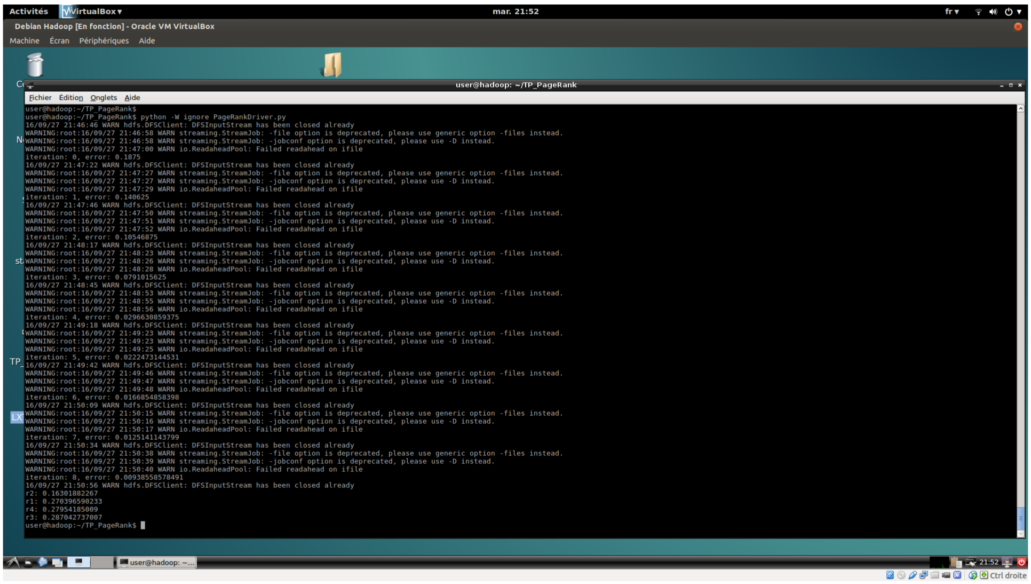
$$\begin{aligned} r_1 &= r_3 \\ r_1 &= r_4 \\ r_2 &= 1/2 * r_1 \\ 1 &= 7/2 * r_1 \end{aligned}$$

$$\begin{aligned} r_1 &= 2/7 \\ r_2 &= 1/7 \\ r_3 &= 2/7 \\ r_4 &= 2/7 \end{aligned}$$

Otherwise, we can solve the system using previous PageRank lab using MapReduce. We just need to adapt *edge_list.txt* to the graph G. In this file, each line represents a page indice and the page indices it links to :

```
1 2 3
2 3
3 4
4 1
```

We have run our python map/reduce implementation of PageRank. We use a teleport coefficient of 0.75, and an error criteria of 0.01. Figure 1 shows correspoding standard output with PageRank scores results at the end.



```
user@hadoop:~/TP_PageRank$
user@hadoop:~/TP_PageRank$ python -W ignore PageRankDriver.py
16/09/27 21:46:46 WARN hdfs.DFSClient: DFSInputStream has been closed already
WARNING:root:16/09/27 21:46:58 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
WARNING:root:16/09/27 21:46:58 WARN streaming.StreamJob: -jobconf option is deprecated, please use -D instead.
WARNING:root:16/09/27 21:47:00 WARN io.ReadaheadPool: Failed readahead on ifile
iteration: 0, error: 0.1875
16/09/27 21:47:22 WARN hdfs.DFSClient: DFSInputStream has been closed already
WARNING:root:16/09/27 21:47:27 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
WARNING:root:16/09/27 21:47:27 WARN streaming.StreamJob: -jobconf option is deprecated, please use -D instead.
WARNING:root:16/09/27 21:47:29 WARN io.ReadaheadPool: Failed readahead on ifile
iteration: 1, error: 0.140625
16/09/27 21:47:46 WARN hdfs.DFSClient: DFSInputStream has been closed already
WARNING:root:16/09/27 21:47:50 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
WARNING:root:16/09/27 21:47:51 WARN streaming.StreamJob: -jobconf option is deprecated, please use -D instead.
WARNING:root:16/09/27 21:47:52 WARN io.ReadaheadPool: Failed readahead on ifile
iteration: 2, error: 0.10546875
16/09/27 21:48:17 WARN hdfs.DFSClient: DFSInputStream has been closed already
WARNING:root:16/09/27 21:48:23 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
WARNING:root:16/09/27 21:48:26 WARN streaming.StreamJob: -jobconf option is deprecated, please use -D instead.
WARNING:root:16/09/27 21:48:28 WARN io.ReadaheadPool: Failed readahead on ifile
iteration: 3, error: 0.0791015625
16/09/27 21:48:45 WARN hdfs.DFSClient: DFSInputStream has been closed already
WARNING:root:16/09/27 21:48:53 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
WARNING:root:16/09/27 21:48:55 WARN streaming.StreamJob: -jobconf option is deprecated, please use -D instead.
WARNING:root:16/09/27 21:48:56 WARN io.ReadaheadPool: Failed readahead on ifile
iteration: 4, error: 0.0296630859375
16/09/27 21:49:18 WARN hdfs.DFSClient: DFSInputStream has been closed already
WARNING:root:16/09/27 21:49:23 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
WARNING:root:16/09/27 21:49:23 WARN streaming.StreamJob: -jobconf option is deprecated, please use -D instead.
WARNING:root:16/09/27 21:49:25 WARN io.ReadaheadPool: Failed readahead on ifile
iteration: 5, error: 0.0222473144531
16/09/27 21:49:42 WARN hdfs.DFSClient: DFSInputStream has been closed already
WARNING:root:16/09/27 21:49:46 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
WARNING:root:16/09/27 21:49:47 WARN streaming.StreamJob: -jobconf option is deprecated, please use -D instead.
WARNING:root:16/09/27 21:49:48 WARN io.ReadaheadPool: Failed readahead on ifile
iteration: 6, error: 0.0166854858398
16/09/27 21:50:09 WARN hdfs.DFSClient: DFSInputStream has been closed already
WARNING:root:16/09/27 21:50:15 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
WARNING:root:16/09/27 21:50:16 WARN streaming.StreamJob: -jobconf option is deprecated, please use -D instead.
WARNING:root:16/09/27 21:50:17 WARN io.ReadaheadPool: Failed readahead on ifile
iteration: 7, error: 0.0125141143799
16/09/27 21:50:34 WARN hdfs.DFSClient: DFSInputStream has been closed already
WARNING:root:16/09/27 21:50:38 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
WARNING:root:16/09/27 21:50:39 WARN streaming.StreamJob: -jobconf option is deprecated, please use -D instead.
WARNING:root:16/09/27 21:50:40 WARN io.ReadaheadPool: Failed readahead on ifile
iteration: 8, error: 0.00938558578491
16/09/27 21:50:56 WARN hdfs.DFSClient: DFSInputStream has been closed already
r2: 0.16301882267
r1: 0.270396590233
r4: 0.27954185009
r3: 0.287042737007
user@hadoop:~/TP_PageRank$
```

FIGURE 1 – Python map/reduce for PageRank scores associated to graph G

Exercise 2 : Suppose we have a file containing integers with theirs indices as follow :

i n
0 256
1 35
2 4122
3 96
...

We present here after a pseudo-code for map/reduce function to find maximum value of this list of integers :

```
map(key=i,value=n)
{
  return 0,n
}

reduce(key,values) // values is an array containing all the integers because
{                  // map has emitted each of them with the same key (0).
  MAX = MIN_POSSIBLE_VALUE
  for n in values
  {
    if n>max
    {
      max = n
    }
  }
  return 0,max
}
```

Exercise 3 :

Exercise 4 :