Data Mining

Apache Giraph Tutorial

Due: xx/12/2016, 23:59

Instructions - Read carefully!

Handing in: You must hand in the homeworks by the due date and time by compressing your answer in a `.zip` file and uploadind it using the following website.

`https://goo.gl/aGwA2m`

The website will ask you to providing an email address to which your delivery will be acked and you will recieve any eventual further comunication.

The solutions must contain the source code and the output generated (to the screen or to files).

You can checkout this repository for a starting base code:

`https://github.com/manuelcoppotelli/giraph-homework.git`

Note that you will be provided with an example input graph but your code will be checked with an unprovided more complex graph.

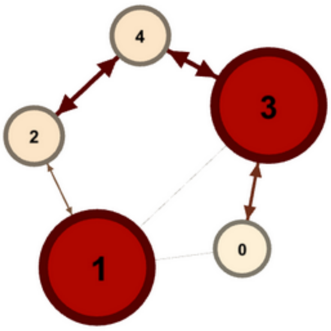
For any other information feel free to contact us.

Task. Imagine that you have a social graph, and you want to find out whom you should ask to introduce you to tennis superstar Roger Federer. If none of your acquaintances knows Roger, chances are they may know somebody who knows him. Or they may know somebody who knows somebody who knows Roger, and so on. Basically, you are looking for a path of "friend of a friend" relationships that allows you to reach Roger.

Often, multiple paths exist between two vertices, and you usually care about the shortest one. In the case of a weighted graph, edge weights represent distances between neighbors. Hence, the length of a path is computed as the sum of the weights of the edges that constitute the path.

You are requested to write a `Giraph Job` aimed to compute the shortest distance from a **fixed** source vertex to any other verteces.

An example of input graph could be:



```
[0,0,[[1,1],[3,3]]]
[1,0,[[0,1],[2,2],[3,1]]]
[2,0,[[1,2],[4,4]]]
[3,0,[[0,3],[1,1],[4,4]]]
[4,0,[[3,4],[2,4]]]
```

The output expected for the above input (starting from vertex 1) is:

```
0    1.0
2    2.0
1    0.0
3    1.0
4    5.0
```

Note: in the above example we've used
`org.apache.giraph.io.formats.JsonLongDoubleFloatDoubleVertexInputFormat`
as input format and
`org.apache.giraph.io.formats.IdWithValueTextOutputFormat`
as output format.