Description

Xiaoyu Zheng

1. Directory structure

In Solution directories, there are three folders, which contain the python one and the scala one separately, and also include the bonus folder. In the python folder, there are two python files, while in the scala folder, there are two scala files and an executable jar. In OutputFiles folder, there also exist three folders for python and scala and the bonus correspondingly, and first two folders contain two output files.

2. Versions

Spark: 2.2.1 Python: 2.7.15 Hadoop: 2.7 Scala: 2.11.8 SBT: 0.13.8

3. Execution rules

1) Put my **executable files** (including all python files and the scala jar file) directly **in the root directory of Spark**.
2) To utilize my work, please **get into the root directory of spark** with terminal or in the command line.
3) Then to execute my work, you need the following commands. Note: These commands will overwrite the files which have the same name as my output files in the root directory of Spark.

   Task1 (Betweenness):
   Python:
   bin/spark-submit Xiaoyu_Zheng_Betweenness.py <rating file path>
   Scala:
   bin/spark-submit --class Betweenness Xiaoyu_Zheng_hw4.jar <rating file path>

   Task2 (Community):
   Python:
   bin/spark-submit Xiaoyu_Zheng_Community.py <rating file path>
   Scala:
   bin/spark-submit --class Community Xiaoyu_Zheng_hw4.jar <rating file path

   Bonus (Community):
   Python:
   bin/spark-submit Xiaoyu_Zheng_bonus.py <rating file path>

4) The output txt files will also be in the root directory of Spark.

4. Bonus

I use the community function of the networkx python package to detect the community. The method applied is called “girvan_newman”, which is the same as we use in this assignment. However, its speed performance is not very good, so I just go to the third level of the communities tree (which only has four communities).