Description

Xiaoyu Zheng

1. Directory structure

In both OutputFiles and Solution directories, there are 2 folders, which contain the python one and the scala one separately. Each folder among python and scala folders includes task1 and task2 directories, to store the corresponding materials to each task. In each task folder, there are 2 folders related to 2 sets of data.

2. Execution rules

To utilize my work, firstly, you should get into the root directory of spark with terminal or in the command line.

Then to execute my work, you need the following commands. Note: These commands will overwrite the folders which have the same path as the output path.

1) If you put my executable files (like Xiaoyu_Zheng_task1.py, or Xiaoyu_Zheng_task1.jar) directly in the root directory of Spark, then:

(1) Task1:

Python:

```
bin/spark-submit Xiaoyu_Zheng_task1.py <input path> <output path>
```

Scala:

```
bin/spark-submit --class Task1 Xiaoyu_Zheng_hw1_task1.jar <input path> <output path>
```

(2) Task2

Python:

```
bin/spark-submit Xiaoyu_Zheng_task2.py <rating path> <tag path> <output path>
```

Scala:

```
bin/spark-submit --class Task1 Xiaoyu_Zheng_hw1_task2.jar <rating path> <tag path> <output path>
```

2) **Or** you can just put the path to these executable files into the commands:

(1) Task1:

Python:

```
bin/spark-submit <path to Xiaoyu_Zheng_task1.py> <input path> <output path>
```

Scala:

```
bin/spark-submit --class Task1 <path to Xiaoyu_Zheng_hw1_task1.jar> <input path> <output path>
```

(2) Task2

Python:

```
bin/spark-submit <path to Xiaoyu_Zheng_task2.py> <rating path> <tag path> <output path>
```

Scala:

```
bin/spark-submit --class Task1 <path to Xiaoyu_Zheng_hw1_task2.jar> <rating path> <tag path> <output path>
```

3. Output files

For I use dataFrame rather than RDD in this assignment, the output path can only specify the names of the result folders rather than the names of the csv files inside. In the result folders created by python codes, there is only one single csv file, while in the result folders created by scala codes, there are two files: one is the csv file, the other is the _SUCCESS file which indicates the execution status result.