INF 553 – Fall 2017 Assignment 1

Overview of the assignment

In this assignment, students will complete two tasks. The goal of these two tasks is to let students get familiar with Spark and do data analysis using Spark. In the assignment description, the first part is about how to configure the environment and data sets, the second part describes the two tasks in details, and the third part is about the files the students should submit and the grading criteria.

Spark Installation

Spark can be downloaded from the official website:

http://spark.apache.org/downloads.html

Spark 1.6.1 combined with Hadoop 2.4 is recommended. The interface of Spark official website is shown in the following figure.

Scala Installation

Please refer to the Spark slides

Python Configuration

You need to add the paths of your Spark (path/to/your/Spark) and Python (path/to/your/Spark/python) folders to the interpreter's environment variables named as SPARK_HOME and PYTHONPATH, respectively.

Data

Please download the data from MovieLen over the following link:
https://grouplens.org/datasets/movielens/
You are required to download data sets. It is ml-1m.zip, which size is 6 MB. The zip file contains three dat files and one readme file. The files users*.dat*, *ratings.dat* and movies.dat are needed for the tasks. The description of the data is provided in the README file.

Task1: (40%)

Students are required to calculate each movie's average rating based on gender of the user. The ratings.dat and users.dat file are needed for this task.

Result format:

1. Save the result as one text file.
2. The result is ordering by *movieId, gender* in ascending order
3. The result file includes three columns *movieId, gender, avg. ratings.*

The following snapshot is an example of result for task 1. It shows the exact format of the result.

```
1,F,4.18781725888
1,M,4.13055181696
2,F,3.27840909091
2,M,3.17523809524
3,F,3.07352941176
3,M,2.99415204678
4,F,2.97647058824
4,M,2.48235294118
5,F,3.21296296296
5,M,2.88829787234
6,F,3.68217054264
6,M,3.90998766954
7,F,3.58823529412
7,M,3.26771653543
8,F,3.35714285714
8,M,2.775
9,F,2.1
9,M,2.71739130435
10,F,3.47014925373
10,M,3.55305039788
```

Task2: (60%)

Students are required to calculate the average rating of each movie genres based on the gender of the user. The *ratings.dat*, *movies.dat and users.dat* files are required for this task.

Result format:

1. Save the result as one text file.
2. There are three columns in the result file. The first column is the genres's name. the second column is the gender and the third column is the avg. ratings. Also, the file should be sorted according to the genres' name in ascending order.

The following snapshots is an example of result for task 2. It shows the exact format of the result.

```
Action,F,3.36747361887
Action,M,3.35299065421
Action|Adventure,F,3.70121334681
Action|Adventure,M,3.67111478507
Action|Adventure|Animation,F,3.84375
Action|Adventure|Animation,M,4.21708185053
Action|Adventure|Animation|Children's|Fantasy,F,3.14634146341
Action|Adventure|Animation|Children's|Fantasy,M,2.51063829787
Action|Adventure|Animation|Horror|Sci-Fi,F,3.42253521127
Action|Adventure|Animation|Horror|Sci-Fi,M,3.56307129799
Action|Adventure|Children's,F,1.25
Action|Adventure|Children's,M,1.325
Action|Adventure|Children's|Comedy,F,2.44715447154
Action|Adventure|Children's|Comedy,M,2.26329113924
Action|Adventure|Children's|Fantasy,F,1.85714285714
Action|Adventure|Children's|Fantasy,M,2.13513513514
Action|Adventure|Children's|Sci-Fi,F,2.16363636364
Action|Adventure|Children's|Sci-Fi,M,1.82033898305
Action|Adventure|Comedy,F,3.14087759815
Action|Adventure|Comedy,M,3.08333333333
Action|Adventure|Comedy|Crime,F,3.17510548523
Action|Adventure|Comedy|Crime,M,3.12863268223
Action|Adventure|Comedy|Horror,F,3.2972972973
Action|Adventure|Comedy|Horror,M,3.90928270042
Action|Adventure|Comedy|Horror|Sci-Fi,F,3.6015625
Action|Adventure|Comedy|Horror|Sci-Fi,M,3.83598531212
Action|Adventure|Comedy|Romance,F,3.97222222222
Action|Adventure|Comedy|Romance,M,3.82441314554
Action|Adventure|Comedy|Sci-Fi,F,3.81784386617
```

What you need to turn in:

1. Source codes for two tasks (you can use either Python or Scala) and name it as *Firstname_Lastname_task1* and *Firstname_Lastname_task2,* respectively. (For example, Priyambada_Jain_task1.py)

2. Result files of two tasks for large and small data sets and name it as *Firstname_Lastname_result_task1.txt,* *Firstname_Lastname_result_task2.txt*

3. Readme documents: please describe how to run your program in this document.

4. If you use Scala, please submit the jar package as well and name them as *Firstname_Lastname_task1.jar* and *Firstname_Lastname_task2.jar.*

5. Zip the above files and name it as *Firstname_Lastname_HW1.zip*

Grading Criteria:

1. Your codes will be run according to your Readme file. If your programs cannot be run with the commands you provide, your submission will be graded based on the result files you submit and **20%** penalty for it.
2. If the file generated by your program is unsorted, there will be **20%** penalty.
3. If your program generates more than one file, there will be **20%** penalty.
4. The deadline for assignment 1 is 09/20 midnight. There will be **20%** penalty for late submission.
5. Also, as described for Scala implementation **10% bonus will be awarded.**