INF 553 – Spring 2017 Assignment 1

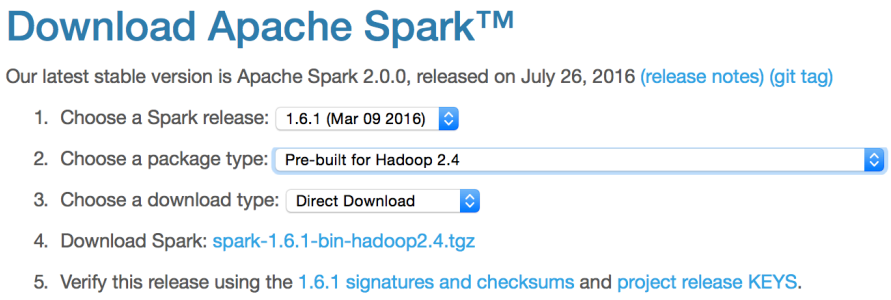
Overview of the assignment

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 two data sets. The first is ml-20m.zip, which size is 190MB, the second is ml-latest-small.zip, which size is 1MB. Each zip file contains five CSV files. The files *tags.csv* and *ratings.csv* are needed for the tasks.

Task1: (40%)

Students are required to calculate each movie's average rating. The ratings.CSV file is needed for this task.

Result format:

1. Save the result as one text file. There is no requirement about the format of the file
2. The result is ordering by *movieId* in ascending order

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

```
1,3.87246963563
2,3.40186915888
3,3.16101694915
4,2.38461538462
5,3.26785714286
6,3.88461538462
7,3.28301886792
8,3.8
9,3.15
10,3.45081967213
```

Task2: (60%)

Students are required to calculate the average rating of each tag. Both the *rating.csv* and *tags.csv* files are required for this task.

Result format:

1. Students are required to save the result in a CSV file
2. There are two columns in the CSV file. The first column is the tag's name, which should be named as *tag*. The second column is the rating, which should be named as *rating_avg*. And the file should be sorted according to the tags' name in descending order

The following two snapshots is an example of result for task 2. The unreadable codes in the first snapshots are because encoding problem. It just shows the format of the result. In the second picture, the data is sorted by first column in descending order.

```
tag,rating_avg
é˜®ä¸€é¸£,3.92123956132
ç»å…¸,4.02900018135
æš´åŠ›,4.17423116922
æµ‹è¯•,3.46666666667
å¥½äºº,3.58919902913
ä½Žä¿—å°è¯´,4.17423116922
Ø§ØØ³Ø§Ø³Ø§ØªÙŠ,2.94611375134
Özgür Yildirim,3.16666666667
Özer Kiziltan,3.82692307692
```

```
zeichnungen als übergang,2.81337103615
zegist,3.70224719101
zef,3.34615384615
zebras,3.36527208894
zebra,3.36527208894
zealots,3.37777777778
zany,2.10344827586
yuppies,3.50327998511
yukon,3.61274509804
yuen woo-ping,3.64606741573
yuen chor,4.0
```

Hints for Task2:

1. Unicode problem: you may encounter problems of text encodings when saving the result as a CSV file. You should save your file with 'uft-8'.
2. You can create Dataframe objects and save the Dataframe objects as CSV file
3. You can learn more about Dataframe by this link: https://spark.apache.org/docs/1.6.0/sql-programming-guide.html#creating-dataframes

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, Weiwei_Duan_task1.py)

2. Result files of two tasks for large and small data sets and name it as
   Firstname_Lastname_result_task1_big,
   Firstname_Lastname_result_task2_big.csv
   Firstname_Lastname_result_task1_small
   Firstname_Lastname_result_task2_small.csv

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. If the CSV file generated in task 2 has more than two columns, there will be **20%** penalty.
5. If the header of the CSV file is missing in task 2, there will be **10%** penalty
6. The deadline for assignment 1 is 02/07 midnight. There will be **20%** penalty for late submission.