INF 553 – Spring 2018 Assignment 1

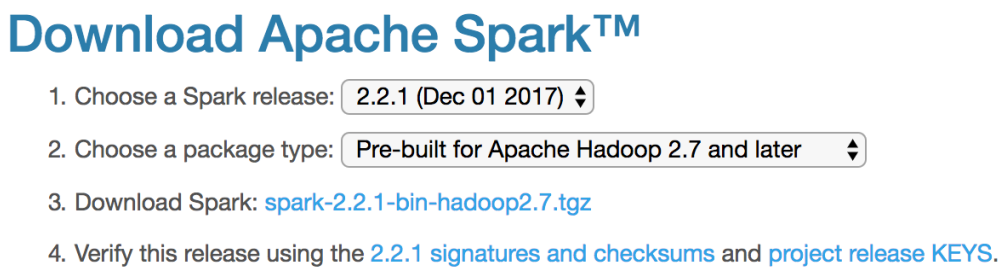
Overview of the assignment

In this assignment, students will complete two tasks. The goal of these two tasks is to let students get familiar with Spark and perform 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 (refer to: link)

Please use Spark 2.2.1 with Hadoop 2.7 for this assignment. The interface of Spark official website is shown in the following figure.



Scala Installation

You can use Intellij if you prefer IDE for creating and debugging projects. And install Scala/SBT plugins for Intellij. You can refer to the tutorial "Setting UP Spark 2.0 environment on intellij community edition".

There will be **10% bonus** if you use Scala for both Task1 and Task2 (*i.e. 10->11; 9->9.9)*.

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.

Environment Requirements

Python: 2.7 Scala: 2.11 Spark: 2.2.1

IMPORTANT: We will use these versions to compile and test your code. If you use other versions, there will be a 20% penalty since we will not be able to grade it automatically.

Data

Please download the Amazon Product data from this link. Detailed introduction of the data can also be found through the link.

You are required to download two data sets. The first is Toys and games (5-core), the second is metadata (3.1gb) – metadata for 9.4 million products. Each zip file contains one JSON file.

Ratings only: These datasets include no metadata or reviews, but only (user,item,rating,timestamp) tuples. Thus they are suitable for use with mymedialite (or similar) packages.

Books	5-core (8,898,041 reviews)	ratings only (22,507,155 ratings)
Electronics	5-core (1,689,188 reviews)	ratings only (7,824,482 ratings)
Movies and TV	5-core (1,697,533 reviews)	ratings only (4,607,047 ratings)
CDs and Vinyl	5-core (1,097,592 reviews)	ratings only (3,749,004 ratings)
Clothing, Shoes and Jewelry	5-core (278,677 reviews)	ratings only (5,748,920 ratings)
Home and Kitchen	5-core (551,682 reviews)	ratings only (4,253,926 ratings)
Kindle Store	5-core (982,619 reviews)	ratings only (3,205,467 ratings)
Sports and Outdoors	5-core (296,337 reviews)	ratings only (3,268,695 ratings)
Cell Phones and Accessories	5-core (194,439 reviews)	ratings only (3,447,249 ratings)
Health and Personal Care	5-core (346,355 reviews)	ratings only (2,982,326 ratings)
Toys and Games	5-core (167,597 reviews)	ratings only (2,252,771 ratings)
Video Games	5-core (231,780 reviews)	ratings only (1,324,753 ratings)
Tools and Home Improvement	5-core (134,476 reviews)	ratings only (1,926,047 ratings)
Beauty	5-core (198,502 reviews)	ratings only (2,023,070 ratings)
Apps for Android	5-core (752,937 reviews)	ratings only (2,638,172 ratings)
Office Products	5-core (53,258 reviews)	ratings only (1,243,186 ratings)
Pet Supplies	5-core (157,836 reviews)	ratings only (1,235,316 ratings)
Automotive	5-core (20,473 reviews)	ratings only (1,373,768 ratings)
Grocery and Gourmet Food	5-core (151,254 reviews)	ratings only (1,297,156 ratings)
Patio, Lawn and Garden	5-core (13,272 reviews)	ratings only (993,490 ratings)
Baby	5-core (160,792 reviews)	ratings only (915,446 ratings)
Digital Music	5-core (64,706 reviews)	ratings only (836,006 ratings)
Musical Instruments	5-core (10,261 reviews)	ratings only (500,176 ratings)
Amazon Instant Video	5-core (37,126 reviews)	ratings only (583,933 ratings)

Metadata

Metadata includes descriptions, price, sales-rank, brand info, and co-purchasing links:

metadata (3.1gb) - metadata for 9.4 million products

Task1: (40%)

Students are required to calculate each product's average rating. The *reviews_Toys_and_Games_5.json* file is needed for this task.

Result format:

1. Save the result as one csv file with header (asin, rating_avg).
2. The result is ordering by *asin (i.e. product id)* in ascending order.

The following snapshot is an example of result for task 1.

```
asin,rating_avg
0439893577,4.352941176470588
048645195X,4.454545454545454
0545496470,3.6666666666666665
0615444172,5.0
0670010936,4.4
0735308365,4.25
0735321396,4.285714285714286
073533305X,5.0
0735333483,4.4
073533417X,4.416666666666667
074242720X,4.2
0786950072,3.6666666666666665
0786955570,4.666666666666667
0786955708,3.8181818181818183
```

Task2: (60%)

Students are required to calculate the average rating of each brand for toys and games products. Both the *reviews_Toys_and_Games_5.json* and *metadata.json* 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 brand's name, which should be named as *brand*. The second column is the overall rating, which should be named as *rating_avg*. And the file should be sorted by the brand's name in alphabetical order.
3. Please remove the null or empty string of brand in your result.

The following snapshot is an example of result for task 2.

```
brand,rating_avg
2 in 1 Space Rocket,4.777777777777778
3M,3.875
4M,4.320388349514563
4Thought Products LLC,4.142857142857143
5K,3.7
A Birthday Place,4.769230769230769
AA,4.555555555555555
AB Gee,4.7272727272727275
ACD Distribution,4.5625
```

Hints for Task2:

1. You can create Dataframe objects and save the Dataframe objects as CSV file.
2. You can learn more about Dataframe by this link:

https://spark.apache.org/docs/2.2.1/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, John_Smith_task1.py)
2. Result files of two tasks for large and small data sets and name it as
   Firstname_Lastname_result_task1.csv
   Firstname_Lastname_result_task2.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 does not use the required Scala/Python/Spark versions, there will be **20%** penalty.
4. If your program generates more than one file, there will be **20%** penalty.
5. If the CSV file generated has more than two columns, there will be **20%** penalty.
6. If the header of the CSV file is missing, there will be **10%** penalty.
7. The deadline for assignment 1 is 02/05 midnight. There will be **20%** penalty for late submission within a week and 0 grade after a week.
8. You can use your free 5-day extension.