CS5344 Lab 1

AY2022/2023 Semester 1

The purpose of this lab is to get you started with Spark, and learn how to write, compile, debug and execute a simple Spark program. You will also be tasked to write and submit your own Spark program individually.

1. **Reference programs and documentation** from Spark release are available at
   https://www.tutorialspoint.com/apache_spark/apache_spark_quick_guide.htm
   https://spark.apache.org/docs/2.2.0/rdd-programming-guide.html

2. A VirtualBox image of Ubuntu with Spark deployment has been configured for you. **Appendix A** gives the instructions on how to download and install it.

3. Alternatively, you can learn to install a stand-alone Spark-2.2.1 instance on Ubuntu by yourself and set up the environment by following the instructions in **Appendix B.**
   For **Mac users,** you can refer to
   https://medium.com/luckspark/installing-spark-2-3-0-on-macos-high-sierra-276a127b8b85
   https://notadatascientist.com/install-spark-on-macos/

4. **Appendix C** is the basic guide to help you get started and run your first Spark *WordCount* program in **Python 3.6**. Remember to change the python driver for PySpark. You can refer to https://stackoverflow.com/questions/30518362/how-do-i-set-the-drivers-python-version-in-spark.

5. **Appendix D** are some problems that the students had last year. Please read it carefully.

6. If you want to **debug with PyCharm**, you can link PyCharm with Spark according to the instruction in
   https://stackoverflow.com/questions/34685905/how-to-link-pycharm-with-pyspark.

   To install PyCharm and run your first project you can refer to
   https://www.jetbrains.com/help/pycharm/installation-guide.html.

Task: Write a Spark program to find the top 10 products based on the number of user reviews and report their average ratings and product price.

Datasets:
Use the Musical Instruments review file (reviews_Musical_Instruments. json) and metadata (meta_Musical_Instruments.json) from the Amazon product dataset **(**http://jmcauley.ucsd.edu/data/amazon/links.html**).** Download both files from the "Per-category files" section.

Algorithm:

Step 1. Find the number of reviews and calculate the average rating for each product from the review file. Use pair RDD, reduceByKey and map function to accomplish this step. The key is the product ID/asin. The value is a tuple (#reviews, average_rating).

Step 2. Create an RDD where the key is the product ID/asin and value is the price of the product. Use the metadata for this step.

Step 3. Join the pair RDD obtained in Step 1 and the RDD created in Step 2.

Step 4. Find the top 10 products with the greatest number of reviews.

Step 5. Output the number of reviews, average rating and price for the top 10 products identified in Step 4.

Input: Review file and metadata.
Output: One line per product in the following format:

<product ID> <the number of reviews> <average rating> <product price>

Notices:

1. As there is none price for some products, we only rank the products with real price at last.

Deliverables: Zip the **Spark program with documentation for important steps** in the code along with **the output file** and upload it to Lab1 folder. The zipped folder should be named as follows, Student ID_Lab1.

Important Notes:

(a) Please specify the python version and the packages used.

(b) Your code should be executable either on the virtual machine configuration given below or on stand-alone Spark configuration

References:

- https://spark.apache.org/docs/2.2.0/rdd-programming-guide.html#transformations
- https://spark.apache.org/examples.html