Homework 1

Task

In this homework, you will run Spark on your local machine and recommend movies to users by predicting movie ratings. The purpose of this part of the assignment is to get everything working before adding the complexities of running on many machines. This assignment refers to Berkeley CS 194-16 homework 3 part1 [1].

Code and data

- All code is included in hw1-part1.ipynb.
- The small version of movielens dataset used in this assignment is part1files.tar.gz.

Dependencies

1. Spark
2. Java
3. Python3
4. pyspark
5. Jupyter Notebook

Install Spark

You can install Spark on Linux/OSX/Windows. There are a number of installation guides on the Web, such as [2]. Please make sure you install Java before you install Spark.

Start Jupyter Notebook

Jupyter Notebook App (formerly IPython Notebook) is an application running inside the browser. If you are unfamiliar with Jupyter Notebook, you can refer to [3].

Basic Spark operations with Python

You can learn some basic Spark operations here [4]. For further usage, look at Spark official API documentations.

Submitting your work

Please submit your code and answers (namely hw1-part1.ipynb) with **output blocks** to learn.tsinghua.edu.cn. Rename hw1-part1.ipynb as hw1-part1-<your_student_id>.ipynb, such as hw1-part1-2016011000.ipynb.

References:

[1] https://nbviewer.jupyter.org/github/amplab/datascience-sp14/blob/master/hw3/hw3-part1.ipynb
[2] https://medium.com/@dvainrub/how-to-install-apache-spark-2-x-in-your-pc-e2047246ffc3
[3] https://jupyter-notebook-beginner-guide.readthedocs.io/en/latest/

[4] https://spark.apache.org/docs/latest/quick-start.html