INF553 Summer 2018 Assignment 1
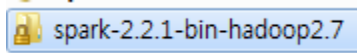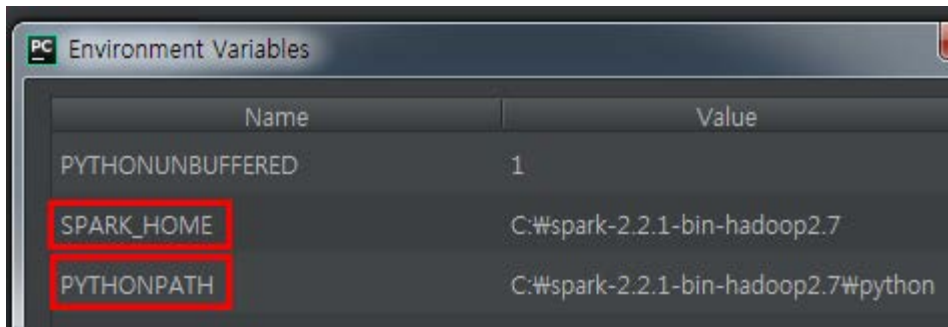
Name: Ye Joo Park (USC# 1128-6851-51)

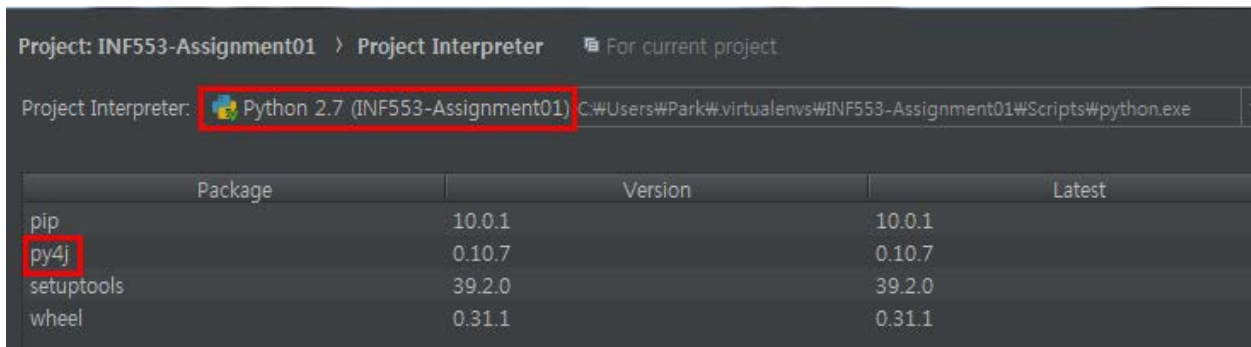## Running the python code

I've used pycharm to run both tasks.

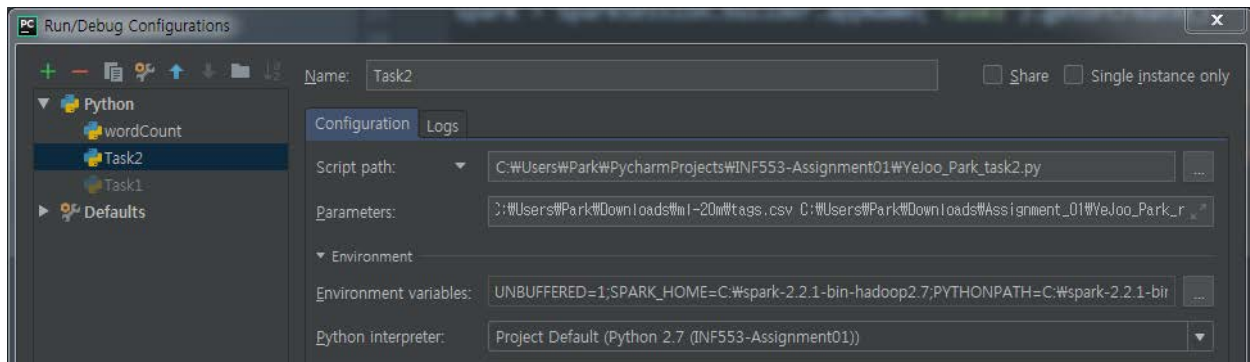- Untar spark-2.2.1 to local drive and add to PATH.

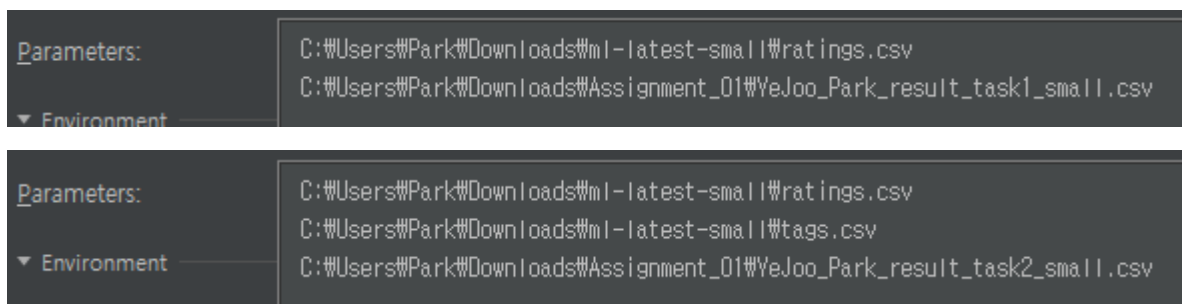- Add SPARK_HOME and PYTHONPATH environment variables in PyCharm.

- Set Python 2.7 interpreter in Settings -> Project -> Project Intrepreter

- In Run/Debug Configurations, to set environment variables and parameters (arguments when running the program)

- Parameters for Run Configuration (for Task1 / Task2)

Running in command line

Task 1

```
$ spark-submit YeJoo_Park_task1.py /path/to/ratings.csv /path/to/output.csv
```
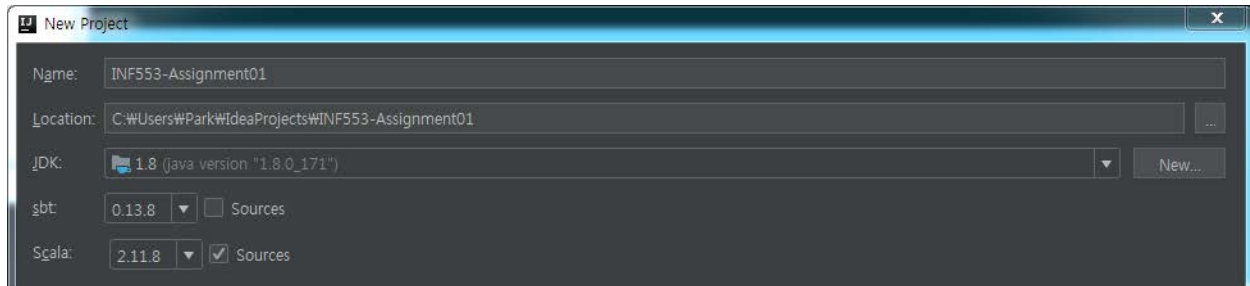
Task 2

```
$ spark-submit YeJoo_Park_task2.py /path/to/ratings.csv /path/to/tags.py /path/to/output.csv
```
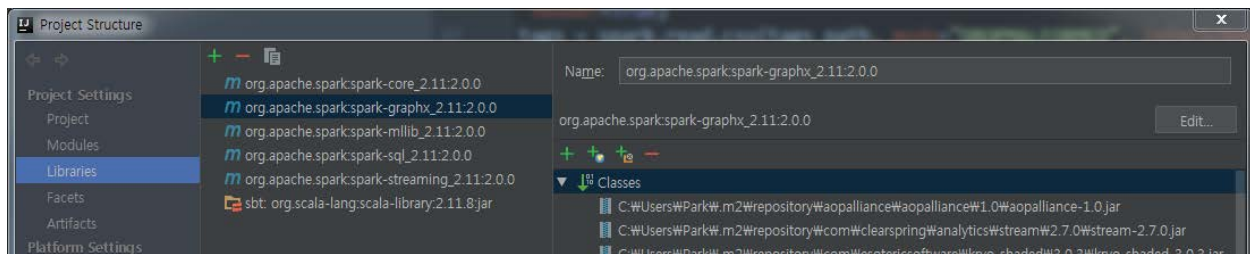
Running the Scala code

Again, I've used intellij IDEA.

- Set sbt to 0.13.8 and Scala to 2.11.8 when creating a Scala project.

- Make sure that the libraries listed in "Setting up spark 2.0 with intellij community edition.pdf" are added.

Running in command line

Task 1

```
$ spark-submit --class Task1 YeJoo_Park_hw1.jar /path/to/ratings.csv
/path/to/output.csv
```

Task 2

```
$ spark-submit --class Task2 YeJoo_Park_hw1.jar /path/to/ratings.csv
/path/to/tags.py /path/to/output.csv
```