School of Computer Science

Self-Test Homework
Tuesday March, 1, 2022
MapReduce Simulator

Homework Task

The self test homework asks you to finish a small MapReduce simulation program that scans a given input file (`data.csv`) and calls two functions to filter and aggregate the data found in there. It tests your Python coding ability, in particular, if you are comfortable writing code under a given framework.

Obtaining the source code

We use university Github to release lab code, you may need to activate your university Github account by login to it from https://github.sydney.edu.au

The homework can be found from the following repository: https://github.sydney.edu.au/COMP5349-Cloud-Computing-2021/python-resources

Format of the input data

In the downloaded archive, you also find an example data set (**`data.csv`**) in CSV (comma separated values) format that contains 100,000 user ratings for films. Users and films are given as numeric IDs. The file format is:

user_id \t film_id \t rating \t timestamp

Note: Be careful parsing this file – there are a few lines which are incomplete, and also make sure that user and film ids are integer values, and ratings should be integers in the range 1 to 5.

Implement the missing RatingFilter and RatingReducer classes

Your homework task is to implement two missing classes: **RatingFilter** and **RatingReducer**, so that the final program determines the average film rating (as float value rounded to 1 decimal place) for those films whose ID is in a given range. You will find the skeleton source code of those classes in the within the respective python files.

RatingFilter

The RatingFilter class is a specialization of the generic Filter class. You have to implement an initialiser that allows to specify the search range of film ids (min and max film_id), as well as a filter() method.
The filter method is called for each individual line of the input file. Each string, which you process in this method, corresponds to the format specified in Section 1.2. The filter method takes a String as input and either returns a tuple in the form of (key, value) as result, if the given input string is about a film id in the search range, or *None* otherwise. The tuple should consist of the film id as key, and the rating for the film in the current input line as value. The structure and methods of the class are defined in the same file.

RatingReducer

The RatingReducer class is a specialization of the generic Reducer class. You have to implement a reduce method which computes the average rating (as Float) of all the given input ratings of the same film. The reduce() method is called by our map_reduce_simulator with a String key and a List of Integer values. For our example data set, this will be the filtered film_id (key) and the list of integer ratings given by various users to this film. The reduce() method shall compute the average rating of all those input ratings and return as result the average rating as a Float value rounded to one decimal place (eg. 3.5).

here are two ways of running the Python program. You may use the notebook version **homework_week1.ipynb**, or the python script version: **map_reduce_simulator.py**.

1.3.1 Python: Running the Python program and example output

The script version can be executed as follows:

```
python3 map_reduce_simulator.py
```

Optionally, you can provide an input file name and different film ids as command line parameters. The program takes three (optional) parameters: an input file path and the range (start and end) of film ids to search for:

```
python3 map_reduce_simulator.py  data.csv  1  2
```

If you code is correct, it should produce the following output for films 1 to 2:

```
1: 3.9
2: 3.2
MapReduce Simulator using mapper:  <mapper.RatingFilter object at … >
MapReduce Simulator using reducer: <reducer.RatingReducer object at … >
Lines in File:    100000 records (should be: 100000)
Filtered records: 582 (should be: 582)
```