Advanced Java – Fall 2021

Assignment 1

Deadline

Due on 23/01/2022

Introduction

In this assignment you will do some basic statistical processing on Covid dataset. This dataset is created and maintained Mathieu, Ritchie, Ortiz-Ospina et al. (Mathieu, E., Ritchie, H., Ortiz-Ospina, E. et al. A global database of COVID-19 vaccinations. Nat Hum Behav (2021) who have kindly open sourced it. Before starting you work on this assignment, you must familiarize first with the dataset, available at https://covid.ourworldindata.org/data/owid-covid-data.csv. Your application should read this data into an appropriate data structure and then process it according to command line parameters. You may work in pairs.

Data

The dataset is being updated daily, and therefore the amount data will certainly change during development period and later during demo/testing time. However you may assume that the structure of the data will not change. The CSV file contains a header line, followed by (as of now) 151K lines of data. Each data line contains COVID information about a particular country on a particular day. While the CSV file contains 60+ columns, you will only focus on the following

- `iso_code (COD)`
- `continent (CNT)`
- `location (LOC)`
- `date (DT)`
- `total_cases (TC)`
- `new_cases (NC)`
- `new_cases_smoothed (NCS)`
- `total_deaths (TD)`
- `new_deaths (ND)`
- `new_deaths_smoothed (NDS)`
- `reproduction_rate (RR)`
- `new_tests (NT)`
- `total_tests (TT)`
- `stringency_index (SI)`
- `population (POP)`
- `median_age (MA)`

The text in parenthesis indicates the code of the field, which will be used in the command line parameters.

Statistics

The statistics that you are going to process are based on the above fields, as well as one additional "computed" field, `new_deaths_per_case (NDPC)`, and it is equal to `new_deaths / new_cases`. The statistics will be of either minimum or maximum type, that is you will be computing the minimum (or maximum) of some quantity, which will be defined shortly. Moreover a limit quantity will be associated with the statistics, say 5, in which case you will be computing the minimum 5 or maximum 5

data items according to given parameter. Along the statistics you will also be given the field in question (via its code in parenthesis above) for which you want to process the given parameter. Using these inputs, one should be able to ask questions like

Command Line Arguments

The command line interface will have the format

```
-file pathToFile -param1 value1 -param2 value2 … -paramN valueN.
```

The parameters (param1 .. paramN) can be `stat`, `limit`, `by` and `display`. We now define the values that of these parameters may assume.

- `file`: path to covid data file
- `stat`: the statistics to be computed, can assume value either min or max
- `limit`: integer in the range [1..100]. In conjunction with the stat value, it indicates how many data items will be displayed, e.g. top 5 (in case of max), bottom 10 (in case of min) etc.
- `by`: the field on which the statistics will be computed. It can have value NC, NCS, ND, NDS, NT, NDPC
- display: The type of data to be displayed. Values can be DATE, COUNTRY, CONTINENT.

Examples

- -file D:/data/owid-covid-data.csv -stat max -limit 5 -by NC -display DATE: top 5 dates with most number of new cases
- -file D:/data/owid-covid-data.csv -stat min -limit 10 -by NC -display COUNTRY: bottom 5 countries with smallest number of new cases

Non-functional requirement

Your program must ideally not contain any loops (for, while, do-while) at all. For iteration needs use streams. Also keep in mind that streams use lambdas and lambdas must be kept short and easy to read.

Data Structure

After reading the data from the input file, you must organize it into an efficient data structure of appropriate entities

- efficient data structure: must support efficient implementation of the functionality outlined above
- appropriate entities: must satisfy data integrity principle as pertaining to 3NF (3rd Normal Form)

Deliverables

You need to submit

- assignment report, detailing your solution, specific implementation and any particularities associated with your program.
- Github repository for your assignment