PV260 Software Quality

Assignment - Initial evaluation

Spring 2021

1 General Information

1.1 Dates

- Assignment start: 4.3.2021
- Assignment deadline: 18.3.2021 16:00

1.2 Submission

The complete solution should be submitted to the Homework vault (`https://is.muni.cz/auth/el/fi/jaro2021/PV260/ode/java_lasaris_initial_assignment/`) in the Information System.

You can submit either txt file containing a link to a remote repository (e.g. github, gitlab, etc.) if you used one or a zip file with all the sources if you didn't use any kind of remote repository.

You should also provide everything necessary to compile and run the solution including instructions on how to execute your solution (e.g. sample command line options), documentation etc.

2 Project

Imagine you want to establish a start-up company focusing on data mining and analysis of various kinds of data. First, you want to develop a prototype of a software tool that would enable automation of the data analysis process. The data analysis process consists of following steps:

1. Parsing of the input options
2. Retrieval of a dataset from repository.
3. Data filtering & analysis
4. Output of results

The prototype is going to have only a limited functionality now but it will serve as a **basis for future development**. It will be essential for your company business, therefore **the quality of architecture and the code is already very important**.

The prototype will focus on analysis of simple datasets containing information about eshop orders. For the individual process steps, the prototype should support following features:

1. The tool will be executed from CMD with following parameters:
   `java -jar da-tool.jar -d [DATASET_LOCATION] -m [MANIPULATION_METHODS] -o [OUTPUT_TYPE] [OUTPUT_FILE]`

   - `DATASET LOCATION` will be a path to remote or local CSV file.

   - `MANIPULATION_METHODS` will contain list of methods names that are going to be applied to the data. The list can contain both filtering and analytical methods. The methods are executed in the provided order.

   - `OUTPUT_TYPE` specifies type of output. The prototype should support either JSON, XML or plain text with corresponding values for this argument - `'json'`, `'xml'`, `'plain'`.

   - `OUTPUT_FILE` is the path of the output file.

2. The dataset will be represented by a single csv file with the following structure:

   `order_id,order_date,customer_email,customer_address,total_price,order_status`

   A sample dataset is available at:

   `https://is.muni.cz/auth/el/fi/jaro2021/PV260/um/seminars/java_groups/`
   `initial_assignment/orders_data.csv`

   For the prototype, it is enough to support at least retrieval of **csv file from local and remote location**.
3. There are two main types of methods for data manipulation. First, the *data filtering methods* take a dataset as input, then they modify the dataset (e.g. by removing invalid entries, removing duplicates, data normalization, etc.) and they output a modified dataset. Second, the analytical methods take dataset as an input and they output results of their analysis, which can be a number, string etc. Multiple methods can be applied to a dataset in a single execution of the tool.

   For the data filtration methods, the prototype should implement **methods for removing items with empty customer's email** or **empty customer's address.** For analytical methods, the prototype should be able to determine **the average price of orders per year (separately for paid and unpaid orders), the total price of orders per year (only paid ones) and the top 3 customers with the highest number of orders**.
4. The results of the analytical methods are stored in specified output format. The prototype should support **at least one of the following: xml, json, plain text.**