DPRPy 2022/2023

Homework assignment no. 3 (max. = 35 p.)

Maximum grade: 35 p.

Deadline: 24.01.2023, 23.59

This task is solved in groups of two or three people.

1 Introduction

Homework should be sent via the `Moodle` platform, i.e., **one archive `.zip`**[1] (for each group) named

`Last-name1_First-name1_Last-name2_First-name2_assignment_3.zip` in which the following files will be placed:

- presentation (slides) containing the results of data analysis (`PDF` or `HTML`);
- `.R` scripts / `py` moduls that allows to read the data;
- all the `.R` scripts / `py` moduls and notebooks that allow to recreate results (figures, tables) contained in the presentation;

  Note: Please **don't** add files containing raw data - the uploaded `.zip` file should be "reasonable" sizes.

2 Data

We will continue working on data from the Stack Exchange network. However you will be using more data for this project - not only simplified data from Travel Stack Exchange forum - but from other forums as well.

At https://archive.org/details/stackexchange an anonymized dump of all user-contributed content on the Stack Exchange network is available. In all cases (except for the StackOverflow - ue to its size) each website is saved as one .7z archive, which contains 8 tables (XML files {`Badges`,`Comments`, `PostHistory`,`PostLinks`, `Posts`, `Tags`,`Users` and `Votes`}). Detailed description can be found on the https://archive.org/27/items/stackexchange/readme.txt and https://meta.stackexchange.com/questions/2677.

You must select at least **three** sites for analysis, one of which must be *not small* (> 100 MB).

3 Task description

This homework is a data science challenge - each group creates interesting (for themselves and the audience) questions and generates answers to them.

We are interested in issues related to specific websites, but also comparisons between sites. The state of "today" and trends over time. Popular stuff and rarities. Differences and similarities. You name it.

The projects that meet the following criteria will receive at least satisfactory grade ($> 50 % $, i.e. 17.5 p.):

1. uses at least 3 data sets - one of which should be of size > 100MB,
2. contain the code that generates at least two (for groups of two people) or three (for groups of three people) interesting results (answers to 'research' questions in the form of charts / tables / etc.),

[1] So not: .rar, .7z etc.

3. present the obtained results during the 15th laboratory class (max. 10 min. long).

   Note: The assessment will be issued mainly based on presentation.

Each additional analysis or non-trivial technique used will have a positive impact on the assessment (e.g. interactive charts, animations, web applications, maps, algorithms and data structures own implementations enabling the improvement of the speed of the analyzes performed methods known from the literature (with author's modifications), etc.).

In particular, the maximum grade (very good) only works that really stand out.