1 Introduction

1.1 Data

We will continue to work 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 - due 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).

2 Homework assignment no. 2

Max. grade: 25 p.

This assignment is to be performed individually by team members.

Code

Prepare the code needed to load data sets. The quality of the code will be taken into account here, e.g., code must be closed in well documented, specialized functions so as to avoid duplication etc. Scripts / moduls must be named clearly and documented as well. Prepared scripts / moduls should allow for automatic data loading (from any forum) based on the given path. Moreover, write code / functions to prepare data for analysis, i.e., the variables containging dates should be transform into correct format (if needed), all missing values should be denoted as `NA`.

Explanatory data analysis

Prepare a short report (in `R - Rmd/html` or in `Python - ipynb/html`) containig explanatory analysis of chosen sets:

1. description of available variables;
2. basic statists concerning them;
3. information about missing values;
4. etc.