Handling Large R Projects

Alex Zolotovitski

Medio Inc.
*Contact author: alex@zolot.us

Keywords: R projects, literate programming

There are many publications on how to solve different data mining, data analysis and modeling tasks in *R*. Typically a task is a specific procedure to analyze a specific data, e.g. k-means clustering of iris data, and requires 10-100 lines of *R* code. Usually the tasks are grouped in projects, e.g. analyze iris data, that could include descriptive statistics, principal component analysis, MDS, clustering, classification, etc., that can require 500-3000 lines of *R* code.

But often an *R* user has to work not with one task, but with a number of data mining projects in the same time, and each of the projects has many tasks. For example he/she might start project *A*, then switch to project *B*, then to project *C*, then after a few weeks or months return back to *A* and continue to work with it and so on. In each project *R* code can consists of many files and have thousands lines of code, which often creates problems in navigation through the code. Some tasks in a project could have long execution time that makes difficult to us usual literate programming methods (as sweave, knitr). This situation is typical for all professional *R* users, but there is no literature how to treat this situation.

We describe how to minimize overhead of switching between number of large projects and deliver results in minimal time.
Author's code snippets, including over fifty R functions in about 1500 lines of code that he developed for this purpose, that considerably (2-3) times increases performance, are available at [1] .

References

[1] Alex Zolotovitski (2014). Handling Large R Projects (HaLaP) http://bit.ly/HaLaP, http://github.com/alexzolot/HaLaP .