A computational analysis of R programming style variations in the last 20 years based on 94 million lines of code from all CRAN packages

by Chung-hong Chan, Chia-Yi Yen, Mia Huai-Wen Chang

Abstract The coexistence of multiple programming styles confuses new users and makes enforcing best practice difficult. This problem is aggravated by the lack of a universally accepted style guide in the R community. To investigate that, we quantified the programming style variation (PSV) in all CRAN packages and observed an emerging consensus in style since 2016, as indicated by the dampened increasing trend in PSV. It seems that a new consensus-based best practice is forming, which is a mixture of various R style guides. Concretely, we summarized the "ins & outs" of different styles based on popularity across years (e.g., rapid rise of underscore_fun_name and fall of dotted.fun.name since 2013) and pointed out the least agreed style elements (e.g., -> v.s. =, space after a comma). Moreover, we identified a source of PSV (the "Naughty, Naughty!") by looking into the style differences between clusters of related packages (e.g., Finance v.s. Biostatistics). Our analysis raises an open question to all stakeholders of the R community, i.e., the R Foundation, opinion leaders, package developers, and ordinary users: should we adopt an official R style guide as in the case of Python's PEP8? The findings from this study validate the R community's effort in reducing PSV and suggest future directions.

Introduction

Introductory section which may include references in parentheses (R Core Team, 2012), or cite a reference such as R Core Team (2012) in the text.

Section title in sentence case

This section may contain a figure such as Figure 1.

Figure 1: The logo of R.

Method

CRAN packages

In Jaunary 2019, a static snapshot of CRAN was archived using the **rsync** method outlined in the CRAN mirror HOWTO/FAQ guide. All CRAN submissions from 1998 to 2018, including active, archived and delisted packages were included for analysis.

The H1 of this study is to analyze the time-related changes in programming style. To this end, we cannot analyze all submissions from our static snapshot. If doing so the analysis would biased towards packages with many submissions. In order to balance the broadness of inclusion and bias, we sampled CRAN submissions using the "one submission per year" approach. For a package, if it has multiple CRAN submissions in a given year, only one submission is randomly selected.

The year of publication of a package is determined by the file time stamp of the package's tarball. This information was extracted with the **fs** package.

Language feature extraction

Summary

This file is only a basic article template. For full details of *The R Journal* style and information on how to prepare your article for submission, see the Instructions for Authors.

Bibliography

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2012. URL http://www.R-project.org/. ISBN 3-900051-07-0. [p1]

Chung-hong Chan
Mannheimer Zentrum für Europäische Sozialforschung
line 1
line 2
chung-hong.chan@mzes.uni-mannheim.de

Chia-Yi Yen
Graduate School of Economic and Social Sciences, Universität Mannheim, Germany
line 1
line 2
author2@work

Mia Huai-Wen Chang
Akelius Residential Property AB, Berlin, Germany
line 1
line 2
author2@work