R package downloads: what does it mean?

by Emi Tanaka

Abstract Abstract

Introduction

Today, R is greatly enhanced by over X R-packages contributed by X of developers all over the world. However, when R originally appeared in August of 1993 with its first official release in June of 1995 (Ihaka 1998), the contributions were managed by only a small group of core developers. In April of 1997, the Comprehensive R Archive Network (CRAN) was established as the official R-packages repository, with 3 mirror sites. Now, the source repositories to install R-packages have expanded to Bioconductor, Gitlab, GitHub, R-Forge and 106 CRAN mirrors in 49 regions. Of all the CRAN mirrors, the daily download counts for each package is only readily available from the RStudio CRAN mirror.

Data

The main source of data used in this report is the download logs from the RStudio CRAN mirror site : `https://cran.rstudio.com/`. These log files are created for every instance of download of an R-package via the RStudio CRAN mirror, then these log files are processed, daily, into CSV files that contain the following variables with the name of header in brackets:

- Date (`date`),
- Time in UTC time zone (`time`),
- Size of the file in bytes (`size`),
- Version of R used to download the package (`r_version`),
- Architecture type for R (i386 = 32 bit, x86_64 = 64 bit) (`r_arch`),
- Operating System (darwin9.8.0 = mac, mingw32 = windows) (`r_os`),
- Package (`package`),
- Country in two letter ISO country code (`country`), and
- Anonymised daily unique id (`ip_id`).

A similar log file is also created for every download of R from the RStudio CRAN mirror with the processed log file generating a CSV file that contains the same variables except `r_arch` and `package`, and `r_version` and `r_os` are named as `version` and `os`. These CSV files are hosted at `http://cran-logs.rstudio.com/` and updated daily with data available from 1st October 2012.

The log files of a particular day is processed and compressed into a single CSV file of about 40 megabytes (file sizes of earlier years are much smaller due to lower number of download logs). As there are over 700,000 CSV files, a simple estimate of the size of the data is 28 terabytes - far exceeding typical portable hard drives which are 1-4 terabytes.

The summarised version of data, where the data show the total daily download counts for each package, is accessible using the `cranlogs` R-package. The `cranlogs` package accesses this summary data through the web application programming interface (API) maintained by r-hub (**?**).

Results

Bibliography

Emi Tanaka
Monash University
Monash University
Clayton campus, VIC 3800, Australia
`http://emitanaka.org/`
ORCiD: 0000-0002-1455-259X
`emi.tanaka@monash.edu`