The Journal of Open Source Software

simpleCache: R caching for reproducible, distributed, large-scale projects

Nathan Sheffield[1], **VP Nagraj**[1], **and Vince Reuter**[1]

1 University of Virginia

DOI: 10.21105/joss.00463

Software

- Review
- Repository
- Archive

Submitted: 01 November 2017
Published: 10 January 2018

Summary

`simpleCache` is an R(R Core Team 2016) package that provides functions for caching R objects. Its purpose is to encourage writing reusable, restartable, and reproducible analysis for projects with large data and computational requirements. Like its name indicates, `simpleCache` is intended to be simple. Users specify a location to store caches, and then provide nothing more than a cache name and instructions (R code) for how to produce an R object. `simpleCache` either creates and saves or simply loads the result as necessary with just a single function call.

In addition to this basic functionality, `simpleCache` has advanced options for assigning objects to specific environments, recreating caches, reloading caches, and even distributing caching operations to cluster computing resources via the `batchools`(Lang, Bischl, and Surmann 2017) interface. These features make the package particularly useful for large-scale data analysis and research projects. `simpleCache` is most helpful for caching objects that are computationally expensive to create, but used in multiple scripts or by multiple users.

`simpleCache` is also useful to enhance performance in a package that relies on large databases. For example, `simpleCache` has been incorporated with the LOLA R package(Sheffield and Bock 2016) to more efficiently cache and retrieve genomic region databases. Similarly, `simpleCache` has been used to store cached baseline statistical tables for faster lookup to determine statistical differences on tables with hundreds of millions of data points (Sheffield et al. 2017).

In summary, `simpleCache` provides a user-friendly interface to help the R programmer manage computationally intensive, repeated data analysis.

References

Lang, Michel, Bernd Bischl, and Dirk Surmann. 2017. "Batchtools: Tools for R to Work on Batch Systems." *The Journal of Open Source Software* 2 (10). https://doi.org/10.21105/joss.00135.

R Core Team. 2016. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Sheffield, N. C., and C. Bock. 2016. "LOLA: Enrichment Analysis for Genomic Region Sets and Regulatory Elements in R and Bioconductor." *Bioinformatics* 32 (4):587–89. https://doi.org/10.1093/bioinformatics/btv612.

Sheffield, N. C., G. Pierron, J. Klughammer, P. Datlinger, A. Schonegger, M. Schuster, J. Hadler, et al. 2017. "DNA Methylation Heterogeneity Defines a Disease Spectrum in Ewing Sarcoma." *Nature Medicine* 23 (3):386–95. https://doi.org/10.1038/nm.4273.