The Journal of Open Source Software

tidyfst: Tidy Verbs for Fast Data Manipulation

Tian-Yuan Huang[1] **and Bin Zhao**[1]

1 School of Life Science, Fudan University

DOI: 10.21105/joss.02388

Software

- Review
- Repository
- Archive

Editor: Mikkel Meyer Andersen

Reviewers:

- @tomsing1
- @rcannood

Submitted: 02 June 2020
Published: 21 August 2020

License
Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License (CC BY 4.0).

Summary

The tidyfst package (Huang, 2020a) is an R package (R Core Team, 2020) for fast data manipulation in tidy syntax. The top-level design is inherited from tidy data structure proposed by Hadley Wickham (Wickham, 2014), which are: (1) each variable is a column; (2) each observation is a row; (3) each type of observational unit is a table. Moreover, the function names as well as parameter settings are very much borrowed from dplyr (Wickham et al., 2020) and tidyr (Wickham & Henry, 2020), reducing the learning cost for tidyverse (Wickham et al., 2019) users. At the bottom, tidyfst is backed by the high performance package data.table (Dowle & Srinivasan, 2019), which is speedy, stable (with little dependency), memory efficient and feature rich.

Sharing similar goals, both data.table and dplyr have gained much popularity among R users. Their features have been compared widely in the community, with pros and cons suggested in ideas and tested in examples (e.g. https://stackoverflow.com/questions/21435339/data-table-vs-dplyr-can-one-do-something-well-the-other-cant-or-does-poorly/27840349#27840349). While some opinions might be subjective, consensus could be made on at least two points: (1) data.table could handle data manipulation in less time than dplyr; (2) dplyr has a possibly more user-friendly syntax for learning and communication than data.table. The tidyfst package is designed to combine the merits of dplyr and data.table, so as to provide a suite of tidy verbs for fast data manipulation.

Note that tidyfst is neither the only nor the first package to make trade-offs between data.table and dplyr. Many similar works have been published on CRAN, including dtplyr (Henry, 2020), maditr (Demin, 2019), table.express (Sarda-Espinosa, 2019), tidyfast (Barrett, 2020), tidytable (Fairbanks, 2020), etc. Nevertheless, tidyfst holds its unique features that no alternative compares so far. One important feature is the support of data manipulation on fst file supported by fst package (Klik, 2020). It means the users could parse the data frames stored in disk first and load the minimum needed subsets to compute on. Other features include convenient column selection in various forms (regular expression, index, etc.), more concise parameter settings and new verbs for frequently-used data operations.

Furthermore, to save memory and lift speed to a higher level, tidyft (Huang, 2020b) has been designed, which utilizes modification by reference feature from data.table whenever possible. The tidyfst and tidyft share similar parameter settings, but function names of tidyft are even simpler (functions in tidyfst usually ends with "_dt", which tidyft does not). Though tidyft has better performance than tidyfst, it is less robust and demands the users to have deeper understanding on the concepts of modification by reference in data.table.

Hopefully, tidyfst could provide some reference for the design of dplyr and bring convenience to even data.table users by wrapping some complicated operations in concise steps.

Acknowledgement

The author of maditr, Gregory Demin and the author of fst, Marcus Klik have helped us a lot in the development of this work. It is so lucky to have them (and many other selfless contributors) in the same open source community of R.

References

Barrett, T. (2020). Tidyfast: Fast tidying of data. Retrieved from https://CRAN.R-project.org/package=tidyfast

Demin, G. (2019). Maditr: Fast data aggregation, modification, and filtering with pipes and 'data.table'. Retrieved from https://CRAN.R-project.org/package=maditr

Dowle, M., & Srinivasan, A. (2019). Data.table: Extension of 'data.frame'. Retrieved from https://CRAN.R-project.org/package=data.table

Fairbanks, M. (2020). Tidytable: Tidy interface to 'data.table'. Retrieved from https://CRAN.R-project.org/package=tidytable

Henry, L. (2020). Dtplyr: Data table back-end for 'dplyr'. Retrieved from https://CRAN.R-project.org/package=dtplyr

Huang, T.-Y. (2020a). Tidyfst: Tidy verbs for fast data manipulation. Retrieved from https://CRAN.R-project.org/package=tidyfst

Huang, T.-Y. (2020b). Tidyft: Tidy verbs for fast data operations by reference. Retrieved from https://CRAN.R-project.org/package=tidyft

Klik, M. (2020). Fst: Lightning fast serialization of data frames for r. Retrieved from https://CRAN.R-project.org/package=fst

R Core Team. (2020). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Sarda-Espinosa, A. (2019). Table.express: Build 'data.table' expressions with data manipulation verbs. Retrieved from https://CRAN.R-project.org/package=table.express

Wickham, H. (2014). Tidy data. *Journal of Statistical Software*, *59*(10), 1–23. doi:10.18637/jss.v059.i10

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., et al. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. doi:10.21105/joss.01686

Wickham, H., François, R., Henry, L., & Müller, K. (2020). Dplyr: A grammar of data manipulation. Retrieved from https://CRAN.R-project.org/package=dplyr

Wickham, H., & Henry, L. (2020). Tidyr: Tidy messy data. Retrieved from https://CRAN.R-project.org/package=tidyr