The Journal of Open Source Software

tabulog: A Language-Agnostic Template System for Parsing Log Files

Austin Nar[1]

1 Department of Statistics, Rochester Institute of Technology, Rochester, NY, USA

DOI: 10.21105/joss.01917

Software

- Review
- Repository
- Archive

Editor: Tania Allard

Reviewers:

- @alexanderfurnas
- @trallard

Submitted: 19 August 2019
Published: 27 November 2019

Summary

Log files (such as Apache's `access.log` files) often have a very regular structure with the same fields in each log entry. The format of these log files is usually optimized for human-readability as opposed to machine-readability. For data scientists and others who wish to perform analytics on thses log files, the ability to easily convert files into a tabular format is necessary to gain any meaningful insight regarding the contents of individual fields of the log file.

While for a single log format most data scientists and endgineers should be able to extract individual data elements from such files, solutions are often a messy patchwork of regex extracts that have poor-to-mediocre code readability. Also, for those who wish to analyze the log files of various formats from different sources, the need to reinvent the wheel for each new format is an unnecessary bottleneck that gets in the way of working on the actual analysis.

`tabulog` is a language-agnostic template syntax for parsing log files, with libraries today for `Python` and `R`. The `tabulog` syntax is influnced by the Jinja2 template library in Python. Parsing a log file is as simple as writing a template, which defines the structure of a line in a log file, and defining the regex patterns that a field in the log record should match. In order to be portable and self-contained, the entire definition of this operation can be stored in a human-readable `YAML` file, which makes it easy to port the parsing logic between `Python` and `R`.

`tabulog` as a syntax is designed to be language agnostic, but is heavily influenced by `Jinja2` (Ronacher, 2019). The `R` (R Core Team, 2019) package is dependent on the package `yaml` (Stephens et al., 2018), and the `Python` (Van Rossum & Drake Jr, 1995) package is dependent on the package `PyYAML` (Simonov, 2019). In both cases `YAML` is used for reading `tabulog` templates which are stored in `YAML` files for use with either `R` or `Python`. The `Python` package is also dependent on Pandas (McKinney, 2019), whose `DataFrame` class is 'tabular format' that is used for the final output in `Python`. The project is designed for data scientists who want a quick, clean, reproducible way of converting semi-structured log files into a tabular format that is easy to use for analysis.

References

McKinney, W. (2019). *Pandas: Powerful data structures for data analysis, time series, and statistics*. Retrieved from https://pandas.pydata.org/

R Core Team. (2019). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Ronacher, A. (2019). *Jinja2: A small but fast and easy to use stand-alone template engine written in pure python*. Retrieved from https://jinja.palletsprojects.com/en/2.10.x/

Simonov, K. (2019). *PyYAML: YAML parser and emitter for python*. Retrieved from https://pyyaml.org/

Stephens, J., Simonov, K., Xie, Y., Dong, Z., Wickham, H., Horner, J., reikoch, et al. (2018). *Yaml: Methods to convert r data to yaml and back*. Retrieved from https://CRAN.R-project.org/package=yaml

Van Rossum, G., & Drake Jr, F. L. (1995). *Python tutorial*. Centrum voor Wiskunde en Informatica Amsterdam, The Netherlands.