Fast, Consistent Tokenization of Natural Language Text

Lincoln A. Mullen\textsuperscript{1}, Kenneth Benoit\textsuperscript{2}, Os Keyes\textsuperscript{3}, Dmitry Selivanov\textsuperscript{4}, and Jeffrey Arnold\textsuperscript{5}

\textsuperscript{1} Department of History and Art History, George Mason University \textsuperscript{2} Department of Methodology, London School of Economics and Political Science \textsuperscript{3} Department of Human Centered Design and Engineering, University of Washington \textsuperscript{4} Open Data Science \textsuperscript{5} Department of Political Science, University of Washington

Computational text analysis usually proceeds according to a series of well-defined steps. After importing texts, the usual next step is to turn the human-readable text into machine-readable tokens. Tokens are defined as segments of a text identified as meaningful units for the purpose of analyzing the text. They may consist of individual words or of larger or smaller segments, such as word sequences, word subsequences, paragraphs, sentences, or lines (Manning, Raghavan, and Schütze 2008, 22). Tokenization is the process of splitting the text into these smaller pieces, and it often involves preprocessing the text to remove punctuation and transform all tokens into lowercase (Welbers, Van Atteveldt, and Benoit 2017, 250–51). Decisions made during tokenization have a significant effect on subsequent analysis (Denny and Spirling 2018; D. Guthrie et al. 2006). Especially for large corpora, tokenization can be computationally expensive, and tokenization is highly language dependent. Efficiency and correctness are therefore paramount concerns for tokenization.

The \texttt{tokenizers} package for R provides fast, consistent tokenization for natural language text (Mullen 2018; R Core Team 2017). (The package is available on GitHub and archived on Zenodo.) Each of the tokenizers expects a consistent input and returns a consistent output, so that the tokenizers can be used interchangeably with one another or relied on in other packages. To ensure the correctness of output, the package depends on the stringi package, which implements Unicode support for R (Gagolewski 2018). To ensure the speed of tokenization, key components such as the \texttt{n}-gram and skip \texttt{n}-gram tokenizers are written using the Rcpp package (Eddelbuettel 2013; Eddelbuettel and Balamuta 2017). The tokenizers package is part of the rOpenSci project.

The most important tokenizers in the current version of the package can be grouped as follows:

- tokenizers for characters and shingled characters
- tokenizers for words and word stems, as well as for Penn Treebank tokens
- tokenizers \texttt{n}-grams and skip \texttt{n}-grams
- tokenizers for tweets, which preserve formatting of usernames and hashtags

In addition the package provides functions for splitting longer documents into sentences and paragraphs, or for splitting a long text into smaller chunks each with the same number of words. This allows users to treat parts of very long texts as documents in their own right. The package also provides functions for counting words, characters, and sentences.

The tokenizers in this package can be used on their own, or they can be wrapped by higher-level R packages. For instance, the tokenizers package is a dependency for the tidytext (Silge and Robinson 2016), text2vec (Selivanov and Wang 2018), and textreuse (Mullen 2016) packages. More broadly, the output of the tokenization functions follows the guidelines set by the text-interchange format defined at an rOpenSci Text Workshop in 2017 (rOpenSci Text Workshop 2017). Other packages which buy into the text-interchange format can thus use the tokenizers package interchangeably.


https://doi.org/10.21105/joss.00655
The tokenizers package has research applications in any discipline which uses computational text analysis. The package was originally created for historical research into the use of the Bible in American newspapers (Mullen forthcoming) and into the borrowing of legal codes of civil procedure in the nineteenth-century United States (Funk and Mullen 2018, Funk and Mullen (2016)). The tokenizers package underlies the tidytext package (Silge and Robinson 2017), and via that package tokenizers has been used in disciplines such as political science (Sanger and Warin, n.d.), social science (Warin, Le Duc, and Sanger, n.d.), communication studies (Xu and Guo 2018), English (Ballier and Lissón 2017), and the digital humanities more generally.

References


