The Journal of Open Source Software

corporaexplorer: An R package for dynamic exploration of text collections

Kristian Lundby Gjerde[1]

1 Research Fellow, Norwegian Institute of International Affairs (NUPI)

DOI: 10.21105/joss.01342

Software

- Review
- Repository
- Archive

Submitted: 17 March 2019
Published: 13 June 2019

Background

Computer technology has profoundly changed the possibilities for doing research with text as data. Terms such as 'computational social science' (see Lazer et al., 2009) and 'digital humanities' (see Gold & Klein, 2016) are now firmly established, indicating how new methods have transformed entire fields of scholarly enquiry, enabling researchers to do things unimaginable even a few years ago. For good reasons – computers excel in quantitative tasks – these advances have been predominantly quantitatively oriented, geared towards 'big data' of various sizes. However, developments in software and hardware have also opened up vast fields of immensely helpful – and often untapped – possibilities for many scholars engaged in various forms of *qualitative* text analysis.

The `corporaexplorer` R package

`corporaexplorer` is an R package that uses the Shiny GUI (graphical user interface) framework for dynamic exploration of text collections. The package is designed for use with a wide range of text collections; one example could be a collection of tens of thousands of documents scraped from a governmental website; another example could be the collected works of a novelist; a third example could be the chapters of a single book.

The intended primary audience are qualitatively oriented researchers in the social sciences and humanities who rely on close reading of textual documents as part of their academic activity. However, the package should also be useful for those doing quantitative textual research and wishing to have convenient access to the texts under study. The chief aim of `corporaexplorer` is to facilitate more powerful, transparent, and efficient workflows. While a typical use case would be an open-ended exploration of text collections in order to identify and extract texts for further scrutiny, the package is above all aimed at flexibility: it does not enforce any given workflow, but may play a small or larger role in many different research designs.

The intention is also to encourage mixed methods in text analysis, by using `corporaexplorer` in combination with the excellent R packages that are available for quantitative text analysis (for a highly useful review, see Welbers, Van Atteveldt, & Benoit, 2017), and to encourage (licences permitting) the sharing of text collections.

The `corporaexplorer` concept is simple, yet has the potential to enhance research work with text collections in a powerful way. Main elements in the interactive apps:

- **Input**: The ability to filter the corpus and/or highlight documents, based on search patterns (in main text or metadata, including date range).
- **Corpus visualisation**: An interactive heat-map of the corpus, based on the search input (calendar heat-map or heat-map where each tile represents one document, optionally grouped by metadata properties).

- **Document visualisation and display**: Easy navigation to and within full-text documents with pattern matches highlighted.
- **Document retrieval**: Extraction of subsets of the corpus in a format suitable for close reading.

The API and GUI of `corporaexplorer`, as well as the inner workings of the package, are documented more thoroughly in the README file included with the package.

While collecting and preparing the text collections to be explored requires some familiarity with R programming, using the Shiny apps for exploring and extracting documents from the corpus should be fairly intuitive also for those with no programming knowledge, once the apps have been set up by a collaborator. Thus, the aim is for the package to be useful for anyone with a rudimentary knowledge of R – or with collaborators who have such knowledge.

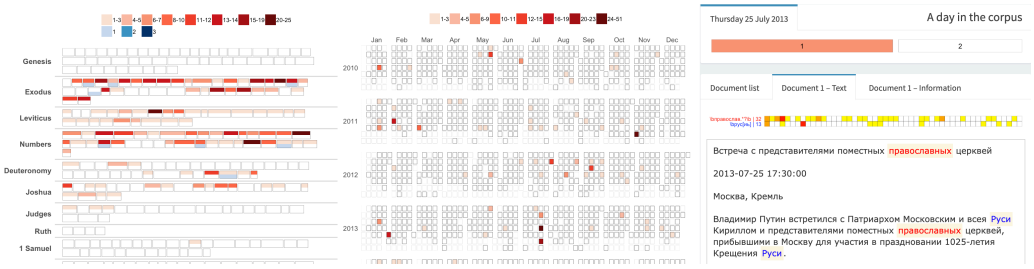

Figure 1: Illustration screenshots. To the left: part of interactive corpus heat-map displaying occurences of Moses and Abraham in the *King James Bible*. In the middle: part of interactive corpus calendar heat-map displaying occurences of 'orthodox' in a collection of transcripts from the Russian president's website. To the right: document view with interactive document map and highlighted pattern matches.

Main dependencies

`corporaexplorer` is an R (R Core Team, 2018) package. The interactive apps are built with the `shiny` (Chang, Cheng, Allaire, Xie, & McPherson, 2018) framework. The plots are built with `ggplot2` (Wickham, 2016). For searches and string operations, `stringi` (Gagolewski, 2019)/`stringr` (Wickham, 2019) and `re2r` (Qin Wenfeng, 2017) are used (details can be found in the package documentation). `data.table` (Dowle & Srinivasan, 2019) is used for fast search operations in the document term matrix. For other data operations, functions from various `tidyverse` (Wickham, 2017) packages are used. Other R packages utilised can be found in `corporaexplorer`'s DESCRIPTION file.

Acknowledgements

`corporaexplorer` has been developed with support from the research project 'Evaluating Power Political Repertoires (EPOS)' (project no. 250419), funded by the Research Council of Norway.

References

Chang, W., Cheng, J., Allaire, J., Xie, Y., & McPherson, J. (2018). *shiny: Web application framework for R*. Retrieved from https://CRAN.R-project.org/package=shiny

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

Gagolewski, M. (2019). *R package stringi: Character string processing facilities*. Retrieved from http://www.gagolewski.com/software/stringi/

Gold, M. K., & Klein, L. F. (Eds.). (2016). *Debates in the digital humanities: 2016*. Minneapolis, MN: University of Minnesota Press.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D., Christakis, N., et al. (2009). Computational social science. *Science*, *323*(5915), 721–723. doi:10.1126/science.1167742

Qin Wenfeng. (2017). *re2r: RE2 regular expression*. Retrieved from https://CRAN.R-project.org/package=re2r

R Core Team. (2018). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Welbers, K., Van Atteveldt, W., & Benoit, K. (2017). Text analysis in R. *Communication Methods and Measures*, *11*(4), 245–265. doi:10.1080/19312458.2017.1387238

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. New York: Springer-Verlag. Retrieved from http://ggplot2.org

Wickham, H. (2017). *tidyverse: Easily install and load the 'tidyverse'*. Retrieved from https://CRAN.R-project.org/package=tidyverse

Wickham, H. (2019). *stringr: Simple, consistent wrappers for common string operations*. Retrieved from https://CRAN.R-project.org/package=stringr