Software paper for submission to the Journal of Open Research Software

To complete this template, please replace the blue text with your own. The paper has three main sections: (1) Overview; (2) Availability; (3) Reuse potential.

Please submit the completed paper to: editor.jors@ubiquitypress.com

---

(1) Overview

Title

mineR: An R Package for Fuzzy Keyword Identification and Quantification in Natural Language

Paper Authors

1. Cole, Christopher B.
2. Patel, Sejal
3. Knight, Joanne

Paper Author Roles and Affiliations

1. Combining Healthcare Informatics, Computation, and Statistics (CHICAS), Lancaster Medical School, Lancaster University, United Kingdom; Data Science Institute (DSI), Faculty of Computing and Communication, Lancaster University, United Kingdom; Campbell Family Mental Health Research Institute, Centre for Addiction and Mental Health, Canada; Department of Biology, University of Ottawa, Canada
2. Campbell Family Mental Health Research Institute, Centre for Addiction and Mental Health, Canada; Institute of Medical Science, University of Toronto, Canada
3. Combining Healthcare Informatics, Computation, and Statistics (CHICAS), Lancaster Medical School, Lancaster University, United Kingdom; Data Science Institute (DSI), Faculty of Computing and Communication, Lancaster University, United Kingdom; Campbell Family Mental Health Research Institute, Centre for Addiction and Mental Health, Canada; Institute of Medical Science, University of Toronto, Canada; Biostatistics Division, Dalla Lana School of Public Health, University of Toronto, Canada

Abstract

Recent growth in the scale and scope of large ontologies has prompted the development of computational methodologies which can best use structured information to further human understanding. However, using structured ”terms” to mine the academic literature has proved difficult; previous efforts have neglected key concepts in natural language processing and efficient computation. In this article we present mineR, an R package capable of identifying co-occuring units within ontological terms to variable confidence, strict quality control, and additional features. mineR is released on Github and allows researchers use information from ontologies to extend and improve text mining in their field.

Keywords
Text mining; natural language processing; R; ontology; quality control

Introduction

As data grows larger and more complex, researchers frequently need to be able to quickly understand and summerize unstructured text in terms of themes and topics. Along with the myriad of complexities and issues that arise from analyzing natural text also allows resarchers the opportunity to perform tasks never before possible.

Implementation and architecture

How the software was implemented, with details of the architecture where relevant. Use of relevant diagrams is appropriate. Please also describe any variants and associated implementation differences.

Quality control

Detail the level of testing that has been carried out on the code (e.g. unit, functional, load etc.), and in which environments. If not already included in the software documentation, provide details of how a user could quickly understand if the software is working (e.g. providing examples of running the software with sample input and output data).

(2) Availability

Operating system

Please include minimum version compatibility.

Programming language

Please include minimum version compatibility.

Additional system requirements

E.g. memory, disk space, processor, input devices, output devices.

Dependencies

E.g. libraries, frameworks, incl. minimum version compatibility.

List of contributors

Please list anyone who helped to create the software (who may also not be an author of this paper), including their roles and affiliations.

Software location:

Archive (e.g. institutional repository, general repository) (required please see instructions on journal website for depositing archive copy of software in a suitable repository)

Name: The name of the archive.
Persistent identifier: e.g. DOI, handle, PURL, etc.
Licence: Open license under which the software is licensed.
Publisher: Name of the person who deposited the software.
Version published: The version number of the software archived.

Date published: dd/mm/yy

Code repository (e.g. SourceForge, GitHub etc.) (required)

Name: The name of the archive.
Persistent identifier: e.g. DOI, handle, PURL, etc.
Licence: Open license under which the software is licensed.
Date published: dd/mm/yy

Emulation environment (if appropriate)

Name: The name of the archive.
Persistent identifier: e.g. DOI, handle, PURL, etc.
Licence: Open license under which the software is licensed.
Date published: dd/mm/yy

Language

Language of repository, software and supporting files.

(3) Reuse potential

Please describe in as much detail as possible the ways in which the software could be reused by other researchers both within and outside of your field. This should include the use cases for the software, and also details of how the software might be modified or extended (including how contributors should contact you) if appropriate. Also you must include details of what support mechanisms are in place for this software (even if there is no support).

Acknowledgements

Please add any relevant acknowledgements to anyone else who supported the project in which the software was created, but did not work directly on the software itself.

Funding statement

If the software resulted from funded research please give the funder and grant number.

Competing interests

If any of the authors have any competing interests then these must be declared. The authors initials should be used to denote differing competing interests. For example: BH has minority shares in [company name], which part funded the research grant for this project. All other authors have no competing interests."
If there are no competing interests, please add the statement: The authors declare that they have no competing interests.

References

Please enter references in the Harvard style and include a DOI where available, citing them in the text with a number in square brackets, e.g.

[1] Piwowar, H A 2011 Who Shares? Who Doesn't? Factors Associated with Openly Archiving Raw Research Data. PLoS ONE 6(7): e18657. DOI: http://dx.doi.org/10.1371/journal.pone.0018657.

Copyright Notice