




text2map: R Tools for Text Matrices

Dustin S. Stoltz^{*1} and Marshall A. Taylor^{†2}

1 Lehigh University 2 New Mexico State University

DOI: [10.21105/joss.03741](https://doi.org/10.21105/joss.03741)

Software

- [Review](#) 
- [Repository](#) 
- [Archive](#) 

Editor: [Chris Hartgerink](#) 

Reviewers:

- [@alexanderfurnas](#)
- [@mbod](#)

Submitted: 25 August 2021

Published: 20 September 2021

License

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

Summary

text2map is an R ([R Core Team, 2021](#)) package that provides several tools for working with text matrices, including document-term matrices, term-context matrices, and word embedding matrices. text2map is published at The Comprehensive R Archive Network (CRAN) at <https://cran.r-project.org/package=text2map>; its source is available at <https://gitlab.com/culturalcartography/text2map/>.

text2map contains a number of vignettes demonstrating basic functionality and more advanced uses of the package.

Statement of Need

text2map offers a consistent set of tools built around representing texts as matrices. This is in contrast to corpus objects ([Perry, 2021](#)) or tidytext's triplet data frame ([Silge & Robinson, 2016](#)). This allows text2map to remain close to the underlying matrix mathematics of contemporary computational text analysis as well as make use of memory-efficient matrix packages—e.g., Matrix ([Bates & Maechler, 2010](#)).

While there are R packages for training word embeddings—e.g., text2vec ([Selivanov et al., 2020](#))—none offer methods for working with embeddings in downstream tasks, in particular, tasks involved in social scientific research. text2map offers functions for creating semantic centroids, semantic regions, semantic directions, and performing concept mover's distance and concept class analysis ([Arseniev-Koehler et al., 2021](#); [Arseniev-Koehler & Foster, 2020](#); [Boutyline et al., 2020](#); [Jones et al., 2020](#); [Stoltz & Taylor, 2019, 2021](#); [Taylor & Stoltz, 2020a, 2020b](#)).

Illustration

Let's consider a short, simple example of text2map in use. Building off some of our previous work ([Stoltz & Taylor, 2019](#); [Taylor & Stoltz, 2020b](#)), say a researcher is interested in examining the extent to which Shakespeare's First Folio plays engage the concept of "death." The text2map package can be used to efficiently convert the raw corpus into a document-term matrix (DTM), compute several different types of summary statistics on that DTM, and then quickly generate the concept mover's distance (CMD) scores using the CMDist() function and a matrix of word vectors. The user can also add "sensitivity intervals" using the package's DTM resampler to assess how robust each production's CMD score is (or is not) to the specific vocabulary frequency distribution of that document.

^{*}co-first author

[†]co-first author

35 The series of summary statistics table and figure below illustrate potential output that text
36 2map can be used to create.

	Measure	Value
1	Total Docs	37
2	Percent Sparse	88.20%
3	Total Types	31124
4	Total Tokens	874393
5	Object Size	3.7 Mb

	Measure	Value
1	Percent Hapax	9.100%
2	Percent Dis	2.300%
3	Percent Tris	1.100%
4	Type-Token Ratio	0.036

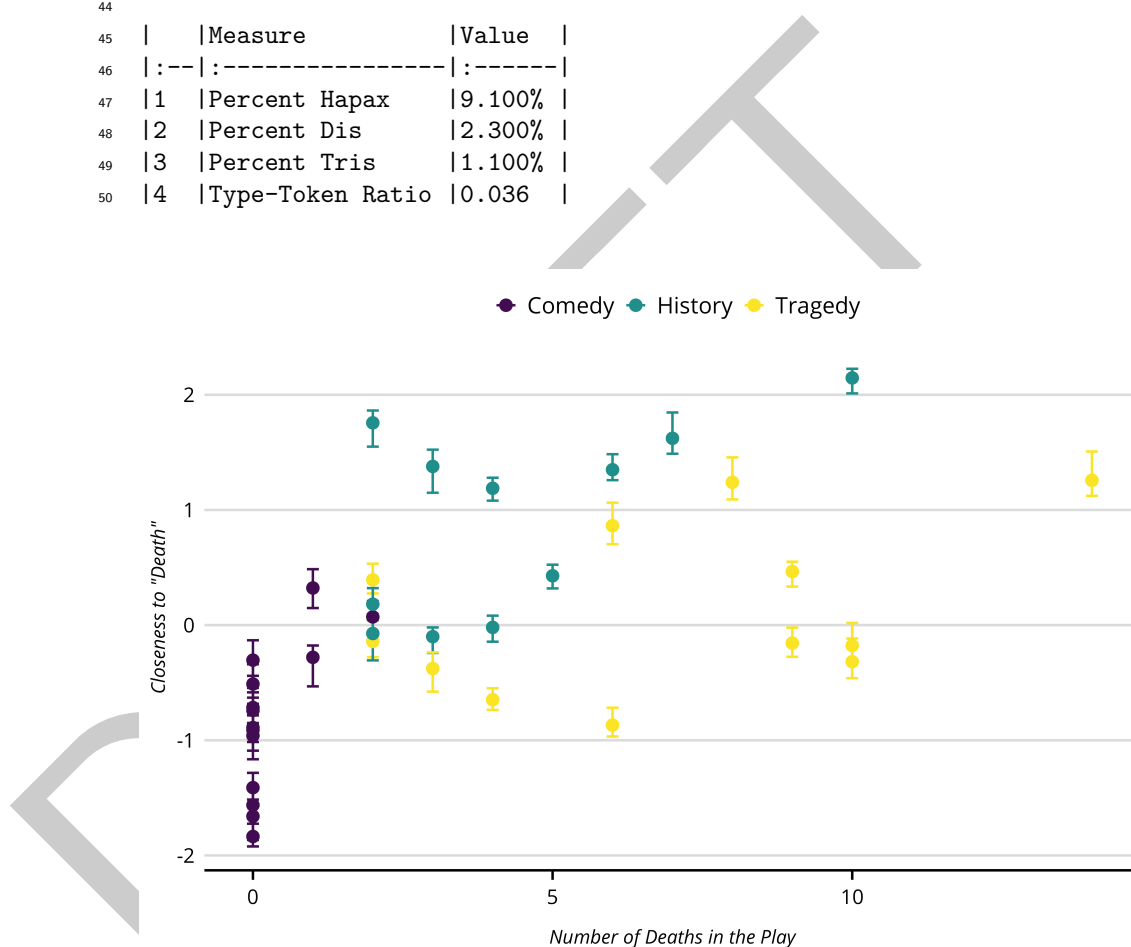


Figure 1: Illustrative figure. Scatterplot of Shakespeare play's CMD scores (y-axis) and the body count in the narrative (x-axis). Bands are sensitivity intervals, which are the CMD scores at the 2.5 and 97.5 percentiles for each document after resampling the vocabulary from the document 20 times.

51 Acknowledgements

52 We would like to thank Michael Lee Wood for helpful advice and test-runs with the software.
53 We would also like to thank Brandon Sepulvado for his assistance in fixing a parallelization
54 error in the CMDist() code.

References

- Arseniev-Koehler, A., Cochran, S. D., Mays, V. M., Chang, K.-W., & Foster, J. G. (2021). *Integrating topic modeling and word embedding to characterize violent deaths*. <https://doi.org/10.31235/osf.io/nkyaq>
- Arseniev-Koehler, A., & Foster, J. G. (2020). *Machine learning as a model for cultural learning: Teaching an algorithm what it means to be fat*. <https://doi.org/10.31235/osf.io/c9yj3>
- Bates, D., & Maechler, M. (2010). *Matrix: Sparse and dense matrix classes and methods*.
- Boutyline, A., Arseniev-Koehler, A., & Cornell, D. (2020). School, studying, and smarts: Gender stereotypes and education across 80 years of american print media, 1930-2009. In *SocArxiv*. <https://doi.org/10.31235/osf.io/bukdg>
- Jones, J. J., Amin, M. R., Kim, J., & Skiena, S. (2020). Stereotypical gender associations in language have decreased over time. *Sociological Science*, 7(1), 1–35. <https://doi.org/10.15195/v7.a1>
- Perry, P. O. (2021). *Corpus: Text corpus analysis*. <https://CRAN.R-project.org/package=corpus>
- R Core Team. (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. <https://www.R-project.org/>
- Selivanov, D., Bickel, M., & Wang, Q. (2020). *text2vec: Modern text mining framework for R*.
- Silge, J., & Robinson, D. (2016). Tidytext: Text mining and analysis using tidy data principles in r. *JOSS*, 1(3). <https://doi.org/10.21105/joss.00037>
- Stoltz, D. S., & Taylor, M. A. (2019). Concept mover's distance. *Journal of Computational Social Science*, 2, 293–313. <https://doi.org/10.1007/s42001-019-00048-6>
- Stoltz, D. S., & Taylor, M. A. (2021). Cultural cartography with word embeddings. *Poetics*, 101567. <https://doi.org/10.1016/j.poetic.2021.101567>
- Taylor, M. A., & Stoltz, D. S. (2020a). Concept class analysis: A method for identifying cultural schemas in texts. *Sociological Science*, 7(23), 544–569. <https://doi.org/10.15195/v7.a23>
- Taylor, M. A., & Stoltz, D. S. (2020b). Integrating semantic directions with concept mover's distance to measure binary concept engagement. *Journal of Computational Social Science*, 4, 231–242. <https://doi.org/10.1007/s42001-020-00075-8>