

# Arabica: A Python package for exploratory analysis of text data

Petr Koráb   and Jitka Poměnková  

1 Zeppelin University in Friedrichshafen, Germany 2 Brno University of Technology, Department of Radio Electronics, Czech Republic ¶ Corresponding author

DOI: [10.21105/joss.06186](https://doi.org/10.21105/joss.06186)

## Software

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

---

Editor: Olivia Guest  

Reviewers:

- [@linuxscout](#)
- [@amitkumarj441](#)

Submitted: 26 August 2023

Published: 05 May 2024

License

Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License ([CC BY 4.0](https://creativecommons.org/licenses/by/4.0/)).

In partnership with



This article and software are linked with research article DOI

[10.3847/xxxxx](https://doi.org/10.3847/xxxxx), published in the Journal of Open Source Software.

## Summary

Research meta-data is typically recorded as a time series with dimensions of cross-sections (e.g., article title, journal, volume, issue, author's names, and affiliations) and time (e.g., publication date). Meta-datasets provide valuable insights into the research trends in a particular field of science. Meta-analysis (a group of methods to analyze research meta-data) currently does not implement text analytics in either programming language. This package aims to fill that need. Arabica offers descriptive analytics, visualization, sentiment classification, and structural break analysis for exploratory analysis of research meta-datasets in easy-to-use Python implementation.

The package operates on three main modules: (1) descriptive and time-series n-gram analysis provides a frequency summarization of the key topics in the meta-dataset, (2) visualization module displays key-term frequencies in a heatmap, line plot, and word cloud, (3) sentiment and structural breakpoint analysis evaluates sentiment from research article titles and identifies turning points in the sentiment of published research. It uses VADER ([Hutto & Gilbert, 2014](#)) and FinVADER ([Koráb, 2023](#)), the updated model with financial lexicons, to classify sentiment. Clustering-based Fisher-Jenks algorithm ([Jenks, 1977](#)) finds break points in the data.

It provides standard cleaning operations for lower-casing, punctuation, numbers, and stopword removal. It allows removing more sets of stopwords from the NLTK corpus at once to clean datasets collected in multilingual regions.

As an example, the package helps analyze trends in economic research published in the leading economic journals (*Econometrica*, *Journal of Political Economy*, *American Economic Review*, *Quarterly Journal of Economics*, and *Review of Economic Studies*) from 1990 - 2017. The meta-data is collected from [Constellate.org](#).

A word cloud in Figure 1 displays the most frequent bigrams (two consecutive words) representing concepts, theories, and models that were central to researchers' attention throughout the period. A heatmap in Figure 2 adds a time dimension and plots the most frequent terms with yearly frequency. The line plot in Figure 3 displays the most frequent concepts in an alternative form for a shorter period, and Figure 4 shows aggregated sentiment with two turning points (the 2007 - 2009 financial crisis and the end of the Great Moderation period in economics in 1999).



Figure 1. Word cloud.

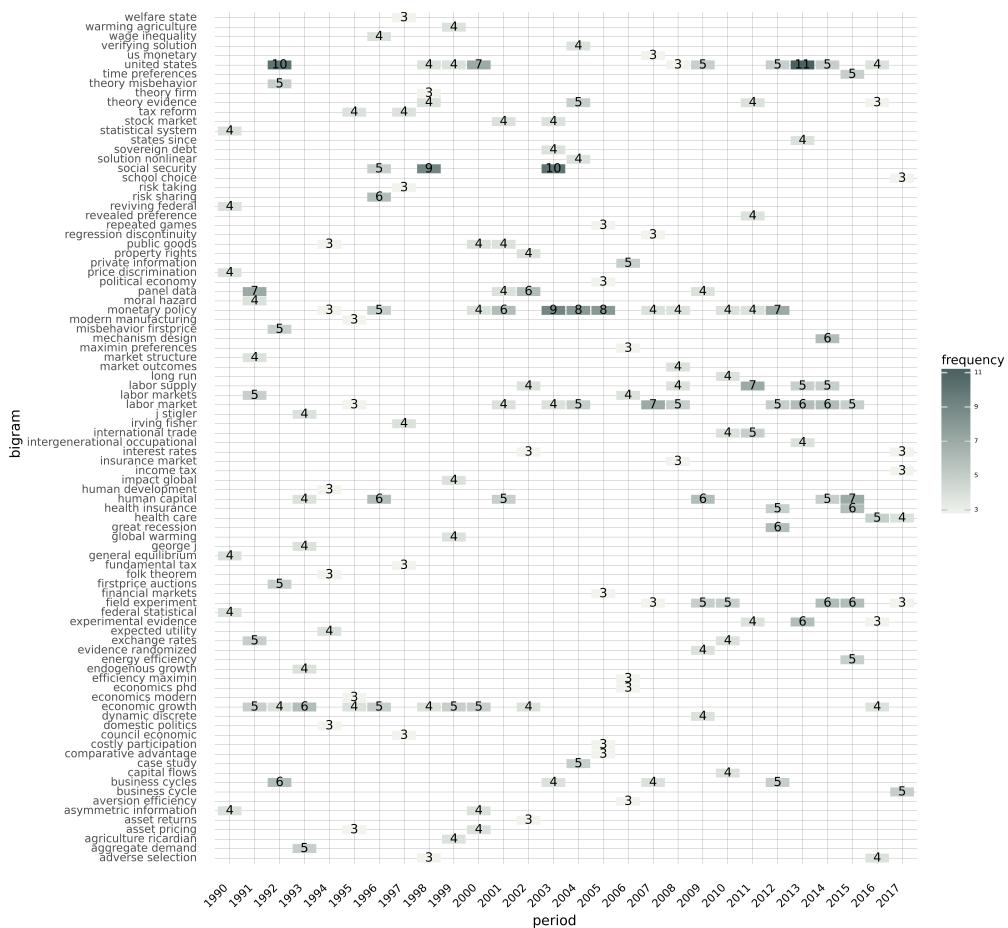


Figure 2. Heatmap.

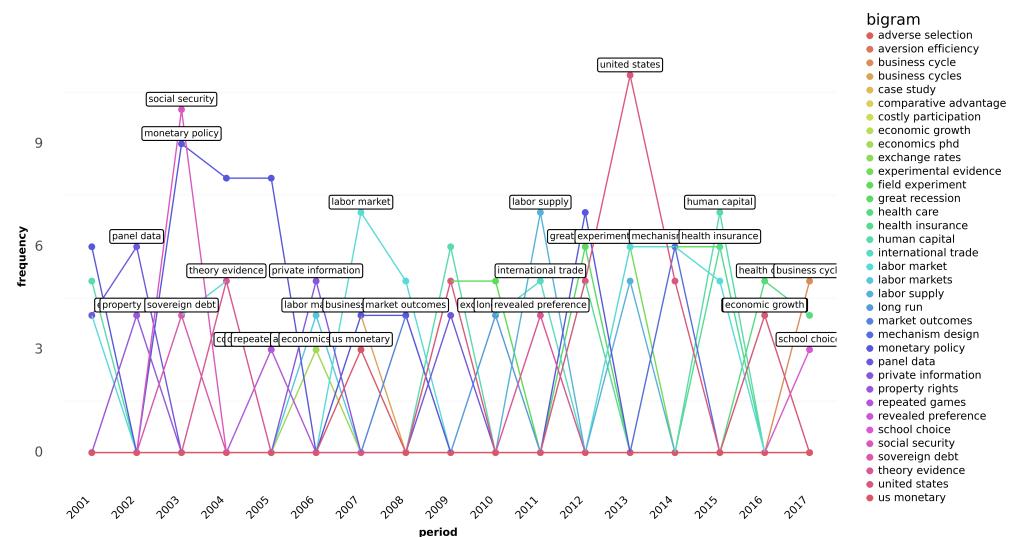


Figure 3. Line plot.

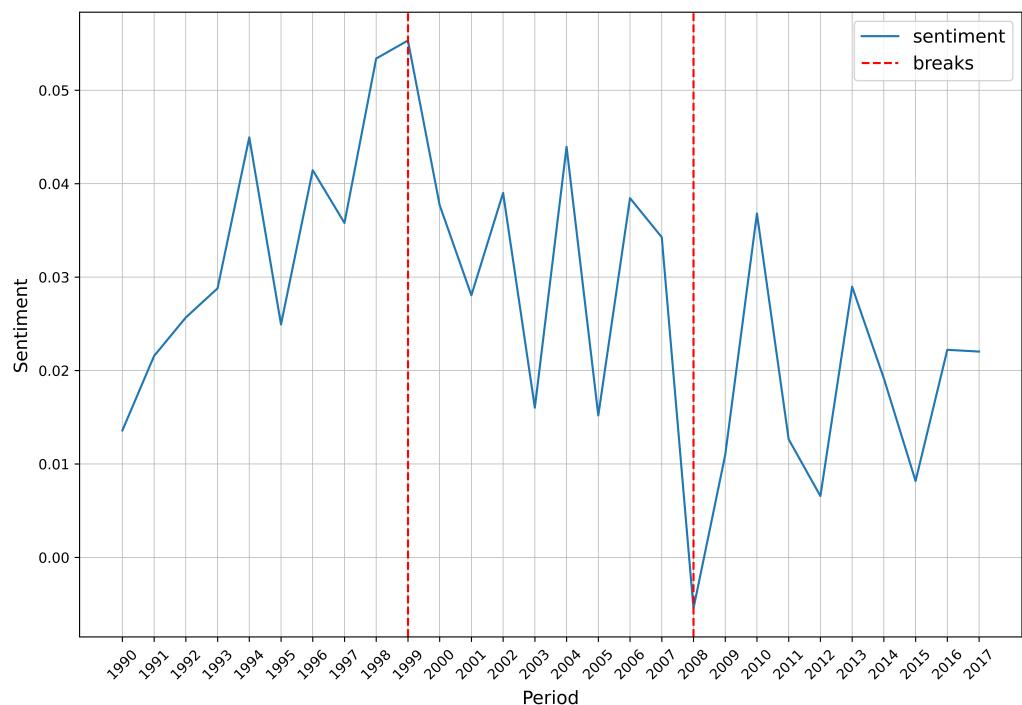


Figure 4. Sentiment and breakpoint analysis.

The package has more general use for exploratory analysis of time-series text datasets, mainly in social sciences. In business economics, it improves customer satisfaction measurement through product reviews analysis. In politology and behavioral economics, it enables detailed text mining of social media interactions. Similarly, in finance, it simplifies financial sentiment analysis of news headlines.

Arabica is well-documented: its API reference and comprehensive tutorials can be found at <https://arabica.readthedocs.io>. For easy installation, the package is included in the Python Package Index. Its code repository and issue tracker are hosted on GitHub at <https://github.com/PetrKorab/Arabica>.

## Statement of need

With Arabica, it is possible to visualize and analyze textual data in novel ways. These are some of the package's distinguishing features:

- Unlike the current packages to perform meta-analysis (Balduzzi et al., 2023; Mikolajewicz & Komarova, 2019; White, 2017), the package leverages text mining methods for in-depth analysis of research meta-data.
- Existing text analysis packages, such as Texthero (Besomi, 2021) and TextBlob (Loria, 2021) provide methods that explicitly focus on cross-sectional datasets and datasets without time variation. This perspective completely omits the time variability in text data. The time-series text approach provides additional insights into the qualitative changes in text datasets that are, as such, generated by human behavior. It involves the package's extension of the word cloud graph and financial sentiment classification for time-series text analysis.
- While some existing text-processing packages, e.g. Zerrouki (2022), focus on a specific group of languages, Arabica offers text-mining methods for all Latin Alphabet languages, including the stopwords removal of 18 lists of stopwords included in the NLTK corpus of stopwords.

## Dependencies

For most processing operations, Arabica uses data structures and methods from Numpy and Pandas (McKinney, 2013). It leverages NLTK for natural language processing (Loper & Bird, 2002) and Cleantext (Guidiva, 2021) for data pre-processing. It uses Plotnine (Wilkinson, 2005) and Matplotlib (Hunter, 2007) for visualization. It also depends on VaderSentiment (Hutto & Gilbert, 2014), FinVADER (Koráb, 2023), and Jenkspy (Viry, 2023) to implement sentiment and breakpoint analysis of general-language and financial texts.

## Acknowledgements

We acknowledge contributions from Jarko Fidrmuc on empirical applications and the visualization design of the library.

## References

- Balduzzi, S., Rücker, G., Nikolakopoulou, A., Papakonstantinou, T., Salanti, G., Efthimiou, O., & Schwarzer, G. (2023). Netmeta: An r package for network meta-analysis using frequentist methods. *Journal of Statistical Software*, 558. <https://doi.org/10.18637/jss.v106.i02>
- Besomi, J. (2021). Texthero — text preprocessing, representation and visualization from zero to hero. In *Python Package Index*. PyPI. <https://pypi.org/project/texthero>
- Guidiva, P. (2021). Cleantext—an open-source python package to clean raw text data. In *Python Package Index*. PyPI. <https://pypi.org/project/cleantext>
- Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 9. <https://doi.org/10.1109/MCSE.2007.55>
- Hutto, C., & Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*, 8. <https://doi.org/10.1609/icwsm.v8i1.14550>
- Jenks, G. F. (1977). *Optimal data classification for choropleth maps*. University of Kansas. [https://openlibrary.org/books/OL22018775M/Optimal\\_data\\_classification\\_for\\_choropleth\\_maps](https://openlibrary.org/books/OL22018775M/Optimal_data_classification_for_choropleth_maps)

- Koráb, P. (2023). FinVADER: VADER sentiment classifier updated with financial lexicons. In *Python Package Index*. PyPI. <https://pypi.org/project/finvader>
- Loper, E., & Bird, S. (2002). NLTK: The natural language toolkit. *ArXiv e-Prints*. <https://doi.org/10.48550/arXiv.cs/0205028>
- Loria, S. (2021). Textblob - simple, pythonic text processing. Sentiment analysis, part-of-speech tagging, noun phrase parsing, and more. In *Python Package Index*. PyPI. <https://pypi.org/project/textblob>
- McKinney, W. (2013). *Python for data analysis*. O'Reilly Media, Inc. <https://www.oreilly.com/library/view/python-for-data/9781491957653/>
- Mikolajewicz, N., & Komarova, S. V. (2019). Meta-analytic methodology for basic research: A practical guide. *Frontiers in Physiology*, 106. <https://doi.org/10.3389/fphys.2019.00203>
- Viry, T. (2023). Jenkspy- compute natural breaks Fisher-Jenks algorithm. In *Python Package Index*. PyPI. <https://pypi.org/project/jenkspy>
- White, I. (2017). NETWORK: Stata module to perform network meta-analysis. In *Boston College Department of Economics Statistical Software Components*. IDEAS/RePEc. <https://ideas.repec.org/c/boc/bocode/s458319.html>
- Wilkinson, L. (2005). *The grammar of graphics*. Princeton University Press. <https://doi.org/10.1007/0-387-28695-0>
- Zerrouki, T. (2022). PyArabic: A python package for arabic text. In *Python Package Index*. PyPI. <https://doi.org/10.21105/joss.04886>