MADStaText: Text Abstracts for Papers in Multi-Attribute Dataset on Statisticians

Zheng Tracy Ke, Pengsheng Ji, Jiashun Jin, and Wanshan Li

June, 2023

The Multi-Attribute Dataset on Statisticians (MADStat) (Ji et al., 2021) contains the bibtex and citation information of over 83K papers published in 36 journals during 1975-2015. MADStaText contains the abstracts of all papers in MADStat and was cleaned and analyzed by Ke et al. (2023).

- *Disclaimer*: Real author names are used, but all information was collected from public sources.
- *Citations*: To use this data set, please cite both Ji et al. (2021) and Ke et al. (2023).
- *Contacts*: Zheng Tracy Ke (zke@fas.harvard.edu), Jiashun Jin (jiashun@stat.cmu.edu).

(1) Text abstracts

The file `TextCorpusFinal.RData` contains the word-document-count matrix after pre-processing. It has four variables.

- `CleanAbstracts`: A list of 83331 items, where each item is the pre-processed abstract of a paper (the pre-processing includes tokenization, stemming, and removal of stop words).
- `dictionary`: A vector with 2106 entries, each corresponding to a word in the dictionary.
- `paperInfo`: A data frame with 4 columns, where *inCorpus* is a categorical variable indicating if this abstract was used in training a topic model (Ke et al., 2023) (*inCorpus==2*), *year* is the publication year, *journal* is the abbreviation of the publishing journal (the full journal names are in the supplementary material of Ke et al. (2023)), and *title* is the paper title.
- `TDM`: A 2106×83331 matrix, whose (j, i)th entry is the count of the jth dictionary word in the ith document.

(2) Citation and bibtex information

The file `AuthorPaperInfo.RData` was copied from Ji et al. (2021). This file contains the paper-paper citation information, as well as author information of all papers. It has two variables.

- `AuPapMat`: This matrix summarizes the bibtex data. It has 4 columns, where *idxAu* is author ID, *idxPap* is paper ID, *year* is publication year, and *journal* is publication journal.
- `PapPapMat`: This matrix summarizes the citation data. It has 5 columns, where *FromPap* and *ToPap* are the paper IDs, *FromYear* and *ToYear* are the publication years of two papers, and *SelfCite* is an indicator whether this is a self citation.

In addition, the file `author_name.txt` contains author names, in the same order as their IDs.

(3) Topic modeling results

Ke et al. (2023) estimated a topic model for paper abstracts, using a spectral method Topic-SCORE (Ke and Wang, 2022). The file `TopicResults.RData` contains the output of topic modeling. It has eight variables.

- `A_hat`: The estimated topic matrix $\widehat{A}$ (dimension: 2106×11).
- `AnchorWords`: The (almost-)anchor words for each topic, where *A_normalized* is the value of $a_j(k) := \widehat{A}_k(j)/[\sum_{\ell=1}^{K} \widehat{A}_\ell(j)]$ for an (almost-)anchor word j of topic k.
- `dictionary`: The 2106 words in the dictionary.
- `paperInTraining`: Topic-SCORE first estimates A (i.e., training) and then uses $\widehat{A}$ to estimate W. Not all abstracts were used in estimating A. This binary vector indicates which abstracts were used in estimating A.
- `paperTitle`: Titles of the 83331 papers.
- `PureDocs`: The (almost-)pure documents, including their titles and estimated weights, for each topic.
- `TopicNames`: The topic names manually assigned by Ke et al. (2023).
- `W_hat`: The estimated weight matrix $\widehat{W}$ (dimension: 11×83331). Not all abstracts were used in training; but after training, a weight vector was estimated for every abstract (see Ke et al. (2023) for details).

In addition, the file `MainTopics80_Authors.RData` contains the estimated topic interests of 80 representative authors (see Section 5.2 of Ke et al. (2023)). It contains one variable:

- `MainTopics_80`: A data frame with 4 columns, where *name* is the author name, *citation* is an author's total citations within the data range, *topics* contain an author's major topic interests.

(4) Raw data

All aforementioned data files were generated from the raw data `Raw-data-2019-12-version.RData`. Especially, this file contains the original abstracts (without any pre-processing). It has eight variables.

- `author_id`: A vector from 1 to 47311.
- `author_name`: A vector with 47311 entries, each being the name of an author.
- `author_pap_list`: A list of 47311 items, where the ith item stores the IDs of all papers written by the ith author.
- `journal`: A data frame containing 36 journals' full names, ISSNs, and their abbreviations in MADStat. When a journal has multiple ISSNs, it is listed as multiple rows in the data frame.
- `paper`: A data frame with all bibtex information, including *mr*, *doi*, *wos*, *title*, *year*, *timesCited*, *sourceURL*, *citingArticlesURL*, *issn*, *isbn*, *volume*, *issue*, *pages*, *abstract*. The column *abstract* contains the original abstracts (without pre-processing).
- `paper_author`: A list of 83331 items, where each item is the author list of a paper.
- `paper_id`: A vector from 1 to 83331.
- `paper_ref`: A list of 83331 items, where each item is the reference list of a paper (each reference is represented by its ID in MADStat, the same as its row number in `paper`).

References

Pengsheng Ji, Jiashun Jin, Zheng Tracy Ke, and Wanshan Li. Co-citation and co-authorship networks of statisticians. *Journal of Business & Economic Statistics*, (just-accepted):1–32, 2021. https://doi.org/10.1080/07350015.2021.1978469.

Zheng Tracy Ke and Minzhe Wang. Using SVD for topic modeling. *Journal of the American Statistical Association*, pages 1–16, 2022.

Zheng Tracy Ke, Pengsheng Ji, Jiashun Jin, and Wanshan Li. Recent advances in text analysis. *Annual Review of Statistics and Its Application*, 11, 2023.