The American Statistician

ISSN: (Print) (Online) Journal homepage: https://amstat.tandfonline.com/loi/utas20

Textual Data Science with R

by Mónica Bécue-Bertaut. Boca Raton, FL: Chapman & Hall/CRC Press, 2019, xvii + 194 pp., $79.95(H), $54.95(e-book), ISBN: 9781138626911(H), 9781315212661(e-book).

Kenneth R. Benoit

To cite this article: Kenneth R. Benoit (2021) Textual Data Science with R, The American Statistician, 75:4, 453-454, DOI: 10.1080/00031305.2021.1985864

To link to this article: https://doi.org/10.1080/00031305.2021.1985864

Published online: 04 Nov 2021.

References

Casella, G., and Berger, R. (2002), *Statistical Inference*, Belmont, CA: Duxbury. [451,452]

DeGroot, M.H., and Schervish, M.J. (2012), *Probability and Statistics*, New York: Addison-Wesley. [452]

Textual Data Science with R, by Mónica Bécue-Bertaut. Boca Raton, FL: Chapman & Hall/CRC Press, 2019, xvii+194 pp., $79.95(H), $54.95(e-book), ISBN: 9781138626911(H), 9781315212661(e-book).

Textual Data Science With R targets an important and relatively understudied area of data science: the statistical analysis of largely unstructured data in the form of natural language text. Using examples spanning fields such as free-form survey responses, bibliographies, and speeches, the book presents multi-dimensional methods for mining patterns and insights from textual data. Beginning with a practical and conceptual overview of textual data and how to pre-preprocess and structure this data, the book proceeds to explain the framework of correspondence analysis and its application to textual data. It then discusses two other major approaches: clustering and a focus on cluster features, including characteristic words, and multiple factor analysis. It finishes with an extensive practical section presenting examples and workflows for bibliographic databases, a rhetorical speech, political speeches, and a corpus of sensory descriptions.

Part of the Chapman & Hall/CRC Computer Science and Data Analysis Series, the book *Textual Data Science With R* is the first in that series to focus on the analysis of textual data. The book therefore points toward an important niche: the intersection of data science and natural language processing, something that poses challenges not just for the application of statistical learning to data with very nonclassical statistical distributions, but also associated with a completely distinct set of data processing needs. As with the other texts in the series, extensive examples accompany the methodological presentation. With *Textual Data Science With R*, the applications are based around the *Xplortext* R package developed by the author (Bécue-Bertaut).

While there is definitely a need for work in textual data science, the book's title is slightly misleading, as the data science in the book involves no machine learning for predictive methods or other models involving the sort of performance metrics usually associated with data science. It does cover unsupervised methods for data mining in the form of correspondence analysis, cluster analysis, and factor analysis. But it makes no attempt to integrate, or even refer in passing to, all of the developments in natural language processing and statistical learning for text that have revolutionized the field in recent decades, including even now classic methods for the same broad applications such as latent Dirichlet allocation-based topic models. Moreover, the text analysis and natural language processing coverage is limited to the first chapter. The unique—and often uniquely challenging—aspects of textual data processing are covered in only the first 16 pages, including the organization of data into documents, the management of metadata, tokenization, lemmatization, and word selection. As the challenges of collecting, organizing, cleaning, and processing textual data are, in my experience, one of the most significant barriers that newcomers to the field face, this section could be more helpful.

The book very strongly reflects the French school of "l'analyse des données" (Benzécri 1981), an approach to statistical data analysis built around the method of correspondence analysis (CA). The introduction and main example data come from Ludovic Lebart, a first-degree disciple of Benzécri. Following that tradition, the book not only adopts the lexicon of CA, such as row and column "profiles", "inertia", "attraction", "repulsion", etc., but also follows this tradition by minting new terms for fairly standard textual data structures. A document-term matrix becomes a "lexical table", for instance, and tokenization becomes a process of identifying spelling-based *graphical forms*, "in the terminology adopted by André Salem", a reference that will be virtually unknown by most students of NLP. The analytic approach also very squarely follows the French tradition, focusing on CA and related approaches as the main task of textual data analysis.

Chapter 2 presents correspondence analysis, using the French approach following the same clear presentation as provided in Greenacre (2017). Using examples from free-form text questions on the *Aspiration Survey*, Chapter 3 shows how to apply CA to produce multidimensional scaling of words, documents, and clusters, using the *Xplortext* package that easily produces results in tables, scatterplots, and biplots. Chapter 4 covers clustering, including hierarchical clustering and ways to combine hierarchical clustering with dimensional reduction from CA.

Chapter 5 explains how to extract characteristic words from clusters, including combining the textual data with external variables (from what is termed an "advanced lexical table") for determining, for instance, words most strongly associated with youngest and oldest age groups. Chapter 6 extends this approach using multiple factor analysis (MFA), an approach that allows multiple document-word matrix representations to be combined to produce both global and local representations of words and documents, as well as to incorporate additional nontextual variables. This offers the intriguing possibility of comparing results across languages, such as the relationship between word usage and gender and age across British, French, and Italian respondents.

Chapter 7 puts the first six chapters to use across five examples. The first uses abstracts from 506 *Medline* articles about lupus to show clusters of characteristic terms, as well as a time analysis of characteristic drug and treatment names. A second application analyzes a speech by Robert Badinter to defend the bill for the abolition of the death penalty in France, split into sequential segments to trace the rhetorical path taken in this hour-long speech as a form of discursive analysis. A third study compares investiture speeches from Spanish presidents since 1978, showing their trajectories across documents when linked by time, as well as word clusters to produce a simple form of topic analysis. In another Iberian example, the last section analyses textual descriptions from tastings of eight Catalonia red wines, using MFA to distinguish clusters of judges based on their sensory descriptions, as well as wine characteristics from the texts.

The presentation of the material both overall and within each chapter is very clear, with effective use of an overview,

conclusions that dovetail nicely with the following chapter, and a final section for how to implement the chapter's methods in *Xplortext*.

The book will find the most comfortable audience among readers already familiar with correspondence analysis and the French tradition of applying this to text (e.g., Lebart et al. 1998). Readers expecting a more comprehensive guide to textual data analysis, or a more modern data science-based combination of NLP and machine learning, may be left scratching their heads at the decision to focus on a very specific approach that has a devoted but limited following. Textual data science as a field is far more broad, and developed, than suggested by this book's contents. Benzécri may have asserted, as quoted in the book, that correspondence analysis is able "*to address any type of issue related to the form, meaning, or style of texts*" but this statement was wrong in the 1980s and is even less true today.

Kenneth R. Benoit
London School of Economics and Political Science/Data Science Institute

References

Bécue-Bertaut, Alvarez-Esteban, R., Sánchez-Espigares, J.-A., and Belchin, K. (2021), *Xplortext: Statistical Analysis of Textual Data*, R package version 1.4.1. *https://CRAN.R-project.org/package=Xplortext* [453]

Benzécri, J.P. (1981), *Pratique de l'Analyse des Données, Tome III, Linguistique et Lexicologie*, Dunod, Paris. [453]

Greenacre, M. (2017), *Correspondence Analysis in Practice*, Boca Raton, FL: CRC Press. [453]

Lebart L., Salem, A., and Berry, L. (1998), *Exploring Textual Data*, Dordrecht: Kluwer. [454]