Natural Language Processing (2025), 1–4
doi:10.1017/nlp.2025.10012

CAMBRIDGE UNIVERSITY PRESS

Book Review

Text Analytics: An Introduction to the Science and Applications of Unstructured Information Analysis by John Atkinson-Abutridy. Boca Raton, Florida: CRC Press, 2022. ISBN 9781032249797 (HB: $140.00), ISBN 9781032245263 (PB: $54.95), ISBN 9781003280996 (eBook: $54.95), xxvii+230 pages.

In the past decade, text analytics technology has rapidly gained popularity among social scientists and researchers in the digital humanities, as well as in the business domain. It is becoming increasingly feasible to mine valuable data from documents as a consequence of abundant data, novel techniques and declining computational costs.

Compared with related books on this topic (e.g. Wilcock 2009; Silge and David 2016; Huang 2021), which focus on the programming aspects and tend to align with their authors' specific interests and research, John Atkinson-Abutridy's Text Analytics: An Introduction to the Science and Applications of Unstructured Information Analysis not only discusses practical and applied aspects of text analytics but also describes the rationale behind these methods and models, that is, the how and the why. In this way, it combines the essential theoretical foundations of these techniques with practical applications. Often, researchers who understand the basics of computational methods and models may nevertheless lack a deep understanding of the theory behind them. As a result, computational linguists might also have difficulty implementing them in other languages or tools. Thus, this book not only provides a guide for scholars and researchers who want to apply text analytics methods to their research, fully understand the background and logic of the computational methods and to use them in other languages or computational tools but also demonstrates new implications for businesses seeking to improve their decision-making and productivity.

Each of the nine chapters in the book consists of two parts. The first section covers the fundamental concepts, paradigms, methods and models; the second provides examples and simple, practical exercises in Python. Each chapter concludes with a brief introduction to the internationally used basic terminology.

The first chapter describes the fundamental concepts, approaches and applications of text analytics. The author distinguishes between structured and unstructured data and between text mining and text analytics. Although these latter two terms are sometimes interchangeable, text mining refers to an automated process that searches for useful and meaningful patterns by examining extensive collections of documents or corpora (Struhl 2015; Zhai and Massung 2016; Ignatow and Mihalcea 2017) and involves hybrid methods combining three different areas: machine learning, natural-language processing and information retrieval (p. 6). Text analytics is a more specific task that involves mining and synthesising text data so that it can be quantified and visualised in a way that supports decision-making and yields actionable insights (p. 4). In addition to describing the process and significant challenges of text analytics, the author considers its applications.

In Chapter 2, the fundamental concepts and the computational and linguistic techniques of natural language processing (NLP) are discussed. The author describes its different processing levels (i.e., phonology, morphology, lexicon, syntax, semantics and pragmatics) and provides several

practical examples of NLP techniques designed to solve a variety of complex problems pertaining to written and spoken human language.

In the third chapter, the author discusses key concepts related to textual analysis and information extraction, showing how they can be performed using NLP techniques such as named entity recognition and relation extraction. These approaches are often based on supervised and unsupervised learning methods.

The fourth chapter presents various concepts, approaches and models of document representation. While Chapter 3 provides effective mechanisms for selecting and representing textual information, this chapter focuses on how this textual information is computationally characterised and represented in the form of documents to be used in textual analytics tasks. It proposes various fundamental approaches based on document indexing and vector space models, such as Boolean representation, term frequency (TF) and inverse document frequency models (TFxIDF). Finally, a practical exercise of this TFxIDF representation model is analysed in depth.

Chapter 5 examines key concepts, methods and problems related to patterns extracted from different documents by means of association rules, which refer to a rule set composed of combinations of frequent itemsets. This process is similar to that of analysing a customer's buying habits in order to discover associations between the items that customers place in their shopping baskets (Tan *et al.* 2018). In the same way, documents can be characterised by containing certain implicit associations between terms or entities mentioned in the documents. The author then proposes three specific metrics (support, confidence and lift) and primary computational methods for assessing the quality of frequent itemsets and association rules (such as the popular APRIORI method). Finally, the author concludes with a practical application of this APRIORI algorithm generating association rules from a large corpus of documents.

The sixth chapter examines the rationale behind the various techniques and models that allow readers to analyse and model the linguistic relationships between words and documents from a training corpus. In these cases, the author outlines various methods for transforming a high-dimensional into a low-dimensional representation model, known as 'word embeddings'. Examples include Latent Semantic Analysis (LSA) (an unsupervised model) and Word2Vec (a supervised model). Both enable us to effectively capture the hidden relationships between words and documents in context. These two models also illustrate a few examples of corpus-based semantic analysis.

Chapter 7 examines the computational concepts and different approaches for document clustering, using machine learning methods. Clustering refers to the process of creating groups of similar objects (Srivastava and Sahami 2009; Ignatow and Mihalcea 2017; Barry and Gurpreet 2022). The author introduces modern grouping principles, metrics and efficient and robust methods, such as K-means clustering and self-organising maps (SOMs) to find the best groupings to uncover hidden patterns. K-means clustering is one of the most popular cluster generation techniques, which searches for non-overlapping clusters based on the distance to their centroids. As an extension of this method, SOM not only creates clusters with distinct topologies but also applies competitive learning to these clusters. The author concludes by providing examples of these two methods in Python applications.

Document clustering methods allow scholars to divide input data into coherent groups, but they are unable to identify any hidden semantic structures within these groups. In order to solve this issue, in chapter eight, a technique is introduced for unsupervised machine learning, topic modelling, that enables researchers to uncover these structures and comprehend the relationships between documents. The main concepts and methods of topic modelling are introduced, including LSA, pLSA and Latent Dirichlet Allocation (LDA), with LDA being the most popular and effective due to its 'robust model for sampling textual data in order to generate efficiently the distributions of topics associated with documents and words associated with topics' (p. 179). On this basis, the author describes how to apply the LDA method to unstructured information data.

https://doi.org/10.1017/nlp.2025.10012 Published online by Cambridge University Press

Chapter 9 focuses on document categorisation, a text mining task involving the automatic labelling of documents based on a variety of predefined categories. It describes related concepts, models and techniques, including Bayesian models, neural network models and maximum entropy methods, some of which make numerous assumptions about distribution, whereas others, such as maximum entropy methods, do not. To evaluate the performance of such models, the author considers typical metrics, such as accuracy, precision and recall. Two examples show how to apply text categorisation.

In his conclusion, the author discusses the increasing popularity of text analytics technology in line with the growth of unstructured data. In addition to discussing techniques and applications of text analytics, he offers suggestions for the future, such as enabling businesses to reach new levels of insight from both positive and negative feedback by analysing sentiment, receiving timely alerts about changes to sentiment and aligning those insights with customer metrics.

A distinguishing feature that sets this book apart from the existing literature is its unified viewpoint, which combines the theoretical and practical aspects together. It provides an introduction to the key concepts and theories and an in-depth analysis of practical topics. This allows professionals, students and technicians to better understand their practice and compare various methods. After opening the 'black boxes' of functionalities, in which technical details are frequently concealed, they will be able to apply these computational models to reduplicate experiences without being limited to particular programming languages.

In contrast to books aimed at advanced professionals or scientists that typically focus on data methods or programming algorithms, this book combines basic foundations of text analytics with practical applications, accompanied by some concrete programming language examples. This enables readers with a basic level of knowledge, such as business professionals or postgraduate students, to comprehend the logic behind the computational methods and apply them in their preferred programming language. At the same time, readers will learn by analogy how to handle decision-making issues arising from textual or documentary sources.

One limitation of this book, however, is that it does not pay explicit attention to the relationship between chapters, and readers may struggle to comprehend the order of the chapters, but this is a minor issue, and in general, the book could be recommended to a wide audience.

Financial support. This research is supported by the project of Henan Philosophy and Social Sciences (Grant No. 2025BYY020), Humanities and Social Sciences, Ministry of Education, People's Republic of China (Grant No. 23YJCZH057) and the project of the National Social Science Fund of China (Grant No. 24BYY153).

Qiuying Zhao
School of Foreign Languages, Xuchang University
Xuchang 461000, P. R. China
Email: tongzhuo321@163.com

References

Barry V. and Gurpreet S.B. (2022). *Text as Data: Computational Methods of Understanding Written Expression Using SAS*. Hoboken, New Jersey: John Wiley & Sons, Inc.

Huang T. (2021). *Text Data Mining*. Beijing: China Machine Press.

Ignatow G. and Mihalcea R. (2017). *An Introduction to Text Mining: Research Design, Data Collection, and Analysis*. New York: SAGE Publications.

Silge J. and David R. (2016). *Texting Mining with R: A Tidy Approach*. Newton, MA: O'Reilly Media.

Srivastava A. and Sahami M. (2009). *Text Mining: Classification, Clustering, and Applications*. Boca Raton, FL: Chapman and Hall/CRC.

Struhl S. (2015). *Practical Text Analytics: Interpreting Text and Unstructured Data for Business Intelligence*. London: Kogan Page.

https://doi.org/10.1017/nlp.2025.10012 Published online by Cambridge University Press

Tan, Q., Lei, X., Wang, X., Wang, H., Wen, X., Ji, Y. and Kang, A. (2018). An adaptive middle and long-term runoff forecast model using EEMD-ANN hybrid approach, *Journal of Hydrology* **567**, 767–780.

Wilcock G. (2009). *Introduction to Linguistic Annotation and Text Analytics*. San Rafael, CA: Morgan & Claypool Publishers.

Zhai C. and Massung S. (2016). *Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining*. San Rafael, CA: ACM and Morgan & Claypool Publishers.

https://doi.org/10.1017/nlp.2025.10012 Published online by Cambridge University Press