Natural Language Engineering (2022), **28**, pp. 137–139
doi:10.1017/S1351324920000649

CAMBRIDGE
UNIVERSITY PRESS

Book Review

Text Mining with R: A Tidy Approach, by Julia Silge and David Robinson. Sebastopol, CA: O'Reilly Media, 2017. ISBN 978-1-491-98165-8. **XI + 184 pages**.

As the title suggests, the volume under review is not only a promotion of the multi-platform, open-source software, R (R Core Team 2020), but is also tailored for text mining. It can be used to address fundamental but prominent issues in text mining, natural language processing, data science, linguistics, etc., based on the tidy principles.

The authors explicate the text mining techniques with the R package **tidytext** they developed (Silge and Robinson 2016). Such techniques are practical to handle natural and unstructured data, which makes it more effective to explore and visualize natural language characteristics. The book also furnishes detailed and practical code examples, not only aiming to facilitate the work of professional analysts or data scientists, but also attempting to help readers develop real insights into human language materials, ranging from literature and social media, to the Internet databases.

Apart from the preface, the bibliography, and the index, the book consists of nine chapters. Chapter 1 serves as an introduction to applying the tidy text principle to text analysis and provides an overview of the most basic but essential feature of texts, **word frequency**. Chapter 2 uses **sentiment analysis** to explore the emotional content of texts. Chapter 3 is devoted to quantifying the most important terms in a document with the **tf-idf** metric. Chapter 4 attempts to investigate relationships and connections between words in documents by **n-grams** and **correlations**. Chapter 5 demonstrates the methods and procedures to convert between tidy and nontidy text formats. Chapter 6 adopts the **topic modeling** algorithm to group document collections. The final three chapters provide step-by-step case studies, digging into text datasets of Twitter archives, NASA metadata, as well as Usenet messages. All chapters are written in a clear fashion to show what the real text mining component is and explain how to carry it out in practice.

At the very beginning, Chapter 1 defines the concept of tidy text format, that is, a table with one-token-per-row. It is noticeable that the token here can be a single word, an n-gram, or other meaningful units of text. The authors then demonstrate how to generate one-token-per-row format by **unnest_tokens()** function and process the output to deal with many tasks, such as removing stop words and calculating word frequencies. In addition, this chapter also introduces two useful literary text datasets accessible via the **gutenbergr** and **janeaustenr** packages, which may provide linguistic materials for text analysis. The authors demonstrate how to run three novels through R packages and compare their word frequencies, followed by plotting and statistical testing. This example gives the readers a clear picture of how to use the tidy data principles to approach questions on the most basic and essential feature of texts, word frequency.

Different from the literal word frequency of words in texts, Chapter 2 focuses on the topic of sentiment analysis, viz., the emotional content of words, such as surprise or disgust, and whether they are positive or negative. The authors consider the sentiment of a text as the incorporation of the sentiments of all individual words, and they provide three general sentiment lexicons, AFINN, Bing, and NRC, in the **tidytext** package for the convenience of users. In this chapter, readers can learn how to figure out the most common words associated with "joy" in a text, analyze the trajectory of sentiment throughout a text, explore which word contributed most to each sentiment,

https://doi.org/10.1017/S1351324920000649 Published online by Cambridge University Press

and plot word clouds of positive and negative words, etc. Finally, the chapter covers larger units of sentiment analysis other than words, namely, sentences and chapters, and gives an example of how to demonstrate the most positive or negative chapters in one novel.

Chapter 3 focuses on quantifying what a document is about. Two approaches are introduced in terms of measuring how frequently a word is used in a text. The first is term frequency(**tf**). Usually, the relationship between rank and frequency can fit Zipf's law well. The other one is inverse document frequency (**idf**), which indicates whether a high-frequency word in one document is shared in a collection of documents. A word decreases its weight if commonly used in other documents and vice versa. Together, these lead to the metric **tf-idf** (term frequency multiplied by inverse document frequency), which is used to find words that are typical of one document within a series of documents. The authors then apply tidy data principles to approach the **tf-idf** analyses with two examples, based on Jane Austen's novels and the classic physics texts retrieved by **gutenbergr**. These operations clarify how to quantify important words in a text compared with the other homogeneous ones through **tf-idf**.

The first three chapters concentrate primarily on the individual words, their relationships to emotions and their relationships to texts, while Chapter 4 focuses on the relationships and connections between words in texts, namely, n-grams and correlations. These are typically of great interest to linguists and data analysts. The authors provide some methods to count and filter n-grams, and to examine pairwise correlations, mostly by analyzing word **networks** in texts using two new packages, **ggraph** and **widyr**. The former is practical and flexible for the visualization of those relationships, while the latter is used to count and correlate pairs of words.

The four chapters above introduce the **tidytext** methods and some applications in informative analyses based on the tidy text format. They are compatible with the popular suite of tidy tools, including readr, tidyr, dplyr, and ggplot2, etc. (Wickham and Grolemund 2017). The combination of these packages can be easily employed in exploring and visualizing text data. However, not all texts under investigation are tidy and not all existing R tools are compatible with the tidy text format. Hence, it is crucial to convert back and forth between tidy and nontidy format. Chapter 5 is devoted to this topic. It gives examples of converting document-term matrices of newspaper articles and document-feature matrices of inaugural speeches into tidy text format by the **tidy()** function from the **broom** package and conducting sentiment analysis, tf-idf analysis on them, respectively. Meanwhile, since some packages expect matrices as input, the authors also show how to cast tidy text into a matrix through the **cast()** function based on an example of creating a document-term matrix of Jane Austen books. The authors then further demonstrate how to tidy a corpus object with document metadata (e.g., ID, date timestamp, title, etc.) using the **tidy()** method to turn it into text data frame with one-row-per-document, which could then be processed into tidy text for further analysis.

After emphasizing the necessity of these conversion techniques, Chapter 6 adds another valuable piece, the concept of topic modeling, to the systematic puzzle of text analysis based on both existing packages and the suite of tidy tools. Similar to clustering, topic modeling is used as a way to perform unsupervised classification on collections of documents. Based on the **topicmodels** package, the authors introduce the foremost popular algorithm for topic modeling, that is, latent Dirichlet allocation (LDA), and demonstrate a case study of restoring disorganized individual chapters to their original books and using the confusion matrix to visualize where LDA assigned the words from each book and what the most commonly wrongly classified words are. It concludes with the introduction of packages besides **tipicmodels** that can be used for text assignment, for example, the **mallet** package.

Finally, the book comes to three case studies that incorporate all the different tidy text techniques adopted and demonstrate comprehensive illustrations of actual studies based on the tidy text format. Specifically, Chapter 7, **Comparing Twitter Archives**, focuses on the authors' own Twitter archives. Employing the metrics of word frequency, log odds ratio, and so on, they give an explicit demonstration of how to study people's habits in social media based on a tidy data

https://doi.org/10.1017/S1351324920000649 Published online by Cambridge University Press

format. Chapter 8, **Mining NASA Metadata**, is devoted to exploring over 32,000 NASA datasets with metadata. It combines network analysis, tf-idf analysis, and topic modeling to further our understanding of how keywords from NASA datasets are related to title and description fields. In the final chapter, **Analyzing Usenet Text**, the authors analyze a set of 20,000 messages sent to the Usenet bulletin boards in 1993. They combine almost every approach introduced, including tf-idf, topic modeling, sentiment analysis, n-gram, etc., to explore each Usenet newsgroup's text features. In general, the last three chapters are clear, comprehensive, and easy to follow. They provide beginning-to-end examples by combining all of the learned concepts and methods into cohesive ways of processing and understanding texts.

To summarize, the authors first introduce the tidy text format and illustrate what we can do with this structure in text analysis, then explain how to convert back and forth the format between tidy and nontidy, which is vital to bridge the **tidytext** package with other existing tools. Finally, three case studies are carried out, bringing multiple tidy approaches together to demonstrate the practical implementation of what the authors have presented.

Based on the **tidytext** package, the book is concise and compact with detailed code and integrated cases. One of the most remarkable contributions is that the authors expand the tidy text principles to data with unstructured and nontidy format, making it a practical and excellent guide about text mining and data processing. However, it assumes that the readers have a general knowledge of the manipulation of the R software and other tidy tools in R. In this case, some quick tutorials (e.g., Adler 2012; Long and Teetor 2019) and the **tidyverse** toolkit reference (Wickham and Grolemund 2017) are highly recommended to be kept at the reader's side while enjoying the book.

Overall, *Text Mining with R: A Tidy Approach* is clear-cut and inspiring for both novices and those experienced in data science to start new projects that require significant text analysis. The tidy data principles could greatly accelerate text processing workflows in real tasks, providing a simple and efficient format to analyze data, thus leading to versatile applications. It is enormously helpful for readers to get through data processing projects with simplicity, speed, and elegance. This book will serve well anyone interested in learning more about text mining, natural language processing, data science, linguistics, etc. in the big data era through excellent hands-on instructions.

Jianwei Yan
Department of Linguistics, Zhejiang University, No. 866,
Yuhangtang Road, Xihu District, Hangzhou 310058, P. R. China
E-mail: jwyan@zju.edu.cn

References

Adler J. (2012). *R in a Nutshell*, 2nd Edn. Sebastopol, CA: O'Reilly Media.

Long J.D. and **Teetor P.** (2019). *R Cookbook*, 2nd Edn. Sebastopol, CA: O'Reilly Media.

R Core Team. (2020). R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing*. Vienna, Austria. Available at https://www.R-project.org/

Silge J. and **Robinson D.** (2016). tidytext: text mining and analysis using tidy data principles in R. *The Journal of Open Source Software* **1**(3), 37. https://doi.org/10.21105/joss.00037

Wickham H. and **Grolemund G.** (2017). *R for Data Science*. Sebastopol, CA: O'Reilly Media.

https://doi.org/10.1017/S1351324920000649 Published online by Cambridge University Press