Book Review of: LEBART L., PINCEMIN B., and POUDAT C., 2019, Analyse des données textuelles [Analysis of textual data], Quebec City, Presses de l'Université du Québec, 510 pages

Bénédicte Garnier

Lebart L., Pincemin B., and Poudat C., 2019, *Analyse des données textuelles* [Analysis of textual data], Quebec City, Presses de l'Université du Québec, 510 pages.

This book is intended for any researcher with an interest in using statistical tools to analyse texts. It is not only an update of the book *Statistique textuelle*, published in 1994 and co-written by Ludovic Lebart and André Salemmais. It also presents new developments in decision statistics as well as open-source languages, such as Python for string manipulation and R for its statistical environment. The book is divided into 10 chapters. It effectively alternates theory and illustrations with diagrams and screenshots that enhance and clarify the presentation of each subject.

Pedagogically, the book is suitable for various types of readers. From sociologists to statisticians, students and scholars from various disciplines will find precise and clear explanations, accompanied both by figures illustrating the applications of the methods and by the mathematical formulae needed to understand the associated calculations. On the practical level, software and scripts are included as appendices. The authors carefully set out the essential elements for understanding the methods of textual analysis, which they situate in relationship to qualitative analysis, natural language processing, and text mining. Drawn from a number of disciplines, such as statistics, linguistics, discourse analysis, and computer science, the analysis of textual data can be used to synthesize the information contained in large corpora of texts and, using different methods, pursue various objectives: keyword searching, comparing corpora, working with metadata, identifying structures, etc.

The book begins with a section on the history of textual data analysis and its evolution with the 'digital revolution' and the rise of Web data. The authors give examples of textual data: answers to open questions, interviews, political speeches, article titles, message contents, etc. The first chapter presents the stages needed to prepare and transform texts into lexical tables that can then be analysed using software: the creation of units for statistical analysis by splitting texts into smaller pieces (context units), automatic word categorization (distinguishing between function words and lexical words), lemmatization. The second chapter presents the identification of words or texts that are representative of subsets of texts, where the calculation of frequencies or specificities allows them to be compared (by author, by date). The third and final chapter in this section highlights the need for complementary qualitative analyses of texts so that the quantitative results can be interpreted.

The second section is devoted to the multidimensional methods that lie at the heart of textual statistics. These exploratory descriptive approaches—so labelled in contrast to the confirmatory inferential approach of methods based on statistical tests or probabilistic models—include principal components analysis, correspondence analysis, and classification techniques. The three chapters on exploratory statistics drawn from algebra and geometry show how these methods can be used to synthesize the information contained in large data tables:

 DOI: 10.3917/popu.2004.0630

tables of measurements, tables of variable values for individuals, contingency tables, or lexical tables in the context of textual statistics. Co-occurrences are visualized on biplots, which reveal collections of words that tend to be found together, or on dendrograms, drawn from automatic classifications. The words that are characteristic of a class can be used to give it a title or a theme. Various types of classifications, such as hierarchical and partitioning, are detailed and illustrated, as are the complexities of their statistical validation.

The book's third section presents analysis strategies implemented using a combination of textual analysis methods and exploratory methods. The detailed and illustrated examples in this section are of particular interest for the less mathematically inclined reader. One chapter sets out the strengths and limitations of exploratory methods for analysing the content of texts. The authors show how the classification of factors resulting from a correspondence analysis can be used in combination with test values from specificity analysis to identify texts that are characteristic of a class. This section is illustrated by analyses of the relationships between responses to open questions and the characteristics of respondents, and by presidential speeches with topic detection.

The last chapter emphasizes the complementarity of exploratory analyses and confirmatory predictive analyses for theme recognition. One application shows how, by analysing the form of texts based on the distribution of vocabulary (stylometry), texts can be assigned to a category, class, or era. In this case, it is used to attribute a poem to an author. An even more advanced process using discriminant analysis allows the use of both the content and the form of texts to determine their membership in a category or a topic. This method is made possible by a learning phase of topic detection and topic modelling. The authors present an application of the method to coding the sociodemographic category of respondents based on their answers to an open question, asked in three different countries and in different languages.

The book closes with a helpful collection of references to additional resources. In addition to an extensive bibliography and an index of terms used, an appendix presents seven software packages chosen for their methodological approach, their availability in open access, and their graphical interface. Readers can thus choose the tools that best meet their needs based on the strengths and specialities of each one.

Bénédicte Garnier