DOI: 10.1111/biom.13181

Textual data science with R

Mónica Bécue-Bertaut

Boca Raton: CRC Press.

Motivated by my interests in methods to understand the effects of built environment exposures on health, over the last several years I have developed an out of the box collaboration involving engineers, epidemiologists, and—the out of the box part—an anthropologist (NSF grant BCS-1744724). I was intrigued by the idea of "making better numbers" by putting together "big" ethnographic data (reams of photographs and textual data collected during long-term, direct observations of participants' interactions with their neighborhood environment), with more typical "big" quantitative data including measures of the environment and participants' use of the environment, for example, GIS- and GPS-based data (Roberts, LFS, personal communication, May 5, 2017). Thus, when the book "*Textual Data Science with R"* came across my *Biometrics Book Reviews Editor* desk, I snatched the opportunity to review it! Admittedly, this review is written from a novice's perspective when it comes to textual data, and therefore, targeted toward other interested beginners.

The book begins by reviewing nomenclature used in textual data science in Chapter 1. The definitions of *corpus* (what I understood to be the "textual dataset"), *document* (the unit of analysis), and *word* (akin to a set of outcomes for each unit) are introduced, as is the concept of *encoding*—the process of converting the textual dataset into a *lexical table* of *documents* (rows) by *words* (columns). The *encoding* step is very important, since during this step the researcher will decide how the *corpus* is divided into *documents* and the *types of words* that will be used (eg, conventionally defined words vs *lemmas* vs *repeated segments*). Once the data preprocessing is complete (eg, keeping only words that occur with at least a certain frequency), quantitative analysis begins. Chapter 2 explains the method of correspondence analysis—a close cousin of principal components analysis—and how it can be applied to *aggregated lexical tables*, which are composed of words (columns) within *groups of documents* (rows). Similar to principal component scores, the result of correspondence analysis is a set of numerical indices that are used to interpret the tendencies or broader concepts represented in the textual data. While Chapter 2 uses somewhat of a toy example to illustrate the method and key ideas and ensure the reader can absorb them, real examples of correspondence analysis are covered in Chapter 3. This chapter also discusses *direct analysis*, that is, analysis of lexical tables formed without grouping *documents*, and continues the discussion started in Chapter 2 on using graphical displays to interpret analysis results.

With the foundations laid out in Chapters 1-3, more advanced approaches are considered in Chapters 4-6. Methods to cluster documents, including direct partitioning and hierarchical clustering among others, are discussed in Chapter 4. All clustering methods discussed use Euclidean distance between documents' coordinates identified using correspondence analysis applied in prior data analysis steps. Chapter 5 focuses on approaches to identify *characteristic words or documents* that can be used to exemplify the meaning of clusters, and other purposes. Given my prior work on latent variable models, I found Chapter 6 most exciting as it covered the development of *multiple factor analysis* for textual data. In the final chapter, the author and a handful of collaborators

explain several more involved, real-data examples that apply the range of methods discussed in the book. The chapter also includes a guide for how to describe the analyses conducted when writing journal articles.

There were several components of the book that I found very useful, and others that I wish had been there. First, as implied by the book title, each chapter includes `R` code using the package `Xplortext` to reproduce the analyses described. Second, almost every chapter ends with a summary of take-home messages, sections that I found necessary to help me retain the large amount of new terminology introduced in the book, as well as check my understanding of the broader concepts covered in each chapter. I wish the book had included references within each chapter to point the reader to further reading on the contents of the chapter. Thankfully, a nice set of essential/seminal works is included at the end of the book. I also wish the book had pointed to open research questions in the field, although I recognize this is getting greedy on my part as clearly this book is a how-to-guide for state-of-the-art methods in the area.

I look forward to using my newly obtained research tools for textual data science. Although it still remains to be told whether "better numbers" will mean better exposure measures, better estimators, better representations of neighborhood causal mechanisms, or all of the above, what is definitely true is that the book has helped me to understand existing approaches of how to quantitatively analyze textual data, and thus, bringing our team closer to making such numbers.

Brisa N. Sánchez

Department of Epidemiology and Biostatistics, Dornsife School of Public Health, Drexel University, Philadelphia, Pennsylvania
Email: bns48@drexel.edu