Journal of Statistical Software

November 2015, Volume 68, Book Review 3. *doi: 10.18637/jss.v068.b03*

Reviewer: Stefano Maria Iacus
University of Milan

Automated Data Collection with R – A Practical Guide to Web Scraping and Text Mining

Simon Munzert, Christian Rubba, Peter Meißner, Dominic Nyhuis
John Wiley & Sons, Chichester, 2014.
ISBN 978-1-118-83481-7. 480 pp. GBP 50.00 (H).
http://www.r-datacollection.com/

Extracting content from diversified web resources, cleaning up the raw data, preparing them for the statistical analysis and actually performing the analysis it is far from being a simple task. A single researcher must have competences in different fields including web technologies and services, authentication strategies, regular expressions and text parsing, different encoding systems, efficient data storage and data base structures, machine learning and advanced statistical techniques to mention a few. Fortunately all of them are available in R today, directly or through additional layers, but this is still not enough. The workflow that works for, e.g. Twitter, does not necessarily work for Facebook. The same workflow can break down at any time when the provider of some service changes its API, the authentication method or the data access policies. And while Twitter, Facebook, etc., are more or less stable, any other digital source from the web may have a very diversified structure, may result in incomplete or malformed data, and so on.

Having deep knowledge of these technologies or, at least, understanding the logic behind them, is important to keep one's software updated, stable and robust. Then comes statistics. There is not yet a standard or definitive method for sentiment analysis, opinion mining or, more generally, text mining. A plethora of approaches exist and they mostly depend on the language, the context (are we talking about politics or movie reviews?), the source (e.g. blogs are lengthy, tweets are short), the goal of the analysis (classification, topic discovery, etc).

This book is then a gentle introduction to widely used web technologies and text mining algorithms for researcher or students in the social sciences with little or no background on these matters. R knowledge, on the contrary, is mandatory.

The book is composed of seventeen chapters further collected into three parts. The first part is about the basics on different web technologies, mark up languages, data bases and string manipulation. The second part focuses on the actual web scraping world and the last six chapters compose the third part on text mining case studies.

The expert R user might find the first eight chapters a bit too verbose, but will still find some valuable hints. More in details, the eight chapters go through HTML: the markup language for web pages; XML and JSON: a key-value system to store content in text format, usually returned from the API of different services like Facebook and Twitter; XPath, not unrelated to XML; HTTP, the internet protocol to send and retrieve data on the web; Ajax and Javascript, for dynamic web pages and finally SQL language for storing content after data retrieval. If you know all of these, you can skip this part and use it as quick reference only.

The second part of the book is really the core. I have to say that some parts are treated a bit too fast. For example, authentication is not necessarily as simple as described in the book especially if one wants to include it in a non-interactive working environment. This does not mean that what is explained is not correct, but that the reader of this book will probably need additional bits of information before really start to work. This topic is somewhat related to Chapter 11 of scripting and batch processing. Another problem concerns the policies of the service providers and API limitations. This may really hurt a productive environment as things may suddenly stop to work if, for example, one does too many requests per unit of time, etc.

This book is already more than 400 pages and so it is impossible for the authors to really describe all techniques, but I expected a bit more on these two points and how to overcome these limitations without infringing the providers' policies. On the other hand, it is very welcome that Chapter 9 takes into account the serious problems of data ownership and privacy, something usually not treated in many other books. This has implication to science of course (think, e.g., about reproducible research) and the authors present this issue correctly before the statistical analysis part.

Chapter 10 is devoted to statistics and the different frameworks and packages available in R. This chapter is not exhaustive for obvious reasons but the basic pillars are there: for example, the notions of stemming, ngrams, document-term matrixes, and sparsity are well discussed. The reader can find something about unsupervised statistical analysis like topics discovery through latent discriminant analysis, and much more on machine learning for sentiment analysis. Chapter 11 concludes the second part of the book exposing some ideas about batch environments and project management.

Most of these techniques are exploited in the last part of the book devoted to case studies. The case studies consider text and opinion mining applications but also the smart use of meta data associated to digital posts like, for example, the network of social connections of Twitter or Facebook users and the geo-tagging of these data. Therefore, the reader can find also information and step by step examples of network analysis and maps drawing in R. Both are nontrivial tasks.

In summary, this book is worth having in one's bookshelf if one wants to approach this world. It is not a gentle introduction to R nor a deep treatment of each subject mentioned in the index, but it is an almost complete guide to what is possible to do in a reasonably short time using R and its huge library of packages. If one does not want to care about data scraping and text parsing but wants to focus on practical analysis, most of the second and the third part of the book can serve the goal provided that someone else prepares the data for the analysis.

Reviewer:

Stefano M. Iacus
Department of Economics, Management and Quantitative Methods
University of Milan
Via Conservatorio 7
I-20123 Milan, Italy
E-mail: stefano.iacus@unimi.it

Journal of Statistical Software http://www.jstatsoft.org/
published by the Foundation for Open Access Statistics http://www.foastat.org/
November 2015, Volume 68, Book Review 3 *Published:* 2015-11-23
doi:10.18637/jss.v068.b03