1.	Record Nr.	UNINA9910823262603321
	Autore	Munzert Simon
	Titolo	Automated data collection with R : a practical guide to web scraping and text mining / / Simon Munzert [and three others]
	Pubbl/distr/stampa	Chichester, England : , : Wiley, , 2015 ©2015
	ISBN	1-118-83480-1 1-118-83473-9 1-118-83478-X
	Descrizione fisica	1 online resource (XXII, 453 p.)
	Classificazione	COM021030
	Disciplina	006.3/12
	Soggetti	Data mining Automatic data collection systems Social sciences - Research - Data processing R (Computer program language)
	Lingua di pubblicazione	Inglese
	Formato	Materiale a stampa
	Livello bibliografico	Monografia
	Note generali	Description based upon print version of record.
	Nota di bibliografia	Includes bibliographical references and index.
	Nota di contenuto	Machine generated contents note: Dedication Table of Contents List of Figures List of Tables Preface 1 Introduction 1.1 Case Study: World Heritage Sites in Danger 1.2 Some Remarks on Web Data Quality 1.3 Technologies for Disseminating, Extracting and Storing Web Data 1.3.1 Technologies for disseminating content on the Web 1.4 Structure of the Book Part One A Primer on Web and Data Technologies 2 HTML 2.1 Browser Presentation and Source Code 2.2 Syntax Rules 2.3 Tags and Attributes 2.4 Parsing Summary Further Reading Problems 3 XML and JSON 3.1 A Short Example XML Document 3.2 XML Syntax Rules 3.3 When Is an XML Document Well-formed or Valid? 3.4 XML Extensions and Technologies 3.5 XML and R in Practice 3.6 A Short Example JSON Document 3.7 JSON Syntax Rules 3.8 JSON and R in Practice Summary Further Reading Problems 4 XPath 4.1 XPath - a Querying Language for Web Documents 4.2 Identifying Node Sets with XPath 4.3 Extracting Node Elements Summary Further Reading Problems 5 HTTP 5.1 HTTP Fundamentals 5.2 Advanced Features of HTTP 5.3 Protocols beyond HTTP 5.4 HTTP in Action Summary Further Reading Problems 6 AJAX

6.1 JavaScript 6.2 XHR 6.3 Exploring AJAX with Web Developer Tools Summary Further Reading Problems 7 SQL and Relational Databases 7.1 Overview and Terminology 7.2 Relational Databases 7.3 SQL: a Language to Communicate with Databases 7.4 Databases in Action Summary Further Reading Problems 8 Regular Expressions and String Functions 8.1 Regular Expressions 8.2 String Processing 8.3 A Word on Character Encodings Summary Further Reading Problems Part Two A Practical Toolbox for Web Scraping and Text Mining 9 Scraping the Web 9.1 Retrieval Scenarios 9.2 Extraction Strategies 9.3 Web Scraping: Good Practice 9.4 Valuable Sources of Inspiration Summary Further Reading Problems 10 Statistical Text Processing 10.1 The running example: classifying press releases of the British government 10.2 Processing Textual Data 10.3 Supervised Learning Techniques 10.4 Unsupervised Learning Techniques Summary Further reading 11 Managing Data Projects 11.1 Interacting with the File System 11.2 Processing Multiple Documents/Links 11.3 Organizing Scraping Procedures 11.4 Executing R Scripts on a Regular Basis Part Three A Bag of Case Studies 12 Collaboration Networks in the U.S. Senate 12.1 Information on the Bills 12.2 Information on the Senators 12.3 Analyzing the network structure 12.4 Conclusion 13 Parsing Information from Semi-Structured Documents 13.1 Downloding Data from the FTP Server 13.2 Parsing Semi-Structured Text Data 13.3 Visualizing station and temperature data 14 Predicting the 2014 Academy Awards using Twitter 14.1 Twitter APIs: Overview 14.2 Twitter-based Forecast of the 2014 Academy Awards 14.3 Conclusion 15 Mapping the Geographic Distribution of Names 15.1 Developing a Data Collection Strategy 15.2 Web Site Inspection 15.3 Data Retrieval and Information Extraction 15.4 Mapping Names 15.5 Automating the Process 15.6 Summary 16 Gathering Data on Mobile Phones 16.1 Page Exploration 16.2 Scraping Procedure 16.3 Graphical Analysis 16.4 Data storage 17 Analyzing Sentiments of Product Reviews 17.1 Introduction 17.2 Collecting the data 17.3 Analyzing the Data 17.4 Conclusion References Bibliography Indices General Index Package Index Function Index .

Sommario/riassunto

"This book provides a unified framework of web scraping and information extraction from text data with R for the social sciences"--