1. Record Nr. UNINA9911020417803321

Autore	Matter Ulrich
Titolo	An introduction to web mining : with applications in R / / Ulrich Matter
Pubbl/distr/stampa	Cham : , : Springer, , [2025] ©2025
ISBN	9783031966385
Descrizione fisica	1 online resource (xxi, 251 pages) : illustrations
Collana	Use R!, , 2197-5744
Disciplina	001.433
Soggetti	Multimedia data mining R (Computer program language) Social sciences - Statistical methods Statistics Methodology of Data Collection and Processing Data Mining and Knowledge Discovery Statistics in Social Sciences, Humanities, Law, Education, Behavorial Sciences, Public Policy Statistics in Business, Management, Economics, Finance, Insurance Mineria de dades R (Llenguatge de programació) Estadística Metodologia de les ciències socials Llibres electrònics
Lingua di pubblicazione	Inglese
Formato	Materiale a stampa
Livello bibliografico	Monografia
Nota di bibliografia	Includes bibliographical references and index.
Nota di contenuto	- Part I: Context, Relevance, and the Basics -- 1. Introduction -- 2. The Internet as a Data Source -- Part II: Web Technologies and Automated Data Extraction -- 3. Web 1.0 Technologies: The Static Web -- 4. Web Scraping: Data Extraction from Websites -- 5. Web 2.0 Technologies: The Programmable/Dynamic Web -- 6. Extracting Data From The Programmable Web -- 7. Data Extraction from Dynamic Websites -- Part III: Advanced Topics in Web Mining -- 8. Web Mining Programs -- 9. Crawler Implementation -- 10. Appearance and Authentication -- 11. Scaling Web Mining in the Cloud -- 12. AI Tools for Web Mining:

Overview and Outlook -- Part IV: Ethical, Legal, and Scientific Rigor -- 13. Ethics and Legal Considerations -- 14. Web Mining and Scientific Rigor.

Sommario/riassunto

This book is devoted to the art and science of web mining — showing how the world's largest information source can be turned into structured, research-ready data. Drawing on many years of teaching graduate courses on Web Mining and on numerous large-scale research projects in web mining contexts, the author provides clear explanations of key web technologies combined with hands-on R tutorials that work in the real world — and keep working as the web evolves. Through the book, readers will learn how to - scrape static and dynamic/JavaScript-heavy websites - use web APIs for structured data extraction from web sources - build fault-tolerant crawlers and cloud-based scraping pipelines - navigate CAPTCHAs, rate limits, and authentication hurdles - integrate AI-driven tools to speed up every stage of the workflow - apply ethical, legal, and scientific guidelines to their web mining activities Part I explains why web data matters and leads the reader through a first “hello-scrape” in R while introducing HTML, HTTP, and CSS. Part II explores how the modern web works and shows, step by step, how to move from scraping static pages to collecting data from APIs and JavaScript-driven sites. Part III focuses on scaling up: building reliable crawlers, dealing with log-ins and CAPTCHAs, using cloud resources, and adding AI helpers. Part IV looks at ethical, legal, and research standards, offering checklists and case studies, enabling the reader to make responsible choices. Together, these parts give a clear path from small experiments to large-scale projects. This valuable guide is written for a wide readership — from graduate students taking their first steps in data science to seasoned researchers and analysts in economics, social science, business, and public policy. It will be a lasting reference for anyone with an interest in extracting insight from the web — whether working in academia, industry, or the public sector.