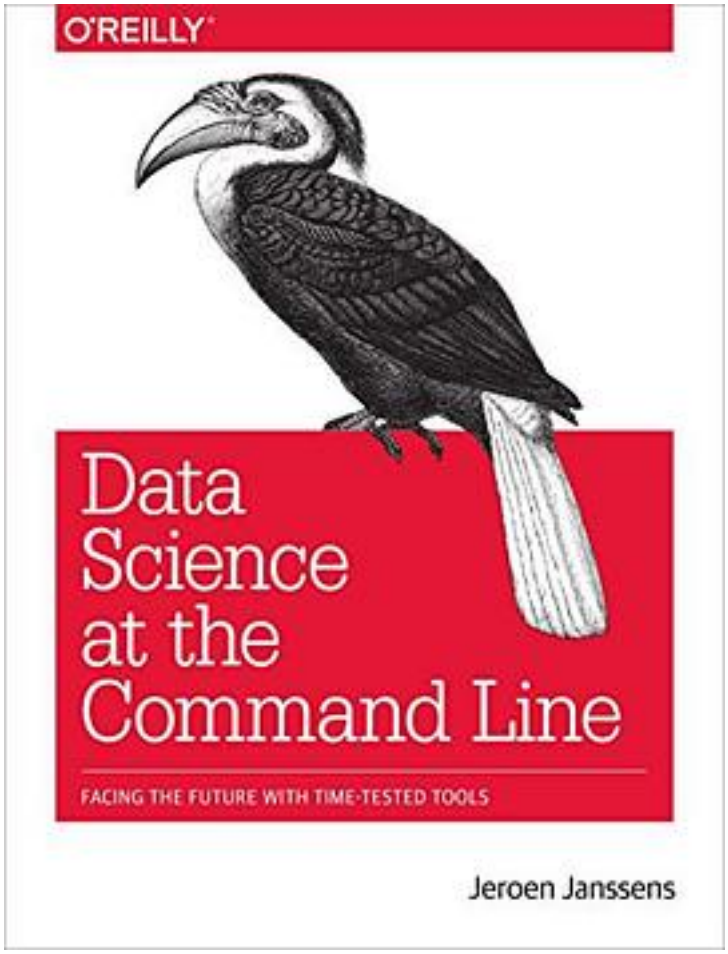
Data Science at the Command Line



Data Science at the Command Line_下载链接1_

著者:Jeroen Janssens

出版者:O'Reilly Media

出版时间:2014-10-20

装帧:Paperback

isbn:9781491947852

This hands-on guide demonstrates how the flexibility of the command line can help you become a more efficient and productive data scientist. You’ ll learn how to combine small, yet powerful, command-line tools to quickly obtain, scrub, explore,

and model your data.

To get you started—whether you’ re on Windows, OS X, or Linux—author Jeroen Janssens introduces the Data Science Toolbox, an easy-to-install virtual environment packed with over 80 command-line tools.

Discover why the command line is an agile, scalable, and extensible technology. Even if you’ re already comfortable processing data with, say, Python or R, you’ ll greatly improve your data science workflow by also leveraging the power of the command line.

●Obtain data from websites, APIs, databases, and spreadsheets

●Perform scrub operations on plain text, CSV, HTML/XML, and JSON

●Explore data, compute descriptive statistics, and create visualizations

●Manage your data science workflow using Drake

●Create reusable tools from one-liners and existing Python or R code

●Parallelize and distribute data-intensive pipelines using GNU Parallel

●Model data with dimensionality reduction, clustering, regression, and classification algorithms

作者介绍:

Jeroen is a Senior Data Scientist at YPlan in New York City. He has an M.Sc. in Artificial Intelligence and a Ph.D. in Machine Learning. He has authored a book titled Data Science at the Command Line, which has just been published by O'Reilly. Jeroen enjoys biking the Brooklyn Bridge, building tools, and eating stroopwafels.

目录: Chapter 1 Introduction
Overview
Data Science Is OSEMN
Intermezzo Chapters
What Is the Command Line?
Why Data Science at the Command Line?
A Real-World Use Case
Further Reading
Chapter 2 Getting Started
Overview
Setting Up Your Data Science Toolbox
Essential Concepts and Tools
Further Reading
Chapter 3 Obtaining Data
Overview
Copying Local Files to the Data Science Toolbox
Decompressing Files
Converting Microsoft Excel Spreadsheets
Querying Relational Databases

Downloading from the Internet
Calling Web APIs
Further Reading
Chapter 4 Creating Reusable Command-Line Tools
Overview
Converting One-Liners into Shell Scripts
Creating Command-Line Tools with Python and R
Further Reading
Chapter 5 Scrubbing Data
Overview
Common Scrub Operations for Plain Text
Working with CSV
Working with HTML/XML and JSON
Common Scrub Operations for CSV
Further Reading
Chapter 6 Managing Your Data Workflow
Overview
Introducing Drake
Installing Drake
Obtain Top Ebooks from Project Gutenberg
Every Workflow Starts with a Single Step
Well, That Depends
Rebuilding Specific Targets
Discussion
Further Reading
Chapter 7 Exploring Data
Overview
Inspecting Data and Its Properties
Computing Descriptive Statistics
Creating Visualizations
Further Reading
Chapter 8 Parallel Pipelines
Overview
Serial Processing
Parallel Processing
Distributed Processing
Discussion
Further Reading
Chapter 9 Modeling Data
Overview
More Wine, Please!
Dimensionality Reduction with Tapkee
Clustering with Weka
Regression with SciKit-Learn Laboratory
Classification with BigML
Further Reading
Chapter 10 Conclusion
Let’s Recap
Three Pieces of Advice
Where to Go from Here?
Getting in Touch
・・・・・・(收起)

Data Science at the Command Line_下载链接1_

标签

数据分析

data

计算机

数据挖掘

编程

cs

Python

计算机科学

评论

一种个人化轻量级的数据处理思路

书中提到的好多tool都不是Linux中原生的，快速翻了大半本，觉得一般

非常精彩，真有种作者在写我内心想法的感觉！推荐给所有人，特别是如果你不习惯命令行操作，这本书真的是提高工作效率的瑞士军刀。

在电脑上细看了前4章，后续是浏览。1.
最新版本已经使用docker来建「虚拟环境」了，2014年的版本是用VirtualBox 2.
数据处理的步骤还是那些：获取，数据清洗，可视化，建模，解释
3.命令行工具很强大，目测可以完成常用的数据操作。4
高级操作要借助别的语言，可视化是用的R，建模是用的Tapkee
5.了解了一些常用的linux命令，这个其实是主要目的，后续可以在实践中遇到了之后再多了解 6. 还是老老实实把python学扎实最实用

不少工具不是linux 原生的..

命令行的强大毋庸置疑，字符界面的简洁高效也叫人觉得异常舒服，然而书中介绍的这个工具不怎么在意，还是喜欢Anaconda，还有书里所采用的环境设置用的是Vagrant，我比较偏好Docker

---
kind of outdated

刚开始读，介绍的全是近年来新开发的工具。手边没电脑，读起来很陌生啊 :(

讲那么多csv、json，我用不上啊……

私有工具差评

最好的数据建模那一章没怎么看懂。本书讲述了如何在命令行进行数据获取和格式化以为建模分析做准备，但在重要的建模分析章节缺很精简不清晰，似乎定位是入门的书却预定读者已经有了很多相关知识。

命令行虽好，但没必要一定都要用，但理解后帮助很大

可操作的展现数据的初步处理

非常好的一本书，特别推荐。和我十几年的数据分析经验非常吻合。不是所有的工具都适合每个人，但是思想非常契合。因为每个人的分析数据差异非常大，完全可以自己定制自己的工具集。经验和思想的力量。

Data Science at the Command Line_下载链接1_

书评

在电脑上细看了前4章。 1.
最新版本已经使用docker来建「虚拟环境」了，2014年的版本是用VirtualBox。最新的在线版本 https://www.datascienceatthecommandline.com/ 2.
数据处理的步骤还是那些：获取，数据清洗，可视化，建模，解释
3.命令行工具很强大，目测可以完成常用的数...

本书集实用性和先进性于一身，为数据分析人员使用命令行这个灵活的工具提供了重要参考。作者讲解了众多实用的命令行工具，以及如何使用它们高效地获取、清洗、探索和建模数据。无论你使用Windows、OS
X，还是Linux，都可以安装包含80多个命令行工具的“数据科学工具箱”，迅速...

Data Science at the Command Line_下载链接1_