

《命令行中的数据科学》

图书基本信息

书名：《命令行中的数据科学》

13位ISBN编号：9787115391688

出版时间：2015-5

作者：[荷] Jeroen Janssens

页数：188

译者：王晓伟,刘 峰

版权说明：本站所提供下载的PDF图书仅提供预览和简介以及在线试读，请支持正版图书。

更多资源请访问：www.tushu111.com

《命令行中的数据科学》

内容概要

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”，迅速建立自己的数据分析环境。无论你是否已经习惯于使用Python或R语言，都能够通过本书体会到使用命令行的快捷、灵活与伸缩自如。

《命令行中的数据科学》

作者简介

Jeroen Janssens

爱思唯尔（世界领先的科技及医学出版公司）首席数据科学家，曾是纽约YPlan公司高级数据科学家。专门从事机器学习、异常检测和数据可视化。在荷兰马斯特里赫特大学获得人工智能硕士学位，在荷兰蒂尔堡大学获得机器学习博士学位。他热衷于创建数据科学的开源工具，个人网站是<http://jeroenjanssens.com/>。

书籍目录

前言	XIII
第1章 简介	1
1.1 概述	1
1.2 数据科学就是OSEMN	2
1.2.1 数据获取	2
1.2.2 数据清洗	2
1.2.3 数据探索	3
1.2.4 数据建模	3
1.2.5 数据解释	3
1.3 插入的几章	4
1.4 什么是命令行	4
1.5 为什么用命令行做数据科学工作	6
1.5.1 命令行的灵活性	6
1.5.2 命令行可增强	6
1.5.3 命令行可扩展	7
1.5.4 命令行可扩充	7
1.5.5 命令行无处不在	7
1.6 一个现实用例	8
1.7 延伸阅读	11
第2章 入门指南	13
2.1 概述	13
2.2 设置数据科学工具箱	13
2.2.1 步骤1：下载和安装VirtualBox	14
2.2.2 步骤2：下载和安装Vagrant	14
2.2.3 步骤3：下载并启动数据科学工具箱	14
2.2.4 步骤4：登录（Linux 和Mac OS X）	16
2.2.5 步骤4：登录（微软Windows）	16
2.2.6 步骤5：关闭或重启	16
2.3 必要的概念和工具	17
2.3.1 环境	17
2.3.2 运行命令行工具	18
2.3.3 五类命令行工具	19
2.3.4 命令行工具的组合	21
2.3.5 输入和输出重定向	22
2.3.6 处理文件	23
2.3.7 寻求帮助	24
2.4 延伸阅读	26
第3章 数据获取	27
3.1 概述	27
3.2 将本地文件复制到数据科学工具箱	28
3.2.1 本地数据科学工具箱	28
3.2.2 远程数据科学工具箱	28
3.3 解压缩文件	29
3.4 微软Excel电子表格的转换	30
3.5 查询关系数据库	32
3.6 从互联网下载	33
3.7 调用Web API	35

3.8 延伸阅读	36
第4章 创建可重用的命令行工具	37
4.1 概述	38
4.2 将单行转变为shell脚本	38
4.2.1 步骤1：复制和粘贴	39
4.2.2 步骤2：添加执行权限	40
4.2.3 步骤3：定义shebang	41
4.2.4 步骤4：删除固定的输入	42
4.2.5 步骤5：参数化	42
4.2.6 步骤6：扩展PATH	43
4.3 用Python和R创建命令行工具	44
4.3.1 移植shell脚本	45
4.3.2 处理来自标准输入的流数据	46
4.4 延伸阅读	47
第5章 数据清洗	49
5.1 概述	50
5.2 纯文本的常见清洗操作	50
5.2.1 行过滤	50
5.2.2 值提取	54
5.2.3 值替换和删除	55
5.3 处理CSV	56
5.3.1 主体、头部和列	56
5.3.2 对CSV执行SQL查询	60
5.4 处理HTML/XML和JSON	61
5.5 CSV的常见清洗操作	65
5.5.1 列的提取和重排序	65
5.5.2 行过滤	66
5.5.3 列合并	67
5.5.4 多个CSV文件的合并	70
5.6 延伸阅读	73
第6章 管理数据工作流	75
6.1 概述	76
6.2 Drake简介	76
6.3 Drake的安装	76
6.4 获取古腾堡计划中下载最多的电子书	78
6.5 所有工作流都从单个步骤开始	79
6.6 具体情况具体对待	81
6.7 重新构建具体目标	82
6.8 讨论	83
6.9 延伸阅读	83
第7章 数据探索	85
7.1 概述	85
7.2 检查数据及其属性	86
7.2.1 确定有无数据头	86
7.2.2 检查所有数据	86
7.2.3 特征名称和数据类型	87
7.2.4 唯一标识、连续变量和因子	89
7.3 计算描述性统计信息	90
7.3.1 使用csvstat	90

7.3.2	在命令行中通过Rio使用R	92
7.4	生成可视化图形	95
7.4.1	介绍Gunplot和feedgnuplot	95
7.4.2	介绍ggplot2	97
7.4.3	直方图	99
7.4.4	条形图	101
7.4.5	密度图	102
7.4.6	箱线图	103
7.4.7	散点图	103
7.4.8	折线图	105
7.4.9	总结	106
7.5	延伸阅读	106
第8章	并行管道	107
8.1	概述	108
8.2	串行处理	108
8.2.1	对数字进行遍历	108
8.2.2	对行进行遍历	109
8.2.3	对文件进行遍历	110
8.3	并行处理	111
8.3.1	GNU Parallel介绍	112
8.3.2	指定输入	113
8.3.3	控制并发任务的个数	114
8.3.4	记录日志和输出	115
8.3.5	创建并行工具	116
8.4	分布式处理	117
8.4.1	获得运行中的AWS EC2实例列表	117
8.4.2	在远程机器上运行命令	118
8.4.3	在远程机器间分发本地数据	119
8.4.4	在远程机器上处理文件	120
8.5	讨论	123
8.6	延伸阅读	123
第9章	数据建模	125
9.1	概述	126
9.2	更多的酒，来吧！	126
9.3	用Tapkee降维	129
9.3.1	介绍Tapkee	130
9.3.2	安装Tapkee	130
9.3.3	线性和非线性映射	130
9.4	用Weka 聚类	132
9.4.1	介绍Weka	132
9.4.2	在命令行里改进Weka	132
9.4.3	在CSV和ARFF格式之间转换	136
9.4.4	比较三种聚类算法	136
9.5	通过SciKit-Learn Laboratory进行回归	139
9.5.1	准备数据	139
9.5.2	运行实验	139
9.5.3	解析结果	140
9.6	用BigML分类	141
9.6.1	生成均衡的训练和测试数据集	141

《命令行中的数据科学》

9.6.2	调用API	143
9.6.3	检查结果	143
9.6.4	小结	144
9.7	延伸阅读	144
第10章	总结	145
10.1	让我们回顾一下	145
10.2	三条建议	146
10.2.1	有耐心	146
10.2.2	有所创新	146
10.2.3	肯于实践	147
10.3	接下来做什么	147
10.3.1	API	147
10.3.2	shell 编程	147
10.3.3	Python、R 和SQL	147
10.3.4	数据解释	148
10.4	联系方式	148
附录A	命令行工具列表	149
附录B	参考文献	167
作者介绍		169
封面介绍		169

《命令行中的数据科学》

精彩短评

- 1、介绍一些数据科学的命令行工具，比较浅，2天就看完了。
- 2、但还是感觉缺乏系统性，不统一.....
- 3、半翻半看的浏览完了。就是命令行操作。2017年1月初离开北京前图书馆看完的。
- 4、我是看完了英文版扫了扫中文版，说实话作者的英文版语气是如此轻松愉快以至于读起来非常舒服和畅快，但是中文版翻译的就和便秘是一个感觉，我知道这是翻译为了所谓的标准而做的妥协或者是什么。我参加过一个翻译，感觉编辑就是个二百五，自己不懂技术，抓着几个小词叽叽哇哇半天，其实根本就是不是重点，重点是内容和思路，可惜碰到不是这个学科的编辑彻底瞎了。我觉得这本书就是这么毁了的。还是推荐看原版。
- 5、每一个人都应该掌握一些基本数据科学分析工具。
- 6、相当于知识索引，面面俱到浅尝则止。推荐的thinking with data很不错
- 7、介绍了很多命令行工具，可以对常见的文本数据和网络数据进行处理 感觉很有意思
- 8、rubbish.....
- 9、看过对应的英文原版。
- 10、很实用
- 11、基本的数据科学分析工具介绍
- 12、介绍了一些数据处理的命令，方法，讲解的不够细。
- 13、稍微看了一下，主要是用命令行的形式来处理分析数据

《命令行中的数据科学》

版权说明

本站所提供下载的PDF图书仅提供预览和简介，请支持正版图书。

更多资源请访问:www.tushu111.com