

《干净的数据：数据清洗入门与实践》

图书基本信息

书名：《干净的数据：数据清洗入门与实践》

13位ISBN编号：9787115420475

出版时间：2016-5

作者：[美] Megan Squire

页数：200

译者：任政委

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.tushu000.com

《干净的数据：数据清洗入门与实践》

内容概要

数据清洗是数据挖掘与分析过程中不可缺少的一个环节，但因为数据类型极其复杂，传统的清洗脏数据工作单调乏味且异常辛苦。如果能利用正确的工具和方法，就可以让数据清洗工作事半功倍。

本书从文件格式、数据类型、字符编码等基本概念讲起，通过真实的示例，探讨如何提取和清洗关系型数据库、网页文件和PDF文档中的数据。最后提供了两个真实的项目，让读者将所有数据清洗技术付诸实践，完成整个数据科学过程。

如果你是一位数据科学家，或者从事数据科学工作，哪怕是位新手，只要对数据清洗有兴趣，那么本书就适合你阅读！

《干净的数据：数据清洗入门与实践》

作者简介

作者简介：

Megan Squire

依隆大学计算科学专业教授，主要教授数据库系统、Web开发、数据挖掘和数据科学课程。有二十年的数据收集与清洗经验。她还是FLOSSmole研究项目的领导者，致力于收集与分析数据，以便研究免费软件、自由软件和开源软件的开发。

译者简介：

任政委

辽宁滨城大连现役程序员一枚，长期从事一线软件开发工作，近年来为成为一名“思路清晰”“视角独特”“不搞办公室政治”“输出有生命力代码”“凭借技术知识普惠初中级IT从业者”的终身制全栈式程序员而不懈努力。曾经翻译《Oracle PL/SQL攻略》一书，并希望这本《干净的数据》能够为奋战在IT前线上的各界小伙伴们带来日常工作之外的另类体验。微信号：KNIGHTRCOM

书籍目录

第1章 为什么需要清洗数据	1
1.1 新视角	1
1.2 数据科学过程	2
1.3 传达数据清洗工作的内容	3
1.4 数据清洗环境	4
1.5 入门示例	5
1.6 小结	9
第2章 基础知识——格式、类型与编码	11
2.1 文件格式	11
2.1.1 文本文件与二进制文件	11
2.1.2 常见的文本文件格式	14
2.1.3 分隔格式	14
2.2 归档与压缩	20
2.2.1 归档文件	20
2.2.2 压缩文件	21
2.3 数据类型、空值与编码	24
2.3.1 数据类型	25
2.3.2 数据类型间的相互转换	29
2.3.3 转换策略	30
2.3.4 隐藏在数据森林中的空值	37
2.3.5 字符编码	41
2.4 小结	46
第3章 数据清洗的老黄牛——电子表格和文本编辑器	47
3.1 电子表格中的数据清洗	47
3.1.1 Excel的文本分列功能	47
3.1.2 字符串拆分	51
3.1.3 字符串拼接	51
3.2 文本编辑器里的数据清洗	54
3.2.1 文本调整	55
3.2.2 列选模式	56
3.2.3 加强版的查找与替换功能	56
3.2.4 文本排序与去重处理	58
3.2.5 Process Lines Containing	60
3.3 示例项目	60
3.3.1 第一步：问题陈述	60
3.3.2 第二步：数据收集	60
3.3.3 第三步：数据清洗	61
3.3.4 第四步：数据分析	63
3.4 小结	63
第4章 讲通用语言——数据转换	64
4.1 基于工具的快速转换	64
4.1.1 从电子表格到CSV	65
4.1.2 从电子表格到JSON	65
4.1.3 使用phpMyAdmin从SQL语句中生成CSV或JSON	67
4.2 使用PHP实现数据转换	69
4.2.1 使用PHP实现SQL到JSON的数据转换	69

《干净的数据：数据清洗入门与实践》

4.2.2	使用PHP实现SQL到CSV的数据转换	70
4.2.3	使用PHP实现JSON到CSV的数据转换	71
4.2.4	使用PHP实现CSV到JSON的数据转换	71
4.3	使用Python实现数据转换	72
4.3.1	使用Python实现CSV到JSON的数据转换	72
4.3.2	使用csvkit实现CSV到JSON的数据转换	73
4.3.3	使用Python实现JSON到CSV的数据转换	74
4.4	示例项目	74
4.4.1	第一步：下载GDF格式的Facebook数据	75
4.4.2	第二步：在文本编辑器中查看GDF文件	75
4.4.3	第三步：从GDF格式到JSON格式的转换	76
4.4.4	第四步：构建D3图	79
4.4.5	第五步：把数据转换成Pajek格式	81
4.4.6	第六步：简单的社交网络分析	83
4.5	小结	84
第5章	收集并清洗来自网络的数据	85
5.1	理解HTML页面结构	85
5.1.1	行分隔模型	86
5.1.2	树形结构模型	86
5.2	方法一：Python和正则表达式	87
5.2.1	第一步：查找并保存实验用的Web文件	88
5.2.2	第二步：观察文件内容并判定有价值的数	88
5.2.3	第三步：编写Python程序把数据保存到CSV文件中	89
5.2.4	第四步：查看文件并确认清洗结果	89
5.2.5	使用正则表达式解析HTML的局限性	90
5.3	方法二：Python和BeautifulSoup	90
5.3.1	第一步：找到并保存实验用的文件	90
5.3.2	第二步：安装BeautifulSoup	91
5.3.3	第三步：编写抽取数据用的Python程序	91
5.3.4	第四步：查看文件并确认清洗结果	92
5.4	方法三：Chrome Scraper	92
5.4.1	第一步：安装Chrome扩展Scraper	92
5.4.2	第二步：从网站上收集数据	92
5.4.3	第三步：清洗数据	94
5.5	示例项目：从电子邮件和论坛中抽取数据	95
5.5.1	项目背景	95
5.5.2	第一部分：清洗来自Google Groups电子邮件的数据	96
5.5.3	第二部分：清洗来自网络论坛的数据	99
5.6	小结	105
第6章	清洗PDF文件中的数据	106
6.1	为什么PDF文件很难清洗	106
6.2	简单方案——复制	107
6.2.1	我们的实验文件	107
6.2.2	第一步：把我们需要数据复制出来	108
6.2.3	第二步：把复制出来的数据粘贴到文本编辑器中	109
6.2.4	第三步：轻量级文件	110
6.3	第二种技术——pdfMiner	111
6.3.1	第一步：安装pdfMiner	111
6.3.2	第二步：从PDF文件中提取文本	111

6.4	第三种技术——Tabula	113
6.4.1	第一步：下载Tabula	113
6.4.2	第二步：运行Tabula	113
6.4.3	第三步：用Tabula提取数据	114
6.4.4	第四步：数据复制	114
6.4.5	第五步：进一步清洗	114
6.5	所有尝试都失败之后——第四种技术	115
6.6	小结	117
第7章	RDBMS清洗技术	118
7.1	准备	118
7.2	第一步：下载并检查Sentiment140	119
7.3	第二步：清洗要导入的数据	119
7.4	第三步：把数据导入MySQL	120
7.4.1	发现并清洗异常数据	121
7.4.2	创建自己的数据表	122
7.5	第四步：清洗&字符	123
7.6	第五步：清洗其他未知字符	124
7.7	第六步：清洗日期	125
7.8	第七步：分离用户提及、标签和URL	127
7.8.1	创建一些新的数据表	128
7.8.2	提取用户提及	128
7.8.3	提取标签	130
7.8.4	提取URL	131
7.9	第八步：清洗查询表	132
7.10	第九步：记录操作步骤	134
7.11	小结	135
第8章	数据分享的最佳实践	136
8.1	准备干净的数据包	136
8.2	为数据编写文档	139
8.2.1	README文件	139
8.2.2	文件头	141
8.2.3	数据模型和图表	142
8.2.4	维基或CMS	144
8.3	为数据设置使用条款与许可协议	144
8.4	数据发布	146
8.4.1	数据集清单列表	146
8.4.2	Stack Exchange上的Open Data	147
8.4.3	编程马拉松	147
8.5	小结	148
第9章	Stack Overflow项目	149
9.1	第一步：关于Stack Overflow的问题	149
9.2	第二步：收集并存储Stack Overflow数据	151
9.2.1	下载Stack Overflow数据	151
9.2.2	文件解压	152
9.2.3	创建MySQL数据表并加载数据	152
9.2.4	构建测试表	154
9.3	第三步：数据清洗	156
9.3.1	创建新的数据表	157
9.3.2	提取URL并填写新数据表	158

《干净的数据：数据清洗入门与实践》

9.3.3	提取代码并填写新表	159
9.4	第四步：数据分析	161
9.4.1	哪些代码分享网站最为流行	161
9.4.2	问题和答案中的代码分享网站都有哪些	162
9.4.3	提交内容会同时包含代码分享URL和程序源代码吗	165
9.5	第五步：数据可视化	166
9.6	第六步：问题解析	169
9.7	从测试表转向完整数据表	169
9.8	小结	170
第10章	Twitter项目	171
10.1	第一步：关于推文归档数据的问题	171
10.2	第二步：收集数据	172
10.2.1	下载并提取弗格森事件的数据文件	173
10.2.2	创建一个测试用的文件	174
10.2.3	处理推文ID	174
10.3	第三步：数据清洗	179
10.3.1	创建数据表	179
10.3.2	用Python为新表填充数据	180
10.4	第四步：简单的数据分析	182
10.5	第五步：数据可视化	183
10.6	第六步：问题解析	186
10.7	把处理过程应用到全数据量（非测试用）数据表	186
10.8	小结	187

《干净的数据：数据清洗入门与实践》

精彩短评

- 1、063. @06142016. 新书，逻辑清晰，但浅尝辄止，略失望，也许是我期望太高了.
- 2、验证一下自己的数据清洗方法。证明数据清洗还是一向初级而琐碎的工作，没有通用的方法。基本上是能行就行。
- 3、对于外行人来说，貌似需要有一定编程经验；对于相关从业人员来说，太多共识型内容。感觉两个方向都没有做好。作者貌似想从一个又一个的例子告诉你怎么做数据清洗，但是这个东西应该跟特定数据特定需求有关系，不太好一次性说清楚。如果做过多次数据清洗工作的，这本书价值不是很大。

《干净的数据：数据清洗入门与实践》

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.tushu000.com