

《数据整理实践指南》

图书基本信息

书名：《数据整理实践指南》

13位ISBN编号：9787115411026

出版时间：2016-3-1

作者：Q.Ethan McCallum

页数：209

译者：魏秀丽,李芳妹

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

《数据整理实践指南》

内容概要

随着数据科学的热门，数据的优化、整理以及如何处理不良数据成为人们关注的重点。本书通过处理不良数据，进行数据清理的案例，向读者展示了处理数据的方法。

本书共有19章，从6部分向读者展示了使用和清理不良数据背后的理论和实践。第1部分是Grubby的动手实践指南，它向读者介绍了驾驭、提取数据的方法，如何处理文本数据中的数据以及Web开发中碰到的数据问题。第2部分是让人充满意外的数据，它向读者介绍了数据也会“撒谎”。第3部分是方法，它向读者介绍了处理不良数据的一些方法。第4部分是数据存储和基础设施，它向读者介绍了如何存储数据。第5部分是数据的商业化，它向读者介绍了如何避免数据处理的一些误差。第6部分是数据策略，它向读者介绍了如何追踪数据、评估数据质量以及构建数据质量相关平台等。

本书适合数据科学家、数据处理和整理相关开发人员阅读。也适合想要进入数据处理领域的读者阅读。

。

《数据整理实践指南》

作者简介

Q . Ethan McCallum , 是一位顾问、作家 , 也是一名科技爱好者。他帮助很多公司在数据和技术方面做出明智的决策 , 他为The O ' Relly Network 和Java.net撰写文章 , 并且为《C/C++Users Journal》《Doctor Dobb ' s Journal》和《Linux Magazine》撰稿。

书籍目录

第1章 从头说起：什么是噪音数据	1
第2章 是我的问题还是数据的问题	4
2.1 理解数据结构	5
2.2 校验	8
2.2.1 字段校验	8
2.2.2 值校验	9
2.2.3 简单统计的物理解释	10
2.3 可视化	11
2.3.1 关键词竞价排名示例	13
2.3.2 搜索来源示例	18
2.3.3 推荐分析	19
2.3.4 时间序列数据	22
2.4 小结	27
第3章 数据是给人看的不是给机器看的	28
3.1 数据	28
3.1.1 问题：数据是给人看的	29
3.1.2 对数据的安排	29
3.1.3 数据分散在多个文件中	32
3.2 解决方案：编写代码	34
3.2.1 从糟糕的数据格式中读取数据	34
3.2.2 从多个文件中读取数据	36
3.3 附言	42
3.4 其他格式	43
3.5 小结	

45	
第4章 纯文本中潜在的噪音数据	
46	
4.1 使用哪种纯文本编码？	
46	
4.2 猜测文本编码格式	
50	
4.3 对文本规范化处理	
53	
4.4 问题：在纯文本中掺入了特定应用字符	
55	
4.5 通过Python处理文本	
59	
4.6 实践练习题	
60	
第5章 重组Web数据	
62	
5.1 你能获得数据吗	
63	
5.1.1 一般工作流程示例	
64	
5.1.2 Robots 协议	
65	
5.1.3 识别数据组织模式	
66	
5.1.4 存储离线版本	
68	
5.1.5 网页抓取信息	
69	
5.2 真正的困难	
73	
5.2.1 下载原始内容	
73	
5.2.2 表单、对话框和新建窗口	
73	
5.2.3 Flash	
74	
5.3 不利情况的解决办法	
75	
5.4 小结	
75	
第6章 检测撒谎者以及相互矛盾网上评论的困惑	
76	
6.1 Weotta公司	
76	
6.2 获得评论	
77	
6.3 情感分类	
77	

6.4 极化语言	78
6.5 创建语料库	80
6.6 训练分类器	81
6.7 分类器验证	82
6.8 用数据设计	84
6.9 经验教训	84
6.10 小结	85
6.11 信息资源	86
第7章 请噪音数据站出来	87
7.1 实例1：在制造业中减少缺陷	87
7.2 实例2：谁打来的电话	90
7.3 实例3：当“典型的”不等于“平均的”	92
7.4 经验总结	95
7.5 到工厂参观能成为试验的一部分吗	96
第8章 血、汗和尿	97
8.1 书呆子戏剧性工作交换	97
8.2 化学家如何整理数字	98
8.3 数据库都是我们的	99
8.4 仔细检查	102
8.5 生命短暂的漂亮代码库	103
8.6 改变化学家（和其他电子表单滥用者）	104
8.7 传递线（tl）和数据记录器（dr）	105
第9章 当数据与现实不匹配	107
9.1 到底是谁的报价机	108
9.2 股票分割、股利和调整	

110	
9.3 糟糕的现实	112
9.4 小结	114
第10章 偏差和误差的来源	115
10.1 估算上的偏差：一般性的问题	117
10.2 报告上的误差：一般性的问题	118
10.3 其他偏差来源	121
10.3.1 顶层编码/底部编码	121
10.3.2 Seam偏差	122
10.3.3 代理报告	123
10.3.4 样本选择	123
10.4 结论	124
参考文献	124
第11章 不要把完美和正确对立起来：噪音数据真是噪音吗	128
11.1 回忆学校生活	128
11.2 向着专业领域前进	129
11.2.1 政府工作	130
11.2.2 政府数据非常真实	131
11.3 应用实例—服务电话	132
11.4 继续前进	133
11.5 经验与未来展望	134
第12章 数据库攻击：什么时候使用文件	135
12.1 历史	135
12.2 建立我的工具箱	136
12.3 数据存储—我的路障	136

12.4 将文件作为数据存储器	137
12.4.1 简单的文件	138
12.4.2 文件处理一切	138
12.4.3 文件可包含任何数据形式	138
12.4.4 局部数据破坏	139
12.4.5 文件拥有很棒的工具	139
12.4.6 没有安装税	139
12.5 文件的概念	140
12.5.1 编码	140
12.5.2 文本文件	140
12.5.3 二进制数据	140
12.5.4 内存映射文件	140
12.5.5 文件格式	140
12.5.6 分隔符	142
12.6 文件支持的网络框架	143
12.6.1 动机	143
12.6.2 实现	145
12.7 反馈	145
第13章 卧库表，隐网络	146
13.1 成本分配模型	147
13.2 组合展开微妙的作用	150
13.3 隐藏网络的浮现	151
13.4 存储图表	151
13.5 利用Gremlin遍历图表	152
13.6 在网络属性里寻找价值	

154	
13.7	从多重数据模型角度考虑并使用正确的工具
155	
13.8	致谢
155	
第14章	云计算神话
156	
14.1	关于云的介绍
156	
14.2	何谓“云”
156	
14.3	云和大数据
157	
14.4	Fred的故事
157	
14.4.1	起初一切都好
157	
14.4.2	基础结构全部放在云端
158	
14.4.3	随着规模增长，最初的扩展很轻松
158	
14.4.4	麻烦出现了
158	
14.4.5	需要提高性能
158	
14.4.6	关键要提高RAID 10性能
158	
14.4.7	重要的局部运行中断引发长期停机
159	
14.4.8	有代价的RAID 10
159	
14.4.9	数据规模增大
160	
14.4.10	地理冗余成为首选
160	
14.4.11	水平扩展并不像想像得那么简单
160	
14.4.12	成本显著增长
160	
14.5	Fred的荒唐事
161	
14.5.1	神话1：云是所有基础设施组件的解决方案
161	
	该神话与Fred故事的联系
161	
14.5.2	神话2：云可以节约成本
161	
	该神话与Fred的故事的联系
162	

14.5.3 神话3：通过RAID可以将cloud 10的性能提高至可接受的水平	163
该神话与Fred故事的联系	163
14.5.4 神话4：云计算使水平扩展轻松	163
该神话与Fred故事的联系	164
14.6 结论和推荐	164
第15章 数据科学的阴暗面	165
15.1 避开这些陷阱	165
15.1.1 对数据一无所知	166
15.1.2 应该只为数据科学家提供一种工具来解决所有问题	167
15.1.3 应该为了分析而分析	169
15.1.4 应该学会分享	169
15.1.5 应该期望数据科学家无所不能	170
15.2 数据学家在机构中的位置	170
15.3 最后的想法	171
第16章 如何雇佣机器学习专家	172
16.1 确定问题	172
16.2 模型测试	173
16.3 创建训练集	174
16.4 选择特征	175
16.5 数据编码	176
16.6 训练集、测试集和解决方案集	176
16.7 问题描述	177
16.8 回答问题	178
16.9 整合解决方案	178
16.10 小结	

179	
第17章 数据的可追踪性	
180	
17.1 原因	
180	
17.2 个人经验	
181	
17.2.1 快照	
181	
17.2.2 保存数据源	
181	
17.2.3 衡量数据源	
182	
17.2.4 逆向恢复数据	
182	
17.2.5 分阶段处理数据并保持各阶段的独立性	
182	
17.2.6 识别根源	
183	
17.2.7 寻找要完善的区域	
183	
17.3 不变性：从函数程序设计借来的理念	
183	
17.4 案例	
184	
17.4.1 网络爬虫	
184	
17.4.2 改变	
185	
17.4.3 聚类	
185	
17.4.4 普及度	
185	
17.5 小结	
186	
第18章 社交媒体：是可抹去的印记吗	
187	
18.1 社交媒体：到底是谁的数据	
188	
18.2 管控	
188	
18.3 商业重组	
190	
18.4 对沟通和表达的期望	
190	
18.5 新的最终用户期望的技术含义	
192	
18.6 这个行业是做什么的	
194	

18.6.1 验证API	195
18.6.2 更新通知API	195
18.7 最终用户做什么	195
18.8 我们怎样一起工作	196
第19章 揭秘数据质量分析：了解什么时候数据足够优质	197
19.1 框架介绍：数据质量分析的4个C	198
19.1.1 完整性	199
19.1.2 一致性	201
19.1.3 准确性	203
19.1.4 可解释性	205
19.2 结论	208

精彩短评

- 1、不看也完全没损失的书，嗯
- 2、比较适合数据分析师
- 3、个别章节有启发，但大多数章节本身的质量和翻译的质量连微信文章都不如。
- 4、翻译的质量感觉不是太好，很多地方读起来拗口；内容上，对于自己体验过的场景，很有共鸣感，学到不少；没有体验过的部分，感觉距离太远，读不进去，以后有经验后可以回头再读；不适合初学者，适合有了一定经验想要进一步提高的相关工作人员

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