TURING 图灵程序设计丛书

[PACKT] PUBLISHING

[美] Megan Squire 著　任政委 译

干净的数据

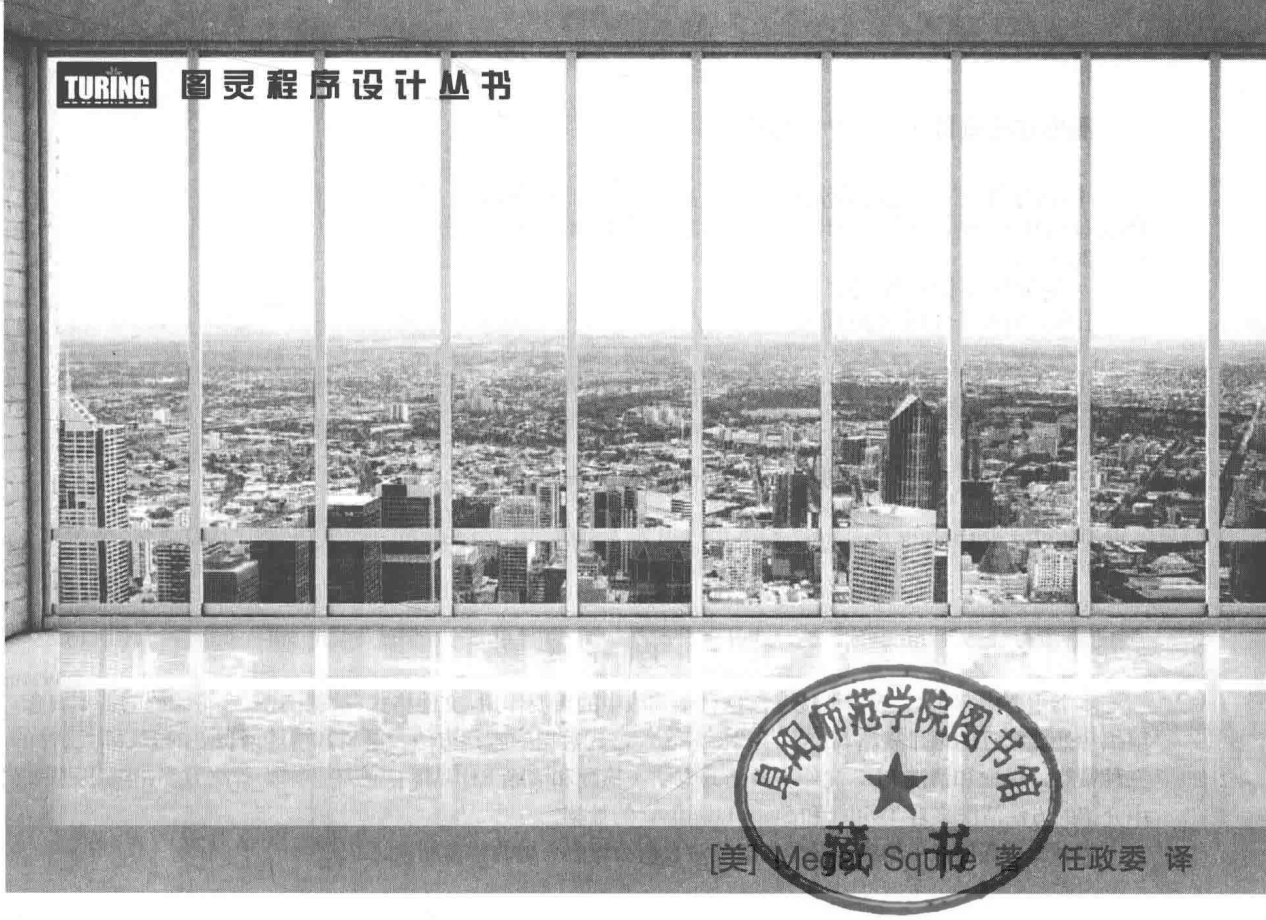
数据清洗入门与实践

Clean Data

掌握高效数据清洗方法，为数据挖掘提供便利，让用户更好地体验大数据价值！

中国工信出版集团

人民邮电出版社
POSTS & TELECOM PRESS



[美] Megan Squire 著 任政委 译

干净的数据

数据清洗入门与实践

Clean Data

人民邮电出版社
北京

图书在版编目（CIP）数据

干净的数据 : 数据清洗入门与实践 / (美) 斯夸尔(Squire,M.) 著 ; 任政委译. -- 北京 : 人民邮电出版社, 2016.5
（图灵程序设计丛书）
ISBN 978-7-115-42047-3

Ⅰ. ①干… Ⅱ. ①斯… ②任… Ⅲ. ①数据处理 Ⅳ. ①TP274

中国版本图书馆CIP数据核字(2016)第062910号

内 容 提 要

本书主要内容包括：数据清洗在数据科学领域中的重要作用，文件格式、数据类型、字符编码的基本概念，组织和处理数据的电子表格与文本编辑器，各种格式数据的转换方法，解析和清洗网页上的 HTML 文件的三种策略，提取和清洗 PDF 文件中数据的方法，检测和清除 RDBMS 中的坏数据的解决方案，以及使用书中介绍的方法清洗来自 Twitter 和 Stack Overflow 的数据。

本书适合任何水平的数据科学家以及对数据清理感兴趣的读者阅读。

◆ 著 [美] Megan Squire
译 任政委
责任编辑 岳新欣
执行编辑 李 敏
责任印制 彭志环

◆ 人民邮电出版社出版发行 北京市丰台区成寿寺路11号
邮编 100164 电子邮件 315@ptpress.com.cn
网址 http://www.ptpress.com.cn
北京鑫正大印刷有限公司印刷

◆ 开本：800×1000 1/16
印张：12.5
字数：296千字 2016年5月第1版
印数：1-3 000册 2016年5月北京第1次印刷
著作权合同登记号 图字：01-2015-7995号

定价：49.00元

读者服务热线：(010)51095186转600 印装质量热线：(010)81055316
反盗版热线：(010)81055315
广告经营许可证：京东工商广字第 8052 号

版权声明

前　言

“请问，巴贝奇先生，如果把错误的数据放进机器中，是否能够得到正确的答案呢？”

——查尔斯·巴贝奇（1864）

“错进，错出。”

——美国国税局（1963）

“压根儿就没有干净的数据集。”

——乔希·沙利文，《财富》杂志收录的博思艾伦咨询公司副总裁语录（2015）

世界上第一台计算机的发明者查尔斯·巴贝奇，在他1864年的随笔文集中记录了这样一件事，他曾经因为有人认为在输入错误数据的情况下计算机依然能够给出正确答案而错愕不止。100年以后，美国税务部门开始耐心地向人们解释“错进，错出”，以此来表达即便是能力再强的税收官，在用计算机处理数据时，依旧要依赖输入数据的质量。又过了50年，到了2015年，在这个看起来超级神奇的时代，有着机器学习、自动纠错、预想接口以及比我们本人更为了解我们自己的各种推荐系统。然而，这一切的算法背后，仍旧需要高质量的数据来保证学习的正确性，而我们往往会叹息道“压根儿就没有干净的数据集”。

本书正是为那些时常需要与数据打交道的人准备的，包括数据科学家、数据新闻记者、软件开发人员以及其他相关人士。无论你从事的是哪种职业，本书都会传授你一套快速而简便的实用策略，用来填补现有数据和期望数据之间的空白。人人都期盼能够拥有高质量的完美数据，但现实中的数据往往与我们的期盼相去甚远。我们是否正在饱受各种折磨呢？数据不是缺失就是格式不对，或者位置错误，还有各种异常情况，而这些问题导致的结果可以借用说唱歌手克里斯托弗·华莱士的一句歌词来表达，那就是“数据越多，麻烦越大”。

在本书中，我们始终把数据清洗当成数据科学过程中有着重要意义和重大价值的一步：轻松改进，不容忽视。我们的目标就是重新定义数据清洗，它不再是开始真正的工作之前所必须做的令人畏惧和乏味的工作。相反，我们会使用久经考验的过程与工具。我们会了解到，就好比在厨房中做菜，如果菜洗干净了，食物的色和味就不会差，我们自己也会倍感愉悦。如果再有良好的

刀工，肉就会鲜嫩可口，菜也会入味均匀。就像手艺高超的大厨们都有他们钟爱的厨具和烹饪手法一样，数据科学家们也想在绝佳的条件下处理最为完美的数据。

本书内容

第1章，*为什么需要清洗数据*。这一章通过说明数据清洗在数据科学过程中的重要作用，激发我们对干净数据的追求。随后用一个简单的例子演示了现实世界中的脏数据。在充分衡量多种清洗过程的优缺点之后，讲述了如何将清洗所带来的变化告诉其他人。

第2章，*基础知识——格式、类型与编码*。这一章介绍关于文件格式、压缩和数据类型的基础知识，同时也讨论了数据缺失和空数据，以及字符编码方面的问题。每一节都配有一个真实的案例。这一章之所以重要，是因为后面几章的学习都以这一章的基本概念为基础。

第3章，*数据清洗的老黄牛——电子表格和文本编辑器*。这一章描述了如何从常见的电子表格和文本编辑器中，尽可能多地发掘出数据清洗功能。我们还介绍了一些常见问题的简便处理方法，包括如何使用函数、搜索和替换、正则表达式来实现数据纠错和转换。在这一章的结尾处，我们将利用已经掌握的技能，使用上述两种工具来完成一个与大学有关的数据清洗任务。

第4章，*讲通用语言——数据转换*。这一章着重讨论了如何把数据从一种格式转换成另一种格式。这是数据清洗工作的重要任务之一，而我们身边各种各样的工具能帮助我们轻松完成该项任务。我们首先在几种常见的格式之间来回进行转换，如逗号分隔值（CSV）、JSON和SQL。为了演示如何在实际中使用这些技术，我们会完成一个实验项目。我们会从Facebook上下载好友关系网络数据，并把它们转换成不同的数据格式，以形象化地展现数据之间的关系。

第5章，*收集并清洗来自网络的数据*。这一章描述了三种专门针对HTML页面的数据清洗方法。其中介绍了如何利用三种流行工具从标记文本中提取数据，同时介绍了一些基本概念，以便理解其他方法。在这一章的例子中，我们会建立起一系列的清洗步骤，专门用于从网络论坛中抽取数据。

第6章，*清洗PDF文件中的数据*。这一章介绍了几种最为常见的数据顽症处理方法：提取Adobe的PDF文件中的数据。我们先采用一些低成本的工具来完成这项任务，然后再选择一些容易上手并且门槛较低的工具，最后采用Adobe自己的付费软件。所有实验都会一如既往地使用真实数据，这能让我们在学会解决问题的同时积累宝贵的经验。

第7章，RDBMS清洗技术。在这一章里，我们使用对公众开放的推文数据来演示多种适用于关系型数据库的数据清洗策略。例子中使用的数据库是MySQL，但其中的许多概念，如基于正则表达式的文本提取和异常检测，可以轻而易举地应用到其他存储系统中。

第8章，*数据分享的最佳实践*。这一章描述了多种分享清洗过的数据的方法，以便他人轻松

地使用你的数据。即使你暂时还没有分享数据的打算，这些方法和策略也会对你以后组织工作中的数据有所帮助。本章具体内容包括如何创建各种格式的理想数据包，如何在文档中对这些数据进行描述，如何为数据选择适合的许可协议，以及如何按需进行数据发布。

第9章，Stack Overflow项目。在这一章中我们将使用真实的数据来指导你完成一个完整的项目。在项目开始前，我们会提出一些与数据集有关的实实在在的问题。在回答这些问题期间，我们将完成第1章中所介绍的整个数据科学过程，并把在前面几章学过的清洗方法应用到其中。除此之外，为了应对庞大的数据量，我们还将采用一些新的技术来创建测试数据集。

第10章，Twitter项目。这一章描述的也是一个完整的项目，其目的是为了演示如何执行最热和变化最快的数据收集和清洗任务：Twitter挖掘。在演示中，我们将会查找并收集与某一时事相关的可公开获取的推文归档数据，同时遵守Twitter服务条款。数据采用的是目前网络API所支持的最为流行的JSON格式。在清洗和提取数据的同时，我们还将解答一个与数据集有关的简单问题。最后，我们将设计一个简单的数据模型，用它来长期存放已经抽取出来并且经过清洗的数据，并做一些简单的可视化实现。

你需要准备些什么

为了完成本书中的项目，你会用到下面这些工具和资源。

- 一款浏览器，互联网接入，现代的操作系统。我们对浏览器和操作系统本身没有什么特别的要求，但最好能支持命令行终端窗口（比如OS X上的Terminal应用）。另外，在第5章里涉及的三项活动中，有一项要依靠Chrome浏览器所提供的一个工具，所以如果你想顺利完成这项活动的话，请务必准备充分。
- 文本编辑器，如Mac OS X上的Text Wrangler，或是Windows上的Notepad++。有的集成开发环境（IDE，如Eclipse）也可以用来充当文本编辑器，但这些应用往往捆绑了太多你根本不需要的特性。
- 电子表格应用程序，如微软的Excel，或者是Google的Spreadsheets。本书中提供的例子会尽量保证在这两种工具下都可以正常运行，但有时也可能只能在一种工具下运行。
- 安装Python开发环境与Python库。我推荐使用Enthought Canopy Python环境，其地址为https://www.enthought.com/products/canopy/。
- 一个能够运行的MySQL，版本需要在5.5以上。
- Web服务器和PHP，PHP版本要求5以上。
- MySQL客户端接口，可以是命令行终端，也可以是MySQL Workbench，或者是phpMyAdmin（前提是你已经装好了PHP）。

本书的目标读者

如果你现在正在读这本书的话，我猜想你可能属于下面两类人之一。第一类是在数据清洗工作上花费了大量时间的数据科学家，希望可以进一步提高工作效率。在面对乏味单调的数据清洗任务时，你也许会觉得有些苦闷，正在寻求能够加快处理速度、提高效率的方法，或者想着是否有些别的什么工具可以立马把工作做完。在厨房的比喻中，你恰好就是那个需要提高刀工本领的大厨。

另一类是从事数据科学工作，但之前从未真正在乎过数据清洗这件事的人。但是现在，你开始琢磨了，如果事先引入清洗过程，最终得出的数据结果也许会有所改善。也许“错进，错出”这句老话让你觉得越发地真实。也许你还是一个乐于与他人分享数据成果的人，但又常常对产出的数据质量缺乏信心。通过本书，你可以学会更多的技巧并养成保持整洁的数据科学环境的习惯，随后自然可以信心满满地“当众献艺”。

但无论你属于哪一类人，本书都将帮助你重塑数据清洗的观念，让数据清洗不再是一件苦差事，而是高质量、有品位、时尚和高效的标志。你只需稍有一些编程背景即可，不要求在这方面特别强，因为在大部分数据科学项目中，发自内心的学习意愿和实验意愿，以及好奇心和对细节的注重，更为重要，也更令人推崇。

本书排版约定

本书中会出现许多用来区分不同类型信息的文本格式。接下来看一下这些格式以及它们所对应的含义。

正文中的代码、数据库表名、用户输入会以等宽字体进行表示，如：“问题是函数`open()`不能处理UTF-8编码的字符。”

代码块的表现形式如下：

```
for tweet in stream:
    encoded_tweet = tweet['text'].encode('ascii','ignore')
    print counter, "-", encoded_tweet[0:10]
    f.write(encoded_tweet)
```

另外，当我们希望你特别注意代码块中的某些部分时，相关的行或者文字会被加粗：

```
First name,birth date,favorite color,salary
"Sally","1971-09-16","light blue",129000
"Manu","1984-11-03","",159960
"Martin","1978-12-10","",76888
```

命令行中的输入和输出内容表示如下：

```
tar cvf fileArchive.tar reallyBigFile.csv anotherBigFile.csv
gzip fileArchive.tar
```

新术语和重点词汇均采用楷体字表示。

这个图标表示警告或需要特别注意的内容。

这个图标表示提示或者技巧。

读者反馈

我们总是欢迎读者的反馈。如果你对本书有些想法，有什么喜欢或是不喜欢的，请反馈给我们。这将有助于我们开发出能够充分满足读者需求的图书。

一般的反馈，请发送电子邮件至feedback@packtpub.com，并在邮件主题中包含书名。

如果你在某个领域有专长，并有意编写一本书或是贡献一份力量，请参考我们的作者指南，地址为http://www.packtpub.com/authors。

客户支持

你现在已经是Packt引以为傲的读者了，为了能让你的购买物有所值，我们还为你准备了以下内容。

彩色图片下载

我们为你准备了一个PDF文件，其中包含了本书用到的屏幕截图和图表的彩色图片。这些彩色图片将有助于你更好地理解书中的内容。该文件的下载地址是https://www.packtpub.com/sites/default/files/downloads/Clean_Data_Graphics.zip。

勘误表

虽然我们已竭尽全力确保本书内容的准确性，但仍可能有所疏漏。如果你发现了书中哪些地方有误——这些问题可能出现在正文或是代码之中——请告知我们，对此我们将万分感谢。你的

这种贡献将让其他读者免受其害，并对本书后续版本的改进有着莫大的帮助。如果你发现了错误，请通过http://www.packtpub.com/submit-errata进行提交，先选择书名，然后点击链接Errata Submission Form，就可以输入详细内容了。勘误一经核实，我们就会把它上传到我们的网站或是添加到现有勘误表中。

如果你需要查看之前已经提交的勘误信息，请访问https://www.packtpub.com/books/content/support，在搜索栏中输入书名，就可以查看现有的勘误表。

关于盗版

受版权保护的材料在互联网上遭到盗版是所有媒体都一直在面对的问题。Packt非常重视保护版权和许可证。如果你发现我们的作品在互联网上被非法复制，不管以什么形式，请立即为我们提供其网络地址或是网站名称，我们将采取相应的应对措施。

请将疑似的盗版材料链接发送至copyright@packtpub.com。

非常感谢你帮助我们保护作者，以及保护我们给你带来有价值内容的能力。

问题反馈

如果你有关于本书任何方面的问题，请联系questions@packpub.com，我们将尽快帮你解决。

目　录

第1章 为什么需要清洗数据

大数据、数据挖掘、机器学习和可视化，近来计算界的几件大事好像总也绕不开数据这个主角。从统计学家到软件开发人员，再到图形设计师，一下子所有人都对数据科学产生了兴趣。便宜的硬件、可靠的处理工具和可视化工具，以及海量的免费数据，这些资源的汇集使得我们能够比以往任何一个时期更加精准地、轻松地发现趋势、预测未来。

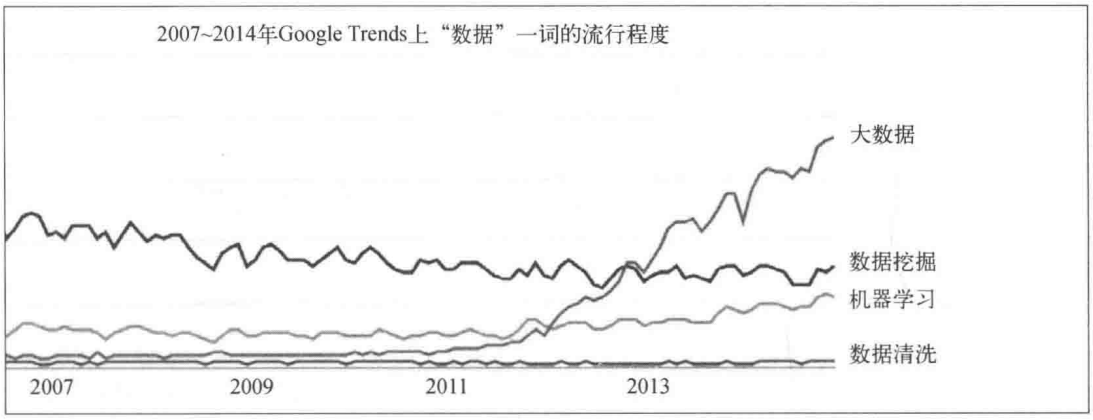
不过，你可能还未听说过的是，数据科学的这些希望与梦想都建立在乱七八糟的数据之上。在正式应用于我们认为是数据科学的核心的算法和可视化之前，这些数据往往需要经过迁移、压缩、清洗、打散、分片、分块以及其他多种转换处理。

本章内容将涵盖以下几个方面：

- 关于数据科学的六个简单处理步骤，包含数据清洗
- 与数据清洗有关的参考建议
- 对数据清洗有帮助的工具
- 一个关于如何将数据清洗融入整个数据科学过程的入门示例

1.1 新视角

最近我们读报时发现《纽约时报》将数据清洗称为看门人工作，并称数据科学家百分之八十的时间都花费在了这些清洗任务上。从下图中我们可以看出，尽管数据清洗是很重要的工作，但它并没有像大数据、数据挖掘或是机器学习那样真正地引起公众的注意。



不会真的有人因为没有见过人们聚众讨论看门人的工作多么有趣、多么酷而开始评头论足吧？说起来还真是惭愧，这工作没比做家务强到哪里去，但话又说回来，与其对它弃之不理、抱怨不断、恶语相加，还不如先把活儿干完，这能让我们过得更好些。

还不相信是吗？那让我们打个比方，你不是数据看门人，而是数据大厨。现在有人交给你一个购物篮，里面装满了你从未见过的各种各样的漂亮蔬菜，每一样都产自有机农场，并在最新鲜的时候经过人工精挑细选出来。多汁的西红柿，生脆的莴苣，油亮的胡椒。你一定激动地想马上开启烹饪之旅，可再看看周围，厨房里肮脏不堪，锅碗瓢盆上尽是油污，还沾着大块叫不出名的东西。至于厨具，只有一把锈迹斑斑的切刀和一块湿抹布。水槽也是破破烂烂的。而恰恰就在此时，你发现从看似鲜美的莴苣下面爬出了一只甲虫。

即使是实习厨师也不可能在这样的地方烹饪。往轻了说，无外乎是暴殄天物，浪费了一篮子精美的食材。如果严重点儿讲，这会使人致病。再说了，在这种地方烹饪根本毫无乐趣可言，也许全天的时间都得浪费在用生锈的破刀切菜上面。

与厨房的道理一样，事先花费些时间清洗和准备好数据科学工作区、工具和原始数据，都是值得的。“错进，错出。”这句源于上20世纪60年代的计算机编程箴言，对如今的数据科学来说亦为真理。

1.2　数据科学过程

数据清洗是如何融入数据科学中的呢？简短的回答就是，清洗工作是关键的一步，它直接影响在它之前和之后的处理工作。

稍微长一些的回答就得围绕数据科学过程的六个步骤来描述了，请看下面的列表。数据清洗正好处于中间的位置，第三步。但是，请不要以纯线性方式看待这些步骤，简单地认为这是一个从头到尾执行的框架，其实在项目的迭代过程中，我们会根据具体情况，反复执行这些步骤。另

外还需要指出的是，并不是每一个项目都会包含列表中所有的步骤。举个例子，有时候我们并不需要数据收集或可视化步骤。这完全取决于项目的实际情况。

(1) 第一步是问题陈述。识别出你要解决的问题是什么。

(2) 接下来要做的是数据收集与存储。数据从何而来？它们在哪里存放？格式又是什么？

(3) 然后是数据清洗。数据需要修改吗？有什么需要删除的吗？数据应该怎么调整才能适用于接下来的分析和挖掘？

(4) 数据分析和机器学习。数据需要哪些处理？需要什么样的转换？使用什么样的算法？运用什么公式？使用什么机器学习算法？顺序又是怎样的呢？

(5) 数据展现和可视化实现。数据处理结果应该怎样呈现出来呢？我们可以用一张或几张数据表来表现，也可以使用图画、图形、图表、网络图、文字云、地图等形式。但这是最佳的可视化方案吗？有没有更好的替代方案呢？

(6) 最后一步是问题决议。你在第一步里所提出的疑问或是问题的答案究竟是什么？数据处理结果还有哪些不足？这个方法能彻底解决问题吗？你还能找出别的什么办法吗？接下来要做的又是什么？

在数据分析、挖掘、机器学习或是可视化实现之前，做好相关的数据清洗工作意义重大。不过，请牢记，这是一个迭代的过程，因为在项目中我们可能需要不止一次地执行这些清洗操作。此外，我们所采用的挖掘或分析方法会影响清洗方式的选取。我们可以认为清洗工作包含了分析方法所能决定的各种任务，这有可能是交换文件的格式、字符编码的修改、数据提取的细节等。

数据清洗与数据收集和存储（第2步）的关系也十分密切。这意味着你得收集原始数据，对它们执行存储和清洗操作，之后再把清洗过的数据保存下来，接下来收集更多的数据，清洗新的数据并把清洗结果与前面处理完的结果数据结合起来，重新进行清洗、保存等操作，反反复复。正因为这个过程非常复杂，所以我们要么选择牢牢记住曾经做过的处理，并记录下那些可以根据需要反复执行的步骤，要么把工作的全部状况告知其他相关人员。

1.3 传达数据清洗工作的内容

六步处理过程是围绕着问题和解决方案这个故事线组织的，因此，在作为报表框架使用时，它的表现十分优秀。如果你已经决定使用六步框架来实现数据科学过程报表，将发现只有到了第三步你才会真正开始进行与清洗有关的工作。

哪怕你并不需要把数据科学过程制成正式的报告文档，你仍然会发现，认真地记录下曾经按什么顺序做了些什么事情，对以后的工作也是极有帮助的。

请记住，哪怕是规模再小、风险再低的项目，你也要面对至少两人规模的受众：现在的你和六个月之后的你。请相信我说的话，因为六个月之后的你基本上不会记得今天的你做过什么样的清洗工作，也不记得其中的缘由，更谈不上如何重新再做一次。

要解决这个问题，最简单的方案就是保留一份工作日志。这个日志应该包含链接、屏幕截图，或是复制粘贴你曾经运行过的具体的命令，并配上为什么要这样做的解释性文字。下面是一个关于小型文本挖掘项目的日志示例，其中记述了每个阶段输出的外部文件链接以及相关的清洗脚本链接。如果你对日志中提到的某些技术不太熟悉的话，也没有关系，因为这个示例的重点只是让你了解一下日志的样子而已。

(1) 我们写了一条SQL查询语句来检索出每条数据及其相关描述。

(2) 为了能在Python中进行词频分析，我们需要把数据调整成JSON格式。因此我们做了一个PHP脚本，用它来循环遍历查询结果，并以JSON格式保存到文件中（第一个版本的数据文件）。

(3) 这个文件里的数据有些格式上的错误，比如包含了没有转义的问号和一些多余的内嵌HTML标签。这些错误可以在第二个PHP脚本中修正。运行第二个脚本之后，我们就可以得到一份干净的JSON文件了（第二个版本的数据文件）。

这里需要注意的是，我们用日志来解释程序做过什么和这样做的原因。日志的内容可以很简短，但要尽可能地包含一些有用的链接。

另外，我们还可以选择许多更复杂的方案来传达信息。例如，如果你对软件项目管理中常用的版本控制系统比较熟悉的话，如Git或是Subversion，就可以好好地规划设计一番，想想怎么使用这些工具来跟踪数据清洗工作。不管你使用什么样的系统，最重要的事情是做好日志，哪怕只有一句话。来吧，学着把它用起来，别耽误进度了。

1.4　数据清洗环境

本书中涉及的数据清洗方法是通用的，适用范围非常广泛。你不需要任何高端专业的数据库产品或是数据分析产品（事实上，这些厂商和产品可能已经提供了数据清洗程序或是解决方法）。我围绕数据处理过程中的常见问题，设计了本书中的清洗教程。而我要展示的都是适用范围较为广泛的开源软件和技术，它们很容易在实际工作中获得和掌握。

下面列出了你需要准备的工具和技术。

- 几乎在每一章中，我们都会用到终端窗口和命令行界面，比如Mac OS X上的Terminal程序或者是Linux系统上的bash程序。而在Windows上，有些命令可以通过Windows的命令提示符运行，但其他的命令则需要通过功能更强的命令行程序来运行，比如CygWin。
- 几乎在每一章中，我们还会用到文本编辑器或者是适合程序员使用的编辑器，如Mac上的