Broadview
www.broadview.com.cn

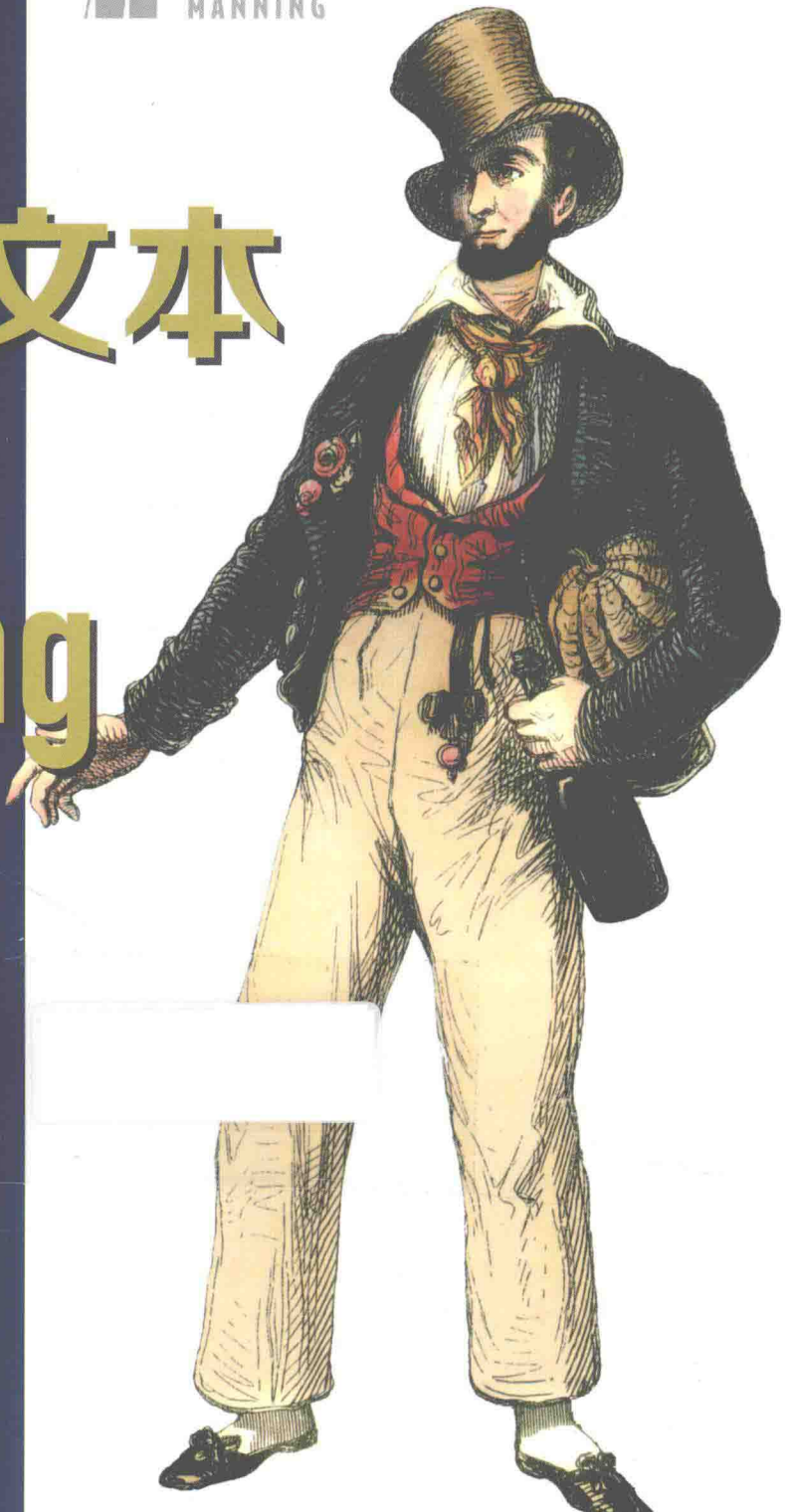
MANNING

驾驭文本

文本的发现、组织和处理

Taming Text

How to Find, Organize, and Manipulate It



Grant S. Ingersoll
[美] Thomas S. Morton 著
Andrew L. Farris

王斌 译

中国工信出版集团

電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

驾驭文本

文本的发现、组织和处理

Taming Text

How to Find, Organize, and Manipulate It

Grant S. Ingersoll
[美] Thomas S. Morton 著
Andrew L. Farris

王斌 译

電子工業出版社
Publishing House of Electronics Industry
北京•BEIJING

内容简介

文本处理是目前互联网内容应用（如搜索引擎、推荐引擎）的关键技术。本书涵盖了文本处理概念和技术的多个方面，包括文本预处理、搜索、字符串匹配、信息抽取、命名实体识别、分类、聚类、标签生成、摘要、问答等。本书的特点在于通过实例来理解文本处理的这些概念和技术，读者利用现有的开源工具就可以自己实现这些实例。本书适合互联网文本内容处理领域的开发人员阅读，也适合有志于加入这一领域的学生、从业人员阅读。即使对于已经从事多年文本处理研究和开发工作的人员来说，本书也不失为一种有益的补充性读物。

版权贸易合同登记号　图字：01-2014-5768

图书在版编目（CIP）数据

驾驭文本：文本的发现、组织和处理 /（美）英格索尔（Ingersoll，G，S.），（美）莫顿（Morton，T.S.），（美）法里斯（Farris，A.L.）著；王斌译. — 北京：电子工业出版社，2015.7
书名原文：Taming text:how to find,organize,and manipulate it
ISBN978-7-121-25230-3

Ⅰ.①驾…　Ⅱ.①英…②莫…③法…④王…　Ⅲ.①自然语言处理—研究　Ⅳ.①TP391
中国版本图书馆CIP数据核字（2014）第302750号

策划编辑：符隆美
责任编辑：徐津平
印　　刷：北京丰源印刷厂
装　　订：三河市华成印务有限公司
出版发行：电子工业出版社
　　北京市海淀区万寿路173信箱　邮编：100036
开　　本：787×980　1/16　印张：21.25　字数：350千字
版　　次：2015 年 7 月第 1 版
印　　次：2015 年 7 月第 1 次印刷
定　　价：79.00元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888。

质量投诉请发邮件至zlts@phei.com.cn，盗版侵权举报请发邮件至dbqq@phei.com.cn。

服务热线：（010）88258888。

译者序

不知不觉，我进入信息内容处理这个领域已经有近20年了。这些年中，我的研究涉及机器翻译、Web搜索、跨语言检索、垃圾邮件过滤、问答、推荐、文本分类、聚类、情感分析等诸多技术或应用，也开发了多个原型以及实用系统。我十分高兴能够在这个有趣的领域不断地学习新技术，了解并开发新应用。与此同时，我也亲眼目睹了很多优秀的技术书籍不断涌现。完全出于兴趣爱好以及与大家分享的个人追求，我先后翻译了《信息检索导论》、《大数据：互联网大规模数据挖掘与分布式处理》、《机器学习实战》、《Mahout实战》等教材或技术书籍。现在，我又推荐给大家手边的这本《驾驭文本》。

文本处理是很多应用的基本技术，包括上面提到的搜索、推荐、问答应用都离不开文本处理。“驾驭”文本对于这些系统至关重要。然而，文本特别是自然语言文本本身的情况十分复杂，处理起来十分烦琐，难度很大。如何利用已有开源工具高效地“驾驭”文本是本书的目标。很显然，对于文本处理开发人员来说，这本书能够提供支撑。当然，由于自然语言文本固有的歧义性，文本处理技术特别是深层“理解”技术还远未成熟，研究人员还在不断努力，全方面真正“驾驭”文本是所有文本处理工作人员的终极梦想。

本书介绍了文本搜索、模糊字符串匹配、命名实体识别、文本聚类分类标注等多种文本处理关键技术，并通过融合上述技术构建了一个简单的事实型问答系统。所有的单项技术都有可供下载使用的数据集和相应的运行代码，读者可以下载这些

代码和数据进行尝试，以便能够更加深入地理解这些技术。

本书作者都是开源社区的重要贡献者，他们在文本处理领域具有丰富的开发经验。这些经验也都体现在本书的内容写作中。

感谢出版社和编辑部的辛勤工作，感谢实验室领导、同事以及译者家人对翻译本书的支持。

因本人各方面水平有限，现有译文中肯定存在许多不足。希望读者能够和我进行联系，以便能够不断改进。来信请联系wbxjj2008@gmail.com。

王 斌

2015年3月15日于中关村

序

在高质量文本处理需求持续指数级增长的年代，很难想象某个部门或业务不依赖某种类型的文本信息。迅速发展的Web经济也明显迅速加大了这种依赖性。与此同时，对高水平技术专家的需求也迅速增加。《驾驭文本》这本书就是应这种形势而出版的一本优秀的实用性书籍，它能够提供大量来自真实世界的经过实际验证的指导性案例。

Grant Ingersoll和Drew Farris是两位优秀的高水平软件工程师，和我一起工作过多年。而Tom Morton是在自然语言处理领域备受尊重的贡献者。他们仨联袂为我们奉献了一本实际课程的教材，该课程可以指导其他有志加入文本处理高级人才行列的技术人员，这些文本处理人才称为自然语言处理工程师。

本书采用学而致用的方法，为一个实际上十分复杂的过程褪去神秘的外衣。通过集中关注已有的工具、可实现的样例和已验证的代码，几位作者带领读者快速学习本来需要修一学期的NLP课程。

作为软件工程师，你已经具备基本能力能够跟进这些样例、代码和书中提到的开源工具，从而能够比预期更快地成为真正的专家，同时也能更快准备好面对来自实际世界的机会。

美国雪城大学信息研究学院院长 LIZ LIDDY

前　言

生活中充满偶然瞬间，它们当中只有极少数会脱颖而出，就像那个确定我（Grant）职业生涯的瞬间一样。那是20世纪90年代末，当时我是一个年轻的软件开发人员，主要从事分布式电磁仿真的工作。有一天我看到一则广告，在纽约雪城（Syracuse）的一家小公司TextWise招聘一个开发职位。看完职位描述之后，我都没想过能获得这份工作，但是当时决定试试运气，就提交了一份简历。莫名其妙地，我获得了这份工作，于是开始了我的搜索和自然语言处理生涯。没想到这么多年以后，我仍然还在做搜索和自然语言处理，更没想到还会写一本这方面的书。

我那时候的第一个任务是开发一个跨语言信息检索（CLIR）系统，要求输入英语查询能够找到法语、西班牙语和日语文档，并将它们自动翻译成英语。回想起来，那个系统触及了我开始喜欢文本处理工作的所有难题：搜索、分类、信息抽取、机器翻译和所有那些奇怪的让每个学习文法的学生都疯狂的语言规则，等等。第一个项目之后，我后来又参与了多个搜索和NLP系统的开发工作，范围从基于规则的分类器到问答系统等。后来在2004年，NLP中心的一份新工作让我开始接触Apache Lucene，这个时代的开源搜索库（无论如何，至少目前还是）。后来我又参与开发一个CLIR系统，不过这次处理的是英语和阿拉伯语。因为需要一些Lucene功能来完成这项任务，我开始提交一些功能和错误的修正补丁。过了一段时间之后，我成为该社区的贡献者。从那之后，开源的“闸门”被轰然打开。我在开源领域涉入更深，并与Isabel Drost和Karl Wettin开始了Apache Mahout机器学习项目，并共同创立了一家

利用Apache Lucene和Solr进行搜索和文本分析的公司Lucid Imagination。

转了一圈之后，我认为搜索和NLP属于计算机科学的定义范围，不论是数据结构还是算法都需要复杂的方法来解决问题。除此之外，还有处理用户生成的大规模Web和社交内容的扩展性需求，这构成你的开发者之梦。这本书由工程师撰写给工程师，特别关注于使用现有、久经考验的开源库来解决文本处理中的疑难问题。个人认为目前这方面的市场还处于空白。我希望本书能够帮助解决当前工作中每天遇到的问题，也能激发你看到带来大量学习机会的文本世界。

GRANT INGERSOLL

我（Tom）在高二时就开始对人工智能感兴趣，本科毕业时选择去读自然语言处理方向的研究生。在宾夕法尼亚大学，我学习了大量文本处理、机器学习、算法和数据结构知识。我也有机会和自然语言处理领域最杰出的一些人共事并从他们身上学到很多东西。

在研究生阶段的课程中，我参加了多个NLP系统的开发工作，并参加了大量DARPA资助的有关共指、摘要和问答的评测。在这些工作中，我熟悉了Lucene和更大的开源运动。我也注意到能够提供高效端对端处理的开源文本处理软件还有较大欠缺。于是在我硕士论文的基础上，我为OpenNLP项目提供了大量贡献代码，并在之后的美国教育测试服务中心（Educational Testing Services）开发自动作文和短答案评分系统时继续学习NLP系统的一些知识。

在开源社区工作教会我很多与其他人一起工作的方法，也使我成为一名更优秀的软件工程师。现在，我在Comcast Cororation工作，与多个软件工程师团队一起使用本书中介绍的工具和技术。我希望本书能够在研究人员的艰难工作（这些工作就像我在研究生阶段学到的那样）与以使用文本处理来解决实际问题为目标的软件工程师之间架起桥梁。

THOMAS MORTON

和Grant一样，我是20世纪90年代中期由Elizabeth Liddy博士、Woojin Paik以及其他一些在TextWise进行研究的人员引入信息检索和自然语言处理领域的。我在完成雪城大学信息研究学院的硕士工作时和这个团队一起工作。那时，TextWise正处于从研

究组转型为创业公司的阶段，主要基于文本处理研究的成果开发商业应用。我在那个公司待了很多年，其间不断地学习和发现新的东西，并与一些优秀的同事一起共事，他们从各个角度来应对“教机器理解语言”这个挑战。

个人而言，我一开始是从软件开发人员的角度切入文本分析这个主题的。我有机会同优秀的研究人员一起工作，将他们的思想从实验转化为功能原型及大规模可扩展的系统。在此过程中，我有机会从事大量现在被称为“数据科学”的工作，发掘出对探索和理解大规模数据以及对它们进行学习的工具和技术的深深热爱。

怎样夸大开源软件对我职业的巨大影响都毫不为过。作为研究的伴随品，可用的开源代码为学习文本分析的新技术和方法以及软件开发提供了一条十分高效的途径。在这里我对所有尽力将知识和经验共享给那些有热情参加学习者的人表示敬意。我特别要感谢Apache软件基金会的那些好伙计们，他们为开源软件、人、处理过程和支持的社区贡献出一个不断成长的生机勃勃的生态系统。

本书中的工具和技术深深扎根于开源软件社区。Lucene、Solr、Mahout和OpenNLP都处于Apache这顶大伞之下。本书只介绍这些工具能实现的一些表面功能。我们的目标是提供对文本处理核心概念的理解，并为本领域的未来探索打下坚实的基础。

祝大家编程愉快！

DREW FARRIS

致　谢

本书经历很长时间完成，倾注了很多人的心血，这里要对他们表示诚挚谢意。

- 感谢Apache Solr、Lucene、Mahout、OpenNLP和其他本书中介绍的工具的用户和开发者
- 感谢Manning出版社，特别是和我们一直密切合作的Douglas Pundick、Karen Tegtmeyer和MarjanBace
- 感谢本书的开发编辑Jeff Bleiel，感谢他在我们疯狂时间表的情况下仍然推进写作过程，感谢他一直以来的优秀反馈，也感谢他将我们这些开发人员转变为作者
- 感谢本书的评阅人，他们提出的问题、评论及批评提高了本书的质量。他们是：Adam Tacy、Amos Bannister、Clint Howarth、CostantinoCerbo、Dawid Weiss、Denis Kurilenko、Doug Warren、Frank Jania、Gann Bierner、James Hatheway、James Warren、Jason Rennie、Jeffrey Copeland、Josh Reed、Julien Nioche、Keith Kim、Manish Katyal、MargrietBruggeman、Massimo Perga、NikanderBruggeman、Philipp K. Janert、Rick Wagner、Robi Sen、SanchetDighe、SzymonChojnacki、Tim Potter、Vaijanath Rao和Jeff Goldschrafe
- 感谢在本书特定章节将专业知识贡献给大家的其他作者，他们是：J. Neal Richter、Manish Katyal、Rob Zinkov、SzymonChojnacki、Tim Potter和Vaijanath Rao

- 感谢Steven Rower，感谢他对本书所进行的全面的技术性评阅，也感谢他在TextWise、CNLP和部分Lucene项目时和我们一起共同度过美好时光
- 感谢Liz Liddy博士，感谢他将Drew和Grant引入到文本分析这个领域，感谢他带来的乐趣和机会，也感谢他为本书写序
- 感谢所有的MEAP读者，感谢他们的耐心和反馈
- 最重要的，要感谢我们的家人、朋友和同事，感谢他们的鼓励、精神支持以及对我们将正常的生活时间投入到本书写作的理解

Grant Ingersoll

感谢我在TextWise和CNLP的同事，他们教会了我太多文本分析的知识。感谢Urdahl让数学那么有趣，感谢Raymond女士让我成为一个更好的人和学生，感谢我的父母Floyd和Delores，感谢我的孩子Jackie和William，感谢我的妻子Robin，她忍受了我经常工作到深夜和写作占去的周末时光——谢谢你一直在那里支持我！

Tom Morton

感谢合作者的辛勤工作和团队合作，感谢我的妻子Thuy和女儿Chloe，感谢她们的耐心、支持和给予的自由时间；感谢我的家庭Mortons和Trans对我的鼓励；感谢我在宾夕法尼亚大学和Comcast的同事的支持和合作，特别要感谢Na-Rae Han、Jason Baldridge、Gann Bierner和Martha Palmer；感谢JörnKottmann为OpenNLP所付出的不懈努力。

Drew Farris

感谢Grant让我参与本书撰写和其他一些有趣的项目，感谢我过去和现在的同事，我从他们身上学到大量的东西并同他们共享文本分析、机器学习和开发优秀软件的乐趣；感谢我的妻子Kristin和孩子们Phoebe、Audrey和Owen，感谢他们对我挤时间写书和参与其他技术工作的耐心和支持；感谢我的大家庭，感谢他们的兴趣和鼓励，特别要感谢我的妈妈，虽然她已无法看到本书的完整版本。

关于本书

本书主要关注软件应用的构建，这些应用使用和处理书面文字的文本内容并从中掘取核心价值。尽管本书用较大篇幅介绍了有关搜索、自然语言处理和机器学习的主题，但是它并非一本有关这些主题的理论著作。我们尽量避免术语和复杂的数学公式，而集中关注当今软件工程师、架构师和从业人员为实现下一代智能文本处理应用时所需的一些概念和示例。本书也使用免费可用、高流行度的开源工具（如Apache Solr、Mahout和OpenNLP）来提供书中一些真实世界中的实例的概念。

本书阅读对象

这书是否适合你？或许是。本书的目标读者是那些完全没有或没有太多搜索、自然语言处理和机器学习背景的软件从业人员。实际上，本书主要面对那些我们在很多公司看到的下面这种场景下的从业人员：某开发团队需要在一个新应用或在已有应用上增加搜索和其他功能，但是大部分开发人员并没有文本处理的经验。他们需要一个很好的入门材料来理解这些概念，同时又不会陷入那些不必要的内容之中。

很多情况下，我们提供容易访问的参考资源，比如维基百科和学术论文，因而本书可以作为读者需要对本领域进行深入探索的初始平台。此外，虽然本书大部分开源工具和样例都是基于Java的，但是本书的概念和思路也很容易移植到其他编程语言，因此Ruby、Python和其他语言的爱好者同样会对本书的内容满意。

尽管本书对需要实现教科书和学术书籍中的概念的学生有所帮助，但是其目标

读者很显然不是那些寻求相关系统中数学解释的用户和学术爱好者。

本书的目标读者也不是那些已经在其职业生涯中构建过很多文本处理应用的经验丰富的从业人员，尽管他们可能也会从本书开源包的使用中发现一些有趣的片段。不止一个有经验的从业人员告诉我们，本书可以加快本领域新人在文本处理应用开发中对思路和代码的理解。

最后，我们希望本书是一本面向现代程序员的最新指导书籍，也希望它成为文本处理应用编程职业道路之初所需要的指导书籍。

本书内容组织

第1章解释文本处理的重要性及其具有挑战性的原因。本章将预览一个基于事实的问答系统，以此来设定利用开源库驾驭文本的一个场景。

第2章介绍文本处理中的一些模块构建：切词、组块、分析及词性标注。之后考察利用Apache Tika开源项目从常见文件格式中抽取文本的过程。

第3章探讨搜索理论及向量空间模型的基本知识，重点介绍Apache Solr搜索服务器并给出利用它进行索引的方法。本章将学习如何对搜索性能（数量和质量）进行评估。

第4章考察基于前缀和*n*元组的模糊字符串匹配方法。我们考察两个字符串重叠度的计算方法：Jaccard和Jaro-Winkler距离，并解释如何利用Solr找到候选匹配并对它们进行排序。

第5章给出了命名实体识别背后的基本概念。我们将展示如何使用OpenNLP寻找命名实体并讨论OpenNLP的性能问题。我们还将介绍如何对OpenNLP进行定制从而在新领域中识别命名实体。

第6章主要介绍文本聚类。这一章会学习到常见文本聚类算法背后的基本概念，并且看到聚类如何提升文本应用的例子。我们也会介绍如何使用Apache Mahout来对整个文档集进行聚类，以及如何使用Carrot2对搜索结果进行聚类。

第7章讨论了分类、归类和标注背后的基本概念。我们会展示分类如何用于文本应用，并且介绍如何利用开源工具来构建、训练和评估分类器。我们还会使用Mahout中的朴素贝叶斯实现来构建文档分类器。

第8章综合前面7章学到的知识构建一个示例QA系统。这个简单的应用利用维基

百科作为知识库，并利用Solr作为基线系统。

第9章探讨搜索和NLP的下一步发展方向及语义、篇章和语用的角色。我们将介绍跨多种语言的搜索、内容中的情感探测，以及新兴的工具、应用和思想。

代码约定及下载

本书包含大量代码样例，所有源代码均采用等宽字体以区分普通文本。代码中的方法名称、类名称和其他元素也采用等宽字体表示。

在很多清单中，代码加以标注以指出重要概念，有时文本中也给出了项目编号来提供代码的额外信息。

本书的源代码样例与在线样例相当接近。但是为简洁起见，书中源码中去掉了像注释一样的内容，以保证代码能够方便地嵌入到文本中。

本书示例的源代码可以从出版社网站www.manning.com/TamingText下载。

作者在线

购买本书的读者能够免费访问Manning出版社管理的一个私有Web论坛，可以在这个论坛上发表对本书的评论、询问技术问题并从作者或其他用户那里得到帮助。你可以通过地址www.manning.com/TammingText访问和订阅该论坛。完成注册后，你可以了解如何访问论坛、该论坛所能提供的帮助及论坛的行为规范。

Manning出版社承诺为读者和作者提供一个进行深入对话的场所，但不对作者的参与程度做任何要求，他们对于该论坛的贡献出于自愿且没有任何报酬。我们建议读者尽量向作者提一些具有挑战性的问题，这样可以让他们保持兴趣！

本书在印期间，读者均可访问作者在线论坛，并查看之前的讨论。

关于封面

封面插图的标题是“Le Marchand”，是商人或店主的意思。该插图取自法国出版的Sylvain Maréchal的一个19世纪版本的四卷地域服饰习俗汇编。汇编中的每幅图都精心描画、手工上色。丰富多样的Maréchal作品生动地告诉我们，200年前的文化差异是如何之大，它将世界上的城镇和地域区分开来。人们彼此远离，说着不同的乡音和语言。不管是在街道或乡村，人们很容易通过服饰就能区分他们居住的位置、交易或购置的物品。

从那之后，服饰的密码逐渐改变，那时地域之间的丰富多样性也逐渐消失。现在很难区分来自不同洲的居民，更别说来自不同城镇或地区的人了。或许我们以文化多样性为代价换来了更多样的个人生活，当然也是更丰富的快节奏的技术生活。

在一个很难分辨两本计算机书籍的年代，Manning出版社通过Maréchal的图画将我们带回过去，封面取材于200年前生活的多样性，借此颂扬计算机行业的创造力和首创精神。

目录