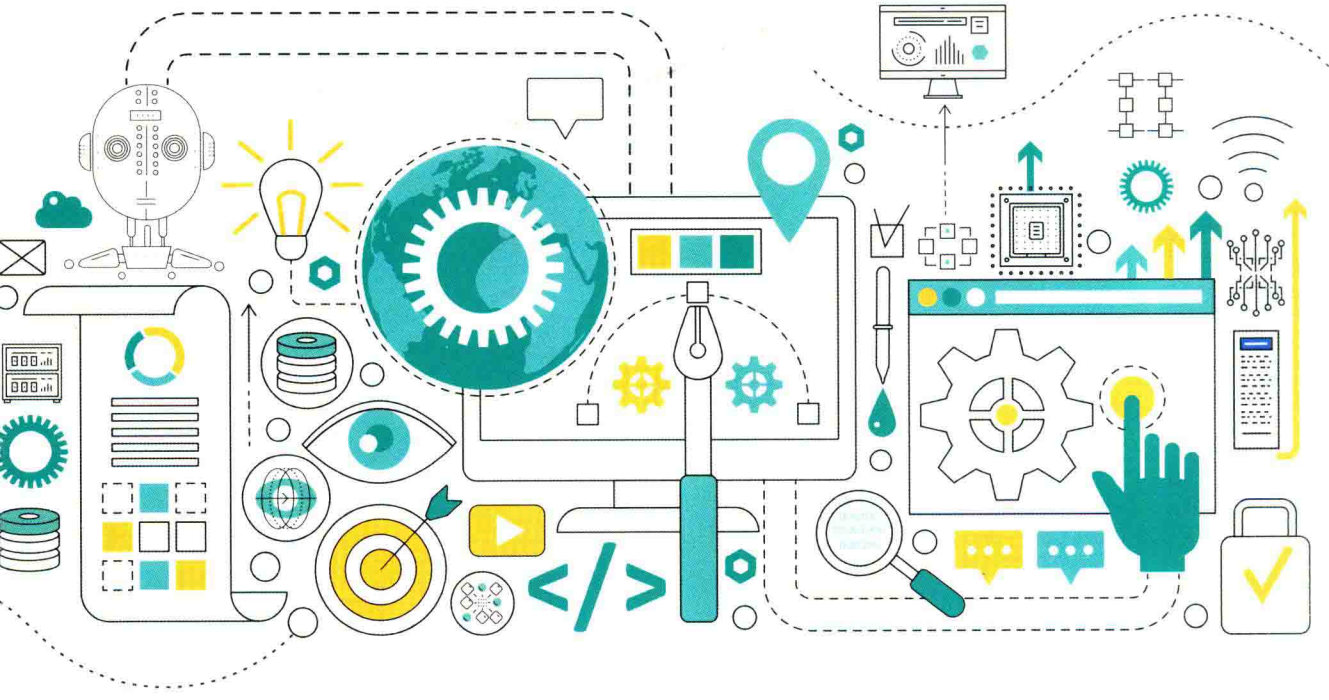
HZ BOOKS 华章IT
[PACKT] PUBLISHING
智能系统与技术丛书

Python Natural Language Processing

Python自然语言处理

[印度] 雅兰·萨纳卡（Jalaj Thanaki）著
张金超 刘舒曼 译



机械工业出版社
China Machine Press

智能系统与技术丛书

Python Natural Language Processing

Python自然语言处理

[印度] 雅兰·萨纳卡（Jalaj Thanaki）著
张金超 刘舒曼 译

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

Python 自然语言处理 /（印）雅兰·萨纳卡（Jalaj Thanaki）著；张金超，刘舒曼译 . —北京：机械工业出版社，2018.8
（智能系统与技术丛书）
书名原文：Python Natural Language Processing

ISBN 978-7-111-60670-3

I. P… II. ①雅… ②张… ③刘… III. 软件工具 – 自然语言处理 IV. ① TP311.56 ② TP391

中国版本图书馆 CIP 数据核字（2018）第 185824 号

本书版权登记号：图字 01-2017-7520

Python 自然语言处理

出版发行：机械工业出版社（北京市西城区百万庄大街 22 号 邮政编码：100037）
责任编辑：张梦玲　　责任校对：李秋荣
印　　刷：三河市宏图印务有限公司　　版　　次：2018 年 9 月第 1 版第 1 次印刷
开　　本：186mm × 240mm　1/16　　印　　张：18.75
书　　号：ISBN 978-7-111-60670-3　　定　　价：79.00 元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换
客服热线：（010）88379426　88361066　　投稿热线：（010）88379604
购书热线：（010）68326294　88379649　68995259　　读者信箱：hzit@hzbook.com

内容简介

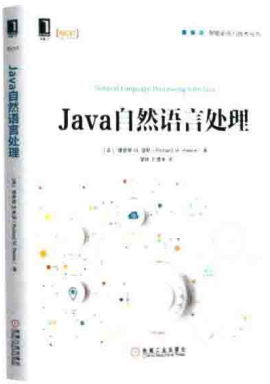
人工智能研究中一个重要的部分就是实现机器设备和人的无障碍交互，而人类最自然常用和最精确的交互方式便是使用语言和文字。因此，从一定程度上说，计算设备对自然语言处理的能力，决定了其人工智能的智力。

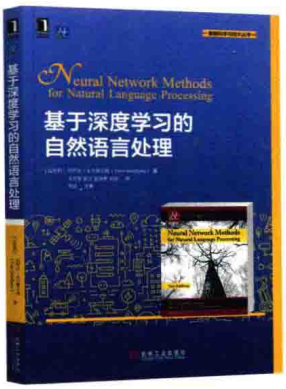
本书致力于总体介绍自然语言处理领域中的一些概念、术语、应用任务、算法和技术、系统搭建方法等，非常适合作为对自然语言处理任务感兴趣的初学者进入该领域的入门书籍。

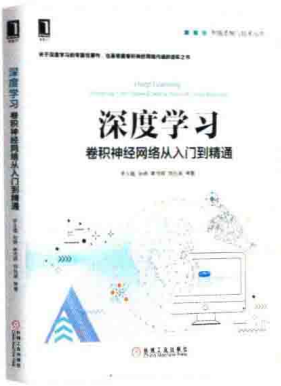
你能学习到：

- 自然语言处理应用开发中的Python编程方法，理解自然语言数据属性和语料分析处理的方法。
- 使用Python类库处理自然语言，像NLTK、Polyglot、SpaCy、Standford CoreNLP等。
- 特征工程中特征抽取和特征选择的方法。
- 深度学习中向量化方法的优势。
- 更好地理解规则式系统的架构。
- 使用自然语言处理中的有监督和无监督机器学习方法进行训练与调优。
- 为自然语言处理和自然语言生成问题找到合适的深度学习方法。

拓展阅读







封面设计　锡　彬

THE TRANSLATOR'S WORDS

译 者 序

人类从蛮荒时代进入文明时代的一个重要标志是文字的出现，即开始以符号的形式记录发生的事件。一方面，文字的出现说明人类的思维方式开始从具象往抽象方向进化，能够从现实当中具体的事情联想到某个或某些由抽象线条构成的符号，也能从这些符号演绎出现实中的事情；另一方面，文字的出现使得人类社会具有了存储信息和传递信息的能力，促成了个体和群体之间跨越地域和时间的交流。

中国人的远古神话传说中，仓颉造字引发了“天雨粟，鬼夜哭”。唐人张彦远对此的解释是：有字之后，“造化不能藏其秘，故天雨粟；灵怪不能遁其形，故鬼夜哭”。文字对于人类文明的贡献远超过像冶炼、排水等这些生产生活中的实用技术。历史和文化以文字为载体，传播千年，横跨世界。当今时代，文字依旧扮演着信息传递载体的作用。不同的是，我们目前日常接触和处理的文本信息大部分存在于计算机、手机等数字设备上。借助高速的计算设备，使用有效的算法对这些数字化的文本信息进行处理分析、抽取语义、完成特定任务，就是**自然语言处理**（Natural Language Processing，NLP）要做的事情。

目前我们所处的时代，被称为**大数据**（Big Data）时代和**人工智能**（Artificial Intelligence）时代。所谓的大数据中最大部分的数据就是文本数据，互联网上每天会产生常人难以想象规模的文本数据，人类社会的绝大部分知识蕴含在文本信息中。人工智能研究中一个重要的部分就是实现机器设备和人的无障碍交互，而人类最自然、最常用和最精确的交互方式便是使用语言和文字。因此可以从一定程度上说，计算设备对自然语言处理的能力，决定了其人工智能的智力。

本书致力于系统介绍自然语言处理领域中的一些概念、术语、应用任务、算法和技术、系统搭建方法等，非常适合作为对自然语言处理任务感兴趣的初学者进入该领域的入门书籍。书中第 1 章对自然语言处理任务进行了总体介绍，第 2 章介绍了语料和数据的获取和构建方法，第 3 章介绍了句子的结构，第 4 章介绍了数据预处理的方法，第 5

章和第 6 章介绍了特征工程和一些算法，第 7 章介绍了规则式自然语言处理系统的搭建方式，第 8 章介绍了自然语言处理中的机器学习方法，第 9 章介绍了自然语言处理中的深度学习方法，第 10 章介绍了一些高级工具，第 11 章给了一些实用的学习建议，以帮助读者进行提升，第 12 章是书中用到的软件和环境的安装指导。

本书具有几个明显的特点：

- 涉及的自然语言处理任务和方法较多，内容比较概述。初学者根据本书所讲内容，对自然语言处理会有一个比较明晰的结构化的图景认识，能够走通自然语言处理系统的搭建流程。
- 书中有大量的例子帮助读者理解概念和算法，同时由于作者拥有较强的工业界背景，书中非常强调系统和算法的实现，提供了大量的可执行代码供读者学习。
- 本书更偏重于教会读者入门自然语言处理领域的方法，提供了大量的学习建议，供读者参考。

最后，非常感谢刘群老师的推荐和机械工业出版社华章公司的信任，让我们承担了本书的翻译工作。在几个月的时间里，我和我的博士生师妹刘舒曼对这本书进行了认真的翻译和校对。在这个过程中，我们也从书中学习到了丰富的知识，看到了作者一些有趣的工业界视角，在此向作者 Jalaj Thanaki 女士致谢！

FOREWORD

推 荐 序

数据科学正极速地改变着这个世界和我们的商业领域，例如零售业、银行和金融服务业、出版业、制药业、制造业等。各种不同格式的数据以指数级别的速度在产生，包括定量的数据、定性的数据、有结构的数据、无结构的数据、语音数据、视频数据等。可以利用这些数据来避免风险和诈骗、改善用户体验、增加营收、精简作业等。

许多机构正在快速地拥抱数据科学并投资了很多高端的数据科学团队。在银行和金融保险领域从业 12 年多，我见证了该领域接受把数据分析当作生意来做，而不再仅仅是用来支持服务。这一点在金融科技和数字贷款领域尤其明显。

在 Jalaj 上大学时，我就已经认识她了，她的活力四射和强大的自我驱动力让我印象深刻。她的研究能力、毅力、投入程度、纪律性和快速掌握复杂概念的能力，让她在短短 4 年的企业工作中，取得了极大的成就。

Jalaj 在数学和统计方面有很好的天赋，并且在面对学习工业界新出现的复杂数据统计分析方法时，表现出了持续的热情。她有丰富的数据科学领域的工作经验，我目睹她完成了很多令人瞩目的项目，这些项目围绕着自然语言处理、机器学习、基础语言学分析、神经网络和深度学习展开。她的工作节奏高效快速，并有很高的工作热情，这些给其所在的组织带来了明确可量化的成果。

Jalaj 最特别的品质之一是，她时刻准备着解决商业中的各种问题，不论是最基本的问题还是最复杂的问题。她是一个很好的团队合作者，能够保证所在机构从她的出色才能中获得最大收益。

在本书中，Jalaj 会带领我们开始一场令人兴奋并深刻的自然语言处理之旅。她会从基本概念讲起，直至最新的概念，例如机器学习方法和深度学习方法是如何应用在自然

语言处理中的。

祝愿 Jalaj 在未来做的更好！

Sarita Arora
SMECorner 首席分析官
印度孟买

ABOUT THE AUTHOR

作者介绍

Jalaj Thanaki 是一名数据科学领域的研究者和科学家。她喜欢解决和数据科学相关的问题。她希望能够使用数据科学和人工智能技术，让这个世界变得更美好。她的研究兴趣包括自然语言处理、机器学习、深度学习和大数据分析。除了是一名数据科学家之外，Jalaj 也是一位社会活动家、旅行家和大自然爱好者。

我想将这本书献给我的丈夫 Shetul Thanaki，因为他给了我持续的支持、鼓励和创造性的建议。

对我的父母、公婆、家庭成员和朋友致以深深的谢意，他们在我人生的每个阶段提供了强大的帮助。也要感谢这些年遇到的所有导师。还十分感谢技术审校人员在审阅这本书时付出的努力。同时，感谢我现在所在的机构——SME Corner 提供的支持。我是很多开源和教育社区的粉丝，所以十分感谢 Kaggel、Udacity、Coursera 等那些直接或间接帮助我理解数据科学各种概念的社区。如果没有从这些社区中学习知识，我不可能有机会做现在做的事情。

感谢 Packt Publishing 和 Aman Singh，是他们找我来写这本书。十分感谢 Packt 编辑团队为了让这本书尽量完美所做的努力。特别要感谢 Aman Singh、Jagruti Babaria、Menka Bohra、Manthan Patel、Nidhi Joshi、Sayli Nikalje、Manisha Sinha、Safis 和 Tania Dutta。

感谢技术编辑团队、策略和管理团队、市场团队、销售团队、图表设计团队、前期制作团队、后期制作团队、排版团队等所做的努力，是他们让我的写作之旅更顺利。

我非常愿意将我的知识传递给那些想要学习的人。

祝读者有一个振奋和愉快的阅读之旅！

ABOUT THE REVIEWERS

关于审校人员

Devesh Raj 是一名在医疗保健、制造、汽车制造、生产等领域具有 10 年算法开发和解决实际问题经验的数据科学家。他的工作是把机器学习方法（有监督和无监督技术）和深度学习方法应用到有结构和无结构数据上（计算机视觉和自然语言处理）。

Gayetri Thakur 是一名自然语言处理领域的语言学家。她参与研发了一些自然语言处理的工具，如自动语法检查、命名实体识别、文本到语音转换、语音到文本转换的系统。她目前在 Google India Pvt.Ltd. 工作。她在 Banaras Hindu 大学完成语言学硕士学位后，正在继续攻读语言学博士学位。

Prabhanjan Tattar 拥有 9 年多的数据分析师工作经验。生存分析和统计推理是他主要的研究兴趣。他曾在相关技术杂志中发表过多篇研究论文，并写了三本书：《R Statistical Application Development by Example》《A Course in Statistics with R》《Practical Data Science Cookbook》。他还维护着 R 包 gpk、RSADBE 和 ACSWR。

Chirag Mahapatra 是一名软件工程师。他致力于将机器学习方法和自然语言处理技术可靠地应用到具体问题中。他现在工作于 Trooly（已被 Airbnb 收购）。之前，他在 A9.com 的广告数据平台工作。

PREFACE

前　言

本书的名字会给你带来广阔的联想。作为读者，你会从头学习关于自然语言处理的方方面面。本书用了非常简单的语言来具体阐述 NLP 中的概念，许多真实有趣的实用样例会加深你对该领域的理解。通过实现这些样例，能提升你的 NLP 技能。

现在，我来回答一些经常会被问起的 NLP 领域的问题。这些问题启发了我来写这样一本书。对我来说，让所有读者能够理解我写这本书的初衷，是一件非常重要的事情。

第一个经常被问起的问题是：什么是自然语言处理？第二个是：为什么在开发自然语言处理程序时主要会用 Python 呢？最后一个重要的问题是：在学习自然语言处理的时候，有哪些资源可用？现在，让我们来看一下答案！

第一个问题的答案是，自然语言，简单来说，就是你所说的、写的、读的或理解的人类自然的语言，同时是交流的媒介。我们使用计算机算法、数学概念、统计技术来处理这些语言，使得机器能够像人一样理解。

现在来回答第二个问题。最简单和直接的原因是 Python 有大量的库，这些库在你开发自然语言处理应用程序的时候，会让事情变得简单。第二个原因是，如果你有 C 或是 C++ 的编程经验，你不用再担心会遇到内存泄漏的问题。Python 的解释器会为你解决这个问题，你要做的仅仅是关注主要的编程过程。除此之外，Python 是一个程序员友好的语言，与其他面向对象的语言比起来，你只需要写少量的代码就可以做更多的事情。因此，所有的这些事实都驱动着人们使用 Python 来开发自然语言处理应用程序以及其他数据科学相关的应用，以进行更快的建模。

最后一个问题对我来说很重要，因为我经常向朋友解释上面的答案，他们听完后会想学 Python，但是有哪些可用的资源呢？我通常会推荐一些书籍、博客、YouTube 上的视频，还有 Udacity、Coursera 等教育平台。但是几天以后，他们还会来问我有没有一个单一的学习资源——书籍或博客。不幸的是，答案是否定的。在那个时刻，我意识到兼顾所有这些资源对他们来说是比较困难的事情。这种痛苦的领悟成了我写作

这本书的动力。

所以在这本书里，我尝试着覆盖大多数自然语言处理中的必要知识，这些对每个人来说都是有用的。一个好消息是我提供了很多实用的 Python 样例，这样读者便既能从理论角度，也能从应用角度出发理解这些概念。阅读、理解、编码是这本书的三个主要部分，会帮助读者轻松学习。

本书内容

第 1 章提供了对自然语言处理和自然语言处理领域其他分支的介绍。我们会看到构建自然语言处理应用程序的各个阶段，并讨论 nltk 安装的问题。

第 2 章展示了语料分析的各个方面。我们将会看到不同类型的语料和语料中展现的数据属性，会接触到各种语料格式，像 CSV、JSON、XML、LibSVM 等。还会看到关于网页爬取的样例。

第 3 章会帮助你理解自然语言里面最基本的东西，也就是语言学。我们会看到词法分析、句法分析、语义分析、消歧等诸多概念。也会使用 nltk 来实际地理解这些概念。

第 4 章会帮助你弄懂各种不同类型的预处理技术以及该怎样定制它们。我们将会看到预处理的各个阶段，像数据准备、数据处理、数据转换。除了这些，你还会从实际应用的角度来理解预处理。

第 5 章是自然语言处理应用程序里最核心的部分。我们会看到不同的算法和工具是怎样用于生成机器学习算法的输入的，它们会被用来开发自然语言处理应用程序。我们也会理解特征工程里的统计概念，然后会对这些工具和算法进行定制化开发。

第 6 章会让你理解处理语义问题时遇到的自然语言处理概念。我们会看到 word2vec、doc2vec、GloVe 等，以及从《权力的游戏》数据集中获得向量的一些 word2vec 的实际应用。

第 7 章会给出很多构建一个规则式系统的细节，以及开发类似自然语言处理系统时，需要牢记的方方面面。我们会看到制定规则的过程和编码规则的过程，也会看到怎样开发一个基于模板的聊天机器人。

第 8 章会提供给你一些新的机器学习技术。我们会看到用于开发自然语言处理应用程序的各种机器学习算法，也会使用机器学习方法实现一些强大的自然语言处理应用程序。

第 9 章会介绍人工智能的很多方面。我们将会看到人工神经网络的基本概念，以及如何才能构建一个人工神经网络。我们将会理解深度学习的核心，研究深度学习的数学

原理，并看一下深度学习是怎样用来做自然语言理解和自然语言生成的。你可以在这里看到很多有意思的实践样例。

第 10 章会简单介绍一些框架，像 Apache Hadoop、Apache Spark 和 Apache Flink。

第 11 章会介绍怎样提高 NLP 技能。

第 12 章会介绍针对必要程序的安装指导。

预备知识

下面来说一下阅读这本书的预备知识。不用紧张，这里不涉及数学或统计学知识，仅仅是 Python 的基本编程语法。除了这些之外，你需要在计算机上安装 Python 2.7.X 或 Python 3.5.X。推荐你使用任意的 Linux 系统。

Python 的一些依赖列表可以在如下链接找到：https://github.com/jalajthanaki/NLPython/blob/master/pip-requirements.txt.

现在来看一下所需的硬件条件。有 4GB 内存和双核 CPU 的电脑足够执行普通代码，但是对于机器学习和深度学习样例，你可能需要更多的内存（8GB 或 16GB）和 GPU 计算资源。

本书的读者对象

本书面向想应用 NLP 技术来使他们的应用程序更智能的 Python 开发者，可作为入门 NLP 领域的资料。

下载样例源码

本书的代码在 GitHub 网站上可找到：https://github.com/PacktPublishing/Python-Natural-Language-Processing。另外还有一些拓展资源和视频可以在如下地址找到：https://github.com/PacktPublishing/。

另外，读者还可在华章公司官网 http://www.hzbook.com / 上搜索本书，下载源代码。

代码文件下载完以后，确保你的解压工具是可用的：

- WinRAR/7-Zip 在 Windows 系统下
- Zipeg/iZip/UnRar 在 Mac 系统下
- 7-Zip/PeaZip 在 Linux 系统下

CONTENTS

目　录