异步图书 www.epubit.com

Packt>

Natural Language Processing: Python and NLTK

Python和NLTK自然语言处理

尼天 • 哈登尼亚（Nitin Hardeniya）

[印度] 雅各布 • 帕金斯（Jacob Perkins） 迪蒂 • 乔普拉（Deepti Chopra） 著

尼什 • 斯乔希（Nisheeth Joshi） 伊提 • 摩突罗（Iti Mathur）

林赐 译

中国工信出版集团

人民邮电出版社 POSTS & TELECOM PRESS

Python和NLTK
自然语言处理

尼天 • 哈登尼亚（Nitin Hardeniya）
[印度] 雅各布 • 帕金斯（Jacob Perkins） 迪蒂 • 乔普拉（Deepti Chopra） 著
尼什 • 斯乔希（Nisheeth Joshi） 伊提 • 摩突罗（Iti Mathur）
林赐 译

人 民 邮 电 出 版 社
北 京

图书在版编目（CIP）数据

Python和NLTK自然语言处理 / （印）尼天·哈登尼亚(Nitin Hardeniya) 等著 ; 林赐译. -- 北京 : 人民邮电出版社, 2019.4
ISBN 978-7-115-50334-3

Ⅰ. ①P… Ⅱ. ①尼… ②林… Ⅲ. ①软件工具－程序设计 Ⅳ. ①TP311.561

中国版本图书馆CIP数据核字(2018)第277916号

版权声明

◆ 著　[印度]尼天·哈登尼亚（Nitin Hardeniya）
[印度]雅各布·帕金斯（Jacob Perkins）
[印度]迪蒂·乔普拉（Deepti Chopra）
[印度]尼什·斯乔希（Nisheeth Joshi）
[印度]伊提·摩突罗（Iti Mathur）
译　林　赐
责任编辑　谢晓芳
责任印制　焦志炜

◆ 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
固安县铭成印刷有限公司印刷

◆ 开本：800×1000　1/16
印张：40.75
字数：810 千字　2019 年 4 月第 1 版
印数：1－2000 册　2019 年 4 月河北第 1 次印刷

著作权合同登记号　图字：01-2017-5038 号

定价：138.00 元

读者服务热线：(010)81055410　印装质量热线：(010)81055316
反盗版热线：(010)81055315
广告经营许可证：京东工商广登字 20170147 号

内容提要

本书旨在介绍如何通过 Python 和 NLTK 实现自然语言处理。本书包括三个模块。模块 1 介绍文本挖掘/NLP 任务中所需的所有预处理步骤，包括文本的整理和清洗、词性标注、对文本的结构进行语法分析、文本的分类等。模块 2 讲述如何使用 Python 3 的 NLTK 3 进行文本处理，包括标记文本、替换和校正单词、创建自定义语料库、词性标注、提取组块、文本分类等。模块 3 讨论了如何通过 Python 掌握自然语言处理，包括统计语言模型、词语形态学、词性标注、解析、语义分析、情感分析、信息检索等。

本书适合 NLP 和机器学习领域的爱好者、Python 程序员以及机器学习领域的研究人员阅读。

译者序

不下雪的日子，渥太华一片喧闹。2018 年 4 月的雪刚刚化去，5 月的郁金香节姗姗来迟，举目望去，大片大片的郁金香花把渥太华装点得五彩缤纷，绚烂多姿，生机盎然。在阳光的照射下，各色的郁金香花芬芳吐蕊、娇艳妩媚，令人陶醉和流连忘返。郁金香的芬芳还未散去，渥太华就迎来了“炎天暑月”的夏季，说是夏季，气温却很少超过 30℃，此时，蛰伏了一个冬天的加拿大人，经受不住太阳的诱惑，蜂拥出动，三五成群地在里多运河的沙滩上玩耍嬉戏，好不自在。门户开放日（Doors Open Ottawa）、爵士音乐节、马拉松，各种活动精彩纷呈。正如鲁迅所说，越是民族的，就越是世界的。世界各地的移民带着自己家乡的文化和节日来到了渥太华。例如，中国的龙舟节成了这里的盛典，同时各个民族的人民也各自庆祝自己的节日，如夏至原住居民节等。

加拿大是个移民国家，在这里，既可以看到不同的民族着装，也可以看到迥异的服饰风格。在不同的文化中，对个性的定义也各不相同，因此即使使用同一个词来形容一个人，也具有不同的韵味。作为中国人，我的思想不偏不倚，因此对各种奇装异服也就抱着听之任之的态度。我既欣赏热情奔放，也不排斥含蓄典雅。但是，对于计算机而言，要体察到语言的深层次含义，可谓“蜀道难，难于上青天”。本书介绍了如何编写能够理解人类语言的系统。

在翻译此书时，我想起了“通天塔”（The Tower of Babel）的故事。在《旧约》中，巴比伦人（Babylonian）想建造一座通天塔，但是上帝对此感到不满，因此借由变乱了人类的语言，使他们互相不能交流，破坏了这一工程。长久以来，人们一直在寻找不同语言之间的沟通方法。在 20 世纪三四十年代，凭借计算机的计算能力，人们开始了机器翻译的伟大尝试。这 80 多年来，人们在自然语言处理方面前赴后继，各个国家你追我赶，都想在这一领域拔得头筹，再造一座通天塔。直至最近，人类才第一次看到了梦想的曙光。市面上

也出现了各种机器翻译的软件，极大提高了人类译者的工作效率。这些软件具有可用性和市场价值。但是，机器翻译只是自然语言处理领域一个非常小的应用部分。自然语言处理其实包罗万象，让人眼花缭乱。其应用还包括问答系统、情感分析、图片题注、语音识别、词性标注、命名实体识别等。

自然语言处理博大精深，是一门融语言学、计算机科学、数学、统计于一体的科学。“纸上得来终觉浅，绝知此事要躬行。”要想完全掌握现在的自然语言技术，靠阅读几本书、几篇文章，是远远不够的。选择学习自然语言处理，并希望成为机器学习工程师，要经过三个阶段：实践、思考、创新。正如《荀子·儒效》中所说，“不闻不若闻之，闻之不若见之，见之不若知之，知之不若行之。学至于行止矣。”但是，“学而不思则罔，思而不学则殆。”在实践中，我们也要注重思考，寻找技术背后的原理，才能触类旁通，举一反三。作为科研工作者，仅仅成为知识的搬运工是不思进取的表现。“删繁就简三秋树，领异标新二月花。”掌握了知识之后，我们还应该学会创新，唯有这样，才可以成为社会的中流砥柱，引领技术的潮流。

在这里，我要特别感谢人民邮电出版社的领导和编辑，感谢他们对我的信任和理解，把这样一本好书交给我翻译。我也要感谢他们为本书的翻译投入了巨大的热情，可谓呕心沥血。没有他们的耐心和帮助，本书不可能顺利付梓。同时，在翻译过程中，加拿大友人 Jack Liu 和 Connie Wang 多次帮我指点迷津，才能使我为读者提供更贴切的译文。

译者才疏学浅，见闻浅薄，译稿中多有疏漏之处，还望读者谅解并不吝指正。读者如有任何意见和建议，请将反馈信息发送到邮箱 cilin2046@gmail.com。本书全由林赐翻译。

林　赐

前言

NLTK 是自然语言处理（Natural Language Processing，NLP）社区中最受欢迎和广泛使用的库之一。NLTK 的优点在于其简单性，其中大多数复杂的 NLP 任务可以使用几行代码实现。本书主要内容包括：如何将文本标记为各个单词，如何使用 WordNet 语言词典，如何以及何时进行词干提取或者词形还原，如何替换单词和纠正拼写，如何创建自己的自定义文本语料库和语料库（包括 MongoDB 支持的语料库）读取器，如何使用词性标注器和部分词性标注单词，如何使用部分解析创建和转换分块短语树，如何进行文本分类的特征提取和情感分析，如何进行并行和分布式文本处理，以及如何在 Redis 中存储单词分布。

这种一边学习一边动手实践的学习方式会教你更多知识。本书有助于你成为使用 NLTK 进行自然语言处理的专家。

本书主要内容

模块 1 讨论文本挖掘/NLP 任务中所需的所有预处理步骤。该模块详细讨论标记化、词干提取、停用词删除和其他文本清理过程，以及如何在 NLTK 中轻松实现这些操作。

模块 2 解释如何使用语料库读取器和创建自定义语料库。它还介绍如何使用 NLTK 附带的一些语料库。它涵盖组块过程（也称为部分分析），组块过程可以识别句子中的短语和命名实体。它还解释如何训练自己的自定义组块器并创建特定的命名实体识别器。

模块 3 讨论如何计算单词频率和实现各种语言建模技术。它还讨论浅层语义分析（即 NER）的概念和应用及使用 Wordnet 的 TSD。

模块 3 有助于你理解和应用信息检索与文本摘要的概念。

学习本书的软硬件配置

在学习模块 1 时，需要满足的软硬件配置如下表所示。

章号	需要的软件	免费/专用	下载软件的网站	硬件规格	需要的操作系统
第 1～5 章	Python/Anaconda 和 NLTK	免费	Python 官网、continuum 官网和 NLTK 官网	通用 UNIX 打印系统	任意
第 6 章	scikit-learn 和 gensim	免费	scikit-learn 官网、radimrehurek 官网	通用 UNIX 打印系统	任意
第 7 章	Scrapy	免费	Scrapy 官网	通用 UNIX 打印系统	任意
第 8 章	NumPy、SciPy、Pandas 和 Matplotlib	免费	Numpy 官网、Scipy 官网、Pandas 官网和 Matplotlib 官网	通用 UNIX 打印系统	任意
第 9 章	Twitter、Python API 和 Facebook API	免费	Twitter 官网和 Facabook 官网	通用 UNIX 打印系统	任意

在学习模块 2 时，需要 Python 3 和列出的 Python 包。在本书中，作者使用了 Python 3.3.5。要安装这些包，可以使用 pip（参见 Python 官网）。以下是学习模块 2 时需要的包列表，以及编写本书时使用的版本号。

- NLTK≥3.0a4
- pyenchant≥1.6.5
- lock file≥0.9.1
- Numpy≥1.8.0
- Scipy≥0.13.0
- scikit-learn≥0.14.1
- execnet≥1.1
- pymongo≥2.6.3

- Redis≥2.8.0
- lxml≥3.2.3
- beautifulsoup4≥4.3.2
- python-dateutil≥2.0
- charade≥1.0.3

你还需要 NLTK-Trainer，可在 GitHub 网站上获得它。

除了 Python 之外，还有使用 MongoDB 和 Redis 两个 NoSQL 的技巧。MongoDB 可以在 MongoDB 官网下载，Redis 可以在 Redis 官网下载。

对于模块 3 的所有章节，使用了 Python 2.7 或 3.2+。NLTK 3.0 必须安装在 32 位计算机或 64 位计算机上。所需的操作系统是 Windows/Mac/UNIX 系统。

本书读者对象

如果你是NLP或机器学习爱好者以及想要快速掌握使用NLTK进行自然语言处理的中级 Python 程序员，那么本书的章节安排将为你带来很多好处。语言学和语义学分析方面的专业人士也会从本书中收益。

读者反馈

欢迎来自读者的反馈。让我们知道你对这本书的看法——你喜欢或不喜欢的内容。读者反馈对我们很重要，因为它可以帮助我们开发真正有用的选题。

要向我们发送一般反馈，请发送电子邮件至 feedback@packtpub.com，并在你的邮件标题中包含本书的书名。

如果你精通某方面的专业知识，并且有兴趣撰写或参与撰写某本书，请参考 packtpub.com 网站上的作者指南。

客户支持

既然你购买了 Packt 图书，那么还有更多配套资源可以帮助获得更大收益。

下载示例代码

可以从你在 packtpub.com 网站上的账号中下载本书的示例代码文件。如果你从其他地方购买了本书，那么请你访问 packtpub.com 网站并注册，之后客服人员会直接通过电子邮件向你发送示例代码文件。

可以按照以下步骤下载示例代码文件。

（1）使用你的电子邮件地址和密码登录 packtpub.com 网站并注册。

（2）将鼠标指针悬停在顶部的 SUPPORT 选项卡上。

（3）单击 Code Downloads & Errata 按钮。

（4）在 Search 框中输入书名。

（5）选择你要下载示例代码文件的书名。

（6）从你已购买书目的下拉菜单中选择对应的书名。

（7）单击 Code Download 按钮。

还可以通过单击 Packt Publishing 网站上本书页面上的 Code Files 按钮来下载代码文件。可以通过在 Search 框中输入本书的书名来访问此页面。请注意，你需要登录你的 Packt 账户。

下载文件后，请确保使用以下最新版本的解压缩软件解压缩文件夹。

- TinRAR/7-Zip（对于 Windows 系统）
- Zipeg/iZip/UnRarX（对于 Mac 系统）
- 7-Zip/PeaZip（对于 Linux 系统）

本书的代码包也托管在 GitHub 网站上。在 GitHub 网站上的书目、视频和课程目录中还提供了其他的代码包。请访问 GitHub 网站确认一下。

勘误表

虽然我们已经尽力确保本书内容的准确性，但是错误在所难免。如果你在我们的一本书中发现了错误，可能是文字或代码中的错误，如果你能向提交勘误，我们将不胜感激。通过这样做，可以避免其他读者少走弯路，并帮助我们进一步提升本书后续版本的质量。

如果你发现了任何勘误，请访问 packtpub 官网，选择书名，单击 Errata Submission Form 链接，并输入详细的勘误信息进行报告。一旦你的勘误表通过了验证，将会接受你的提交，并且将勘误信息上传到我们的网站或添加到本书勘误部分下现有的勘误表中。

要查看以前提交的勘误表，请访问 packtpub 官网并在 Search 框中输入本书的名称。所需信息将显示在 Errata 部分下。

盗版行为

盗版因特网上受版权保护的内容是所有媒体上层出不穷的问题。在 Packt，我们非常重视保护我们的版权和许可。如果你在因特网发现以任何形式抄袭我们作品的行为，请立即向我们提供网络地址或网站名称，以便我们采取补救措施。

如果你发现了可疑的盗版内容，请过 copyright@packtpub.com 联系我们。

感谢你帮助保护作者的版权，能够为你提供有价值的内容，我们也感到非常欣慰。

问题

如果你对本书的任何方面有疑问，欢迎通过 questions@packtpub.com 联系我们，我们将尽力解决问题。

资源与支持

本书由异步社区出品，社区（https://www.epubit.com/）为您提供相关资源和后续服务。

配套资源

本书配套资源包括书中示例的源代码。

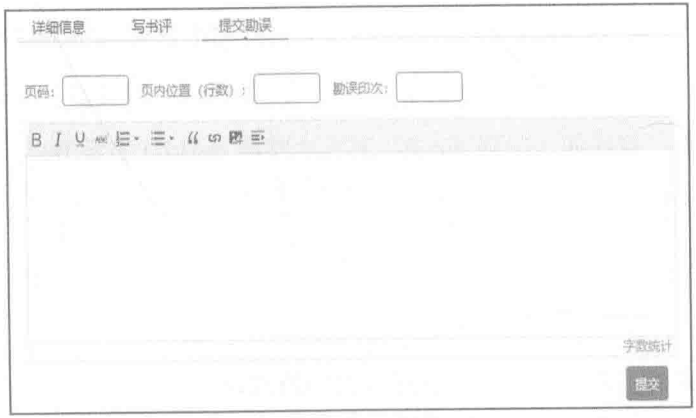
要获得以上配套资源，请在异步社区本书页面中单击 配套资源 ，跳转到下载界面，按提示进行操作即可。注意，为保证购书读者的权益，该操作会给出相关提示，要求输入提取码进行验证。

如果您是教师，希望获得教学配套资源，请在社区本书页面中直接联系本书的责任编辑。

提交勘误

作者和编辑尽最大努力来确保书中内容的准确性，但难免会存在疏漏。欢迎您将发现的问题反馈给我们，帮助我们提升图书的质量。

当您发现错误时，请登录异步社区，按书名搜索，进入本书页面，点击“提交勘误”，输入勘误信息，点击“提交”按钮即可。本书的作者和编辑会对您提交的勘误进行审核，确认并接受后，您将获赠异步社区的 100 积分。积分可用于在异步社区兑换优惠券、样书或奖品。



扫码关注本书

扫描下方二维码，您将会在异步社区微信服务号中看到本书信息及相关的服务提示。

与我们联系

我们的联系邮箱是 contact@epubit.com.cn。

如果您对本书有任何疑问或建议，请您发邮件给我们，并请在邮件标题中注明本书书名，以便我们更高效地做出反馈。

如果您有兴趣出版图书、录制教学视频，或者参与图书翻译、技术审校等工作，可以发邮件给我们；有意出版图书的作者也可以到异步社区在线提交投稿（直接访问 www.epubit.com/selfpublish/submission 即可）。

如果您是学校、培训机构或企业，想批量购买本书或异步社区出版的其他图书，也可以发邮件给我们。

如果您在网上发现有针对异步社区出品图书的各种形式的盗版行为，包括对图书全部或部分内容的非授权传播，请您将怀疑有侵权行为的链接发邮件给我们。您的这一举动是对作者权益的保护，也是我们持续为您提供有价值的内容的动力之源。

关于异步社区和异步图书

“异步社区”是人民邮电出版社旗下 IT 专业图书社区，致力于出版精品 IT 技术图书和相关学习产品，为作译者提供优质出版服务。异步社区创办于 2015 年 8 月，提供大量精品 IT 技术图书和电子书，以及高品质技术文章和视频课程。更多详情请访问异步社区官网 https://www.epubit.com。

“异步图书”是由异步社区编辑团队策划出版的精品 IT 专业图书的品牌，依托于人民邮电出版社近 30 年的计算机图书出版积累和专业编辑团队，相关图书在封面上印有异步图书的 LOGO。异步图书的出版领域包括软件开发、大数据、AI、测试、前端、网络技术等。

异步社区

微信服务号

目录

模块 1　NLTK 基础知识

模块 2 使用 Python 3 的 NLTK 3 进行文本处理