# 前言

这是一本介绍NLTK，以及NLTK如何与其它Python库搭配运用的书。NLTK是目前自然语言处理（NLP）社区中最为流行、使用最为广泛的库之一。NLTK的设计充分体现了简单的魅力。也就是说，对于大多数复杂的NLP任务，它都可以用寥寥几行代码来实现。

本书的前半部分将从介绍Python和NLP开始。在这本书中，您将会学到一些通用的预处理技术，譬如断词处理（tokenization）、词干处理（stemming）、停用词（stop word）去除，以及一些专属于NPL领域的预处理技术，譬如词性标注（part-of-speech tagging），以及大多数文本相关的NLP任务都会涉及到的命名实体识别（Named-entity recognition，简称NER）等技术。然后，我们还会逐步将焦点转到更为复杂的NLP任务上，譬如语法解析（parsing）、以及其它NLP应用。

而本书的后半部分将会更侧重于介绍如何构建一些NLP应用，譬如对于文本分类，我们可以用NLTK搭配scikit-learn库来进行。我们还会谈到一些其它的Python库，您应该了解一下这些与文本挖掘或自然语言处理任务相关的库。另外，我们也会带您看看如何从网页和社交媒体中采集数据，以及如何用NLTK对本章进行一次大规模的处理。

## 本书所涵盖的内容

**第1章，自然语言处理简介**：本章会涉及一些NPL中的基本概念，并对NLTK和Python做一些介绍。这一章的重点是让您快速了解一下NLTK，并介绍如何安装所需要的库，以便开始构建一个非常基本的单词云实例。

**第2章，文本的歧义及其清理**：本章将谈论在任何文本挖掘/NLP任务中所需的所有预处理步骤。在这一章中，我们将会具体讨论断词处理、词干处理、停用词去除等技术。并且，我们还会为您详细介绍一些别的文本清理技术，以及如何用NLTK来简化它们的实现。

**第3章，词性标注**：本章将重点对词性标注进行概述。在这一章中，我们将会为您介绍如何将NLTK运用到一些标注器中，并讨论NLTK中有哪些不同的NLP标注器可用。

**第4章，文本结构解析**：本章将会带您继续深入NLP，讨论不同的语法解析方法，并介绍如何用NLTK来实现这些方法。在此过程中，我们会讨论到语法解析在NLP语境中的，以及一些常见的信息提取技术（譬如实体提取）中的重要性。

**第5章，NLP应用**：本章将会谈及各种不同的NLP应用，我们将会带领您利用一些当前已掌握的知识来构建出一个简单的NLP应用实例。

**第6章，文本分类**：本章将会介绍一些机器学习领域中常见的分类方法。我们的讨论重点将主要集中在文本语料库，以及如何用NLTK和scikit来构建管道，从而实现一个文本分类器的问题上。当然，我们也会讨论到与文本聚类和主题模型相关的内容。

**第7章，Web爬虫**：本章将讨论NLP、数据科学和数据收集中其它方面的处理任务，以及如何从最大的文本数据源之一：Web中获取相关的数据。在这里，我们将学习如何用Python库，Scrapy来建立一只运作良好的Web爬虫（crawler）。

**第8章，NLTK与其它Python库的搭配运用**：本章将会谈及一些骨干的Python库，譬如NumPy和SciPy。另外，我们也会简单地介绍一下，用于数据处理的panda和用于可视化处理的matplotlib。

**第9章，Python中的社交媒体挖掘**：本章将致力于数据采集相关的内容。在这里，我们将会讨论社交媒体，以及与社交媒体相关的其它问题。当然，我们也会讨论具体应该如何收集、分析并可视化社交媒体中的数据。

**第10章，大规模文本挖掘**：在这一章我们将讨论如何扩展NLTK，并配合一些别的Python库，使其适应大数据时代规模化执行的需要。我们将会给出一个简短的演示，以说明NLTK和scikit是如何与Hadoop搭配使用的。

## 前期准备

在阅读这本书之前，我们建议您应该准备好下列软件：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 章节数 | 所需软件（及其版本） | 自由软件/专有软件 | 软件下载链接 | 硬件技术指标 | 所需操作系统 |
| 1-5 | Python/ Anaconda、NLTK | 自由软件 | https://www.python.org/  http://continuum.io/downloads  http://www.nltk.org/ | 通用Unix打印系统 | 不限 |
| 6 | scikit-learn、gensim | 自由软件 | http://scikit-learn.org/stable/  https://radimrehurek.com/gensim/ | 通用Unix打印系统 | 不限 |
| 7 | Scrapy | 自由软件 | http://scrapy.org/ | 通用Unix打印系统 | 不限 |
| 8 | NumPy、SciPy、pandas以及 matplotlib | 自由软件 | http://www.numpy.org/  http://www.scipy.org/  http://pandas.pydata.org/  http://matplotlib.org/ | 通用Unix打印系统 | 不限 |
| 9 | Twitter Python API与Facebook Python API | 自由软件 | https://dev.twitter.com/overview/api/twitter-libraries  https://developers.facebook.com | 通用Unix打印系统 | 不限 |

## 本书的适用读者

只要您是一个NLP和机器学习领域的爱好者，无论之前有没有文本处理方面的经验，这本书都是为您准备的。当然，这本书也非常适合那些想要快速学习一下NLTK的资深Python程序员。

## 一些约定

在本书中，我们会用不同的文本样式来突显不同类型信息之间的区别。下面，我们就通过几个例子来介绍一下这些样式，以及它们所代表的含义。

对于正文当中说涉及到的代码、数据库表名、文件夹名、文件名、文件扩展名、路径名、伪URL、用户输入以及Twitter句柄，我们将采取如下形式：“我们需要创建一个名为NewsSpider.py文件，并将其路径设置为/tutorial/spiders。”

接下来是Python代码块：

>>>import nltk

>>>import numpy

还有一般性的代码块：

add FILE vectorizer.pkl;

add FILE classifier.pkl;，

另外，到了第7章，Web信息采集这一章中，我们还将会用到Scrapy shell中的IPython 记法，其样式如下：

In [1] : sel.xpath('//title/text()')

Out[1]: [<Selector xpath='//title/text()' data=u' Google News'>]

最后是所有命令行输入或输出信息的样式：

# cp /usr/src/asterisk-addons/configs/cdr\_mysql.conf.sample

/etc/asterisk/cdr\_mysql.conf

|  |
| --- |
| 这种形式表达的是一些需要读者警惕的或需要重点关注的内容。 |
| 这种形式所提供的是一些提示或小技巧。 |

## 读者反馈

我们始终欢迎任何来自读者的反馈信息。它能让我们了解您对于这本书的看法——无论是喜欢还是不喜欢。这些反馈对于我们的选题开发来说都是至关重要的。

对于一般的反馈，您只需简单地给feedback@packtpub.com发一份电子邮件，并在邮件的标题中注明这本书的书名即可。

如果您对某一话题有专长，并且有兴趣写（或奉献）一本这方面的书，请参考我们的作者指南：www.packtpub.com/authors。

## 客户支持

您一直都是Packt图书的主人，我们将会尽一切努力来帮助您获取最好的图书资讯。

### 实例代码的下载

您可以在<http://www.packtpub.com>自己的账户页面中找到所有已购买Packt书籍，并下载相关的实例代码。如果您是在别处购买了我们的图书，也可以通过访问<http://www.packtpub.com/support>，注册有关文件，我们会通过电子邮件将其直接发给您。

### 勘误

尽管我们已经尽了最大的努力来确保书中内容的正确性，但错误始终是存在的。如果您在我们的书中发现了错误——无论是关于文字的还是代码的——只要您能告诉我们，我们都将不胜感激。因为这样可以大大减少其他读者在阅读方面所遇到的困难。因此，当您发现错误时，只需要访问http://www.packtpub.com/submit-errata，选择相应的书名，然后单击“errata submission form”链接并输入相关错误的详细信息即可。一旦您提供的信息获得了确认，相关的内容就被更新到我们的网站或对应书籍勘误章节下面现有的勘误表中。

如果想要查看先前已提交的勘误信息，您只需访问https://www.packtpub.com/books/ content/support，并在其搜索域中输入相关书籍的名称，所需信息就会出现在下面的勘误部分中。

### 版权

在互联网上，版权对于所有媒介一直是一个很大的问题。在Packet，我们向来对于版权许可非常重视。如果您在网络上发现我们出版过的作品的任何形式，都请马上将网址或网站名称告知我们，以便于我们采取补救措施。

请将您怀疑有侵权行为的文档链接发送到：copyright@packetpub.com。

您付出的帮助是对作者权利的保护，我们也由此才能继续为您带来有价值的内容。

### 如有疑问

如果您对本书有任何疑问，也可以通过questions@packtpub.com跟我们联系，我们会竭尽所能地帮您解决问题。