

# 语料库与python应用



[语料库与python应用\\_下载链接1](#)

著者:管新潮

出版者:上海交通大学出版社

出版时间:2018-8-1

装帧:平装

isbn:9787313197481

本书以如何在语料库的教与学及其应用、语料库科研中习得Python能力的逻辑关系为线索，描述了Python的价值、意义和作用，并将内容组合成可有效助力于Python能力习得的三个层次。第一层次是掌握与语料库相关的基础性代码；第二层次是活学活用这

些基础性代码；第三层次是以创新方式运用这些代码去解决与语料库相关的较为复杂的问题。Python是语料文本处理的利器，需要在一定的理念指导下方可充分理解其在特定领域内所呈现的特征，而本书的首要目标就是帮助读者去运用这一“语言+技术”理念，其次才是Python技术本身。本书的适用读者是那些设想从语料库中挖掘出更多信息的文科生、文科教师或相关研究人员。

作者介绍:

管新潮，职业译者，长期从事德英汉翻译实践，至今已累计翻译和审校德英汉字数达3000万（包括审校）；主要翻译领域涉及海洋工程与船舶制造（英语）、医学（英语）、法律（德语+英语）、机电（德语）等；建有各类相关语料库，如英汉医学平行语料、英汉海洋工程平行语料库、英汉法律平行语料库、德汉合同文本平行语料库、马克思《资本论》德汉平行语料库（百年）、德语法院判决书语料库等。曾经或正在为国际知名企业提供服务解决方案，如德国劳氏船级社、挪威船级社、艾斯维尔出版社、施普林格出版社、华为技术公司、毕马威咨询公司等。主要研究方向：语料库翻译学、翻译管理与技术、法律翻译、语料数据分析（Python）。

现任上海交通大学外国语学院MTI导师。主持国家级项目3个，发表论文15篇，出版专著2部、译著10部，拥有专利2项、软件著作权2项。

目录: 目录

第1章 绪论

1.1 语料库与Python

1.1.1 语料库的若干维度

1.1.2 语料库的技术实现

1.2 本书概要

上篇 语料文本的基础性代码

第2章 语料文本的读取及其运行结果的输出

2.1 概述

2.2 语料文本的读取

2.2.1 读取NLTK固有语料库

2.2.2 读取自制语料库

2.2.3 读取非独立存储的语料文本

2.2.4 读取docx格式的语料文本

2.2.5 读取xlsx格式的语料文本

2.3 语料文本运行结果的输出

2.3.1 操作界面直接输出结果

2.3.2 输出txt文件格式

2.3.3 输出xlsx文件格式

2.4 中文语料文本的读取和结果输出

2.4.1 自制语料库

2.4.2 非独立存储的语料文本

第3章 语料库应用的基础性代码

3.1 概述

3.2 停用词的使用

3.2.1 不同语种的停用词

3.2.2 自有停用词的设置

3.3 文本降噪代码

3.3.1 具体代码的功用

3.3.2 组合使用代码的功用

3.3.3 降噪与文本计数

3.4 语料文本的语言学处理代码

- 3.4.1 字母大小写转换
- 3.4.2 词形还原
- 3.4.3 文本分句或分词
- 3.4.4 词性标注
- 3.5 语料库词频排序
- 3.5.1 简单词频排序
- 3.5.2 降噪处理后词频排序
- 3.5.3 清除停用词后排序
- 3.6 语料库检索与统计
- 3.6.1 上下文关键词检索
- 3.6.2 类符形符比
- 3.6.3 N连词提取
- 3.6.4 指定词检索与统计
- 3.7 中文语料文本的处理方法
- 3.7.1 上下文关键词检索
- 3.7.2 中文停用词
- 第4章 数据可视化
- 4.1 概述
- 4.2 表格绘制
- 4.3 图形绘制
- 4.3.1 词频图形绘制
- 4.3.2 柱状图和点状图绘制
- 4.4 词云图绘制
- 4.4.1 英文文本词云图
- 4.4.2 中文文本词云图
- 第5章 代码运行错误分析
- 5.1 概述
- 5.2 错误分析案例
- 5.2.1 输入输出错误 (IOError)
- 5.2.2 对象属性错误 (AttributeError)
- 5.2.3 数据类型错误 (TypeError)
- 5.2.4 变量名称错误 (NameError)
- 5.2.5 索引错误 (IndexError)
- 5.2.6 缩进错误 (Indentati)
- 5.2.7 参数类型错误 (ValueError)
- 5.2.8 语法错误 (SyntaxError)
- 5.2.9 Unicode解码错误 (UnicodeDecodeError)
- 5.2.10 关键字错误 (KeyError)
- 中篇 基础性代码的组合使用
- 第6章 算法、代码与编程
- 6.1 篇章结构
- 6.2 算法和代码
- 6.2.1 算法
- 6.2.2 代码
- 6.3 选择不同代码的影响
- 6.3.1 分词处理方式对后续文本分析的影响
- 6.3.2 不同的降噪效果
- 6.3.3 链表、字符串、元组和字典对比
- 6.3.4 停用词的功用
- 6.4 Python与既有语料库工具的关系
- 第7章 基础性代码的语料库组合应用
- 7.1 以Excel文件格式输出术语 (类符)
- 7.1.1 简单输出术语
- 7.1.2 按词频输出术语

7.2 以Excel文件格式输出表格  
7.3 语篇词汇密度的计算  
7.4 语篇词汇复杂性的计算  
7.5 语篇词长分布的计算  
7.6 NLTK固有语料库  
7.6.1 总统就职演说语料库  
7.6.2 华尔街杂志语料库  
7.6.3 其他相关语料库介绍  
下篇 Python探索路径  
第8章 Python的语料库拓展应用  
8.1 概述  
8.2 单语语料导入Excel工作簿  
8.3 KWIC检索功能的拓展  
8.4 语篇词形还原  
8.5 术语提取效果的改进  
8.6 语篇段落对齐  
8.7 应用语言学文献计量研究的数据提取  
8.8 专业通用词的提取路径探索  
附录1 与本书相关的加载模块与函数命令对应表  
附录2 Python2 和Python3部分代码对比  
附录3 部分NLTK固有语料库  
附录4 汉英对照术语表  
索引  
• • • • • (收起)

[语料库与python应用\\_下载链接1](#)

## 标签

语料库

python

计算机科学

tobuy

TC

## 评论

填补了这类书中文版的空缺，对文科生和第一次接触代码的人还是很友好的，因为复制代码就能用…实际上就是挑了点儿NLTK的功能讲了讲，既然做语料库，英语能力过关，直接看NLTK的相关书籍或者文档更好。此外python2太老旧了，虽然列表给出了2&3的部分语法差异。第六章过于简略，不过重点不在此，可以理解，给三星是依旧存在各种各样的不足，百度google一定程度上完全可以替代此书，不过总体上推荐给不知道从何处入门的初学者，如果有一点点python基础就可以随意按需翻阅了。

-----  
[语料库与python应用\\_下载链接1](#)

## 书评

-----  
[语料库与python应用\\_下载链接1](#)