目录

[分类一 2](#_Toc14506621)

[1. Python网页爬虫工具集 2](#_Toc14506622)

[1.1 Scrapy 2](#_Toc14506623)

[1.2 Beautiful Soup 2](#_Toc14506624)

[1.3 Python-Goose 2](#_Toc14506625)

[2. Python文本处理工具集 2](#_Toc14506626)

[2.1 NLTK — Natural Language Toolkit 2](#_Toc14506627)

[2.2 Pattern 3](#_Toc14506628)

[2.3 TextBlob: Simplified Text Processing 3](#_Toc14506629)

[2.4 MBSP for Python 3](#_Toc14506630)

[2.5 Gensim: Topic modeling for humans 3](#_Toc14506631)

[2.6 langid.py: Stand-alone language identification system 3](#_Toc14506632)

[2.7 Jieba: 结巴中文分词 4](#_Toc14506633)

[3. Python科学计算工具包 4](#_Toc14506634)

[3.1 NumPy 4](#_Toc14506635)

[3.2 SciPy：Scientific Computing Tools for Python 4](#_Toc14506636)

[3.3 Matplotlib 4](#_Toc14506637)

[4. Python 机器学习 & 数据挖掘 工具包 5](#_Toc14506638)

[4.1 scikit-learn: Machine Learning in Python 5](#_Toc14506639)

[4.2 Pandas: Python Data Analysis Library 5](#_Toc14506640)

[4.3 mlpy – Machine Learning Python 5](#_Toc14506641)

[4.4 PyBrain 5](#_Toc14506642)

[4.5 Theano 5](#_Toc14506643)

[4.6  Pylearn2 5](#_Toc14506644)

[分类二 6](#_Toc14506645)

[网络 6](#_Toc14506646)

[网络爬虫框架 6](#_Toc14506647)

[HTML/XML解析器 6](#_Toc14506648)

[文本处理 6](#_Toc14506649)

[自然语言处理 6](#_Toc14506650)

[浏览器自动化与仿真 7](#_Toc14506651)

[多重处理 7](#_Toc14506652)

[异步 7](#_Toc14506653)

[异步网络编程库 7](#_Toc14506654)

[队列 7](#_Toc14506655)

[云计算 7](#_Toc14506656)

[网页内容提取 8](#_Toc14506657)

[WebSocket 8](#_Toc14506658)

[DNS解析 8](#_Toc14506659)

[计算机视觉 8](#_Toc14506660)

[代理服务器 8](#_Toc14506661)

分类一

1. Python网页爬虫工具集

一个真实的项目，一定是从获取数据开始的。无论文本处理，机器学习和数据挖掘，都需要数据，除了通过一些渠道购买或者下载的专业数据外，常常需要大家自己动手爬数据，这个时候，爬虫就显得格外重要了，幸好，Python提供了一批很不错的网页爬虫工具框架，既能爬取数据，也能获取和清洗数据，也就从这里开始了：

1.1 Scrapy

鼎鼎大名的Scrapy，相信不少同学都有耳闻，课程图谱中的很多课程都是依靠Scrapy抓去的，这方面的介绍文章有很多，推荐大牛pluskid早年的一篇文章：《Scrapy 轻松定制网络爬虫》，历久弥新。

官方主页：<http://scrapy.org/>  
Github代码页: <https://github.com/scrapy/scrapy>

1.2 Beautiful Soup

客观的说，Beautifu Soup不完全是一套爬虫工具，需要配合urllib使用，而是一套HTML / XML数据分析，清洗和获取工具。

官方主页：<http://www.crummy.com/software/BeautifulSoup/>

1.3 Python-Goose

Goose最早是用Java写得，后来用Scala重写，是一个Scala项目。Python-Goose用Python重写，依赖了Beautiful Soup。前段时间用过，感觉很不错，给定一个文章的URL, 获取文章的标题和内容很方便。

Github主页：<https://github.com/grangier/python-goose>

2. Python文本处理工具集

从网页上获取文本数据之后，依据任务的不同，就需要进行基本的文本处理了，譬如对于英文来说，需要基本的tokenize，对于中文，则需要常见的中文分词，进一步的话，无论英文中文，还可以词性标注，句法分析，关键词提取，文本分类，情感分析等等。这个方面，特别是面向英文领域，有很多优秀的工具包，我们一一道来。

2.1 NLTK — Natural Language Toolkit

搞自然语言处理的同学应该没有人不知道NLTK吧，这里也就不多说了。不过推荐两本书籍给刚刚接触NLTK或者需要详细了解NLTK的同学: 一个是官方的《Natural Language Processing with Python》，以介绍NLTK里的功能用法为主，同时附带一些Python知识，同时国内陈涛同学友情翻译了一个中文版，这里可以看到：推荐《用Python进行自然语言处理》中文翻译-NLTK配套书；另外一本是《Python Text Processing with NLTK 2.0 Cookbook》，这本书要深入一些，会涉及到NLTK的代码结构，同时会介绍如何定制自己的语料和模型等，相当不错。

官方主页：<http://www.nltk.org/>  
Github代码页：<https://github.com/nltk/nltk>

2.2 Pattern

Pattern由比利时安特卫普大学CLiPS实验室出品，客观的说，Pattern不仅仅是一套文本处理工具，它更是一套web数据挖掘工具，囊括了数据抓取模块（包括Google, Twitter, 维基百科的API，以及爬虫和HTML分析器），文本处理模块（词性标注，情感分析等），机器学习模块(VSM, 聚类，SVM）以及可视化模块等，可以说，Pattern的这一整套逻辑也是这篇文章的组织逻辑，不过这里我们暂且把Pattern放到文本处理部分。我个人主要使用的是它的英文处理模块Pattern.en, 有很多很不错的文本处理功能，包括基础的tokenize, 词性标注，句子切分，语法检查，拼写纠错，情感分析，句法分析等，相当不错。

官方主页：<http://www.clips.ua.ac.be/pattern>

2.3 TextBlob: Simplified Text Processing

TextBlob是一个很有意思的Python文本处理工具包，它其实是基于上面两个Python工具包NLKT和Pattern做了封装（TextBlob stands on the giant shoulders of NLTK and pattern, and plays nicely with both），同时提供了很多文本处理功能的接口，包括词性标注，名词短语提取，情感分析，文本分类，拼写检查等，甚至包括翻译和语言检测，不过这个是基于Google的API的，有调用次数限制。TextBlob相对比较年轻，有兴趣的同学可以关注。

官方主页：<http://textblob.readthedocs.org/en/dev/>  
Github代码页：<https://github.com/sloria/textblob>

2.4 MBSP for Python

MBSP与Pattern同源，同出自比利时安特卫普大学CLiPS实验室，提供了Word Tokenization, 句子切分，词性标注，Chunking, Lemmatization，句法分析等基本的文本处理功能，感兴趣的同学可以关注。

官方主页：<http://www.clips.ua.ac.be/pages/MBSP>

2.5 Gensim: Topic modeling for humans

Gensim是一个相当专业的主题模型Python工具包，无论是代码还是文档，我们曾经用《如何计算两个文档的相似度》介绍过Gensim的安装和使用过程，这里就不多说了。

官方主页：<http://radimrehurek.com/gensim/index.html>  
github代码页：<https://github.com/piskvorky/gensim>

2.6 langid.py: Stand-alone language identification system

语言检测是一个很有意思的话题，不过相对比较成熟，这方面的解决方案很多，也有很多不错的开源工具包，不过对于Python来说，我使用过langid这个工具包，也非常愿意推荐它。langid目前支持97种语言的检测，提供了很多易用的功能，包括可以启动一个建议的server，通过json调用其API，可定制训练自己的语言检测模型等，可以说是“麻雀虽小，五脏俱全”。

Github主页：<https://github.com/saffsd/langid.py>

2.7 Jieba: 结巴中文分词

“结巴”中文分词：做最好的Python中文分词组件 “Jieba” (Chinese for “to stutter”) Chinese text segmentation: built to be the best Python Chinese word segmentation module.  
好了，终于可以说一个国内的Python文本处理工具包了：结巴分词，其功能包括支持三种分词模式（精确模式、全模式、搜索引擎模式），支持繁体分词，支持自定义词典等，是目前一个非常不错的Python中文分词解决方案。

Github主页：<https://github.com/fxsjy/jieba>

3. Python科学计算工具包

说起科学计算，大家首先想起的是Matlab，集数值计算，可视化工具及交互于一身，不过可惜是一个商业产品。开源方面除了GNU Octave在尝试做一个类似Matlab的工具包外，Python的这几个工具包集合到一起也可以替代Matlab的相应功能：NumPy+SciPy+Matplotlib+iPython。同时，这几个工具包，特别是NumPy和SciPy，也是很多Python文本处理 & 机器学习 & 数据挖掘工具包的基础，非常重要。最后再推荐一个系列《用Python做科学计算》，将会涉及到NumPy, SciPy, Matplotlib，可以做参考。

3.1 NumPy

NumPy几乎是一个无法回避的科学计算工具包，最常用的也许是它的N维数组对象，其他还包括一些成熟的函数库，用于整合C/C++和Fortran代码的工具包，线性代数、傅里叶变换和随机数生成函数等。NumPy提供了两种基本的对象：ndarray（N-dimensional array object）和 ufunc（universal function object）。ndarray是存储单一数据类型的多维数组，而ufunc则是能够对数组进行处理的函数。

官方主页：<http://www.numpy.org/>

3.2 SciPy：Scientific Computing Tools for Python

“SciPy是一个开源的Python算法库和数学工具包，SciPy包含的模块有最优化、线性代数、积分、插值、特殊函数、快速傅里叶变换、信号处理和图像处理、常微分方程求解和其他科学与工程中常用的计算。其功能与软件MATLAB、Scilab和GNU Octave类似。 Numpy和Scipy常常结合着使用，Python大多数机器学习库都依赖于这两个模块。”—-引用自“Python机器学习库”

官方主页：<http://www.scipy.org/>

3.3 Matplotlib

matplotlib 是python最著名的绘图库，它提供了一整套和matlab相似的命令API，十分适合交互式地进行制图。而且也可以方便地将它作为绘图控件，嵌入GUI应用程序中。Matplotlib可以配合ipython shell使用，提供不亚于Matlab的绘图体验，总之用过了都说好。

官方主页：<http://matplotlib.org/>

4. Python 机器学习 & 数据挖掘 工具包

4.1 scikit-learn: Machine Learning in Python

首先推荐大名鼎鼎的scikit-learn，scikit-learn是一个基于NumPy, SciPy, Matplotlib的开源机器学习工具包，主要涵盖分类，回归和聚类算法，例如SVM， 逻辑回归，朴素贝叶斯，随机森林，k-means等算法，代码和文档都非常不错，在许多Python项目中都有应用。例如在我们熟悉的NLTK中，分类器方面就有专门针对scikit-learn的接口，可以调用scikit-learn的分类算法以及训练数据来训练分类器模型。

官方主页：<http://scikit-learn.org/>

4.2 Pandas: Python Data Analysis Library

Pandas也是基于NumPy和Matplotlib开发的，主要用于数据分析和数据可视化，它的数据结构DataFrame和R语言里的data.frame很像，特别是对于时间序列数据有自己的一套分析机制，非常不错。这里推荐一本书《Python for Data Analysis》，作者是Pandas的主力开发，依次介绍了iPython, NumPy, Pandas里的相关功能，数据可视化，数据清洗和加工，时间数据处理等，案例包括金融股票数据挖掘等，相当不错。

官方主页：<http://pandas.pydata.org/>

4.3 mlpy – Machine Learning Python

官方主页：<http://mlpy.sourceforge.net/>

4.4 PyBrain

“PyBrain(Python-Based Reinforcement Learning, Artificial Intelligence and Neural Network)是Python的一个机器学习模块，它的目标是为机器学习任务提供灵活、易应、强大的机器学习算法。（这名字很霸气）

PyBrain正如其名，包括神经网络、强化学习(及二者结合)、无监督学习、进化算法。因为目前的许多问题需要处理连续态和行为空间，必须使用函数逼近(如神经网络)以应对高维数据。PyBrain以神经网络为核心，所有的训练方法都以神经网络为一个实例。”

4.5 Theano

“Theano 是一个 Python 库，用来定义、优化和模拟数学表达式计算，用于高效的解决多维数组的计算问题。Theano的特点：紧密集成Numpy；高效的数据密集型GPU计算；高效的符号微分运算；高速和稳定的优化；动态生成c代码；广泛的单元测试和自我验证。自2007年以来，Theano已被广泛应用于科学运算。theano使得构建深度学习模型更加容易，可以快速实现多种模型。PS：Theano，一位希腊美女，Croton最有权势的Milo的女儿，后来成为了毕达哥拉斯的老婆。”

4.6  Pylearn2

“Pylearn2建立在theano上，部分依赖scikit-learn上，目前Pylearn2正处于开发中，将可以处理向量、图像、视频等数据，提供MLP、RBM、SDA等深度学习模型。”

官方主页：<http://deeplearning.net/software/pylearn2/>

分类二

网络

urllib -网络库(stdlib)。

requests -网络库。

grab – 网络库（基于pycurl）。

pycurl – 网络库（绑定libcurl）。

urllib3 – Python HTTP库，安全连接池、支持文件post、可用性高。

httplib2 – 网络库。

RoboBrowser – 一个简单的、极具Python风格的Python库，无需独立的浏览器即可浏览网页。

MechanicalSoup -一个与网站自动交互Python库。

mechanize -有状态、可编程的Web浏览库。

socket – 底层网络接口(stdlib)。

网络爬虫框架

grab – 网络爬虫框架（基于pycurl/multicur）。

scrapy – 网络爬虫框架。

pyspider – 一个强大的爬虫系统。

cola – 一个分布式爬虫框架。

HTML/XML解析器

lxml – C语言编写高效HTML/ XML处理库。支持XPath。

cssselect – 解析DOM树和CSS选择器。

pyquery – 解析DOM树和jQuery选择器。

BeautifulSoup – 低效HTML/ XML处理库，纯Python实现。

html5lib – 根据WHATWG规范生成HTML/ XML文档的DOM。该规范被用在现在所有的浏览器上。

feedparser – 解析RSS/ATOM feeds。

MarkupSafe – 为XML/HTML/XHTML提供了安全转义的字符串。

文本处理

用于解析和操作简单文本的库。

difflib – （Python标准库）帮助进行差异化比较。

Levenshtein – 快速计算Levenshtein距离和字符串相似度。

fuzzywuzzy – 模糊字符串匹配。

esmre – 正则表达式加速器。

ftfy – 自动整理Unicode文本，减少碎片化。

自然语言处理

处理人类语言问题的库。

NLTK -编写Python程序来处理人类语言数据的最好平台。

Pattern – Python的网络挖掘模块。他有自然语言处理工具，机器学习以及其它。

TextBlob – 为深入自然语言处理任务提供了一致的API。是基于NLTK以及Pattern的巨人之肩上发展的。

jieba – 中文分词工具。

SnowNLP – 中文文本处理库。

loso – 另一个中文分词库。

浏览器自动化与仿真

selenium – 自动化真正的浏览器（Chrome浏览器，火狐浏览器，Opera浏览器，IE浏览器）。

Ghost.py – 对PyQt的webkit的封装（需要PyQT）。

Spynner – 对PyQt的webkit的封装（需要PyQT）。

Splinter – 通用API浏览器模拟器（selenium web驱动，Django客户端，Zope）。

多重处理

threading – Python标准库的线程运行。对于I/O密集型任务很有效。对于CPU绑定的任务没用，因为python GIL。

multiprocessing – 标准的Python库运行多进程。

celery – 基于分布式消息传递的异步任务队列/作业队列。

concurrent-futures – concurrent-futures 模块为调用异步执行提供了一个高层次的接口。

异步

异步网络编程库

asyncio – （在Python 3.4 +版本以上的 Python标准库）异步I/O，时间循环，协同程序和任务。

Twisted – 基于事件驱动的网络引擎框架。

Tornado – 一个网络框架和异步网络库。

pulsar – Python事件驱动的并发框架。

diesel – Python的基于绿色事件的I/O框架。

gevent – 一个使用greenlet 的基于协程的Python网络库。

eventlet – 有WSGI支持的异步框架。

Tomorrow – 异步代码的奇妙的修饰语法。

队列

celery – 基于分布式消息传递的异步任务队列/作业队列。

huey – 小型多线程任务队列。

mrq – Mr. Queue – 使用redis & Gevent 的Python分布式工作任务队列。

RQ – 基于Redis的轻量级任务队列管理器。

simpleq – 一个简单的，可无限扩展，基于Amazon SQS的队列。

python-gearman – Gearman的Python API。

云计算

picloud – 云端执行Python代码。

dominoup.com – 云端执行R，Python和matlab代码

网页内容提取

提取网页内容的库。

HTML页面的文本和元数据

newspaper – 用Python进行新闻提取、文章提取和内容策展。

html2text – 将HTML转为Markdown格式文本。

python-goose – HTML内容/文章提取器。

lassie – 人性化的网页内容检索工具

WebSocket

用于WebSocket的库。

Crossbar – 开源的应用消息传递路由器（Python实现的用于Autobahn的WebSocket和WAMP）。

AutobahnPython – 提供了WebSocket协议和WAMP协议的Python实现并且开源。

WebSocket-for-Python – Python 2和3以及PyPy的WebSocket客户端和服务器库。

DNS解析

dnsyo – 在全球超过1500个的DNS服务器上检查你的DNS。

pycares – c-ares的接口。c-ares是进行DNS请求和异步名称决议的C语言库。

计算机视觉

OpenCV – 开源计算机视觉库。

SimpleCV – 用于照相机、图像处理、特征提取、格式转换的简介，可读性强的接口（基于OpenCV）。

mahotas – 快速计算机图像处理算法（完全使用 C++ 实现），完全基于 numpy 的数组作为它的数据类型。

代理服务器

shadowsocks – 一个快速隧道代理，可帮你穿透防火墙（支持TCP和UDP，TFO，多用户和平滑重启，目的IP黑名单）。

tproxy – tproxy是一个简单的TCP路由代理（第7层），基于Gevent，用Python进行配置