常用 Python 爬虫库汇总

Python 爬虫，全称 Python 网络爬虫，是一种按照一定的规则，自动地抓取万维网信息的程序或脚本，主要用于抓取证券交易数据、天气数据、网站用户数据和图片数据等，Python 为支持网络爬虫正常功能实现，内置了大量的库，主要有以下类型：

一、Python 爬虫网络库

Python 爬虫网络库主要包括：urllib 、requests 、grab 、pycurl 、urllib3 、httplib2 、RoboBrowser 、MechanicalSoup 、mechanize 、socket 、Unirest for Python 、hyper 、PySocks、treq 以及 aiohttp 等。

二、Python 网络爬虫框架

Python 网络爬虫框架主要包括：grab 、scrapy 、pyspider 、cola 、portia 、restkit 以及 demiurge 等。

三、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；

xmltodict ： 一个可以让你在处理 XML 时感觉像在处理 JSON 一样的 Python 模块；

xhtml2pdf ： 将 HTML/CSS 转换为 PDF；

untangle ： 轻松实现将 XML 文件转换为 Python 对象；

Bleach ： 清理 HTML 文本(需要 html5lib) ；

四、文本处理

difflib ： 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；

unidecode ： 将 Unicode 文本转为 ASCII；

uniout ： 打印可读字符，而不是被转义的字符串；

chardet ： 兼容 Python 的 2/3 的字符编码器；

xpinyin ： 一个将中国汉字转为拼音的库；

pangu.py ： 格式化文本中 CJK和字母数字的间距。

awesome-slugify ： 一个可以保留 unicode 的 Python slugify 库；

python-slugify ： 一个可以将 Unicode 转为 ASCII 的 Python slugify 库；

unicode-slugify ： 一个可以将生成 Unicode slugs 的工具；

pytils ： 处理俄语字符串的简单工具(包括 pytils.translit.slugify) ；

PLY： lex 和 yacc 解析工具的 Python 实现；

pyparsing ： 一个通用框架的生成语法分析器；

python-nameparser ：解析人的名字的组件；

phonenumbers：解析，格式化，存储和验证国际电话号码；

python-user-agents ：浏览器用户代理的解析器；

HTTP Agent Parser ：Python 的 HTTP代理分析器。

五、特定格式文件处理

tablib ：一个把数据导出为 XLS、CSV、JSON、YAML等格式的模块；

textract ：从各种文件中提取文本，比如 Word、PowerPoint 、PDF等；

messytables ：解析混乱的表格数据的工具；

rows：一个常用数据接口，支持的格式很多，目前支持 CSV、HTML、XLS、TXT；

python-docx ：读取，查询和修改的 Microsoft Word2007/2008 的 docx 文件；

xlwt / xlrd ：从 Excel 文件读取写入数据和格式信息；

XlsxWriter ：一个创建 Excel.xlsx 文件的 Python 模块；

xlwings ：一个 BSD许可的库，可以很容易地在 Excel 中调用 Python，反之亦然；

openpyxl ：一个用于读取和写入的 Excel2010 XLSX/ XLSM/ xltx/ XLTM 文件的库；

Marmir：提取 Python 数据结构并将其转换为电子表格；

PDFMiner：一个从 PDF文档中提取信息的工具；

PyPDF2 一个能够分割、合并和转换 PDF页面的库；

ReportLab ：允许快速创建丰富的 PDF文档；

pdftables ：直接从 PDF文件中提取表格；

Python-Markdown：一个用 Python 实现的 John Gruber 的 Markdown；

Mistune：速度最快，功能全面的 Markdown 纯 Python 解析器；

markdown2：一个完全用 Python 实现的快速的 Markdown；

PyYAML 一个 Python 的 YAML 解析器；

cssutils ：一个 Python 的 CSS 库；

feedparser ：通用的 feed 解析器；

sqlparse ：一个非验证的 SQL 语句分析器；

http-parser ：C 语言实现的 HTTP 请求/响应消息解析器；

opengraph：一个用来解析 Open Graph 协议标签的 Python 模块；

pefile ：一个多平台的用于解析和处理可移植执行体(即 PE)文件的模块；

psd-tools ：将 Adobe Photoshop PSD(即 PE)文件读取到 Python 数据结构。

六、自然语言处理

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；

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；

genius：基于条件随机域的中文分词；

langid.py ：独立的语言识别系统；

Korean：一个韩文形态库；

pymorphy2：俄语形态分析器(词性标注+词形变化引擎)；

PyPLN：用 Python 编写的分布式自然语言处理通道。这个项目的目标是创建一种简单的方法使用 NLTK 通过网络接口处理大语言库。

七、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)。

八、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。

九、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。

十、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；

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 。

十一、云计算

picloud ： 云端执行 Python 代码；

dominoup.com：云端执行 R，Python 和 matlab 代码。

十二、电子邮件

flanker ： 电子邮件地址和 Mime解析库；

Talon：Mailgun 库用于提取消息的报价和签名。

十三、网址和网络地址操作

furl ： 一个小的 Python 库，使得操纵 URL简单化；

purl ： 一个简单的不可改变的 URL以及一个干净的用于调试和操作的 API；

urllib.parse ：用于打破统一资源定位器(URL)的字符串在组件之间的隔断，为了结合组件到一个 URL字符串，并将“相对 URL”转化为一个绝对 URL，称之为“基本 URL”；

tldextract ：从 URL的注册域和子域中准确分离 TLD，使用公共后缀列表；

etaddr ：用于显示和操纵网络地址的 Python 库。

十四、网页内容提取

ewspaper：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；

micawber：一个从网址中提取丰富内容的小库；

sumy：一个自动汇总文本文件和 HTML网页的模块；

Haul：一个可扩展的图像爬虫；

python-readability ：arc90 readability 工具的快速 Python 接口；

scrapely ：从 HTML网页中提取结构化数据的库；

youtube-dl ：一个从 YouTube下载视频的小命令行程序；

you-get ：Python3 的 YouTube、优酷/ Niconico 视频下载器；

WikiTeam：下载和保存 wikis 的工具。

十五、WebSocket

Crossbar ：开源的应用消息传递路由器；

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。

十六、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。

十七、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；

mahotas：快速计算机图像处理算法，完全基于 numpy 的数组作为它的数据类型。

十八、代理服务器

shadowsocks：一个快速隧道代理，可帮你穿透防火墙；

tproxy ：tproxy 是一个简单的 TCP路由代理，基于 Gevent，用 Python 进行配置。

您的评论 *感谢支持，给文档评个星吧！

写点评论支持下文档贡献

240

发布评论

用户评价

暂无评论

评价文档：

分享到：

QQ空间 新浪微博 微信

扫二维码，快速分享到微信朋友圈

文档可以转存到百度网盘啦！

转为pdf格式

转为其他格式 >

VIP专享文档格式自由转换

/

下载券

立即下载

加入VIP

免券下载