**爬虫库**

通用：

1. urllib -网络库(stdlib)。
2. requests -网络库。
3. grab – 网络库（基于pycurl）。
4. pycurl – 网络库（绑定libcurl）。
5. urllib3 – Python HTTP库，安全连接池、支持文件post、可用性高。
6. httplib2 – 网络库。
7. 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。
8. MechanicalSoup -一个与网站自动交互Python库。
9. mechanize -有状态、可编程的Web浏览库。
10. socket – 底层网络接口(stdlib)。
11. Unirest for Python – Unirest是一套可用于多种语言的轻量级的HTTP库。
12. hyper – Python的HTTP/2客户端。
13. PySocks – SocksiPy更新并积极维护的版本，包括错误修复和一些其他的特征。作为socket模块的直接替换。

网络爬虫框架

1. 功能齐全的爬虫
2. grab – 网络爬虫框架（基于pycurl/multicur）。
3. scrapy – 网络爬虫框架（基于twisted），不支持Python3。
4. pyspider – 一个强大的爬虫系统。
5. cola – 一个分布式爬虫框架。
6. 其他
7. portia – 基于Scrapy的可视化爬虫。
8. restkit – Python的HTTP资源工具包。它可以让你轻松地访问HTTP资源，并围绕它建立的对象。
9. demiurge – 基于PyQuery的爬虫微框架。
10. HTML/XML解析器

**HTML/XML解析器**

1. 通用
2. lxml – C语言编写高效HTML/ XML处理库。支持XPath。
3. cssselect – 解析DOM树和CSS选择器。
4. pyquery – 解析DOM树和jQuery选择器。
5. BeautifulSoup – 低效HTML/ XML处理库，纯Python实现。
6. 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。
7. feedparser – 解析RSS/ATOM feeds。
8. MarkupSafe – 为XML/HTML/XHTML提供了安全转义的字符串。
9. 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。
10. xhtml2pdf – 将HTML/CSS转换为PDF。
11. untangle – 轻松实现将XML文件转换为Python对象。
12. 清理
13. Bleach – 清理HTML（需要html5lib）。
14. sanitize – 为混乱的数据世界带来清明。

**文本处理**

用于解析和操作简单文本的库。

1. 通用
2. difflib – （Python标准库）帮助进行差异化比较。
3. Levenshtein – 快速计算Levenshtein距离和字符串相似度。
4. fuzzywuzzy – 模糊字符串匹配。
5. esmre – 正则表达式加速器。
6. 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。

1. NLTK -编写Python程序来处理人类语言数据的最好平台。
2. 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。
3. 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。
4. jieba – 中文分词工具。
5. SnowNLP – 中文文本处理库。
6. loso – 另一个中文分词库。

**浏览器自动化与仿真**

1. 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）。
2. Ghost.py – 对PyQt的webkit的封装（需要PyQT）。
3. Spynner – 对PyQt的webkit的封装（需要PyQT）。
4. Splinter – 通用API浏览器模拟器（selenium web驱动，Django客户端，Zope）。

**多重处理**

1. 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。
2. multiprocessing – 标准的Python库运行多进程。
3. celery – 基于分布式消息传递的异步任务队列/作业队列。
4. 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

1. 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。
2. Twisted – 基于事件驱动的网络引擎框架。
3. Tornado – 一个网络框架和异步网络库。
4. pulsar – Python事件驱动的并发框架。
5. diesel – Python的基于绿色事件的I/O框架。
6. gevent – 一个使用greenlet 的基于协程的Python网络库。
7. eventlet – 有WSGI支持的异步框架。
8. Tomorrow – 异步代码的奇妙的修饰语法。

**队列**

1. celery – 基于分布式消息传递的异步任务队列/作业队列。
2. huey – 小型多线程任务队列。
3. mrq – Mr. Queue – 使用redis & Gevent 的Python分布式工作任务队列。
4. RQ – 基于Redis的轻量级任务队列管理器。
5. simpleq – 一个简单的，可无限扩展，基于Amazon SQS的队列。
6. python-gearman – Gearman的Python API。

**云计算**

1. picloud – 云端执行Python代码。
2. 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。

1. HTML页面的文本和元数据
2. newspaper – 用Python进行新闻提取、文章提取和内容策展。
3. html2text – 将HTML转为Markdown格式文本。
4. python-goose – HTML内容/文章提取器。
5. lassie – 人性化的网页内容检索工具

**WebSocket**

用于WebSocket的库。

1. Crossbar – 开源的应用消息传递路由器（Python实现的用于Autobahn的WebSocket和WAMP）。
2. AutobahnPython – 提供了WebSocket协议和WAMP协议的Python实现并且开源。
3. WebSocket-for-Python – Python 2和3以及PyPy的WebSocket客户端和服务器库。
4. DNS解析
5. dnsyo – 在全球超过1500个的DNS服务器上检查你的DNS。
6. 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**

1. OpenCV – 开源计算机视觉库。
2. 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）。
3. 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**

1. 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）。
2. 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。