# 常用库

## 请求库：

### 1.requests：

发送请求的一个库

### 2.Selenium：

Selenium 是一个自动化测试工具，利用它我们可以驱动浏览器执行特定的动作，如点击、下拉等操作。对于一些 JavaScript 渲染的页面来说，这种抓取方式非常有效。

### 3.ChromeDriver

chrome的驱动软件，主要是使用自动化测试工具时，驱动Chrome浏览器

### 4.GeckoDriver：

驱动火狐浏览器的驱动软件

### 5.PhantomJS

PhantomJS 是一个无界面的、可脚本编程的 WebKit 浏览器引擎，它原生支持多种 Web 标准：DOM 操作、CSS 选择器、JSON、Canvas 以及 SVG。

Selenium 支持 PhantomJS，这样在运行的时候就不会再弹出一个浏览器了。而且 PhantomJS 的运行效率也很高，还支持各种参数配置，使用非常方便。

### 6.aiohttp：

提供异步web服务的库。requests 库是一个阻塞式 HTTP 请求库，当我们发出一个请求后，程序会一直等待服务器响应，直到得到响应后，程序才会进行下一步处理。其实，这个过程比较耗费时间。

## 解析库：

### 1. lxml

lxml 是 Python 的一个解析库，支持 HTML 和 XML 的解析，支持 XPath 解析方式，而且解析效率非常高。

### 2. Beautiful Soup

Beautiful Soup 是 Python 的一个 HTML 或 XML 的解析库，我们可以用它来方便地从网页中提取数据。它拥有强大的 API 和多样的解析方式

### 3. pyquery

pyquery 同样是一个强大的网页解析工具，它提供了和 jQuery 类似的语法来解析 HTML 文档，支持 CSS 选择器，使用非常方便。

### 4. tesserocr

在爬虫过程中，难免会遇到各种各样的验证码，而大多数验证码还是图形验证码，这时候我们可以直接用 OCR 来识别。

tesserocr 是 Python 的一个 OCR 识别库，但其实是对 tesseract 做的一层 Python API 封装，所以它的核心是 tesseract。因此，在安装 tesserocr 之前，我们需要先安装 tesseract。

在 Windows 下，首先需要下载 tesseract，它为 tesserocr 提供了支持。

 tesserocr GitHub：<https://github.com/sirfz/tesserocr>

 tesserocr PyPI：<https://pypi.python.org/pypi/tesserocr>

 tesseract 下载地址：<http://digi.bib.uni-mannheim.de/tesseract>

 tesseract GitHub：<https://github.com/tesseract-ocr/tesseract>

 tesseract 语言包：<https://github.com/tesseract-ocr/tessdata>

 tesseract 文档：<https://github.com/tesseract-ocr/tesseract/wiki/Documentation>

勾选 Additional language data (download) 选项来安装 OCR 识别支持的语言包，这样 OCR 便可以识别多国语言。

接下来，再安装 tesserocr 即可，此时直接使用 pip 安装：

pip3 install tesserocr pillow

# 技巧

## 1.抓取的数据中有 \u003C 等字符

是因为是unicode编码。进行重新编码再解码即可。

s1='\u003Cdiv\u003E\u003Cp\u003E众所周知'  
s = s1.encode('utf8').decode('utf8')  
print(s)

输出：<div><p>众所周知