链滴

Python3 网络爬虫实战：10、爬虫框架的安装：PySpider、Scrapy

作者：zhaolixiang

原文链接：https://ld246.com/article/1566383608762

来源网站：链滴

许可协议：署名-相同方式共享 4.0 国际 (CC BY-SA 4.0)

我们直接用 Requests、Selenium 等库写爬虫，如果爬取量不是太大，速度要求不高，是完全以满足需求的。但是写多了会发现其内部许多代码和组件是可以复用的，如果我们把这些组件抽离出，将各个功能模块化，就慢慢会形成一个框架雏形，久而久之，爬虫框架就诞生了。
利用框架我们可以不用再去关心某些功能的具体实现，只需要去关心爬取逻辑即可。有了它们，以大大简化代码量，而且架构也会变得清晰，爬取效率也会高许多。所以如果对爬虫有一定基础，上框架是一种好的选择。
本书主要介绍的爬虫框架有 PySpider 和 Scrapy，本节我们来介绍一下 PySpider、Scrapy 以它们的一些扩展库的安装方式。
<h2 id="PySpider的安装">PySpider 的安装</h2>
PySpider 是国人 binux 编写的强大的网络爬虫框架，它带有强大的 WebUI、脚本编辑器、任监控器、项目管理器以及结果处理器，同时它支持多种数据库后端、多种消息队列，另外它还支持 Jav Script 渲染页面的爬取，使用起来非常方便，本节介绍一下它的安装过程。
<h2 id="1--相关链接">1. 相关链接</h2>

官方文档：http://docs.pyspider.org/
PyPi：<a href="https://ld246.com/forward?goto=https%3A%2F%2Fpypi.python.org%2Fpi%2Fpyspider" target="_blank" rel="nofollow ugc">https://pypi.python.org/pypi/pyspider</>
GitHub：https://github.com/binux/pyspider
官方教程：http://docs.pyspider.org/en latest/tutorial
在线实例：http://demo.pyspider.org

<h2 id="2--准备工作">2. 准备工作</h2>
PySpider 是支持 JavaScript 渲染的，而这个过程是依赖于 PhantomJS 的，所以还需要安装 Ph ntomJS，所以在安装之前请安装好 PhantomJS，安装方式在前文有介绍。
<h2 id="3--Pip安装">3. Pip 安装</h2>
推荐使用 Pip 安装，命令如下：
<pre><code class="highlight-chroma">pip3 install pyspider
</code></pre>
命令执行完毕即可完成安装。
<h2 id="4--常见错误">4. 常见错误</h2>
Windows 下可能会出现这样的错误提示：Command "python setup.py egg_info" failed with rror code 1 in /tmp/pip-build-vXo1W3/pycurl
这个是 PyCurl 安装错误，一般会出现在 Windows 下，需要安装 PyCurl 库，下载链接为：http://www.lfd.uci.edu/~gohlke/pyth nlibs/#pycurl ，找到对应 Python 版本然后下载相应的 Wheel 文件。
如 Windows 64 位，Python3.6 则下载 pycurl‑7.43.0‑cp36‑cp36m‑win_amd64.whl，随后用 Pip 装即可，命令如下：
<pre><code class="highlight-chroma">pip3 install pycurl‑7.43.0‑cp36‑cp36m‑win_amd64.whl
</code></pre>
Linux 下如果遇到 PyCurl 的错误可以参考本文：https://iml nghao.com/19.html
Mac 遇到这种情况,执行下面操作：
<pre><code class="highlight-chroma"><span class="highlight

cl">brew install openssl

openssl version
查看版本
find /usr/local -name ssl.h

可以看到形如：
usr/local/Cellar/openssl/1.0.2s/include/openssl/ssl.h

添加环境变量
export PYCURL_SSL_LIBRARY=openssl
export LDFLAGS=-L/usr/local/Cellar/openssl/1.0.2s/lib
export CPPFLAGS -I/usr/local/Cellar/openssl/1.0.2s/include

 pip3 install pyspider
</code></pre>
<h2 id="5--验证安装">5. 验证安装</h2>
安装完成之后，可以直接在命令行下启动 PySpider：
<pre><code class="highlight-chroma">pyspider all
</code></pre>
控制台会有类似如下输出，如图 1-75 所示：

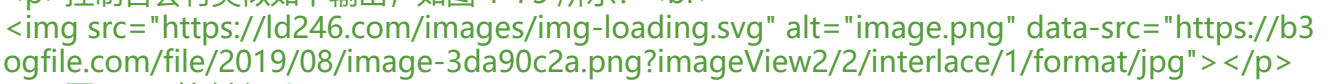



图 1-75 控制台
这时 PySpider 的 Web 服务就会在本地 5000 端口运行，直接在浏览器打开：http://localhost:5000/ 即可进入 PySpider 的 WebUI 管理页面，如图 1-76 所示：

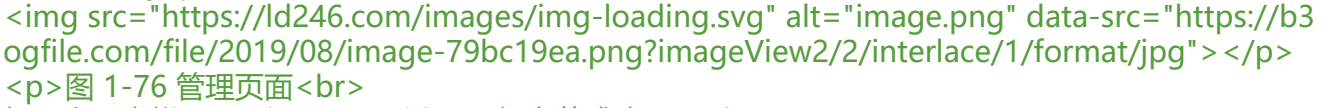



图 1-76 管理页面
如果出现类似页面那证明 PySpider 已经安装成功了。
在后文会介绍 PySpider 的详细用法。
这里有一个深坑，PySpider 在 Python3.7 上运行时会报错
<pre><code class="highlight-chroma">File "/usr/local/lib/python3.7/site-packages/pyspider/run.py", line 231
    async=True, getobject=False, no_input=False):
          ^
SyntaxError: invalid syntax
</code></pre>
原因是 python3.7 中 async 已经变成了关键字。因此出现这个错误。
修改方式是手动替换一下
<pre><code class="highlight-chroma">下面位置的async改为mark_async

/usr/local/lib/python3.7/site-packages/pyspider/run.py 的231行、245行（两个）、365行

/usr/local/lib/python3.7/site-packages/pyspider/webui/app.py 的95行

/usr/local/lib/python3.7/site-packages/pyspider/fetcher/tornado_fetcher.py 的81行、89行（两个）、95行、117行

</code></pre>
<h2 id="Scrapy的安装">Scrapy 的安装</h2>
Scrapy 是一个十分强大的爬虫框架，依赖的库比较多，至少需要依赖库有 Twisted 14.0，lxml 3.4，pyOpenSSL 0.14。而在不同平台环境又各不相同，所以在安装之前最好确保把一些基本库安装好。本节介绍一下 Scrapy 在不同平台的安装方法。
<h2 id="1--相关链接-">1. 相关链接</h2>

官方网站：https://scrapy.org
官方文档：https://docs.scrapy.org
PyPi：https://pypi.python.org/pypi/Scrapy
GitHub：https://github.com/scrapy/scrapy
中文文档：http://scrapy-chs.readthedocs.io

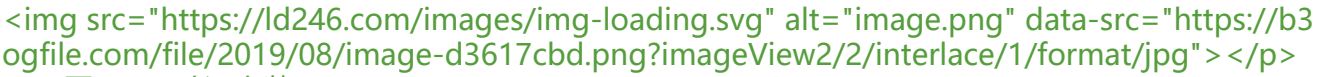
<h2 id="3--Mac下的安装">3. Mac 下的安装</h2>
在 Mac 上构建 Scrapy 的依赖库需要 C 编译器以及开发头文件，它一般由 Xcode 提供，运行如命令安装即可：
<pre><code class="highlight-chroma">xcode-select --install
</code></pre>
随后利用 Pip 安装 Scrapy 即可，运行如下命令：
<pre><code class="highlight-chroma">pip3 install Scrapy
</code></pre>
运行完毕之后即可完成 Scrapy 的安装。
<h2 id="4--验证安装">4. 验证安装</h2>
安装之后，在命令行下输入 scrapy，如果出现类似下方的结果，就证明 Scrapy 安装成功，如图-80 所示：


图 1-80 验证安装
<h2 id="5--常见错误">5. 常见错误</h2>
<pre><code class="highlight-chroma">pkg_resources.VersionConflict: (six 1.5.2 (/usr/lib/python3/dist-packages), Requirement.parse('six>=1.6.0'))
</code></pre>
six 包版本过低，six 包是一个提供兼容 Python2 和 Python3 的库，升级 six 包即可：
<pre><code class="highlight-chroma">sudo pip3 install -U six
</code></pre>

<hr>
<pre><code class="highlight-chroma"><span class="highlight
cl">c/_cffi_backend.c:15:17: fatal error: ffi.h: No such file or directory
</code></pre>
这是在 Linux 下常出现的错误，缺少 Libffi 这个库。什么是 libffi？“FFI” 的全名是 Foreign Fu
ction Interface，通常指的是允许以一种语言编写的代码调用另一种语言的代码。而 Libffi 库只提供
最底层的、与架构相关的、完整的”FFI”。
安装相应的库即可。
Ubuntu、Debian：
<pre><code class="highlight-chroma"><span class="highlight
cl">sudo apt-get install build-essential libssl-dev libffi-dev python3-dev
</code></pre>
CentOS、RedHat:
<pre><code class="highlight-chroma"><span class="highlight
cl">sudo yum install gcc libffi-devel python-devel openssl-devel
</code></pre>
<hr>
<pre><code class="highlight-chroma"><span class="highlight
cl">Command "python setup.py egg_info" failed with error code 1 in /tmp/pip-build/cryptog
aphy/
</code></pre>
这是缺少加密的相关组件，利用 Pip 安装即可。
<pre><code class="highlight-chroma"><span class="highlight
cl">pip3 install cryptography
</code></pre>
<hr>
<pre><code class="highlight-chroma"><span class="highlight
cl">ImportError: No module named 'packaging'
</code></pre>
缺少 packaging 这个包，它提供了 Python 包的核心功能，利用 Pip 安装即可。
<hr>
<pre><code class="highlight-chroma"><span class="highlight
cl">ImportError: No module named '_cffi_backend'
</code></pre>
缺少 cffi 包，使用 Pip 安装即可：
<pre><code class="highlight-chroma"><span class="highlight
cl">pip3 install cffi
</code></pre>
<hr>
<pre><code class="highlight-chroma"><span class="highlight
cl">ImportError: No module named 'pyparsing'
</code></pre>
缺少 pyparsing 包，使用 Pip 安装即可：
<pre><code class="highlight-chroma"><span class="highlight
cl">pip3 install pyparsing appdirs
</code></pre>