Python爬虫

1. 对于https的链接 爬虫需要

import ssl

ssl.\_create\_default\_https\_context = ssl.\_create\_unverified\_context

1. BeautifulSoup 解析网页 高级匹配

外部包 需要安装

pip/pip3 Install beautifulsoup4

<https://www.crummy.com/software/BeautifulSoup/bs4/doc.zh/> 中文手册

安装解析器 lxml

在pip安装后仍无法成功，在mac解决的办法为：

sudo easy\_install lxml

sudo easy\_install beautifulsoup4

1. BeautifulSoup 解析网页：css

{“class”:”month”}

1. BeautifulSoup 解析网页：Regex 正则表达式
2. 联系：爬百度百科

a = {"a","b","c"}  #这是元组，不支持索引

a = ["a","b","c"]  #这是列表，支持索引，只要把花括号改成方括号就行

print(soup.find('h1').get\_text(), ' url: ', his[-1])

1. request 外部包 post ／get 请求

需要安装包

webbrowser.open(r.url) 运行之后自动打开链接

使用cookies时，每一次请求都要使用cookie，session可以在对话中直接使用

1. 下载文件 图片
2. 多进程分布式爬虫

爬取地址：<https://morvanzhou.github.io/>

多进程，多线程的课程学习

多进程爬虫，用了所有的进程；异步爬取，单线程的进行。

1. 加速爬虫：异步加载 asyncio

单线程做这个东西

安装aiohttp

1. 高级爬虫：让selenium控制浏览器帮助爬虫

打开默认浏览器，在浏览器中爬取

安装selenium pip3 之后再浏览器中安装driver 配置路径

火狐浏览器插件： Katalon Recorder

<http://selenium-python.readthedocs.io/> selenium官网

1. Scrapy爬虫库

<https://scrapy.org/> 官网

安装pip3 异步爬去的一个过程

详见：scripy\_demo.ipnby