

目录

前言	1.1
PySpider简介	1.2
PySpider安装与基本用法	1.3
PySpider安装	1.3.1
PySpider基本用法	1.3.2
PySpider的高级用法	1.4
self.crawl	1.4.1
config.json	1.4.2
data目录	1.4.3
phantomjs	1.4.4
PySpider经验与心得	1.5
PySpider的心得	1.5.1
PySpider常见的坑	1.5.2
PySpider案例	1.6
附录	1.7
参考资料	1.7.1

Python爬虫框架：PySpider

- 最新版本： v0.2
- 更新时间： 20180920

简介

PySpider是一个Python的主流爬虫框架。下面就来总结PySpider的安装，基本的使用操作，常见的配置和参数的含义，常见的坑，等等。

源码+浏览+下载

本书的各种源码、在线浏览地址、多种格式文件下载如下：

Gitbook源码

- [crifan/python_spider_pyspider](#)

如何使用此Gitbook源码去生成发布为电子书

详见：[crifan/gitbook_template: demo how to use crifan gitbook template and demo](#)

在线浏览

- Python爬虫框架： PySpider [book.crifan.com](#)
- Python爬虫框架： PySpider [crifan.github.io](#)

离线下载阅读

- Python爬虫框架： PySpider PDF
- Python爬虫框架： PySpider ePub
- Python爬虫框架： PySpider Mobi

crifan.com, 使用[知识署名-相同方式共享4.0协议](#)发布 all right reserved, powered by Gitbook该文件修订时间： 2018-09-20 18:47:51

PySpider简介

PySpider 对比 Scrapy

对于两个流行的Python的爬虫框架，PySpider和Scrapy，常常会被拿来对比。

对此，之前简单总结如下：

- PySpider：简单易上手，带图形界面（基于浏览器页面）
 - 一图胜千言：在WebUI中调试爬虫代码
 -
 -
- Scrapy：可以高级定制化实现更加复杂的控制
 - 一图胜千言：Scrapy一般是在命令行界面中调试页面返回数据：
 -
 -

详见： [【整理】pyspider vs scrapy](#)

crifan.com, 使用[知识署名-相同方式共享4.0协议](#)发布 all right reserved, powered by Gitbook该文件修订时间： 2018-09-20 18:20:03

PySpider安装与基本用法

crifan.com, 使用[知识署名-相同方式共享4.0协议](#)发布 all right reserved, powered by Gitbook该文件修订时间: 2018-09-20 14:11:38

PySpider安装

下面介绍，如何安装 PySpider。

Mac中安装PySpider

此处以Mac中为例，去介绍如何安装PySpider。

如果只是简单的直接安装的话，则可以去：

```
pip install pyspider
```

即可。

不过，Python的开发中，一般避免不同开发环境的影响，推荐使用虚拟环境工具，比如 pipenv。

此处使用 pipenv 虚拟环境去安装 PySpider：

安装（虚拟环境同时安装）PySpider： pipenv install pyspider

即可。

安装期间出现SSL或pycurl错误

如果安装期间出错：

```
__main__.ConfigurationError: Curl is configured to use SSL, but we have not been able to determine which
SSL backend it is using
```

或运行期间出错：

```
ImportError pycurl libcurl link-time ssl backend (openssl) is different from compile-time ssl backend (none/other)
```

则解决办法是：

```
pip uninstall pycurl
export PYCURL_SSL_LIBRARY=openssl
export LDFLAGS=-L/usr/local/opt/openssl/lib;export CPPFLAGS=-I/usr/local/opt/openssl/include;pip install pycurl --compile
--no-cache-dir
```

详见：

- 【记录】 Mac中安装和运行pyspider
- 【已解决】 pipenv虚拟环境中用pip安装pyspider出错： `__main__.ConfigurationError: Curl is configured to use SSL, but we have not been able to determine which SSL backend it is using`
- 【已解决】 pyspider运行出错： `ImportError pycurl libcurl link-time ssl backend (openssl) is different from compile-time ssl backend (none/other)`

上述命令要在对应虚拟环境下运行

如果本身你前面是用虚拟环境，比如 pipenv 中去安装的pyspider，则上述命令要先去进入虚拟环境： pipenv shell 之后再去运行，否则是无效的。

09-20 17:01:09

PySpider基本用法

使用PySpider的基本步骤

下面来介绍一下PySpider的使用的步骤和操作：

运行PySpider

在某个目录下的终端命令行中输入 `pyspider` 即可启动运行。

注：

- 如果是用虚拟环境安装的PySpider，记得先进去虚拟环境后再运行PySpider
 - 比如用的 `pipenv`，则是先 `pipenv shell`，再 `pyspider`
- `pyspider` 等价于 `pyspider all`

进入 WebUI

然后去用浏览器打开：

<http://0.0.0.0:5000/>

即可进入爬虫的管理界面了，此界面一般称为：`WebUI`

新建爬虫项目

点击 `Create`，去新建一个爬虫项目

输入：

- 爬虫名称：
- 入口地址：自动生成的代码中，会作为起始要抓取的url
 - 也可以不填
 - 后续也可以去代码中修改的

然后再点击新建的爬虫项目，进入调试页面

新建出来的项目，默认状态是 `TODO`

点击新建出来的项目名，直接进入调试界面

然后右边是编写代码的区域

左边是调试的区域，用于执行代码，显示输出信息等用途

调试爬虫代码

编写代码，调试输出信息，保存代码

调试代码期间，对于想要返回上一级：

先说之前不熟悉的时候的操作：

之前调试运行时，不知道还有回到上一级，在想要返回上一级时，都直接是点击左上角的项目名字

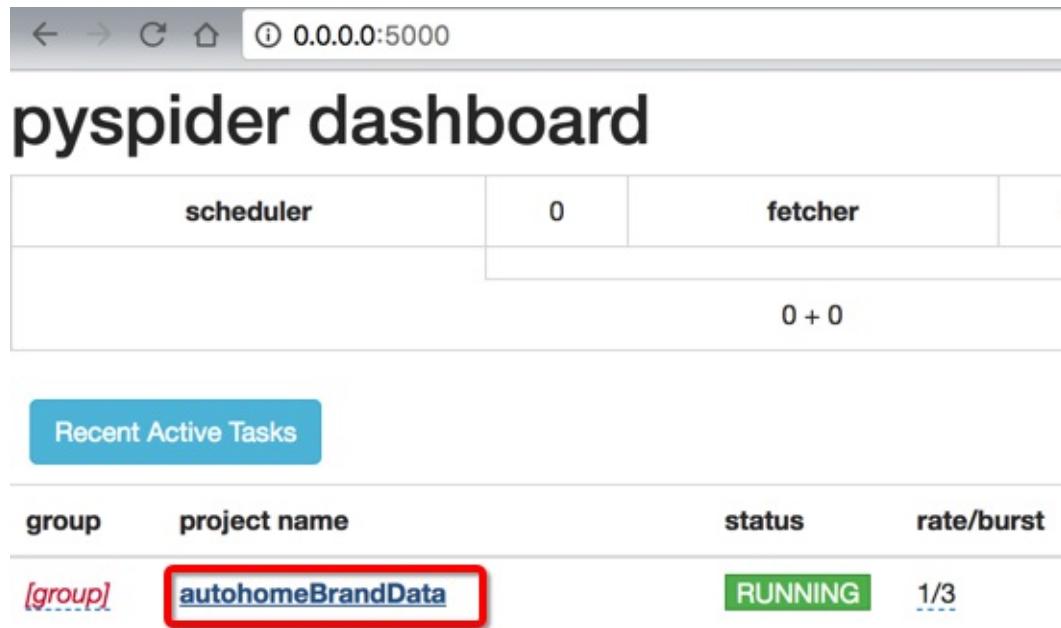


```

{
  "process": {
    "callback": "on_start"
  },
  "project": "autohomeBrandData",
  "taskid": "data:,on_start",
  "url": "data:,on_start"
}
gradCarHtmlPage > https://www.autohome.com.cn/grade/carhtml/a.html ...
gradCarHtmlPage > https://www.autohome.com.cn/grade/carhtml/b.html ...
gradCarHtmlPage > https://www.autohome.com.cn/grade/carhtml/c.html ...

```

返回项目列表：



group	project name	status	rate/burst
[group]	autohomeBrandData	RUNNING	1/3

然后重新进去，重新点击Run，直到跑到对应的层级，去继续调试。

再说后来知道了PySpider内置支持这种逻辑操作：

PySpider对在调试期间所需要在上一个连接和下一个连接之间切换的操作，支持的很好：

点击 < | > 的 < 或 >，则可以 返回上一级 或 进入下一级

实际效果演示：

The screenshot shows the PySpider IDE interface. On the left, a list of URLs for 'gradCarHtmlPage' is displayed, each with a green 'run' button. A red box highlights the '< | >' navigation buttons at the top of this list. On the right, the Python code for the Handler class is shown, with a red arrow pointing to the line where 'gradCarHtmlPage' is called.

```

1 #!/usr/bin/env python
2 # -*- encoding: utf-8 -*-
3 # Created on 2018-04-27 21:53:02
4 # Project: autohomeBrandData
5
6 from pyspider.libs.base_handler import *
7 import string
8
9 class Handler(BaseHandler):
10     crawl_config = {
11         }
12
13     # @every(minutes=24 * 60)
14     def on_start(self):
15         for eachLetter in list(string.ascii_lowercase):
16
17             self.crawl("https://www.autohome.com.cn/grade/carhtml/%s.html" % eachLetter, callback=self.gradCarHtmlPage)
18
19             gradCarHtmlPage(self, response):
20                 picSeriesItemList = response.doc('.rank-list-ul li div a[href^="/pic/series"]').items()
21                 # print("len(picSeriesItemList)=%s" % (len(picSeriesItemList)))
22                 for each in picSeriesItemList:
23                     self.crawl(each.attr.href, callback=self.detail_page)
24
25             # @config(age=10 * 24 * 60 * 60)
26             def picSeriesPage(self, response):
27                 # for each in response.doc('.rank-list-ul li div a[href^="//car.autohome.com.cn/pic/series/]').items():
28                 # for each in response.doc('.rank-list-ul li div a[href^="/pic/series"]').items():
29                     self.crawl(each.attr.href, callback=self.detail_page)
30
31             # config(priority=2)
32             def detail_page(self, response):
33                 # <a href="/pic/series-t/66.html">查看停产车型</a>
34                 # <a class="ckmore" href="/pic/series/588.html">查看在售车型
35                 # <span class="fn-right">&ampnbsp</span>
36                 fnRightPicSeries = response.doc('.search-pic-bar .fn-right a[href^="/pic/series"]')
37                 print("fnRightPicSeries", fnRightPicSeries)
38                 if fnRightPicSeries:
39
40
    
```

想要返回上一级的爬取函数的话，点击 左箭头

The screenshot shows the PySpider IDE interface after a run. The results of the previous run are displayed on the left, showing a list of items for a car model. A red box highlights the '< | >' navigation buttons at the top of the results list. On the right, the Python code for the Handler class is shown, with a red arrow pointing to the line where 'gradCarHtmlPage' is called.

```

1 #!/usr/bin/env python
2 # -*- encoding: utf-8 -*-
3 # Created on 2018-04-27 21:53:02
4 # Project: autohomeBrandData
5
6 from pyspider.libs.base_handler import *
7 import string
8
9 class Handler(BaseHandler):
10     crawl_config = {
11         }
12
13     # @every(minutes=24 * 60)
14     def on_start(self):
15         for eachLetter in list(string.ascii_lowercase):
16
17             self.crawl("https://www.autohome.com.cn/grade/carhtml/%s.html" % eachLetter, callback=self.gradCarHtmlPage)
18
19             gradCarHtmlPage(self, response):
20                 picSeriesItemList = response.doc('.rank-list-ul li div a[href^="/pic/series"]').items()
21                 # print("len(picSeriesItemList)=%s" % (len(picSeriesItemList)))
22                 for each in picSeriesItemList:
23                     self.crawl(each.attr.href, callback=self.detail_page)
24
25             # @config(age=10 * 24 * 60 * 60)
26             def picSeriesPage(self, response):
27                 # for each in response.doc('.rank-list-ul li div a[href^="//car.autohome.com.cn/pic/series/]').items():
28                 # for each in response.doc('.rank-list-ul li div a[href^="/pic/series"]').items():
29                     self.crawl(each.attr.href, callback=self.detail_page)
30
31             # config(priority=2)
32             def detail_page(self, response):
33                 # <a href="/pic/series-t/66.html">查看停产车型</a>
34                 # <a class="ckmore" href="/pic/series/588.html">查看在售车型
35                 # <span class="fn-right">&ampnbsp</span>
36                 fnRightPicSeries = response.doc('.search-pic-bar .fn-right a[href^="/pic/series"]')
37                 print("fnRightPicSeries", fnRightPicSeries)
38                 if fnRightPicSeries:
39
40
    
```

然后再点击Run：

```


← → C ⌂ ⓘ 0.0.0.0:5000/debug/autohomeBrandData



pyspider > autohomeBrandData



```

{
 "fetch": {},
 "process": {
 "callback": "gradCarHtmlPage"
 },
 "project": "autohomeBrandData",
 "schedule": {},
 "taskid": "40aa2a3ad7c7a8973043d60a32df38f0",
 "url": "https://www.autohome.com.cn/grade/carhtml/c.html"
}

fnRightPicSeries= <a href="https://car.autohome.com.cn/pic/series-t/2567.h
fullPicSeriesUrl= https://car.autohome.com.cn/pic/series-t/2567.html
eachADict= {'text': '汽车图片', 'href': 'https://car.autohome.com.cn/pic/'}
eachADict= {'text': '长安', 'href': 'https://car.autohome.com.cn/pic/brand'}
eachADict= {'text': '长安汽车', 'href': 'https://car.autohome.com.cn/pic/br'}
eachADict= {'text': '悦翔V3', 'href': 'https://car.autohome.com.cn/pic/seri'}
aDictList= [{text: '汽车图片', href: 'https://car.autohome.com.cn/pic/'}
mainBrandDict= {'text': '长安', 'href': 'https://car.autohome.com.cn/pic/br'}
dtTextList= ['2015款']
groupCount= 1
ddUlEltList= [[]]
-----[0] 2015款
curModelUrl= https://car.autohome.com.cn/pic/series-s25858/2567.html#pvare
curmodelName= 1.4L 手动美满型 国V

```



[



```

 [
 "autohomeBrandData",
 {
 "品牌": "长安",
 "子品牌": "长安汽车",
 "车型": "2015款 1.4L 手动美满型 国V",
 "车系": "悦翔V3"
 },
 "https://car.autohome.com.cn/pic/series-s25858/2567.html#pvareaid=2042220
],
 [
 "autohomeBrandData"
]

```



enable css selector helper web html follows 1 messages 2


```

运行爬虫去爬取数据

调试完毕后，返回项目，status改为DEBUG或RUNNING，点击Run

想要暂停运行：status改为STOP

保存已爬取的数据

当爬取完毕数据，需要保存下来时，可以有多种保存方式：

- mysql数据库
- MongoDB数据库
- CSV或Excel文件

保存到csv或Excel文件

基本思路：确保自己代码中，最后return返回的字段是你要的字段

如何得到CSV文件：在任务运行期间或完毕后，去 `Results ->` 点击下载 `csv`，即可得到你要的csv格式的数据文件。

结果：PySpider会自动在已有字段中加上额外的 `url` 字段

用VSCode编辑csv文件

- 如果想要去除多余的不需要的 `url` 字段，则可以通过文本编辑器，比如 `vscode` 去列编辑模式，批量删除，或者查找和替换，都可以实现
- 最后会多余一列，标题是 `...`，内容全是 `,`，所以直接用编辑器比如 `VSCode` 去替换为空以清空，即可

详见：

[【已解决】PySpider如何把json结果数据保存到csv或excel文件中 – 在路上](#)

Excel去打开CSV文件结果乱码

csv文件编码默认为UTF8（是好事，通用的），但是如果用（不论是Mac还是Win中的）excel去打开，结果（估计对于中文系统，都是）会默认以GKB（或GB18030）打开，所以会乱码

解决办法：[【已解决】Mac或Win中用Excel打开UTF8编码的csv文件显示乱码](#)

PySpider中选择html中的元素和内容

PySpider中的html的选择，或者说 `css选择器`，默认是用的 `PyQuery`。

确切的说是：PySpider针对html的响应 `response`，默认提供了一个 `doc` 属性，其内置了 `PyQuery` 解析后结果，所以你可以用 `response.doc("your_css_selector")` 去选择你要的html中的内容了。

而具体的`your_css_selector`的写法，则就变成 `PyQuery` 的写法了。

举例：

想要提取：

```
<ul class="rank-list-ul" ...>
  <li id="s3170">
    ...
    <div>
      ...
      <a id="atk_3170" href="//car.autohome.com.cn/pic/series/3170.html#pvareaid=103448">图库</a>
    ...
  </div>
</li>
...
</ul>
```

中的`href`的值，则：

PyQuery的写法是：`.rank-list-ul li div a[href*="/pic/series"]`

PySpider的代码是：

```
for each in response.doc('.rank-list-ul li div a[href*="/pic/series"]').items():
```

详见：[【已解决】pyspider中如何写规则去提取网页内容 – 在路上](#)

PyQuery

`response.doc` 返回后的 PyQuery对象，之后可以继续用PyQuery去操作

此处列出PyQuery的一些典型的操作函数：

- `PyQuery.filter(selector)`
- `PyQuery.find(selector)`
- `PyQuery.items(selector=None)`
- `PyQuery.siblings(selector=None)`

另外，常见的一些属性来说：

- `PyQuery.text(value=<NoDefault>)`：当前节点的text文本值
- `PyQuery.html(value=<NoDefault>, **kwargs)`：当前节点的html值

另外：

- 关于css操作可以参考：[CSS — pyquery 1.2.4 documentation](#)

关于PyQuery，详见另外教程：

【整理Book】Python心得：PyQuery

crifan.com, 使用[知识署名-相同方式共享4.0协议](#)发布 all right reserved, powered by Gitbook该文件修订时间：2018-09-20 18:38:55

PySpider的高级用法

下面介绍PySpider中，除了基本用法之外的，可以算作是高级，稍微更加复杂一些的用法。

crifan.com, 使用[知识署名-相同方式共享4.0协议](#)发布 all right reserved, powered by Gitbook该文件修订时间: 2018-09-20 15:55:26

self.crawl详解

官方文档

此处要介绍的 PySpider 的获取网络请求的函数是： `self.crawl`， 功能很强大。

具体详细的解释，可以参考：

- 官网的英文文档：
 - 比如：[crawl参数 params](#)
- 某热心网友整理的中文文档：
 - [self.crawl - pyspider中文文档 - pyspider中文网](#)

给 GET 的请求添加 查询参数

给 `self.crawl` 中给 `params` 传递对应字典变量，PySpider 内部会自动把字典编码为url的查询参数 query string.

官方实例：

```
self.crawl('http://httpbin.org/get', callback self.callback, params={'a': 123, 'b': 'c'})
```

等价于：

```
self.crawl('http://httpbin.org/get?a=123&b=c', callback self.callback)
```

自己之前用的例子有：

```
topSignTopParam = {
    "start": 0,
    "rows": 20
}
self.crawl(TopSignTopUrl,
           callback self.getMoreUserCallback,
           params topSignTopParam,
           save {
               "baseUrl": TopSignTopUrl,
               "isNeedCheckNextPage": True,
               "curPageParam": topSignTopParam
           }
)
```

给callback函数加上额外的参数

使用 `self.crawl` 的 `save` 参数即可，然后callback中用 `response.save` 获取传入的值

举例：

```
def getUserDetail(self, userId):
    self.crawl(UserDetailUrl,
               callback self.userDetailCallback,
               params {"member_id": userId},
               save userId
)

def userDetailCallback(self, response):
```

```
userId = response.save  
print("userId=%s" % userId)
```

当请求出错时也执行callback回调函数

需要给callback回调函数加上修饰符 `@catch_status_code_error`

举例：

```
def picSeriesPage(self, response):  
    ...  
    self.crawl(curSerieDict["url"], callback self.carModelSpecPage, save curSerieDict)  
  
@catch_status_code_error  
def carModelSpecPage(self, response):  
    curSerieDict = response.save  
    print("curSerieDict=%s", curSerieDict)  
    ...  
    return curSerieDict
```

crifan.com, 使用[知识署名-相同方式共享4.0协议](#)发布 all right reserved, powered by Gitbook该文件修订时间： 2018-09-20 16:53:43

独立的配置文件: config.json

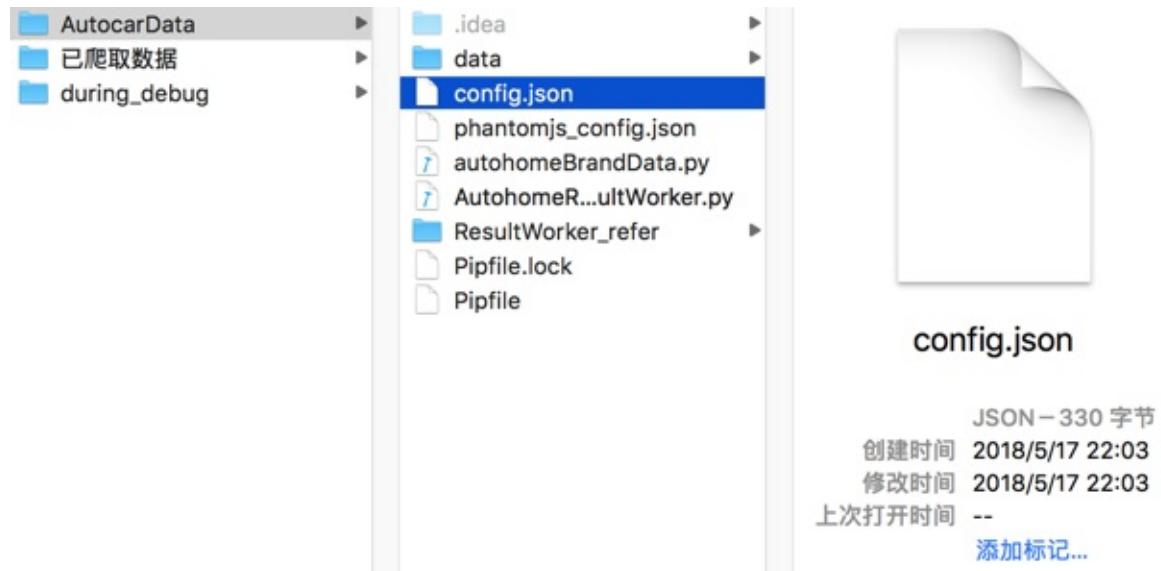
如果需要用到复杂一点的配置，比如 `result worker`，则是需要单独写配置文件。

PySpider的配置文件一般叫做 `config.json`

比如用如下内容：

```
{
  "taskdb": "mysql+taskdb://root:crifan_mysql@127.0.0.1:3306/AutohomeTaskdb",
  "projectdb": "mysql+projectdb://root:crifan_mysql@127.0.0.1:3306/AutohomeProjectdb",
  "resultdb": "mysql+resultdb://root:crifan_mysql@127.0.0.1:3306/AutohomeResultdb",
  "result_worker": {
    "result_cls": "AutohomeResultWorker.AutohomeResultWorker"
  }
}
```

将 `config.json` 保存在 `pyspider` 命令运行所在的当前目录下：



然后去 `-c` 指定配置文件：

```
pyspider -c config.json
```

PySpider所在目录下的 data 目录

在你运行 `pyspider` 后，自动会在命令执行路径下生成 `data` 文件夹，其中包含几个（SQLite）文件：

- `project.db`：保存了用户的爬虫项目相关信息，包括项目的 Python代码
 - 比如用（SQLite）工具去查看，可以看到详细数据
 - 比如Mac中的 `DB Browser for SQLite` 查看的效果：
 - Python代码：
 -
 - 对应数据库结构字段：
 -
- `result.db`：项目运行的结果数据
- `task.db`：项目相关的任务信息
 - 其中如果开始运行爬虫，还会出现相关的调度信息：
 - `scheduler.all`, `scheduler.id`, `scheduler.1h`：保存了任务执行后所有，1天，1小时内相关的信息，和WebUI

中的 progress 中的 all , 1d , 1h 对应:

-

如何清除之前的或正在运行的任务

对于一个写好的爬虫，且已经点击 Run 运行，或者运行了一段时间后，主动停止了。

接着想要去删除之前下载的数据，则：

[官网的解释](#)是：

设置 group 为 delete，以及 status 为 STOP 后，过了(默认) 24小时 后，会自动删除该项目所有信息。

但是往往没法满足我们需求：

我不想要等待，只想现在就去：删除掉所有的信息，包括之前已经爬取的数据，之前的调度的任务等等数据。

经过一番研究后，发现了解决方案：

- 先去停止项目

- WebUI中设置 status 为 STOP

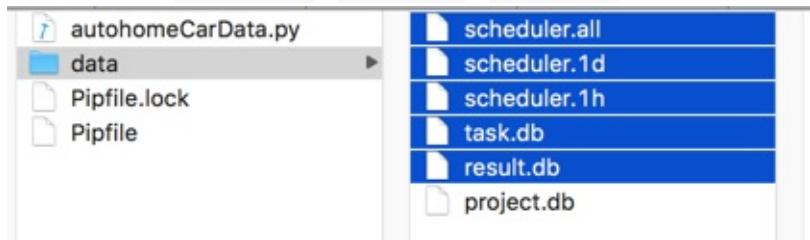
- 终端中用 Control+C 强制停止 pyspider 的运行

```
I 180428 10:34:47 scheduler:647] scheduler starting...
[I 180428 10:34:47 scheduler:126] project autohomeBrandData updated, status:DEBUG, paused:False, 0 tasks
[I 180428 10:34:47 scheduler:965] select autohomeBrandData:_on_get_info data:,_on_get_info
[I 180428 10:34:47 scheduler:586] in Sm: new:0,success:0,retry:0,failed:0
[I 180428 10:34:47 scheduler:782] scheduler.xmlrpc listening on 127.0.0.1:23333
[I 180428 10:34:47 tornado_fetcher:188] [200] autohomeBrandData:_on_get_info data:,_on_get_info 0s
[I 180428 10:34:47 app:76] webui running on 0.0.0.0:5000
[D 180428 10:34:47 project_module:145] project: autohomeBrandData updated.
[I 180428 10:34:47 processor:202] process autohomeBrandData:_on_get_info data:,_on_get_info -> [200] len:12 -> result:None fol:0 msg:0 err:None
[I 180428 10:34:47 scheduler:360] autohomeBrandData on_get_info {'min_tick': 0, 'retry_delay': {}, 'crawl_config': {}}
^C[I 180428 10:35:20 app:84] webui exiting...
[I 180428 10:35:20 processor:229] processor exiting...
[I 180428 10:35:20 scheduler:663] scheduler exiting...
[I 180428 10:35:20 result_worker:66] result_worker exiting...
[I 180428 10:35:20 tornado_fetcher:671] fetcher exiting...

aborted!
  ➔ AutocarData
  ➔ AutocarData
```

- 再去删除文件： result.db 和 task.db

- 如果还有任务相关的 scheduler.all , scheduler.1d , scheduler.1h ，则一并删除



不要轻易在没备份代码情况下删除project.db

注意不要删除，保存了项目（配置和）代码的： project.db，否则代码就没了。（我最开始就这么干过，）

之后去重新运行pyspider，再去刷新WebUI界面：

<http://0.0.0.0:5000/>

即可看到干净的项目，没有了之前的任务和数据了。

crifan.com, 使用[知识署名-相同方式共享4.0协议](#)发布 all right reserved, powered by Gitbook该文件修订时间: 2018-09-20 18:22:30

Phantomjs

如何解决部分页面内部不显示，无法抓取的问题？

折腾 [【已解决】PySpider中页面部分内容不显示 – 在路上](#)，遇到个问题：

页面中的部分内容不显示，所以无法抓取。

经过研究发现，其实是：

这部分不显示的内容，是原网页中通过后续调用js去生成和获取的，所以可以通过：

给 `self.crawl` 添加

```
fetch_type 'js'
```

会使得内部调用 `phantomjs`，模拟js，渲染生成页面内容。

从而，此处 `PySpider`，在这种需要显示js加载的页面内容时，可以利用 `phantomjs`。

用了 `phantomjs` 后又出错： `FETCH_ERROR HTTP 599 Connection timed out after milliseconds`

后续继续运行，加了 `fetch_type='js'` 的代码，去爬取页面数据，结果遇到了：

[【未解决】pyspider运行出错：FETCH_ERROR HTTP 599 Connection timed out after milliseconds](#)

尝试了多种办法，都无法解决此问题。

所以目前的情况是：

如果加了 `phantomjs`，结果在大量爬取页面期间，又会导致出错 `FETCH_ERROR HTTP 599 Connection timed out after milliseconds`，而暂时找不到解决办法。

给 `Phantomjs` 添加额外参数

之前折腾过：

[【未解决】pyspider中如何给phantomjs传递额外参数 – 在路上](#)

基本上没有实现想要的效果。但是可供参考。

phantomjs中的proxy是什么意思

对于pyspider来说，`phantomjs-proxy`参数指的是：

你另外所运行的 `phantomjs`的实例 = `host:port`

比如：

在一个终端中运行：

```
pyspider phantomjs --port 23450 --auto-restart true
```

然后去另外一个终端中运行pyspider:

```
pyspider -c config.json all
```

其中 config.json 包含了:

```
"phantomjs-proxy": "127.0.0.1:23450"
```

就可以使得此处的pyspider在启动时不另外启动phantomjs了。

而是去在需要用到phantomjs时，去连接本地电脑127.0.0.1的23450端口中的phantomjs去处理，去加载页面了。

而对于phantomjs本身来说：

proxy，指的是代理，比如翻墙的代理，等等。

具体相关设置，可以参考：

[Command Line Interface | PhantomJS](#)

中的：

- `-proxy=address:port` specifies the proxy server to use (e.g. `-proxy=192.168.1.42:8080`)
- `-proxy-type=[http|socks5|none]` specifies the type of the proxy server (default is http).
- `-proxy-auth` specifies the authentication information for the proxy, e.g. `-proxy-auth=username:password`)

详见： [【已解决】pyspider中phantomjs中的proxy是什么意思 – 在路上](#)

crifan.com, 使用[知识署名-相同方式共享4.0协议](#)发布 all right reserved, powered by Gitbook该文件修订时间: 2018-09-20 16:45:17

PySpider经验与心得

折腾了一些PySpider项目，有些经验和心得，整理如下：

crifan.com, 使用[知识署名-相同方式共享4.0协议](#)发布 all right reserved, powered by Gitbook该文件修订时间: 2018-09-20 17:21:43

PySpider的心得

对于加载更多内容，除了想办法找js或api，也可以换个其他的思路

问题：想要获取单个页面的更多的内容，一般页面都是向下滚动，加载更多。内部往往是js实现，调用额外的api获取更多数据，加载更多数据。

思路：所以一般往往会去研究和抓包，搞清楚调用的api。但是其实有思路多去看看网页中与之相关的其他内容，往往可以通过其他途径，比如另外有个单独的页面，可以获取我所需要的所有的车型车系的数据。就可以避免非要去研究和抓包api了。

详见：【已解决】pyspider中如何加载汽车之家页面中的更多内容

调试界面中的 enable css selector helper

点击web后可以看到html页面内容

再点击 enable css selector helper 后

之后点击某个页面元素，则可以直接显示出对应的css的selector

```

"fetch": {},
"process": {
  "callback": "picSeriesPage"
},
"project": "autohomeBrandData",
"schedule": {
  "priority": 2
},
"taskId": "085c986e0e9dd74d8525784d63823ee2",
"url": "https://car.autohome.com.cn/pic/series/2429.html#pvareaid=103448"
}
curlModelId=$s, curSeriesId=$s 34045 2429
curlModelSpecUrl= https://www.autohome.com.cn/spec/34045/#pvareaid=2042128
curlModelName= 1.6L GDI 自动时尚型
curlFullModelName= 2018款 1.6L GDI 自动时尚型
curlModelPicUrl= https://car.autohome.com.cn/pic/series-s34047/2429.html#pvare
foundModelSeriesId= < sre.SRE_Match object; span=(28, 55), match='pic/series-
curlModelId=$s, curSeriesId=$s 34046 2429
curlModelSpecUrl= https://www.autohome.com.cn/spec/34046/#pvareaid=2042128
curlModelName= 1.6L GDI 自动领尚型
curlFullModelName= 2018款 1.6L GDI 自动领尚型
curlModelPicUrl= https://car.autohome.com.cn/pic/series-s34047/2429.html#pvare
foundModelSeriesId= < sre.SRE_Match object; span=(28, 55), match='pic/series-
curlModelId=$s, curSeriesId=$s 34047 2429
curlModelSpecUrl= https://www.autohome.com.cn/spec/34047/#pvareaid=2042128
curlModelName= 1.6L GDI 自动尊尚型
curlFullModelName= 2018款 1.6L GDI 自动尊尚型

```

```

1 #!/usr/bin/env python
2 # -*- encoding: utf-8 -*-
3 # Created on 2018-04-27 21:53:02
4 # Project: autohomeBrandData
5
6 from pyspider.libs.base_handler import *
7 import string
8 import re
9
10
11 class Handler(BaseHandler):
12     crawl_config = {
13     }
14
15     # @every(minutes=24 * 60)
16     def on_start(self):
17         for eachLetter in list(string.ascii_lowercase):
18             self.crawl("https://www.autohome.com.cn/grade/carhtml/%s.html" % eachLetter, callback=self.gradCarHtmlPage)
19
20     @catch_status_code_error
21     def gradCarHtmlPage(self, response):
22         print("gradCarHtmlPage: response:", response)
23         picSeriesItemList = response.doc('.rank-list-ul li div
a[href*="/pic/series"]').items()
24         print("picSeriesItemList:", picSeriesItemList)
25         # print("len(picSeriesItemList)=%s" % (len(picSeriesItemList)))
26         for each in picSeriesItemList:
27             self.crawl(each.attr.href, callback=self.picSeriesPage)
28
29     @config(priority=2)
30     def picSeriesPage(self, response):
31         # <a href="/pic/series-t-66.html">查看停产车型</a>
32         # <a class="ckmore" href="/pic/series/588.html">查看在售车型
33         # <span class="fn-right">&ampnbsp</span>
34         fnRightPicSeries = response.doc('.search-pic-tbar .fn-right
a[href*="/pic/series"]')
35         print("fnRightPicSeries:", fnRightPicSeries)
36         if fnRightPicSeries:
37             # hrefValue = fnRightPicSeries.attr.href
38             # print("hrefValue=", hrefValue)
39             # fullPicSeriesUrl = "https://car.autohome.com.cn" +

```

不过话说我个人调试页面期间，很少用到。

都是直接去Chrome浏览器中调试页面，查看html源码，寻找合适的css selector。

crifan.com, 使用[知识署名-相同方式共享4.0协议](#)发布 all right reserved, powered by Gitbook该文件修订时间: 2018-09-20 18:45:03

PySpider常见的坑

关于折腾PySpider期间，遇到很多或大或小的坑，常见和具体细节相关的坑，已记录到对应部分中了。

此处再继续整理出，其他的一些常见的坑。

css的选择器不工作

背景：网页中的源码本来是：

```
<a href="//car.autohome.com.cn/pic/series/3170.html#pvareaid=103448">
```

或者类似的：

```
href="/pic/series-t/3170.html"
```

The screenshot shows a search interface for Audi models on AutoHome. The search bar contains '大众' (Dongfeng) and the results page displays various Audi models like A3, A4L, A6L, Q3, etc. On the right, there's a sidebar for filtering by color and year. Below the sidebar, a dropdown menu is open, showing a list of items including '查看停产车型' (View discontinued models). The browser's developer tools are overlaid on the page, specifically the Elements tab, which highlights the same 'a' element from the dropdown menu. The CSS selector used is 'a[href = "//car.autohome.com.cn/pic/series/"]'. The developer tools also show the computed styles for this element.

然后去写css选择器：

```
a[href = "//car.autohome.com.cn/pic/series/"]
```

但是却无法匹配

原因： PySpider 内部的css选择器用的是 `jQuery`，其默认把href的路径，加上了对应的host，所以此时获取到的html实际上变成了：

```
<a href="https://car.autohome.com.cn/pic/series/3170.html#pvareaid=103448">
```

详见：

[response.doc](#)

Reponse.doc() 返回的就是一个PyQuery的对象 **Links have made as absolute by default**

猜测：估计是为了方便小白用户，所以默认加上了host，但是坑了其他人啊。

解决办法：此处被逼的css选择器写法只能改为：

```
a[href = "pic/series/"]
```

或类似的代码：

```
fnRightPicSeries = response.doc('.search-pic-tbar .fn-right a[href*="/pic/series"]')
fullPicSeriesUrl = fnRightPicSeries.attr.href
```

已经得到的是，加了host/domain的绝对路径了：

```
fullPicSeriesUrl https://car.automobile.com.cn/pic/series/t/3170.html
```

```

run 28         self.crawl(each.attr.href, callback=self.detail_page)
29
30     @config(priority=2)
31     def detail_page(self, response):
32         # <a href="/pic/series-t/66.html">查看停产车型<br>;</a>
33         # <a href="/pic/series/588.html">查看在售车型
34         # <span class="ckmore" href="/pic/series/588.html">查看在售车型
35         fnRightPicSeries = response.doc('.search-pic-tbar .fn-right
36         a[href*="/pic/series"]')
37         print("fnRightPicSeries=", fnRightPicSeries)
38         if fnRightPicSeries:
39             hrefValue = fnRightPicSeries.attr.href
40             print("hrefValue=", hrefValue)
41             fullPicSeriesUrl = "https://car.automobile.com.cn/" +
42             hrefValue
43             fullPicSeriesUrl = fnRightPicSeries.attr.href
44             print("fullPicSeriesUrl=", fullPicSeriesUrl)
45             self.crawl(fullPicSeriesUrl, callback=self.detail_page)
46
47         # continue parse brand data
48         abDictList = []
49         # for eachA in response.doc('.breadnav a[href="/"]').items():
50         for eachA in response.doc('.breadnav a[href*="/pic/"]').items():
51             eachADict = {
52                 "text": eachA.text(),
53                 "href": eachA.attr.href
54             }
55             print("eachADict=", eachADict)
56             abDictList.append(eachADict)
57
58             print("abDictList=", abDictList)
59
60             mainBrandDict = abDictList[-1]
61             subBrandDict = abDictList[-2]
62             brandSerieDict = abDictList[-3]
63             print("mainBrandDict=%s, subBrandDict=%s, brandSerieDict=%s" %
64             (mainBrandDict, subBrandDict, brandSerieDict))
65
66             dtTextList = []
67             for eachDt in response.doc("dl.search-pic-card1 dt").items():
68                 dtTextList.append(eachDt.text())
69             print("dtTextList=", dtTextList)

```

详见：[【已解决】pyspider中的css选择器不工作 – 在路上](#)

Error Could not create web server listening on port 25555

原因：对应的25555端口被占用了

根本原因：之前的PySpider没有正常的彻底的被关闭，所以残留了。

解决办法：彻底 kill 干掉之前的PySpider的进程即可。

举例：

普通Linux类系统，用：

- 找到占了25555端口的进程的id: `ps aux | grep 25555`
- 再去杀掉进程: `kill process_id -9`

即可。

如果是Mac中，则用 `lsof`

```
→ AutocarData lsof -i:25555
COMMAND      PID  USER   FD   TYPE             DEVICE SIZE/OFF NODE NAME
phantomjs 46971 crifan    12u  IPv4  0xe4d24cdcaf5e481f      0t0    TCP *:25555 (LISTEN)
→ AutocarData kill 46971
```

crifan.com, 使用[知识署名-相同方式共享4.0协议](#)发布 all right reserved, powered by Gitbook该文件修订时间: 2018-09-20 18:20:25

PySpider案例

crifan.com, 使用[知识署名-相同方式共享4.0协议](#)发布 all right reserved, powered by Gitbook该文件修订时间: 2018-09-20 14:12:50

附录

下面列出相关参考资料。

crifan.com, 使用[知识署名-相同方式共享4.0协议](#)发布 all right reserved, powered by Gitbook该文件修订时间: 2018-09-20 13:57:58

参考资料

- pyspider是开源强大的python爬虫系统 - [pyspider中文网](#)
- 【已解决】PySpider中保存数据到mysql
- 【已解决】PySpider中如何清空之前运行的数据和正在运行的任务 – 在路上
- 【未解决】pyspider中如何给phantomjs传递额外参数 – 在路上
- 【已解决】PySpider中页面部分内容不显示 – 在路上
- 【未解决】pyspider运行出错: `FETCH_ERROR HTTP 599 Connection timed out after milliseconds`
- 【记录】Mac中安装和运行pyspider
- 【整理】pyspider vs scrapy
- 【已解决】pyspider中phantomjs中的proxy是什么意思 – 在路上
- 【已解决】pyspider中运行result_worker出错: `ModuleNotFoundError No module named mysql`
- 【已解决】PySpider中传递参数给下一级且当下一级失败时也可以执行
- 【已解决】pyspider中出错: `TypeError __init__() got an unexpected keyword argument resultdb` – 在路上
- 【已解决】pyspider运行出错: `ImportError pycurl libcurl link-time ssl backend (openssl) is different from compile-time ssl backend (none/other)`
- 【已解决】pyspider中pymysql中insert失败且except部分代码没有执行
- 【已解决】pyspider运行出错: `Error Could not create web server listening on port 25555` – 在路上
- 【已解决】pyspider中的css选择器不工作 – 在路上
- 【已解决】pyspider中如何写规则去提取网页内容 – 在路上
- 【已解决】PySpider如何把json结果数据保存到csv或excel文件中 – 在路上
- 【已解决】PySpider中如何单个页面返回多个json数据结果 – 在路上
- 【已解决】Mac或Win中用Excel打开UTF8编码的csv文件显示乱码
-

crifan.com, 使用[知识署名-相同方式共享4.0协议](#)发布 all right reserved, powered by Gitbook该文件修订时间: 2018-09-20 18:25:44