```
LinkExtractor
from scrapy.linkextractors import LinkExtractor as LE
le=LE(restrict_xpaths='//div[@class="pg"]',tags='a',attrs='href',allow='bbs\.vivo\.com\.cn/thread-\d+-\d+\.html')
for _ in le.extract_links(response):
    yield Request(_.url,callback=self.parse)
# 这样写更健壮,不容易断
scrapy_redis
spiders.py和connection.py实现分布式(手动推进去地址)
pipelines.py和connection.py实现把数据存储到redis数据库
scheduler.py, dupefilter.py, queue.py和connection.py实现断点续爬,分布式
要想实现分布式,只需将该spider复制一份,然后再运行即可
item_key(self, item, spider)此处可以自定义item_key
SCHEDULER_PERSIST = True,中断程序,dmoz:dupefilter和dmoz:requests会存在,程序执行完后还会保存dmoz:dupefilter(此时已不存在
dmoz:requests)
SCHEDULER_PERSIST = False,中断程序,dmoz:dupefilter和dmoz:requests会被清空,程序执行完后会删除dmoz:dupefilter(此时已不存在
dmoz:requests)
type dmoz:dupefilter返回set
type dmoz:requests返回zset
采集坑:
import requests
r=requests.get(
  url='https://tsvideo.oss-cn-hangzhou.aliyuncs.com/崔庆才Python三大案例/405003.mp4',
  headers={
    'Accept-Encoding':'identity;q=1, *;q=0',
    'Range':'bytes=0-',
    'Referer': 'https://edu.hellobi.com/course/157/play/lesson/2575',
    'User-Agent':'Mozilla/5.0 (Windows NT 6.1; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/59.0.3071.115
Safari/537.36',
  },
  stream=True,
with open('16.mp4','wb') as f:
                                  # 注意头信息Range,必须要从0开始,对于大文件下载一般要stream=True
  for chunk in r.iter_content(1024 * 100):
  f.write(chunk)
print(r.status_code)
 import requests
 import re
import json
r=requests.get('http://commoncgi.wepiao.com/data/v5/movies/cities/227/movies_city_227.json')
 _=r.content.decode('unicode_escape').replace('\/','/') #注意
_=re.search(r'"info":(.+?), "announcement":',_,re.S).group(1)
print(_)
亚马逊基本信息页采集ASIN时xpath('//*[@id="atfResults"]/ul/li/@data-asin')会出现采集不全(只有前4个)的情况
是应为某些标签被JS更改了,而爬虫只能获取未被JS处理的网页源代码
解决办法是Chrome设置->禁用亚马逊执行JS,然后再F12观察其源码结构
                                                 Cookie
                                                   允许设置本地数据(推荐)
                                                                                                    ×
                                                    JavaScript 例外情况
                                                      主机名(可包含通配符)
                                                                                      行为
                                                                                      禁止
                                                      https://www.amazon.com
                                                 图片
                                                      [*.]example.com
                                                                                       允许
```

```
了解详情
                                                                                                                                                                                                                                                                                                                              完成
                                                                                                                                                                处理程序
                                                                                                                                                                     ● 分准网让更求成为协议的野门办理程度/推荐
www.amazon.com/s?rh=k%3Aiphone%2Cn%3A2335752011%2Cn%3A7072561011%2Cn%3A2407749011&page=3
                                                                                                                                                            prime
                                                                                                                                                                                                                        ▲ Elements Console Sources Network Performance Memory Application
                                                                                                                                                                                                                                                     <div id="rhsAiax"></div>
                                                                                                                                                                                                                                                  ▼ <div id="resultsCol" class="showRightCol">
                                                                                                                                                                                                                                                    ▼ <div id="centerMinus" class="leftCol";
                                                                                                                                                                                                                                                       ▼ <div id="atfResults" class="a-row s-result-list-parent-
                                                       Apple iPhone 4S 16GB 3G WiFi White Smartphone
                                                       AT&T
                                                                                                                                                                                                                                                          ▼ <ul id="s-results-list-atf" class="s-result-list s-col-1 s-col-ws-1 s-result-list-hgrid s-height-equalized s-list-view s-
                                                       by Hyperion EA
                                                       $44.95 (78 used & new offers)
                                                                                                               ★★★★☆ ▼ 865
                                                                                                                                                                                                                                                             ▶ id="result_48" data-asin="B010GNNXSG" class="s-result-
                                                                                                                                                                                                                                                            item celwidget ">...
▶

    Operating System: iOS

                                                                                                                                                            Charger, ASUMI 4.2A 21W Fast Ch...

    Display Size: 3.5 inches

                                                                                                                                                            $12<sup>99</sup> $25.99 \( \text{prime} \)
                                                                                                                                                                                                                                                            item celwidget ">...
lid="result_50" data-asin="B0055SB0YO" class="s-result-item celwidget ">...

                                                                                                              · Special Feature: Includes
                                                                                                                                                            金金金金金 62
                                                                                                                                                                                                                                                            item celwidget ">...

| Class="s-result-
| id="result_51" | data-asin="B06XGKQT4L" | class="s-result-
| id="result_51" | data-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-asin="bata-as
                                                                                                                 White LifeProof Case
                                                                                                              . Display Dimensions: 4.5 in
                                                                                                              . Display Type: LED-backlit
                                                                                                                                                                                                                                                             item celwidget ">...
                                                                                                                                                                                                                                                             ::after
                                                       Apple iPhone 7 Plus Factory Unlocked GSM
                                                                                                                                                                                                                                                        </div>
                                                                                                                                                                                                                                                     Smartphone - (Certified Refurbished) (32, Black)
                                                                                                                                                                                                                                                     ▼<div id="centerBelowPlus" class="leftCol">
                                                       by Apple
                                                                                                                                                                                                                                                      ▶ <div id="search-js-btf">...</div>
▶ <div id="raw-search-js-btf" class="searchUndoAUIHacks">...</div>
                                                       AmazonBasics Clear Case for iPho...
                                                                                                                                                                                                                                                        ▼ <div id="btfResults" class="a-row s-result-list-parent-
                                                       Get it by Tomorrow, Jun 14
                                                                                                              Product Features
                                                                                                                                                            $348 $8.99 prime
                                                                                                                                                                                                                                                       container">
                                                                                                                                                                                                                                                            ::before
                                                       FREE Shipping on eligible orders
                                                                                                                 The brightest, most colorfu
                                                                                                                Phone display ever. The
                                                                                                                                                                                                                                                          ▼<ul class="s-result-list s-col-1 s-col-ws-1 s-result-list-
                                                       More Buying Choices
                                                                                                                                                                                                                                                          hgrid s-height-equalized s-list-view s-text-condensed">
                                                       $699.99 (2 new offers)
                                                                                                                                                                                                                                                             ▶ <li id="result_52" data-asin="B014KHBS7W" class="s-result-
                                                                                                                                                                                                                                                            item celwidget ">...
                                                                                                                                                                                                                                                             ▶ id="result_53" data-asin="B01EVPCFH0" class="s-result-
   \textbf{ul.}s - \textbf{result-list.}s - \textbf{col-l.}s - \textbf{col-ws-l.}s - \textbf{result-list-hgrid.}s - \textbf{height-equalized.}s - \textbf{list-view.}s - \textbf{text-condensed} \mid \textbf{650} \times \textbf{5125} 
                                                                                                                                                                                                                                                            item celwidget ">...
                                                                                                                                                                                                                                                             ▶ id="result_54" data-asin="B01FS0AAZG" class="s-result-
                                                                                                                                                                                                                                                            Apple iPhone 6 64GB Space Gray - (Verizon
                   \li id="result_56" data-asin="B01I48LT50" class="s-result-
                                                                                                                                                                                                                                                             item celwidget ">...
                                                                                                        ★★★★☆ → 387
                                                                                                                                                                                                                                                             ▶ <li id="result_57" data-asin="B01BI2VZEI" class="s-result-
                                                      $45900 prime
                                                                                                                                                                                                                                                             item celwidget ">...
                                                      FREE Shipping on eligible orders

Operating System: iOS

Instaction lune 20, 2017

Display Size: 4.7 inches
                                                                                                                                                                                                                                                             ▶ <li id="result_58" data-asin="B00F3IY76M" class="s-result-
                                                      In stock on June 20, 2017
                                                                                                                                                                                                                                                             item celwidget ">...
                                                                                                                                                                                                                                                             ▶ id="result_59" data-asin="B015E8W1RC" class="s-result-

    Special Feature: Ite

    Display Dimensions: 4.7

                                                                                                                                                                                                                                                             item celwidget ">...
                                                       More Buying Choices
                                                                                                                                                                                                                                                             \(\text{li id="result_60" data-asin="B00SHYBN70" class="s-result-
                                                                                                                                                                                                                                                             item celwidget ">...
                                                                                                                                                                                                                                                             ▶id="result_61" data-asin="B00F3IYJZ6" class="s-result-
                                                                                                                                                                                                                                                             item celwidget ">...
                                                                                                                                                                                                                                                            Apple iPhone 6s Plus a1687 16GB Space Gray
                                                                                                                                                                                                                                                            \blacktriangleright <1i id="result_63" data-asin="B00SHYBVCW" class="s-result-item celwidget ">...
                                                       Smartphone Unlocked (Certified Refurbished)
                      \di id="result_64" data-asin="B015T8K69C" class="s-result-
                                                                                                                                                                                                                                                             item celwidget ">...
                                                     $43900 $499.99 prime

Get it by Tomorrow, Jun 14

FREE Shipping on eligible orders

Operating System: IOS 10

Display Size: 5.5 inches
                                                                                                                                                                                                                                                             \di id="result_65" data-asin="B00NP691RY" class="s-result-
                   item celwidget ">...
                                                                                                                                                                                                                                                             \( \li id="result_66" data-asin="B01BYETZ20" class="s-result-
                                                                                                                                                                                                                                                             item celwidget ">...
                                                                                                                                                                                                                                                             ▶ id="result_67" data-asin="B01CR1SRTA" class="s-result-
                                                      More Buying Choices
                                                                                                                                                                                                                                                             item celwidget ">...
```

copied. response.meta['item'].update({'data':data}) #因为response.meta['item']是可变对象,so这样写存在BUG yield Request(url,callback = self.parse,meta={'item':response.meta['item']}) 新浪微博为了让自己的结果呈现在搜索引擎中,对来自搜索引擎的爬虫是"来者不拒" import requests headers = {'User-agent': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/58.0.3029.110 Safari/537.36'} r = requests.get("http://weibo.com/rmrb",headers=headers) with open('bad.html','wb') as f: #没有内容 f.write(r.content)

meta (dict) – the initial values for the Request.meta attribute. If given, the dict passed in this parameter will be shallow

爬虫不要瞎鸡巴加头信息,采集亚马逊商品基本信息时遇到此坑,一般按useragent,host,referer,cookie等顺序添加,一旦有效则不再往后添加

headers = {'User-agent': 'Baiduspider'} #把自己伪装成搜索引擎爬虫,具体用什么随意啦~谷歌、必应都可以,或者仅仅用spider也行!

```
f.write(r.content)
 Google
              http://weibo.com/rmrb
                     图片
                            地图
                                  视频
              网页
                                         新闻
                                                更多▼
                                                       搜索工具
              找到约 20,200,000 条结果 (用时 0.45 秒)
              人民日报的微博 - 新浪微博
              weibo.com/rmrb ▼
              人民日报 http://weibo.com/rmrb. 人民日报法人微博。参与、沟通、记录时代。 《人民日
              报》法人微博, 举报身份 申请认证 · 主页; 微博; 人民日报博客; 新闻客户端; 文章 ...
              人民日报的微博 - 新浪微博
              weibo.com/rmrb?entry=hkweibo ▼
              人民日报 http://weibo.com/rmrb. 人民日报法人微博。 ... 请网友认准"张静杰",千万别上
              当。http://t.cn/R7hykGX 转起提醒身边人! (577)| 转发(1176)| 收藏| 评论(451).
              人民日报的微博 - 新浪微博
              weibo.com/rmrb?topnav=1&wvr=5&topsug=1 ▼
              人民日报 http://weibo.com/rmrb. 人民日报法人微博。参与、沟通、记录时代 ... 嫁妆没用
              一根钉子,全是限量版! http://t.cn/RhsklCF. (2285)| 转发(667)| 收藏| 评论(431).
UserAgent
有些网站必须把头信息设置成手机头信息,eg: https://www.amazon.com/gp/aw/d/B00YD53YQU
Noscript标签的使用
```

<noscript>标签是在浏览器(或者用户浏览标识),没有启动脚本支持的情况下触发的标签,在低级爬虫中,基本都没有配置js引擎,通常这种方式和 Ajax异步加载同时使用。用于保护自己不想让爬虫接触的信息。

天眼查绕过登录拿页面数据

访问https://m.tianyancha.com/company/2323965264 ● 访问https://m.baidu.com/s?word=广州奥利门窗有限公司 site:www.tianyancha.com,获取到公司在企查查对应地址url及Referer信息,再访问

import requests,re from lxml.html import fromstring

html=fromstring(r.content)

```
headers = {
    'Referer': 'https://m.baidu.com/from=0/bd_page_type=1/ssid=0/uid=0',
```

print(r.json()['mods']['listItems'])

r = requests.get("http://weibo.com/rmrb",headers=headers)

with open('good.html','wb') as f: #有内容

'User-Agent':'Mozilla/5.0 (Windows NT 6.1; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/66.0.3359.170 Safari/537.36',

data=html.xpath('//*[@id="results"]/div[1]/@data-log')[0] url=re.findall(r'https://www\.tianyancha\.com/company/\d+',data)[0] r=requests.get(url, headers=headers) # 头信息中的Referer非常重要 print(r.text) Lazada 必须要加cookie,不然每次请求都只能拿每个品类的第一页数据

不能并行查看品类信息,不然每次请求都只能拿每个品类的第一页数据,只能一个品类一个品类查看

url=f'https://www.lazada.co.id/beli-wanita-emas-murni/?ajax=true&page={page}'

r=requests.get('<u>https://m.baidu.com/s?word=</u>广州奥利门窗有限公司 site:<u>www.tianyancha.com</u>')

```
验证如下:
 import requests,re,pymysql,time,random
 headers = {
     'accept': '*/*',
     'content-type': 'application/json',
     'accept-encoding': 'gzip, deflate, br',
     'accept-language': 'en-US,en;q=0.9',
     'cookie':'t_uid=c4929da6-ca29-4143-83e9-e996fcac4189;',
     'user-agent': 'Mozilla/5.0 (Windows NT 6.1; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/66.0.3359.117
 Safari/537.36',
 for page in range(1,3): # OK
     url=f'https://www.lazada.co.id/beli-handphone/?ajax=true&page={page}'
     r=requests.get(url,headers=headers,timeout=18)
     print(r.json()['mods']['listItems'])
 for page in range(1,3):
     url=f'https://www.lazada.co.id/beli-wanita-emas-murni/?ajax=true&page={page}'
```

r=requests.get(url,headers=headers,timeout=18) print(r.json()['mods']['listItems']) for page in range(1,3): # Error url=f'https://www.lazada.co.id/beli-handphone/?ajax=true&page={page}' r=requests.get(url,headers=headers,timeout=18)

r=requests.get(url,headers=headers,timeout=18) print(r.json()['mods']['listItems']) 也可以用浏览器分别打开https://www.lazada.co.id/beli-handphone/?page=1和https://www.lazada.co.id/beli-wanita-emas-murni/?page=1进行 验证

```
import requests
res = requests.get("https://114.1688.com/sitemap.html?")
res.encoding = 'gb2312'
print(res.text)
```