

案例2: 阳光热线问政平台

一、想一想

回顾案例1，发现花费了较多的功夫在处理寻找下一页的url地址或者内容的url地址，能不能简化这个过程？

- 思路点拨

1.从response中提取所有的a标签对应的url地址

2.自动构造request请求，发送给引擎

能不能再优化？

满足需求的url地址，才发给引擎，同时能指定callback函数解析响应

二、数据提取需求

网站首页: <http://wz.sun0769.com/html/top/report.shtml>

共需提取4个字段信息：编号、标题、url、投诉内容

对比列表页和详情页的页面结构，发现需要提取的数据在详情页也可以获取到



那么，可不可以只在列表页获取链接(包括详情页和下一页)，而只在详情页做数据解析？可以，这种思想使用的是Scrapy框架中的CrawlSpider类爬虫

三、新建项目

创建阳光热线问政平台爬虫项目

```
$ scrapy startproject Sun
```

四、编写 items.py

```
import scrapy

class SunItem(scrapy.Item):
    # define the fields for your item here like:
    number = scrapy.Field()      # 编号
    detail_url = scrapy.Field()  # 详情页url
    title = scrapy.Field()       # 标题
    context = scrapy.Field()     # 投诉内容
```

五、创建CrawlSpider爬虫

```
$ cd Sun # 进入项目目录
$ scrapy genspider -t crawl sun 'sun.com'
```

上面的命令参数

-t t代表template，即使用模板

sun 代表爬虫名

'sun.com' 是起始的url

- 修改allowed_domains、start_urls

```
allowed_domains = ['wz.sun0769.com']
start_urls =
['http://wz.sun0769.com/index.php/question/questionType?type=4']
```

六、url提取规则

- 提取下一页url

[分析] 才页面上可以观察到列表url都包含路径questionType

The screenshot shows a web browser displaying a list of complaints on the website wz.sun0769.com. The browser's developer tools are open, showing the HTML structure of the page. A red box highlights the pagination links, which include URLs like <http://wz.sun0769.com/index.php/question/questionType?type=4&page=30>, <http://wz.sun0769.com/index.php/question/questionType?type=4&page=31>, <http://wz.sun0769.com/index.php/question/questionType?type=4&page=32>, <http://wz.sun0769.com/index.php/question/questionType?type=4&page=33>, <http://wz.sun0769.com/index.php/question/questionType?type=4&page=34>, <http://wz.sun0769.com/index.php/question/questionType?type=4&page=35>, <http://wz.sun0769.com/index.php/question/questionType?type=4&page=36>, <http://wz.sun0769.com/index.php/question/questionType?type=4&page=37>, <http://wz.sun0769.com/index.php/question/questionType?type=4&page=38>, <http://wz.sun0769.com/index.php/question/questionType?type=4&page=39>, <http://wz.sun0769.com/index.php/question/questionType?type=4&page=40>, <http://wz.sun0769.com/index.php/question/questionType?type=4&page=41>, <http://wz.sun0769.com/index.php/question/questionType?type=4&page=42>, <http://wz.sun0769.com/index.php/question/questionType?type=4&page=43>, <http://wz.sun0769.com/index.php/question/questionType?type=4&page=44>, <http://wz.sun0769.com/index.php/question/questionType?type=4&page=45>, <http://wz.sun0769.com/index.php/question/questionType?type=4&page=46>, <http://wz.sun0769.com/index.php/question/questionType?type=4&page=47>, <http://wz.sun0769.com/index.php/question/questionType?type=4&page=48>, <http://wz.sun0769.com/index.php/question/questionType?type=4&page=49>, <http://wz.sun0769.com/index.php/question/questionType?type=4&page=50>, <http://wz.sun0769.com/index.php/question/questionType?type=4&page=51>, <http://wz.sun0769.com/index.php/question/questionType?type=4&page=52>, <http://wz.sun0769.com/index.php/question/questionType?type=4&page=53>, <http://wz.sun0769.com/index.php/question/questionType?type=4&page=54>, <http://wz.sun0769.com/index.php/question/questionType?type=4&page=55>, <http://wz.sun0769.com/index.php/question/questionType?type=4&page=56>, <http://wz.sun0769.com/index.php/question/questionType?type=4&page=57>, <http://wz.sun0769.com/index.php/question/questionType?type=4&page=58>, <http://wz.sun0769.com/index.php/question/questionType?type=4&page=59>, <http://wz.sun0769.com/index.php/question/questionType?type=4&page=60>, <http://wz.sun0769.com/index.php/question/questionType?type=4&page=61>, <http://wz.sun0769.com/index.php/question/questionType?type=4&page=62>, <http://wz.sun0769.com/index.php/question/questionType?type=4&page=63>, <http://wz.sun0769.com/index.php/question/questionType?type=4&page=64>, <http://wz.sun0769.com/index.php/question/questionType?type=4&page=65>, <http://wz.sun0769.com/index.php/question/questionType?type=4&page=66>, <http://wz.sun0769.com/index.php/question/questionType?type=4&page=67>, <http://wz.sun0769.com/index.php/question/questionType?type=4&page=68>, <http://wz.sun0769.com/index.php/question/questionType?type=4&page=69>, <http://wz.sun0769.com/index.php/question/questionType?type=4&page=70>, <http://wz.sun0769.com/index.php/question/questionType?type=4&page=71>, <http://wz.sun0769.com/index.php/question/questionType?type=4&page=72>, <http://wz.sun0769.com/index.php/question/questionType?type=4&page=73>, <http://wz.sun0769.com/index.php/question/questionType?type=4&page=74>, <http://wz.sun0769.com/index.php/question/questionType?type=4&page=75>, <http://wz.sun0769.com/index.php/question/questionType?type=4&page=76>, <http://wz.sun0769.com/index.php/question/questionType?type=4&page=77>, <http://wz.sun0769.com/index.php/question/questionType?type=4&page=78>, <http://wz.sun0769.com/index.php/question/questionType?type=4&page=79>, <http://wz.sun0769.com/index.php/question/questionType?type=4&page=80>, <http://wz.sun0769.com/index.php/question/questionType?type=4&page=81>, <http://wz.sun0769.com/index.php/question/questionType?type=4&page=82>, <http://wz.sun0769.com/index.php/question/questionType?type=4&page=83>, <http://wz.sun0769.com/index.php/question/questionType?type=4&page=84>, <http://wz.sun0769.com/index.php/question/questionType?type=4&page=85>, <http://wz.sun0769.com/index.php/question/questionType?type=4&page=86>, <http://wz.sun0769.com/index.php/question/questionType?type=4&page=87>, <http://wz.sun0769.com/index.php/question/questionType?type=4&page=88>, <http://wz.sun0769.com/index.php/question/questionType?type=4&page=89>, <http://wz.sun0769.com/index.php/question/questionType?type=4&page=90>, <http://wz.sun0769.com/index.php/question/questionType?type=4&page=91>, <http://wz.sun0769.com/index.php/question/questionType?type=4&page=92>, <http://wz.sun0769.com/index.php/question/questionType?type=4&page=93>, <http://wz.sun0769.com/index.php/question/questionType?type=4&page=94>, <http://wz.sun0769.com/index.php/question/questionType?type=4&page=95>, <http://wz.sun0769.com/index.php/question/questionType?type=4&page=96>, <http://wz.sun0769.com/index.php/question/questionType?type=4&page=97>, <http://wz.sun0769.com/index.php/question/questionType?type=4&page=98>, <http://wz.sun0769.com/index.php/question/questionType?type=4&page=99>, <http://wz.sun0769.com/index.php/question/questionType?type=4&page=100>. The page also shows a list of complaints with columns for ID, Title, Status, and Assignee.

立即导出，下一页url提取规则

```
# 下一页url
# http://wz.sun0769.com/index.php/question/questionType?
type=4&page=30
# http://wz.sun0769.com/index.php/question/questionType?
type=4&page=60
# http://wz.sun0769.com/index.php/question/questionType?
type=4&page=90
Rule(LinkExtractor(allow=r'questionType'), follow=True),
```

其中:

LinkExtractor为链接解析器，接收了2个参数

allow: 值的类型为正则表达式字符串

follow: 值的类型为布尔值

链接解析器的作用：从response中获取链接并解析出url地址，并判断是否满足allow设定的正则规则，如果满足就交给引擎，对该url发起请求，否则就过滤掉，follow参数是设定，请求该url获取到响应后，是否对响应继续做链接解析器过滤。

- 提取详情页url

先观察详情页url

The screenshot shows the 'Sunshine Hotline Question and Answer Platform' (阳光热线问政平台) website. The page displays a list of questions under the '投诉' (Complaints) section. The questions are listed in a table with columns for '编号' (Number), '标题' (Title), '状态' (Status), and '网友' (User). The questions are as follows:

编号	标题	状态	网友
208437	[投诉] 关于中国移动元宵活动相关事宜(图)	待处理	kathyl
208435	[投诉] 东莞长安马沙环西路南向北成停车场了吗?	待处理	dgcayh
208432	[投诉] 餐饮店油烟直排!!! 市环保局	已受理	kirin_new
208431	[投诉] 大朗洋坑塘新围路铭美汽车美容噪音废气扰民 太闹	待处理	yanglei8_8
208430	[投诉] 举报东莞市石碣碧桂园房地产开发有限公司 石碣	待处理	开成
208426	[投诉] 挂门牌号的万江社区工作人员素质低劣, 暴力执法!	待处理	csewei
208425	[投诉] 出租房饲养狗群噪声扰民 严重(图)	待处理	3188943
208419	[投诉] 咨询电话是否是摆设 市社保局	待处理	lefan10
208417	[投诉] 简沙洲的模具厂噪音废气污染大, 严重扰民 市环保局	已受理	流出汁

Red arrows point from the '标题' column to the HTML source code on the right. The arrows point to the following URLs in the code:

- From the first question (208437) to the URL: <http://wz.sun0769.com/html/question/201902/482402.shtml>
- From the second question (208435) to the URL: <http://wz.sun0769.com/html/question/201902/482398.shtml>

总结得出详情页提取的url

```
# 详情页url
# http://wz.sun0769.com/html/question/201902/402402.shtml
# http://wz.sun0769.com/html/question/201902/402398.shtml
Rule(LinkExtractor(allow=r'http://wz.sun0769.com/html/question/\nd+/\d+.shtml'),callback='parse_item', follow=False),
```

这里出现了链接解析器的第三个参数——callback，没错了，这就是指定解析数据的函数，但是注意了，其参数类型是字符串！！！！还需要注意的是这里的follow参数为False，即解析完数据后，直接将数据交给引擎，不必再在对当前页面做链接解析器过滤了。

七、解析详情页

编号

```
# 解析编号
item['number'] =
response.xpath('//tr/td[2]/span[2]/text()').extract_first().split(':')[
    -1].strip()
```

标题

```
# 解析标题
item['title'] =
response.xpath('/html/head/title/text()').extract_first().split(
    '_')[0]
```

url

```
# 获取详情url
item["detail_url"] = response.url
```

投诉内容

注意，这里有坑了，投诉内容可能会包含图片，那么解析规则就会不同

```

# 获取是否有图片
img = response.xpath('//td[@class="txt16_3"]/div/img')
if img:
# 有图片时, 解析投诉内容
text =
response.xpath('//div[@class="contenttext"]/text()').extract()
else:
# 无图片时, 解析投诉内容
text = response.xpath('//td[@class="txt16_3"]/text()').extract()
item["context"] = "".join([x.strip() for x in text if not
x.isspace()])

```

爬虫部分 `sun.py` 完整代码

```

from scrapy.linkextractors import LinkExtractor
from scrapy.spiders import CrawlSpider, Rule

from Sun.items import SunItem

class SunSpider(CrawlSpider):
    name = 'sun'
    allowed_domains = ['wz.sun0769.com']
    start_urls =
['http://wz.sun0769.com/index.php/question/questionType?type=4']

    rules = (
        # 下一页url
        # http://wz.sun0769.com/index.php/question/questionType?
type=4&page=30
        # http://wz.sun0769.com/index.php/question/questionType?
type=4&page=60
        # http://wz.sun0769.com/index.php/question/questionType?
type=4&page=90
        Rule(LinkExtractor(allow=r'questionType'), follow=True),
        # 详情页url

```

```

#
http://wz.sun0769.com/html/question/201902/402402.shtml
#
http://wz.sun0769.com/html/question/201902/402398.shtml

Rule(LinkExtractor(allow=r'http://wz.sun0769.com/html/question/
\d+/\d+.shtml'), callback='parse_item',
      follow=False),
)

def parse_item(self, response):
    # 创建item对象
    item = SunItem()
    # 解析编号
    item['number'] =
response.xpath('//tr/td[2]/span[2]/text()').extract_first().spli
t(':')[
        -1].strip()
    # 解析标题
    item['title'] =
response.xpath('/html/head/title/text()').extract_first().split(
'_')[0]
    # 获取详情url
    item["detail_url"] = response.url
    # 获取是否有图片
    img = response.xpath('//td[@class="txt16_3"]/div/img')
    if img:
        # 有图片时，解析投诉内容
        text =
response.xpath('//div[@class="contenttext"]/text()').extract()
    else:
        # 无图片时，解析投诉内容
        text =
response.xpath('//td[@class="txt16_3"]/text()').extract()
        item["context"] = "".join([x.strip() for x in text if
not x.isspace() and x])
        print("---" * 50)

```

```
print(item)
print("---" * 50)
yield item
```

八、数据持久化

数据存储到MongoDB，请同学们自行完成

九、启动爬虫

```
$ scrapy crawl sun
```

