

基于 scrapy 框架的爬虫

22920212204359 陈新

采集网页的地址: <https://honglouloumeng.5000yan.com/>

网页描述格式: HTML 格式

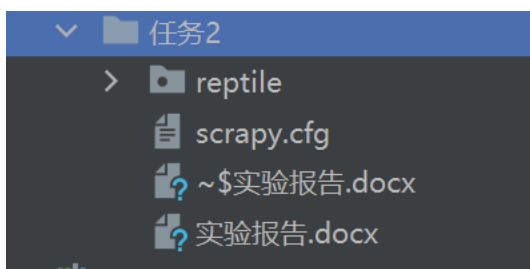
采集数据: 获取红楼梦每个章节的标题, 保存到指定的文本文件 1 中; 之后获取各章节的原文, 保存到文本文件 2 中

主要步骤:

1. 建立工程文件

```
\Desktop\陈新 > scrapy startproject 任务2
```

在终端输入命令 scrapy startproject 任务2, 此时生成一个项目文件夹任务2



之后在 reptile\spiders 文件夹下面创建一个 spiders.py 文件, 用于之后我们写入爬取网站的 url 和进行数据解析

2. 修改设置文件 settings.py

因为我们需要将文件存入 txt 文件, 使用终端命令无法满足要求, 故在该文件里开启 pipelines 用来后续输出

```
# See https://docs.scrapy.org/en/latest/topics/item-pipeline.html
ITEM_PIPELINES = {
    'reptile.pipelines.ReptilePipeline': 300,
}
```

同时我们修改对应的请求头，并且将 ROBOTSTXT_OBEY 修改为 False 否则无法进行爬虫。

```
BOT_NAME = 'reptile'

SPIDER_MODULES = ['reptile.spiders']
NEWSPIDER_MODULE = 'reptile.spiders'

# Crawl responsibly by identifying yourself (and your website) on the user-agent
USER_AGENT = 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) \
' Chrome/108.0.0.0 Safari/537.36 Edg/108.0.1462.42'

# Obey robots.txt rules
ROBOTSTXT_OBEY = False

LOG_LEVEL = 'ERROR'
```

为了让终端的输出简洁，我们仅允许打印错误日志

```
LOG_LEVEL = 'ERROR'
```

3. 修改设置文件 items.py

```
class ReptileItem(scrapy.Item):
    # define the fields for your item here like:
    title = scrapy.Field()
    url = scrapy.Field()
    content = scrapy.Field()
```

这里我们爬取的数据有标题、详情页 url 和原文内容，所以我们利用 scrapy.Field 方法实例化三个对象用于后续填入获取的变量

4. 修改设置文件 spiders.py

```
class SpidersSpider(scrapy.Spider):
    name = 'spiders'
    allowed_domains = ['hongloumeng.5000yan.com']
    start_urls = ['http://hongloumeng.5000yan.com/']

    def parse(self, response):
        li_list = response.xpath('//div[@class="sidamingzhu-list-mulu"]/ul/li')
        all_data = []
        for li in li_list:
            title = li.xpath('./a/text()')[0].extract()
            url = li.xpath('./a/@href')[0].get()
            item = ReptileItem()
            item['title'] = title.replace(u'\xa0', ' ') + '\n'
            item['url'] = url
            yield scrapy.Request(url=url, meta={'item': item}, callback=self.parse_detail)

    def parse_detail(self, response):
        item = response.meta['item']
        content = ''
        div_list = response.xpath('//div[@class="grap"]/div')
        for div in div_list:
            content += div.xpath('./text()')[0].extract().strip() + '\n'
        item['content'] = content.replace(u'\xa0', ' ') + '\n\n'
        print(item['title'].strip() + ' 已完成')
        yield item
```

在这个文件里我们填入需要爬取网站的 url，之后进行数据解析，利用 xpath 进行定位并获取相关的数据，之后实例化一个 item 对象，将需要爬取的数据写入 item 之中，这里为了能够爬取详细页面的内容，我们在 parse 方法中嵌套另一个 scrapy.request 方法，通过 meta 传递 item 的值到 parse_detail 方法之中，同理进行数据解析与获取存入 item 对象，最后返回 item 对象

5. 修改设置文件 pipelines.py

pipelines 会接收到刚刚返回 item 对象，在这个文件内我们进行数据的持久化存储，通过新建需要的文本文件，将爬取的数据写入其中，完成对数据的爬取与存储

```

class ReptilePipeline:
    fp1 = None
    fp2 = None

    def open_spider(self, spider):
        print('*****开始爬虫程序*****')
        self.fp1 = open('./1.txt', 'w', encoding='utf-8')
        self.fp2 = open('./2.txt', 'w', encoding='utf-8')

    def process_item(self, item, spider):
        title = item['title']
        content = item['content']
        self.fp1.write(title)
        self.fp2.write(title)
        self.fp2.write(content)

        return item

    def close_spider(self, spider):
        print('*****爬虫结束*****')
        self.fp1.close()
        self.fp2.close()

```

6. 运行程序，获取爬取的数据

```

D:\Desktop\陈新 > cd 任务2
D:\Desktop\陈新\任务2> 

```

我们先输入 cd 指令进入对应的工程文件夹，之后调用 crawl 指令运行爬虫程序
获得输出

```

scrapy crawl spiders

```

(终端输出的日志信息)

```
*****开始爬虫程序*****
第十八回  林黛玉误剪香袋囊  贾元春归省庆元宵  已完成
第二十一回  贤袭人娇嗔箴宝玉  俏平儿软语救贾琏  已完成
第二十回  王熙凤正言弹妒意  林黛玉俏语谑娇音  已完成
第二十三回  西厢记妙词通戏语  牡丹亭艳曲警芳心  已完成

第一回  甄士隐梦幻识通灵  贾雨村风尘怀闺秀  已完成
第六回  贾宝玉初试云雨情  刘姥姥一进荣国府  已完成
第八回  薛宝钗小恙梨香院  贾宝玉大醉绛芸轩  已完成

*****爬虫结束*****
```

(输出的 1.txt)

```
任务一.py × spiders.py × 1.txt × 2.txt × __init__.py × settings.py × items.py × pipelines.py ×
1 第十八回  林黛玉误剪香袋囊  贾元春归省庆元宵
2 第二十一回  贤袭人娇嗔箴宝玉  俏平儿软语救贾琏
3 第二十回  王熙凤正言弹妒意  林黛玉俏语谑娇音
4 第二十三回  西厢记妙词通戏语  牡丹亭艳曲警芳心
5 第十七回  大观园试才题对额  怡红院迷路探曲折
6 第十九回  情切切良宵花解语  意绵绵静日玉生香
7 第二十二回  听曲文宝玉悟禅机  制灯迷贾政悲谶语
8 第十五回  王凤姐弄权铁槛寺  秦鲸卿得趣馒头庵
9 第十三回  秦可卿死封龙禁尉  王熙凤协理宁国府
10 第十二回  王熙凤毒设相思局  贾天祥正照风月鉴
11 第九回  恋风流情友入家塾  起嫌疑顽童闹学堂
12 第十四回  林如海捐馆扬州城  贾宝玉路谒北静王
13 第十一回  庆寿辰宁府排家宴  见熙凤贾瑞起淫心
14 第十六回  贾元春才选凤藻宫  秦鲸卿夭逝黄泉路
15 第十回  金寡妇贪利权受辱  张太医论病细穷源
16 第一百二十回甄士隐详说太虚情  贾雨村归结红楼梦
17 第二十四回  醉金刚轻财尚义侠  痴女儿遗帕惹相思
18 第一百十九回中乡魁宝玉却尘缘  沐皇恩贾家延世泽
```

(输出的 2.txt)

```
1 第十三回  秦可卿死封龙禁尉  王熙凤协理宁国府
2 话说凤姐儿自贾琏送黛玉往扬州去后，心中实在无趣。每到晚间，不过和平儿说笑一回，就胡乱睡了。
3
4 这日夜间，正和平儿灯下拥炉倦绣，早命浓熏绣被，二人睡下，屈指算行程该到何处，不知不觉已交三鼓。平儿已睡熟了。凤姐方觉星眼微朦，恍惚只见秦
5
6 凤姐听了，恍惚问道：“有何心愿？你只管托我就是了。”秦氏道：“婶婶，你是个脂粉队里的英雄，连那些束带顶冠的男子也不能过你，你如何连两句俗语
7
8 凤姐便问何事。秦氏道：“目今祖莹虽四时祭祀，只是无一定的钱粮；第二，家塾虽立，无一定的供给。依我想来，如今盛时固不缺祭祀、供给，但将来败
9
10 彼时合家皆知，无不纳罕，都有些疑心。那长一辈的想她素日孝顺，平一辈的想她素日和睦亲密，下一辈的想她素日慈爱，以及家中仆从老小想她素日怜
11
12 闲言少叙，却说宝玉因近日林黛玉回去，剩得自己孤凄，也不和人顽耍，每到晚间，便索然睡了。如今从梦中听见说秦氏死了，连忙翻身爬起来，只觉心中
13
14 一直到了宁国府前，只见府门洞开，两边灯笼照如白昼，乱烘烘人山人海，里面哭声撼山振岳。宝玉下了车，忙奔至停灵之室，痛哭一番。然后见过尤氏
15
16 正说着，只见秦业、秦钟并尤氏的几个眷属、尤氏姊妹也都来了。贾珍便命贾琏、贾琛、贾璉、贾蔷四个人去陪客，一面吩咐去请钦天监阴阳司来择日，择
17
```

问题思考：

利用 scrapy 爬取的数据中明显存在顺序混乱，爬取章节顺序、所打印出来的标题顺序和原文顺序都发生错乱，且每次运行的时候结果都存在不同

原因：查阅资料我们知道 scrapy 框架在爬取数据时为提高运行速度采用了异步处理，也就是说 Scrapy 发送请求之后，不会等待这个请求的响应（也就是不会阻塞），而是可以同时发送其他请求或者做别的事情。而我们知道服务器对于请求的响应是由很多方面的因素影响的，如网络速度、解析速度、资源抢占等等，其响应的顺序是难以预测的。所以导致了我们的对于每个章节完成的时间是不同的，最后输出的顺序也发生不同。