基于 requests 和 BeautifulSoup 的爬虫

22920212204359 陈新

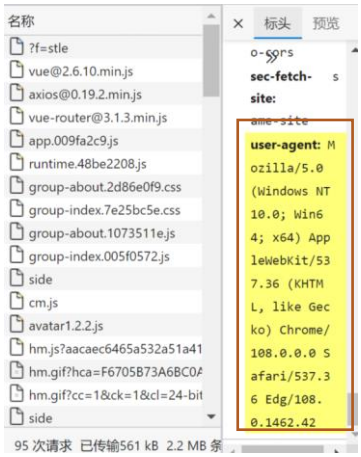
采集网页的地址：https://hongloumeng.5000yan.com/
网页描述格式：HTML 格式
采集数据：获取红楼梦每个章节的标题，保存到指定的文本文件 1 中；之后获取各章节的原文，保存到文本文件 2 中

主要步骤；

1.导入对应的包，并且利用 requests.get 打开 url 获取对应的响应数据

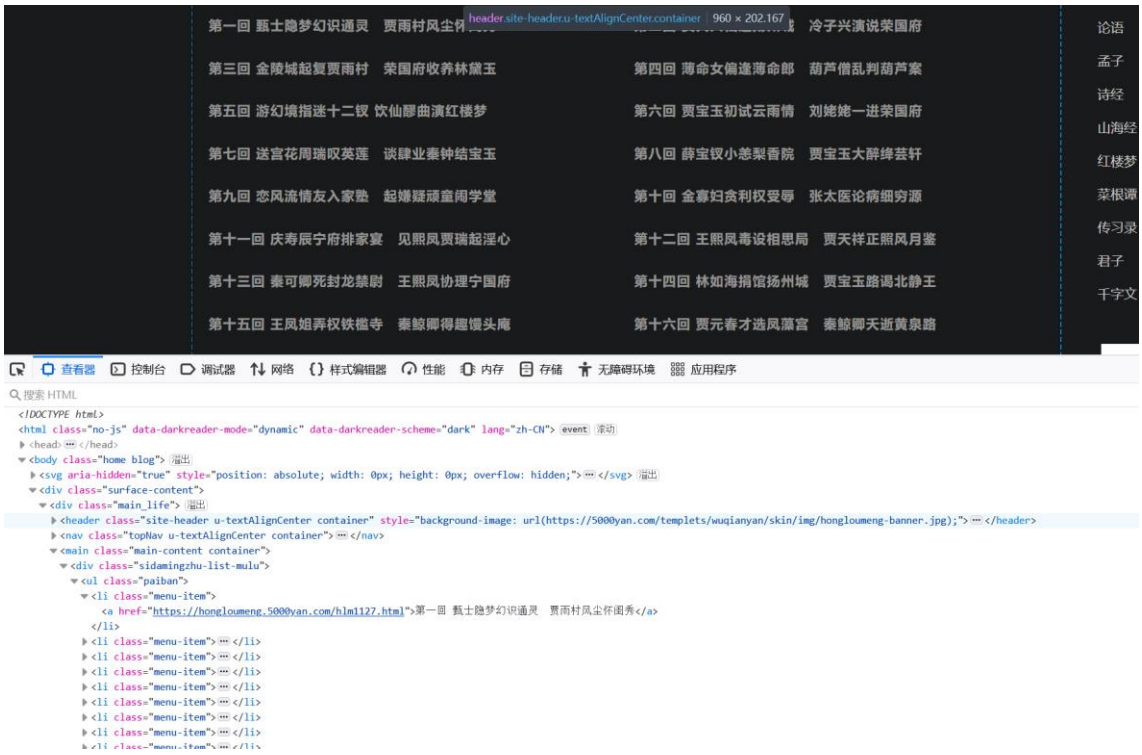
```
def request_page(url):
    page = requests.get(url=url, headers=headers)
    page.encoding = 'utf-8'
    return page.text
```

2.打开对应网站，右键检查打开抓包工具，先在网络->XHR 中找到自己的请求头并作为全局变量保存在代码中方便后续使用

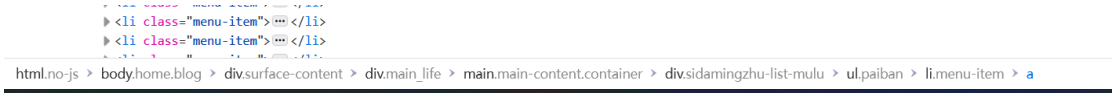



```
headers = {
    'user-agent': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/108.'
                  '0.0.0 Safari/537.36 Edg/108.0.1462.42'
}
```

之后定位到标题所在的层级



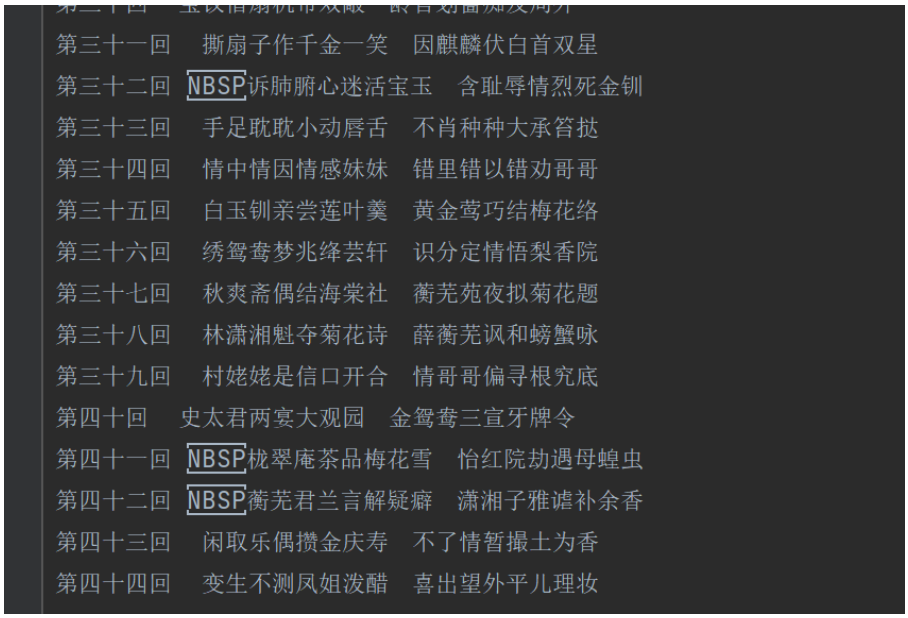
于是定位到章节名称所在的层级依次为



3.实例化一个 BeautifulSoup 对象，这里使用 lxml 解析器，通过刚刚获取的层级关系定位到需要爬取的信息，在这里我们需要获取一个章节信息和章节详细页面的 url

```
def parse_page(page):
    soup = BeautifulSoup(page, 'lxml')
    data_list = soup.select('.sidamingzhu-list-mulu>ul>li')
    for data in data_list:
        title = data.a.string
        title_url = data.a['href']
        yw_page = yw_page_requests(title_url)
        yw_page_parse(yw_page,title)
        print(title)
        fp1.write(title.replace(u'\xa0', '') + '\n')
```

在这里由于编码格式问题，我们需要对 title 进行一些处理，需要将'x\a0'替换为‘ ’，否则代码运行时输出的文本会出现如下情况



经查阅资料得知这个 nbsp 是 unicode 字符集当中的空白符，文本文件中以 utf-8 编码输出，所以导致出现显示异常，我们利用 python 内置方法 replace 进行替换即可

成功爬取标题后我们进行数据的持久化存储，将其存储于 1 文本文件内，完成对标题的爬取和记录

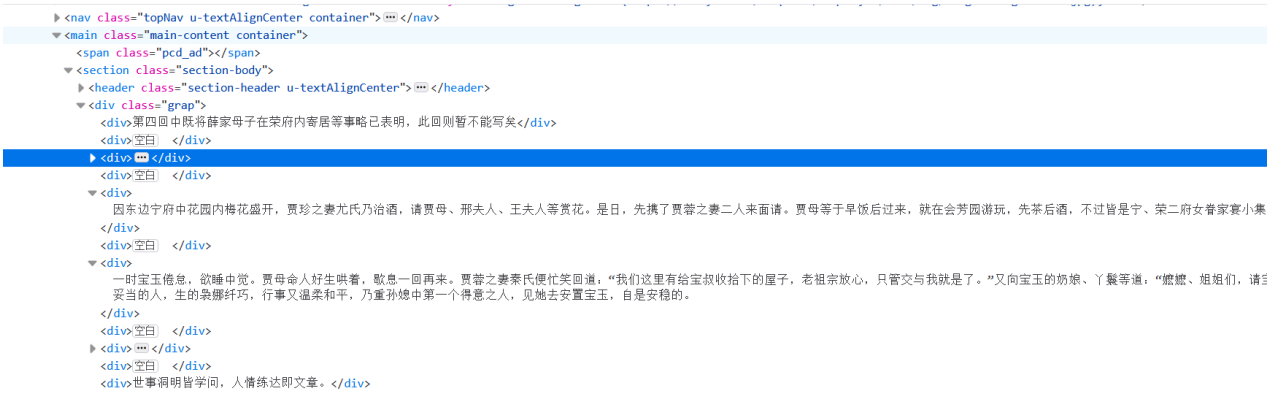
另外我们注意到 a 标签内的 href 即为详细页面的 url，因此我们直接获取这个 url 之后进行爬取详细页面的信息

4.详细页面处理

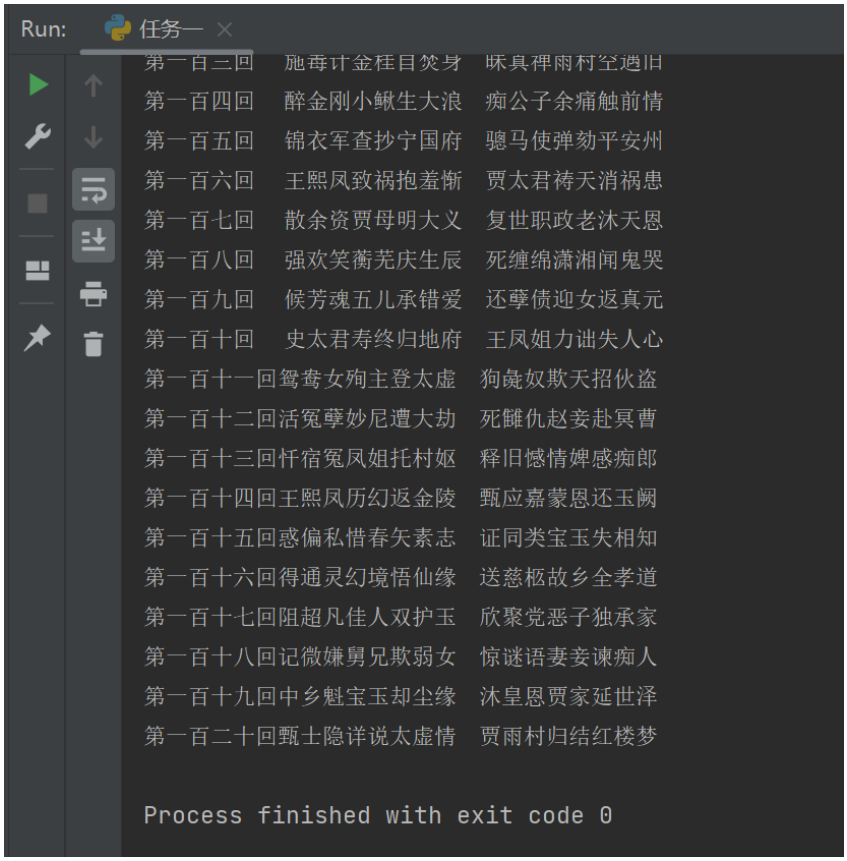
详细页面的处理也是同理，在数据解析时我们先定位到 div.grap，之后获取其下所有的 div 列表

```
29
30      def yw_page_requests(url):
31          yw_page = requests.get(url=url, headers=headers)
32          yw_page.encoding = yw_page.apparent_encoding
33          return yw_page.text
34
35
36      def yw_page_parse(page,title):
37          fp2.write(title+'\n')
38          soup2 = BeautifulSoup(page, 'lxml')
39          content = soup2.find('div', class_='grap')
40          cons = content.find_all('div')
41          for con in cons:
42              text = con.text.replace(u'\xa0', '').strip()
43              fp2.write(text)
44          fp2.write('\n\n')#分隔用
45
```

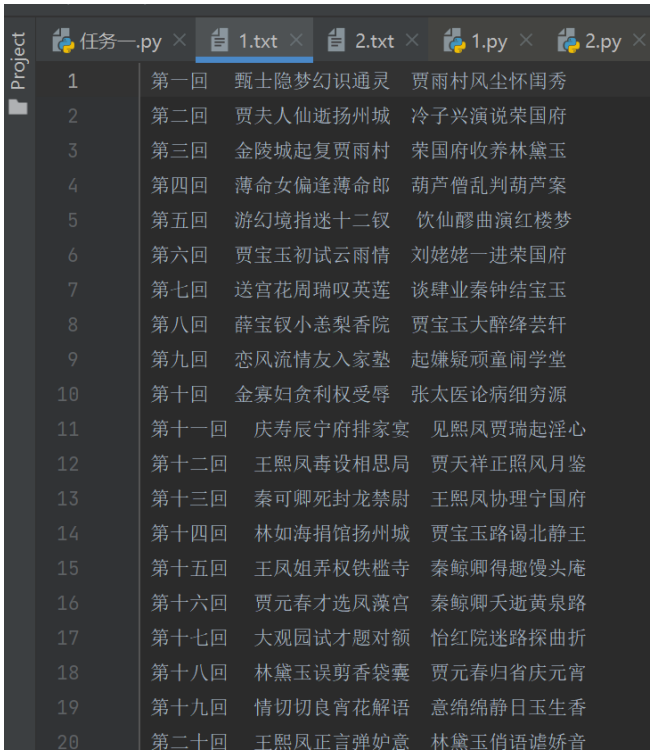
观察层次结构知道我们需要爬取的信息位于 div 标签内，我们遍历刚刚获取的列表，逐个向文本 2 写入内容即可。



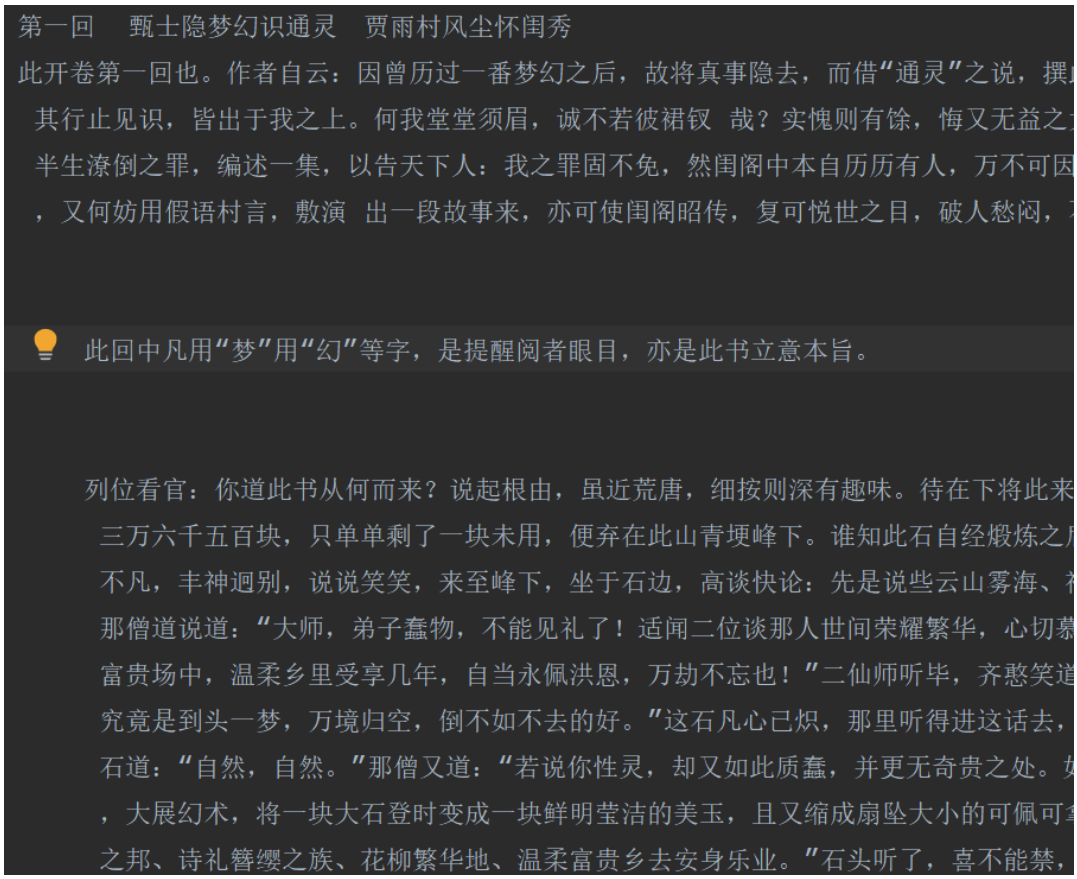
运行 python 程序，完成对数据的爬取
（日志打印）



（打印的 1.txt 文本，即我们所爬取的标题）



（打印 2.txt 文本，是我们所爬取的原文内容）



第一回　甄士隐梦幻识通灵　贾雨村风尘怀闺秀

此开卷第一回也。作者自云：因曾历过一番梦幻之后，故将真事隐去，而借“通灵”之说，撰
其行止见识，皆出于我之上。何我堂堂须眉，诚不若彼裙钗　哉？实愧则有馀，悔又无益之
半生潦倒之罪，编述一集，以告天下人：我之罪固不免，然闺阁中本自历历有人，万不可因
，又何妨用假语村言，敷演　出一段故事来，亦可使闺阁昭传，复可悦世之目，破人愁闷，

此回中凡用“梦”用“幻”等字，是提醒阅者眼目，亦是此书立意本旨。

列位看官：你道此书从何而来？说起根由，虽近荒唐，细按则深有趣味。待在下将此来
三万六千五百块，只单单剩了一块未用，便弃在此山青埂峰下。谁知此石自经煅炼之
不凡，丰神迥别，说说笑笑，来至峰下，坐于石边，高谈快论：先是说些云山雾海、
那僧道说道：“大师，弟子蠢物，不能见礼了！适闻二位谈那人世间荣耀繁华，心切慕
富贵场中，温柔乡里受享几年，自当永佩洪恩，万劫不忘也！”二仙师听毕，齐憨笑道
究竟是到头一梦，万境归空，倒不如不去的好。”这石凡心已炽，那里听得进这话去，
石道：“自然，自然。”那僧又道：“若说你性灵，却又如此质蠢，并更无奇贵之处。
，大展幻术，将一块大石登时变成一块鲜明莹洁的美玉，且又缩成扇坠大小的可佩可
之邦、诗礼簪缨之族、花柳繁华地、温柔富贵乡去安身乐业。”石头听了，喜不能禁，