

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

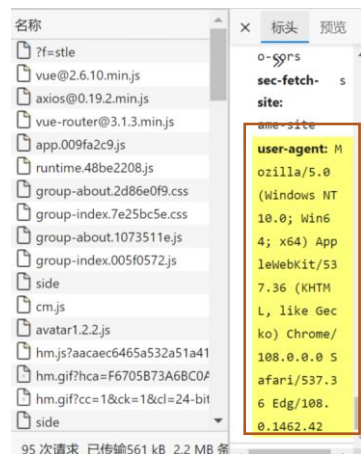
采集数据: 获取红楼梦每个章节的标题, 保存到指定的文本文件 1 中; 之后获取各章节的原文, 保存到文本文件 2 中

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\xa0'替换为' '，否则代码运行时输出的文本会出现如下情况

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

```
28
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。

```
<nav class="topNav u-textAlignCenter container">...</nav>
<main class="main-content container">
  <span class="pcd_ad"></span>
  <section class="section-body">
    <header class="section-header u-textAlignCenter">...</header>
    <div class="grap">
      <div>第四回中既将薛家母子在荣府内寄居等事略已表明，此回则暂不能写矣</div>
      <div>空白</div>
    </div>
    <div>空白</div>
    <div>
      因东边宁府中花园内梅花盛开，贾珍之妻尤氏乃治酒，请贾母、邢夫人、王夫人等赏花。是日，先挑了贾蓉之妻二人来面请。贾母等于早饭后过来，就在会芳园游玩，先茶后酒，不过皆是宁、荣二府女眷家宴小集
    </div>
    <div>空白</div>
    <div>
      一时宝玉倦怠，欲睡中觉。贾母命人好生哄着，歇息一回再来。贾蓉之妻秦氏便忙笑道：“我们这里有给宝叔收拾下的屋子，老祖宗放心，只管交与我就是了。”又向宝玉的奶娘、丫鬟等道：“嬷嬷、姐姐们，请宝
      妥当的人，生的袅娜纤巧，行事又温柔和平，乃重孙媳中第一个得意之人，见他去安置宝玉，自是安稳的。
    </div>
    <div>空白</div>
    <div>...</div>
    <div>空白</div>
    <div>世事洞明皆学问，人情练达即文章。</div>
```

运行 python 程序，完成对数据的爬取
(日志打印)

Run: 任务一 ×

▶

⬆

⚙

■

⌵

🔍

⬆

⬇

⏮

⏪

⏩

⏭

🗑

第一百二回	施毒计金桂自焚身	昧真禪雨村空遇旧
第一百四回	醉金刚小鳊生大浪	痴公子余痛触前情
第一百五回	锦衣军查抄宁国府	驢马使弹劾平安州
第一百六回	王熙凤致祸抱羞慚	贾太君禱天消祸患
第一百七回	散余资贾母明大义	复世职政老沐天恩
第一百八回	强欢笑蘅芜庆生辰	死缠绵潇湘闻鬼哭
第一百九回	候芳魂五儿承错爱	还孽债迎女返真元
第一百十回	史太君寿终归地府	王凤姐力诤失人心
第一百十一回	鸳鸯女殉主登太虚	狗彘奴欺天招伙盗
第一百十二回	活冤孽妙尼遭大劫	死讎仇赵妾赴冥曹
第一百十三回	怀宿冤凤姐托村姬	释旧憾情婢感痴郎
第一百十四回	王熙凤历幻返金陵	甄应嘉蒙恩还玉阙
第一百十五回	惑偏私惜春矢素志	证同类宝玉失相知
第一百十六回	得通灵幻境悟仙缘	送慈柩故乡全孝道
第一百十七回	阻超凡佳人双护玉	欣聚党恶子独承家
第一百十八回	记微嫌舅兄欺弱女	惊谜语妻妾谏痴人
第一百十九回	中乡魁宝玉却尘缘	沐皇恩贾家延世泽
第一百二十回	甄士隐详说太虚情	贾雨村归结红楼梦

Process finished with exit code 0

(打印的 1.txt 文本，即我们所爬取的标题)

Project 任务一.py × 1.txt × 2.txt × 1.py × 2.py ×

Project

1

2

3

4

5

6

7

8

9

10

11

12

13

14

15

16

17

18

19

20

第一回

甄士隐梦幻识通灵

贾雨村风尘怀闺秀

第二回

贾夫人仙逝扬州城

冷子兴演说荣国府

第三回

金陵城起复贾雨村

荣国府收养林黛玉

第四回

薄命女偏逢薄命郎

葫芦僧乱判葫芦案

第五回

游幻境指迷十二钗

饮仙醪曲演红楼梦

第六回

贾宝玉初试云雨情

刘姥姥一进荣国府

第七回

送宫花周瑞叹英莲

谈肄业秦钟结宝玉

第八回

薛宝钗小恙梨香院

贾宝玉大醉绛芸轩

第九回

恋风流情友入家塾

起嫌疑顽童闹学堂

第十回

金寡妇贪利权受辱

张太医论病细穷源

第十一回

庆寿辰宁府排家宴

见熙凤贾瑞起淫心

第十二回

王熙凤毒设相思局

贾天祥正照风月鉴

第十三回

秦可卿死封龙禁尉

王熙凤协理宁国府

第十四回

林如海捐馆扬州城

贾宝玉路谒北静王

第十五回

王凤姐弄权铁槛寺

秦鲸卿得趣馒头庵

第十六回

贾元春才选凤藻宫

秦鲸卿天逝黄泉路

第十七回

大观园试才题对额

怡红院迷路探曲折

第十八回

林黛玉误剪香袋囊

贾元春归省庆元宵

第十九回

情切切良宵花解语

意绵绵静日玉生香

第二十回

王熙凤正言弹妒意

林黛玉俏语谑娇音

(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

此回中凡用“梦”用“幻”等字，是提醒阅者眼目，亦是此书立意本旨。