



3.4 HTML解析与Python实现

3.4.1 HTML与CSS要素

3.4.2 BeautifulSoup库及对象

3.4.3 BeautifulSoup库遍历文档树

3.4.4 BeautifulSoup库搜索文档树

3.4.5 BeautifulSoup库查找CSS过滤器

3.4.6 Python解析网页实例



3.4.6 Python解析网页实例

- 大多数情况下，我们使用聚焦爬虫，爬取页面中指定部分的数据值，而不是整个页面的数据。因此，在聚焦爬虫中使用数据解析。
- 数据爬取的流程：
 - 指定url
 - 基于requests模块发起请求
 - 获取响应中的数据
 - 数据解析：解析标签之间或者标签对应的属性中的数据
 - 进行持久化存储

■ 例1.爬取影视网站“最新上线”电影信息简介（2021）

<https://blog.csdn.net/dick633/article/details/79638336>

■ 2345影城网址：<http://dianying.2345.com/>（20210430网站主页）

最新上线

更多>



双面妖姬

8.7

童璇三世虐恋终成正果



鲁班四杰之伏龙海眼

8.8

奇门异士护墓守宝斩龙伏妖



人潮汹涌

9.2

刘德华肖央身份爆笑互换



熊出没狂野大陆

7.5

狂野冒险现在开启!



九叔之夜行病魔

8.8

九叔归来! 斩僵尸、除妖魔



蛇王2021

盗猎队丛林激战



2021

9.2

2021



鲁班四杰之伏龙海眼

8.8

奇门异士护墓守宝斩龙伏妖



人潮汹涌

9.2

刘德华肖央身份爆笑互换



黄飞鸿

8.4

黄飞鸿之英雄林世荣



九叔之夜行病魔

8.8

九叔归来! 斩僵尸、除妖魔



蛇王2021

盗猎队丛林激战

- 1.查看“最新上线”对应代码（202104网站主页源码）（页面选中“最新上线”，点击右键菜单“检查”）
- 2.查找对应标签（可借助Ctrl+F查找）

<!-- 最新上线 -->

```
▼<div id="sideNavRowDy2" class="row sideNavRow row-dongman row-line2 clearfix" data-ajax25wrap="栏目" data-ajax25location="最新  
上线">
```

▼<div class="v mod">

```
▶ <div class="v th">...</div>
```

▼<div class="v tb">

▼<div class="tab-plugin-con">

▼<div class="con">

▼<div class="v picConBox v picConBoxNoDesc">

▼<ul class="hoverPopTarget clearfix">


▼<li data-hover="{\"hover_actor\":[\"\\u8463\\u74b7\", \"\\u5c01\\u5c01\\u737d\", \"\\u90d1\\u67f3\"], \"hover_type\":

"\u5947\u5e7b,\u7231\u60c5", "hover intro":

"\u79c0\u574a\u505a\u5de5\u7684\u5973\u5b69\u6625\u5c71\u5728\u4e34\u8fd118\u5c81\u65f6\u7ecf\u5386\u4e86\u6570\u6b

▼<div class="pic">

图标签: `pic`, `img`, `src`, `data-src`



► `<div class="season">...</div>`

0\u82e6\u82e6\u7b49\u5f85\u2026.", "hover_year": "2021", "dataForm": "\u7535\u5f71" "dataTab":

"\u5185\u5730|\u5947\u5e7b-\u7231\u60c5|2021","dataSource":"qq","id":"210501

"\u53cc\u9762\u5996\u59ec", "pic": "\\vingshi-stream.2345cdn.net/dvpcimg/i

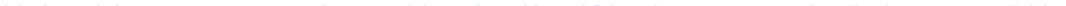
```
"8.7","url":"\\dianying.2345.com\\detail\\210501.html"}" data-form="电影"
```

```
source="qq">
```

▼<a class="v picTxt" href="//dianying.2345.com/detail/210501.html" target=" blank" data-ajax25="双面妖姬" data-

```
ajax25index="1" data-ajax25form="电影" data-ajax25tab="内地|奇幻-爱情|2021" data-ajax25source="qq">
```

▼<div class="pic">



```
▶ <div class="season">...</div>
```

信息太简单
无法满足我们的需求

3.寻找我们需要的电影简介以及包含它们的标签2021

每部电影超链接都在这个标签下，
可以遍历得到每部电影的超链接

```
<div class="con" style="display: none;">
  <div class="pic">
    <a href="//dianying.2345.com/detail/210501.html" data-ajax25="双面妖姬" data-ajax25index="3" data-ajax25location="轮播海报" data-ajax25form="电影" target="_blank" data-ajax83="ys_dy_index_jdt_3">...</a> == $0
  </div>
</div>
```

超链接，可以跳转到电影信息页面

▲ 不安全 | dianying.2345.com/detail/210501.html

5影视 首页 电影 电视剧 动漫 综艺 玩游戏 美女秀场 锦心似玉

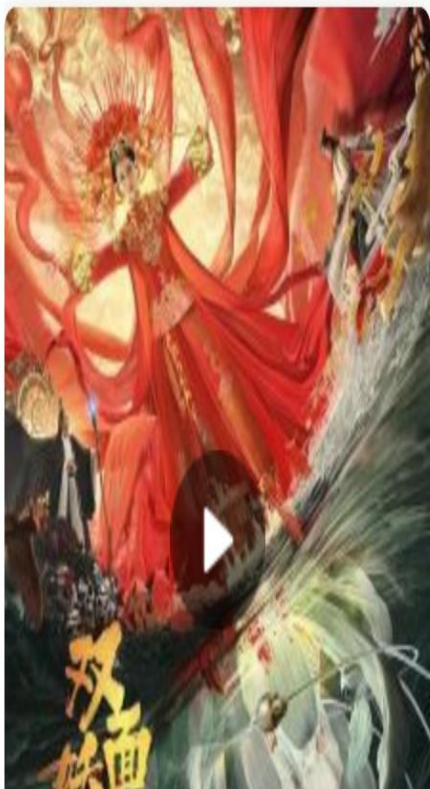


双面妖姬 8.9分

主演：董璇 小小白 郑柳 导演：赵世尧 类型：奇幻 爱情

简介：秀坊做工的女孩春山在临近18岁时经历了数次妖怪袭击后偶遇了捉妖师沉玄，沉玄送给春山一个铃铛，每当遇到危机春山便摇动铃铛，几经危机，两人感情日渐深厚，直到春山18岁，她变成了一个半人半妖的“怪物”，夜晚会被妖性控制无意识杀人。原来落霞镇有一个祭祀的秘密，祭师为了平复河神每隔三年便杀掉活人封印血气到春山体。沉玄几经波折找到了仪式的根源，拼死搏斗后为救春山献出生命变成渡桥，春山在两人约定之地苦苦等待.... [收起全部](#)

3.4.6 Python解析网页实例



双面妖姬 8.9分

主演：董璇 小小白 郑柳

导演：赵世尧

类型：奇幻 爱情

简介：秀坊做工的女孩春山在临近18岁时经历了数次妖怪袭击后偶遇了捉妖师沉玄，沉玄送给春山一个铃铛，每当遇到危机春山便摇动铃铛，几经危机，两人感情日渐深厚，直到春山18岁，她变成了一个半人半妖的“怪物”，夜晚会被妖性控制无意识杀人。原来落霞镇有一个祭祀的秘密，祭师为了平复河神每隔三年便杀掉活人封印血气到春山体。沉玄几经波折找到了仪式的根源，拼死搏斗后为救春山献出生命变成渡桥，春山在两人约定之地苦苦等待.... [收起全部](#) ^

- 点击简介页源码，展开代码，辅以Ctrl+F，查找相应代码和标签

■ 展开代码：所需标签和信息2021

```
<!-- 海报 -->
<!-- 介绍 -->
<div class="txtIntroCon">
  <div class="tit">
    <h1>双面妖姬</h1>
    <p class="pTxt">
      <em class="emScore">8.9分</em>
    </p>
  </div>
  <div class="wholeTxt">
```

电影名

电影信息

```
    <ul class="txtList clearfix">
      <li class="liActor li_3">
        <em class="emTit">主演: </em>
        <a data-ajax83="ys_dy_2015_detail_zhy_1" title="董璇" href="//www.baidu.com/s?word=董璇
        &tn=25017023_3_pg" target="_blank" rel="nofollow">董璇</a>
        <a data-ajax83="ys_dy_2015_detail_zhy_2" title="小小白" href="//www.baidu.com/s?word=小小白
        &tn=25017023_3_pg" target="_blank" rel="nofollow">小小白</a>
        <a data-ajax83="ys_dy_2015_detail_zhy_3" title="郑柳" href="//www.baidu.com/s?word=郑柳
        &tn=25017023_3_pg" target="_blank" rel="nofollow">郑柳</a>
      </li>
      <li class="li_3">
        <em class="emTit">导演: </em>
        <a data-ajax83="ys_dy_2015_detail_daoy" title="赵世尧" href="//www.baidu.com/s?word=赵世尧
        &tn=25017023_3_pg" target="_blank" rel="nofollow">赵世尧</a>
      </li>
      <li class="li_3">
        <em class="emTit">类型: </em>
        <a title="奇幻电影" data-ajax83="ys_dy_2015_detail_leix_1" href="/list/qihuan-----.html"
        target="_blank">奇幻</a>
        <a title="爱情电影" data-ajax83="ys_dy_2015_detail_leix_2" href="/list/aiqing-----.html"
        target="_blank">爱情</a>
      </li>
    </ul>
    <ul class="txtList clearfix newIntro">
      <li class="extend">
        <em class="emTit">简介: </em>
        <p class="pIntro pHide" style="display: none;">...</p>
        <p class="pIntro pShow" style="display: block;">
          <span> == $0
```

"秀坊做工的女孩春山在临近18岁时经历了数次妖怪袭击后偶遇了捉妖师沉玄，沉玄送给春山一个铃铛，每当遇到危机春山便摇动铃铛，几经危机，两人感情日渐深厚，直到春山18岁，她变成了一个半人半妖的“怪物”，夜晚会被妖性控制无意识杀人。原来落霞镇有一个祭祀的秘密，祭师为了平复河神每隔三年便杀掉活人封印血气到春山体内存。沉玄几经波折找到了仪式的根源，拼死搏斗后为救春山献出生命变成渡桥，春山在两人约定之地苦苦等待...

- 编写代码bs_3.4.6_2021.py
- 功能：1.下载海报图片，存文件；2.提取电影信息，打印输出

```
from bs4 import BeautifulSoup as bs # as起到改名作用，以便后面书写
import os
import requests
import urllib
```

函数：提取电影信息，
打印电影信息

```
def turn(newhtml):
```

解决乱码

```
    page = requests.get(newhtml)
    html = page.content.decode('gbk')
    soup = bs(html)
```

BeautifulSoup解析

提取电影名，见前面分析标签过程

```
# 获取电影信息
```

```
text = soup.select('.tit') # css选择用法： '.' 代表class， '#' 代表id
```

```
print(text[0].text)
```

提取电影信息，见前面分析标签过程

```
div2 = soup.select('div.wholeTxt')
```

```
for em in div2[0].find_all('li'):
```

提取li列表

```
    text1 = ''.join(em.text.strip().split())
```

```
    print(text1)
```

连接字符串数组

主程序

```
i = 0
src = 'https://dianying.2345.com/' # 爬取入口: 网站首页url
page = requests.get(src)
html = page.content.decode('gbk')

soup = bs(html, 'html.parser')
uls = soup.find_all('ul', class_='hoverPopTarget clearfix')
# print(div)
# 保存图片
for li in uls[0].find_all('div', class_='pic'):
    a = li.find('img')
    a = a.get('data-src')
    inforhtml = 'https:' + a #生成电影信息页面的url
    if not os.path.exists('img'):
        os.mkdir('img')
    filename = 'img/image' + str(i) + '.jpg'
    urllib.request.urlretrieve(inforhtml, filename)
    i +=1
# 显示电影信息
for li in uls[0].find_all('a', class_='v_picTxt'):
    a = li.get('href')
    newhtml = 'https:' + a
    turn(newhtml)
```

网站首页源码

爬取入口: 网站首页url

查找最新上线树根

查找电影图片类,
pic、img、data-
src标签内容

判断img文件夹是否存在, 在当前目录
创建img文件夹, 用于存放图片

构建图片文件名

下载指定url内容到本地

什么标签?

为您推荐



鲁班四杰之伏龙海眼
王润泽 宁心



如果声音不记得
章若楠 孙晨竣

p.pRightBottom 38.97 × 20

8.9分

9.3分

```
<div class="col_a" id="dyDetail_uniqFlow">
  <!--为您推荐-->
  <div class="v_mod v_recommend mt10">
    <div class="v_th">...</div>
    <div class="v_tb">
      <div class="v_picConBox">
        <ul class="v_picTxt pic167_223 clearfix">
          <li>
            <div class="pic">
              
            <p class="pRightBottom">
              <em>8.9分</em>
            </p>
            <a class="aPlayBtn" href="//dianying.2345.com/detail/210481.html" target="_blank" title="鲁班四杰之伏龙海眼" data-ajax83="ys_dy_2015_detail_cnxh_cnxh_1">
              <i>...</i>
            </a>
          </li>
        </ul>
      </div>
    </div>
  </div>
</div>
```

V_picTxt标签

```
<div class="col_a" id="dyDetail_uniqFlow">
  <!--为您推荐 start-->
  <div class="v_mod v_recommend mt10">
    <div class="v_th">...</div>
    <div class="v_tb">
      <div class="v_picConBox">
        <ul class="v_picTxt pic167_223 clearfix">
          <li>
            <div class="pic">
              
            <p class="pRightBottom">
              <em>8.9分</em>
            </p>
            <a class="aPlayBtn" href="//dianying.2345.com/detail/210481.html" target="_blank" title="鲁班四杰之伏龙海眼" data-ajax83="ys_dy_2015_detail_cnxh_cnxh_1">
              <i>...</i>
            </a>
          </li>
        </ul>
      </div>
    </div>
  </div>
</div>
```

■ urllib.request.urlretrieve()函数

https://blog.csdn.net/pursuit_zhangyu/article/details/80556275

urllib.request.urlretrieve

(url, filename=None, reporthook=None, data=None)

■ 功能：

- 将URL表示的网络对象复制到本地文件；如果URL指向本地文件，则对象将不会被复制，除非提供文件名。

■ 参数

- url: 外部或者本地url
- filename: 保存到本地的路径（如果未指定，urllib会生成一个临时文件来保存数据）；
- reporthook: 一个回调函数，当连接上服务器、以及相应的数据块传输完毕的时候会触发该回调。我们可以利用这个回调函数来显示当前的下载进度。
- data: 指post到服务器的数据
- 返回值：一个包含两个元素的元组(filename, headers)
 - Filename: 保存到本地的路径
 - Header: 服务器的响应头

双面妖姬

8.9分

人潮汹涌

主演:

导演:

9.3分

类型:

简介:

主演: 刘德华 肖央 万茜 程怡

送给

山18岁

一个祭

临近18

机的春

的“怪

河神每

后为救

鲁班四

8.9分

主演:

导演:

类型:

简介:

熊出没·狂野大陆

熊出没·狂野大陆

7.5分

主演: 熊大 熊二 光头强

导演: 丁亮

九叔之夜行疯魔

8.9分

主演: 刘铁柱 田沛菲 曹旭 雷琨

导演: 姚锐

类型: 动作 喜剧 爱情

简介: 故事发生在一个古镇，往日里平静的小镇突然惊现诡异事件！九叔白天经营一家药房，晚上给枉死之人送葬。小镇近来时常半夜出现惨叫，有人离奇死亡都有被撕咬痕迹疑是神秘生物所为。警队频繁接到报案但都一无所获。九叔觉得事情蹊跷，便自己暗中调查。之后又连续发生多起命案，发现死者的死法大径相同... 展开全部 故事发生在一个古镇，往日里平静的小镇突然惊现诡异事件！九叔白天经营一家药房，晚上给枉死之人送葬。小镇近来时常半夜出现惨叫，有人离奇死亡都有被撕咬痕迹疑是神秘生物所为。警队频繁接到报案但都一无所获。九叔觉得事情蹊跷，便自己暗中调查。之后又连续发生多起命案，发现死者的死法大径相同。九叔为钱老爷做法送葬。和徒弟们说万一处理不好会造成无法想象的后果。徒弟阿武和阿果晚上回去无意中偷看到师妹无双正在洗澡被师父九叔发现动静，他们俩急忙逃跑半路遇见一群“僵尸”。九叔闻风赶来。展开一场搏斗。九叔根据每次案发现场的一丝丝线索顺藤摸瓜找到了凶手的故事，原来比鬼神更可怕的是…… 收起全部

3.4.6 Python解析网页实例

处理2021-笔记本 > python代码 > img

存在文件中的电影海报



image0.jpg



image1.jpg



image2.jpg



image3.jpg



image4.jpg



image5.jpg



image6.jpg



image7.jpg



image8.jpg



image9.jpg



image10.jpg



image11.jpg



image12.jpg



image13.jpg

例2.将“诗词名句网站” (<http://www.shicimingju.com/book/sanguoyanyi.html>) 中三国演义小说的每一章的内容爬取到本地磁盘进行存储 (Python爬虫 | BeautifulSoup解析html页面 <https://www.cnblogs.com/Summer-skr--blog/p/11397434.html>)

学网—思想海洋学网—思想海洋... 百度翻译 欢迎来到中国多模... EF English Centers 教育部学位与研究... Google Links 北京邮

诗词名句网 www.ShiCiMingJu.Com 首页 分类 作者 排行榜 课本古诗 词牌名 藏头诗 合称 古籍

主 页 > 史书典籍 > 三国演义

《三国演义》

年代：元末明初
作者：罗贯中

《三国演义》中国古典四大名著之一。元末明初小说家罗贯中所著，是中国第一部长篇章回体历史演义的小说。描写了从东汉末年到西晋初年之间近100年的历史风云。全书反映了三国时代的政治军事斗争，反映了三国时代各类社会矛盾的渗透与转化，概括了这一时代的历史巨变，塑造了一批叱咤风云的英雄人物。

全文检索

第一回·宴桃园豪杰三结义 斩黄巾英雄首立功
第二回·张翼德怒鞭督邮 何国舅谋诛宦竖
第三回·议温明董卓叱丁原 馈金珠李肃说吕布
第四回·废汉帝陈留践位 谋董卓献刀
第五回·发矫诏诸镇应曹公 破关兵三英战吕布
第六回·焚金阙董卓行凶 匿玉玺孙坚背约
第七回·袁绍磐河战公孙 孙坚跨江击刘表
第八回·王司徒巧使连环计 董太师大闹凤仪亭
第九回·除暴凶吕布助司徒 犯长安李傕听贾诩

推荐阅读

四大名著

《三国演义》
《西游记》

二十四史

《史记》
《后汉书》
《晋书》
《南齐书》
《陈书》
《北齐书》
《隋书》
《北史》
《新唐书》
《新五代史》
《辽史》
《元史》

四书

《论语》
《大学》

五经

《尚书》

■ 分析源代码

```

85     </div>
86     <div class="line"></div>
87     <div class="book-mulu">
88         <ul>
89             <li><a href="/book/sanguoyanyi/1.html">第一回·宴桃园豪杰三结义 斩黄巾英雄首
立功</a></li><li><a href="/book/sanguoyanyi/2.html">第二回·张翼德怒鞭督邮 何国舅谋诛宦竖
</a></li><li><a href="/book/sanguoyanyi/3.html">第三回·议温明董卓叱丁原 馈金珠李肃说吕布</a>
</li><li><a href="/book/sanguoyanyi/4.html">第四回·废汉帝陈留践位 谋董卓献刀</a></li>
<li><a href="/book/sanguoyanyi/5.html">第五回·发矫诏诸镇应曹公 破关兵三英战吕布</a></li><li>
<a href="/book/sanguoyanyi/6.html">第六回·焚金阙董卓行凶 匿玉玺孙坚背约</a></li><li><a
href="/book/sanguoyanyi/7.html">第七回·袁绍磐河战公孙 孙坚跨江击刘表</a></li><li><a
href="/book/sanguoyanyi/8.html">第八回·王司徒巧使连环计 董太师大闹凤仪亭</a></li><li><a
href="/book/sanguoyanyi/9.html">第九回·除暴凶吕布助司徒 犯长安李傕听贾诩</a></li><li><a
href="/book/sanguoyanyi/10.html">第十回·勤王室马腾举义 报父仇曹操兴师</a></li><li><a
href="/book/sanguoyanyi/11.html">第十一回·刘皇叔北海救孔融 吕温侯濮阳破曹操</a></li><li><a
href="/book/sanguoyanyi/12.html">第十二回·陶恭祖三让徐州 曹孟德大战吕布</a></li><li><a
href="/book/sanguoyanyi/13.html">第十三回·李傕郭汜大交兵 杨奉董承双救驾</a></li><li><a
href="/book/sanguoyanyi/14.html">第十四回·曹孟德移驾幸许都 吕奉先乘夜袭徐郡</a></li><li><a
href="/book/sanguoyanyi/15.html">第十五回·太史慈酣斗小霸王 孙伯符大战严白虎</a></li><li><a
href="/book/sanguoyanyi/16.html">第十六回·吕奉先射戟辕门 曹孟德败师涓水</a></li><li><a
href="/book/sanguoyanyi/17.html">第十七回·袁公路大起七军 曹孟德会合三将</a></li><li><a
href="/book/sanguoyanyi/18.html">第十八回·贾文和料敌决胜 夏侯惇拔矢啖睛</a></li><li><a
href="/book/sanguoyanyi/19.html">第十九回·下邳城曹操鏖兵 白门楼吕布殒命</a></li><li><a
href="/book/sanguoyanyi/20.html">第二十回·曹阿瞒许田大宴群臣 吕奉先白河堤单挑吕布</a></li><li><a
href="/book/sanguoyanyi/21.html">第二十一回·曹操煮酒论英雄 关公斩颜良诛文丑</a></li><li><a
href="/book/sanguoyanyi/22.html">第二十二回·吕布辕门射戟 曹操青梅煮酒论英雄</a></li><li><a
href="/book/sanguoyanyi/23.html">第二十三回·曹操小宴铜雀台 周瑜大宴逍遥津</a></li><li><a
href="/book/sanguoyanyi/24.html">第二十四回·国贼曹操杀小贼 关公斩华雄</a></li><li><a
href="/book/sanguoyanyi/25.html">第二十五回·屯土山关公约三军 救白马曹操解重围</a></li><li><a
href="/book/sanguoyanyi/26.html">第二十六回·袁本初败兵折将 关云长挂印封金</a></li><li><a
href="/book/sanguoyanyi/27.html">第二十七回·美髯公千里走单骑 关云长挂印封金</a></li><li><a
href="/book/sanguoyanyi/28.html">第二十八回·关公斩华雄 曹操献金</a></li><li><a
href="/book/sanguoyanyi/29.html">第二十九回·曹操败走华容道 关公斩华雄</a></li><li><a
href="/book/sanguoyanyi/30.html">第三十回·曹操败走华容道 关公斩华雄</a></li><li><a
href="/book/sanguoyanyi/31.html">第三十一回·曹操败走华容道 关公斩华雄</a></li><li><a
href="/book/sanguoyanyi/32.html">第三十二回·曹操败走华容道 关公斩华雄</a></li><li><a
href="/book/sanguoyanyi/33.html">第三十三回·曹操败走华容道 关公斩华雄</a></li><li><a
href="/book/sanguoyanyi/34.html">第三十四回·曹操败走华容道 关公斩华雄</a></li><li><a
href="/book/sanguoyanyi/35.html">第三十五回·曹操败走华容道 关公斩华雄</a></li><li><a
href="/book/sanguoyanyi/36.html">第三十六回·曹操败走华容道 关公斩华雄</a></li><li><a
href="/book/sanguoyanyi/37.html">第三十七回·曹操败走华容道 关公斩华雄</a></li><li><a
href="/book/sanguoyanyi/38.html">第三十八回·曹操败走华容道 关公斩华雄</a></li><li><a
href="/book/sanguoyanyi/39.html">第三十九回·曹操败走华容道 关公斩华雄</a></li><li><a
href="/book/sanguoyanyi/40.html">第四十回·曹操败走华容道 关公斩华雄</a></li><li><a
href="/book/sanguoyanyi/41.html">第四十一回·曹操败走华容道 关公斩华雄</a></li><li><a
href="/book/sanguoyanyi/42.html">第四十二回·曹操败走华容道 关公斩华雄</a></li><li><a
href="/book/sanguoyanyi/43.html">第四十三回·曹操败走华容道 关公斩华雄</a></li><li><a
href="/book/sanguoyanyi/44.html">第四十四回·曹操败走华容道 关公斩华雄</a></li><li><a
href="/book/sanguoyanyi/45.html">第四十五回·曹操败走华容道 关公斩华雄</a></li><li><a
href="/book/sanguoyanyi/46.html">第四十六回·曹操败走华容道 关公斩华雄</a></li><li><a
href="/book/sanguoyanyi/47.html">第四十七回·曹操败走华容道 关公斩华雄</a></li><li><a
href="/book/sanguoyanyi/48.html">第四十八回·曹操败走华容道 关公斩华雄</a></li><li><a
href="/book/sanguoyanyi/49.html">第四十九回·曹操败走华容道 关公斩华雄</a></li><li><a
href="/book/sanguoyanyi/50.html">第五十回·曹操败走华容道 关公斩华雄</a></li><li><a
href="/book/sanguoyanyi/51.html">第五十一回·曹操败走华容道 关公斩华雄</a></li><li><a
href="/book/sanguoyanyi/52.html">第五十二回·曹操败走华容道 关公斩华雄</a></li><li><a
href="/book/sanguoyanyi/53.html">第五十三回·曹操败走华容道 关公斩华雄</a></li><li><a
href="/book/sanguoyanyi/54.html">第五十四回·曹操败走华容道 关公斩华雄</a></li><li><a
href="/book/sanguoyanyi/55.html">第五十五回·曹操败走华容道 关公斩华雄</a></li><li><a
href="/book/sanguoyanyi/56.html">第五十六回·曹操败走华容道 关公斩华雄</a></li><li><a
href="/book/sanguoyanyi/57.html">第五十七回·曹操败走华容道 关公斩华雄</a></li><li><a
href="/book/sanguoyanyi/58.html">第五十八回·曹操败走华容道 关公斩华雄</a></li><li><a
href="/book/sanguoyanyi/59.html">第五十九回·曹操败走华容道 关公斩华雄</a></li><li><a
href="/book/sanguoyanyi/60.html">第六十回·曹操败走华容道 关公斩华雄</a></li><li><a
href="/book/sanguoyanyi/61.html">第六十一回·曹操败走华容道 关公斩华雄</a></li><li><a
href="/book/sanguoyanyi/62.html">第六十二回·曹操败走华容道 关公斩华雄</a></li><li><a
href="/book/sanguoyanyi/63.html">第六十三回·曹操败走华容道 关公斩华雄</a></li><li><a
href="/book/sanguoyanyi/64.html">第六十四回·曹操败走华容道 关公斩华雄</a></li><li><a
href="/book/sanguoyanyi/65.html">第六十五回·曹操败走华容道 关公斩华雄</a></li><li><a
href="/book/sanguoyanyi/66.html">第六十六回·曹操败走华容道 关公斩华雄</a></li><li><a
href="/book/sanguoyanyi/67.html">第六十七回·曹操败走华容道 关公斩华雄</a></li><li><a
href="/book/sanguoyanyi/68.html">第六十八回·曹操败走华容道 关公斩华雄</a></li><li><a
href="/book/sanguoyanyi/69.html">第六十九回·曹操败走华容道 关公斩华雄</a></li><li><a
href="/book/sanguoyanyi/70.html">第七十回·曹操败走华容道 关公斩华雄</a></li><li><a
href="/book/sanguoyanyi/71.html">第七十一回·曹操败走华容道 关公斩华雄</a></li><li><a
href="/book/sanguoyanyi/72.html">第七十二回·曹操败走华容道 关公斩华雄</a></li><li><a
href="/book/sanguoyanyi/73.html">第七十三回·曹操败走华容道 关公斩华雄</a></li><li><a
href="/book/sanguoyanyi/74.html">第七十四回·曹操败走华容道 关公斩华雄</a></li><li><a
href="/book/sanguoyanyi/75.html">第七十五回·曹操败走华容道 关公斩华雄</a></li><li><a
href="/book/sanguoyanyi/76.html">第七十六回·曹操败走华容道 关公斩华雄</a></li><li><a
href="/book/sanguoyanyi/77.html">第七十七回·曹操败走华容道 关公斩华雄</a></li><li><a
href="/book/sanguoyanyi/78.html">第七十八回·曹操败走华容道 关公斩华雄</a></li><li><a
href="/book/sanguoyanyi/79.html">第七十九回·曹操败走华容道 关公斩华雄</a></li><li><a
href="/book/sanguoyanyi/80.html">第八十回·曹操败走华容道 关公斩华雄</a></li><li><a
href="/book/sanguoyanyi/81.html">第八十一回·曹操败走华容道 关公斩华雄</a></li><li><a
href="/book/sanguoyanyi/82.html">第八十二回·曹操败走华容道 关公斩华雄</a></li><li><a
href="/book/sanguoyanyi/83.html">第八十三回·曹操败走华容道 关公斩华
```

例2.从“诗词名句网站”爬取三国演义

```
import requests
from bs4 import BeautifulSoup
import time

headers = {
    'User-Agent': 'Mozilla/5.0 (Windows NT 6.1; Win64; x64) AppleWebKit/537.36 (
}
url = 'https://www.shicimingju.com/book/sanguoyanyi.html'

page = requests.get(url=url, headers=headers)
html = page.content.decode('utf-8')
soup = BeautifulSoup(html, 'html.parser')

a_list = soup.select('.book-mulu > ul > li > a')
print(a_list)
```

解决乱码

例2.从“诗词名句网站”爬取三国演义

```
with open('sanguo.txt', 'w', encoding='utf-8') as f:
    for a in a_list:
        title = a.string
        detail_url = 'http://www.shicimingju.com'+a['href']
        detail_page_text = requests.get(url=detail_url, headers=headers)

        html = detail_page_text.content.decode('utf-8')
        soup = BeautifulSoup(html, 'html.parser')
        content = soup.find('div', class_='chapter_content').text

        f.write(title+':'+content+'\n')
        print(title, '保存成功!')
        time.sleep(1)

print('over!')
```

#把a标签当soup对象使用，因为它也是源码

解决乱码

bs4中，把text提取出来的列表直接转换成字符串

第一回·宴桃园豪杰三结义
第二回·张翼德怒鞭
宦竖第三回·议
金珠李肃说吕布
位 谋董贼孟德献刀</p>

\2021示例\三国演义提取.py =====

Squeezed text (99 lines).

[illegible]

程序bs_sanguoyanyi1.py
运行结果2021年

■ 程序bs_sanguoyanyi1.py写得文件sanguo.txt内容

sanguo - 记事本

文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H)

第一回·宴桃园豪杰三结义 斩黄巾英雄首立功:

滚滚长江东逝水，浪花淘尽英雄。是非成败转头空。青山依旧在，几度夕阳红。 白发渔樵江渚上，惯看秋月春风。一壶浊酒喜相逢。古今多少事，都付笑谈中。

——调寄《临江仙》

话说天下大势，分久必合，合久必分。周末七国分争，并入于秦。及秦灭之后，楚、汉分争，又并入于汉。汉朝自高祖斩白蛇而起义，一统天下，后来光武中兴，传至献帝，遂分为三国。推其致乱之由，殆始于桓、灵二帝。桓帝禁锢善类，崇信宦官。及桓帝崩，灵帝即位，大将军窦武、太傅陈蕃，共相辅佐。时有宦官曹节等弄权，窦武、陈蕃谋诛之，机事不密，反为所害，中涓自此愈横。

建宁二年四月望日，帝御温德殿。方升座，殿角狂风骤起。只见一条大青蛇，从梁上飞将下来，蟠于椅上。帝惊倒，左右急救入宫，百官俱奔避。须臾，蛇不见了。忽然大雷大雨，加以冰雹，落到半夜方止，坏却房屋无数。建宁四年二月，洛阳地震；又海水泛滥，沿海居民，尽被大浪卷入海中。光和元年，雌鸡化雄。六月朔，黑气十余丈，飞入温雄殿中。秋七月，有虹现于玉堂；五原山岸，尽皆崩裂。种种不祥，非止一端。帝下诏问群臣以灾异之由，议郎蔡邕上疏，以为蜺堕鸡化，乃妇寺干政之所致，言颇切直。帝览奏叹息，因起更衣。曹节在后窃视，悉宣告左右；遂以他事陷邕于罪，放归田里。后张让、赵忠、封谡、段珪、曹节、侯览、蹇硕、程旷、夏惲、郭胜十人朋比为奸，号为“十常侍”。帝尊信张让，呼为“阿父”。朝政日非，以致天下人心思乱，盗贼蜂起。

时巨鹿郡有兄弟三人，一名张角，一名张宝，一名张梁。那张角本是个不第秀才，因入山采药，遇一老人，碧眼童颜，手执藜杖，唤角至一洞中，以天书三卷授之，曰：“此名《太平要术》，汝得之，当代天宣化，普救世人；若萌异心，必获恶报。”角拜问姓名。老人曰：“吾乃南华老仙也。”言讫，化阵清风而去。角得此书，晓夜攻习，能呼风唤雨，号为“太平道人”。中平元年正月内，疫气流行，张角散施符水，为人治病，自称“大贤良师”。角有徒弟五百余人，云游四方，皆能书符念咒。次后徒众日多，角乃立三十六方，大方万余人，小方六七千，各立渠帅，称为将军；讹言：“苍天已死，黄天当立；岁在甲子，天下大吉。”令人各以白土，书“甲子”二字于家中大门上。青、幽、徐、冀、荆、扬、兖、豫八州之人，家家侍奉大贤良师张角名字。角遣其党马元义，暗赍金帛，结交中涓封谡，以为内应。角与二弟商议曰：“至难得者，民心也。今民心已顺，若不乘势取天下，诚为可惜。”遂一面私造黄旗，约期举事；一面使弟子唐周，驰书报封谡。唐周乃径赴省中告变。帝召大将军何进调兵擒马元义，斩之；次收封谡等一千人下狱。张角闻知事露，星夜举兵，