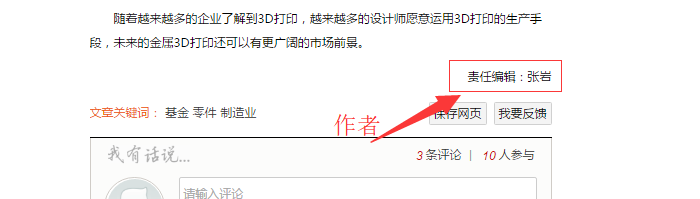
# 第五课 抓取新闻的内容页

内容页结构：



...



以这篇文章为例子：

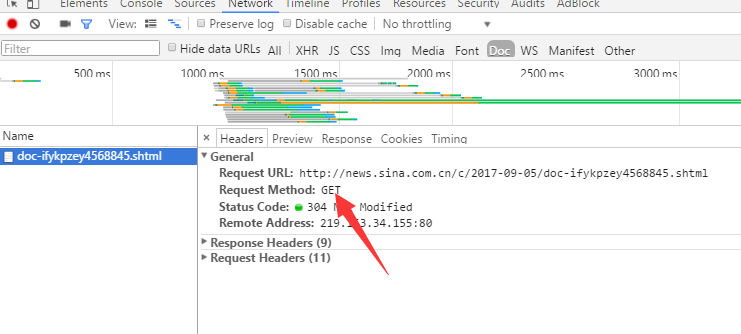
<http://news.sina.com.cn/c/2017-09-05/doc-ifykpzey4568845.shtml>

因为简短，方便观看演示：

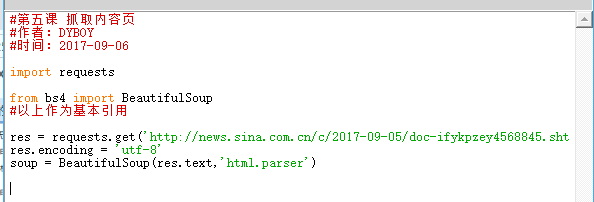


GET请求：

<http://news.sina.com.cn/c/2017-09-05/doc-ifykpzey4568845.shtml>

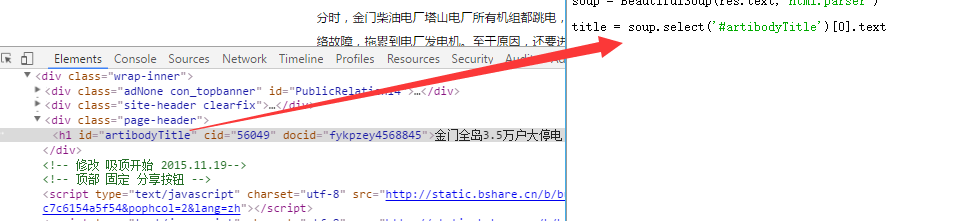


现在的代码：



抓取标题：

title = soup.select('#artibodyTitle')[0].text



抓取时间：

涉及到拆分的过程：

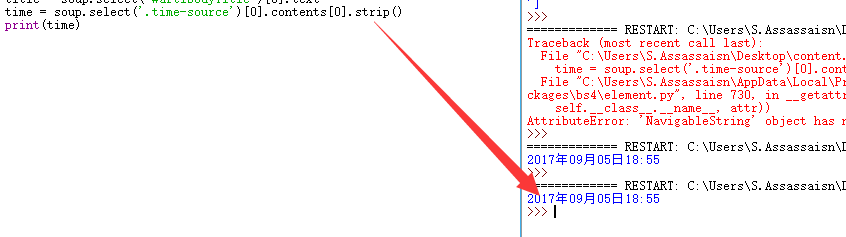
.contents的用法：



拆分成两个部分

.strip()的用法：

移除特殊字符



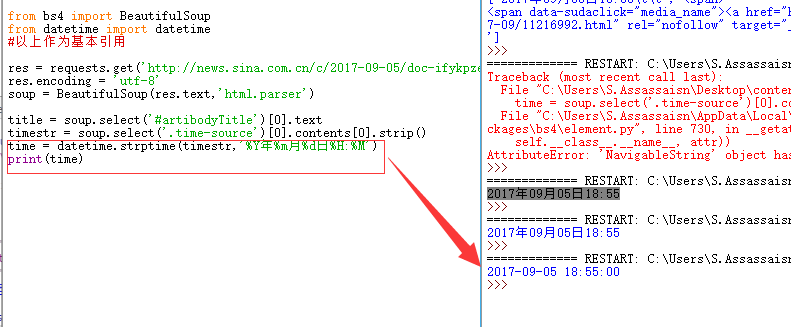
时间格式转化：

from datetime import datetime

字符串转时间： ----strptime

Dt = datetime.strptime(timesource,’%Y年%m月%d日%H:%M’)

Dt



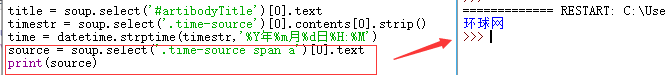
这样我们就可以将数据转化并以正确的格式存储到数据库。

时间转字符串：

Dt.strftime(‘%Y-%m-%d’)

抓取文章来源：

source = soup.select('.time-source span a')[0].text



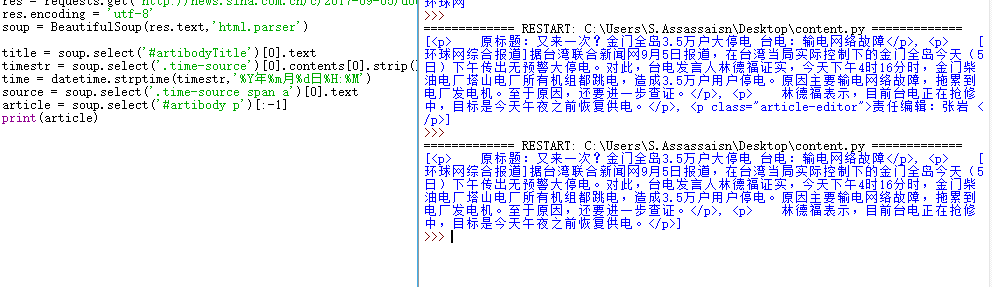
抓取文章正文：



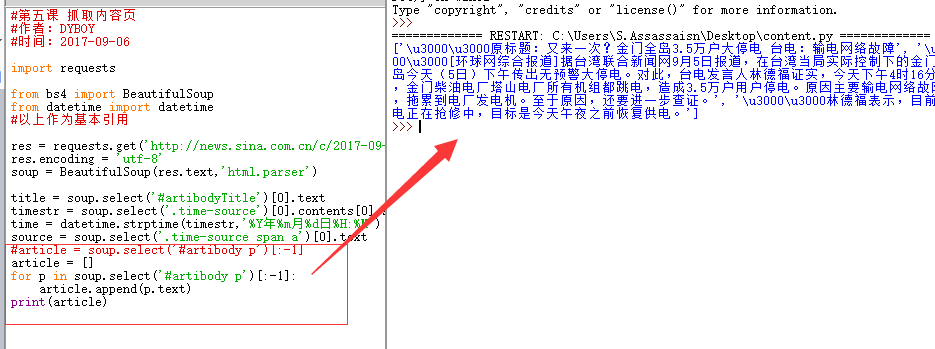
正文抓取去除最后的“责任编辑”

后缀加上[:-1]即可：

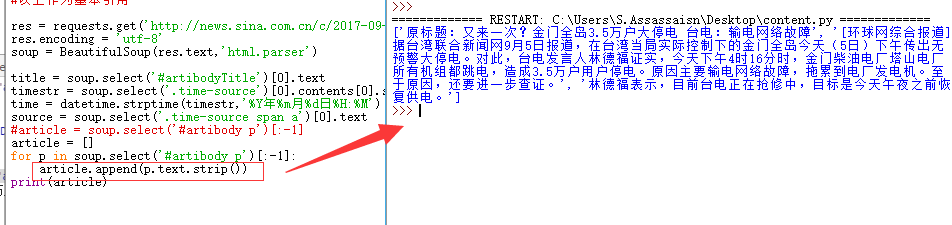
article = soup.select('#artibody p')[:-1]



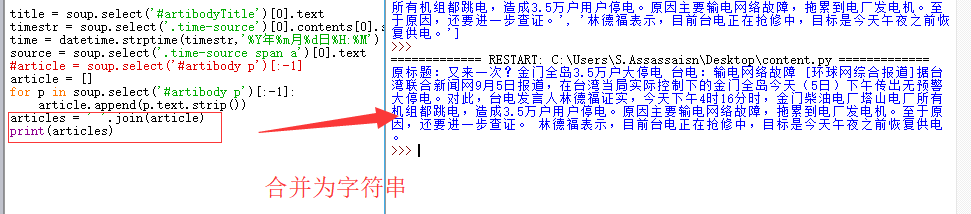
为了将所有的文字内容合并：



去除‘\u3000’,只需要使用.strip()



.join()合并列表为字符串：



上图是以一个空格连接内容

至此代码：

article = []

for p in soup.select('#artibody p')[:-1]:

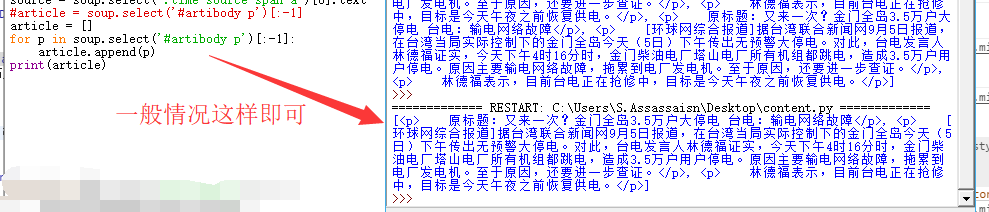
article.append(p.text.strip())

articles = ' '.join(article)

print(articles)

说明：

为啥搞得那么复杂，其实是为了大家更充分理解各种正则操作，在实际的抓取网页时，一般是会将HTML标签等一同抓取。

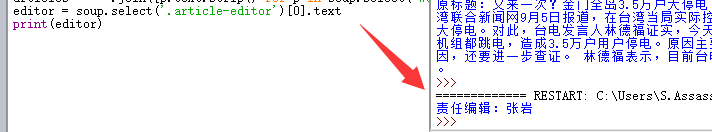


更简单的写法：



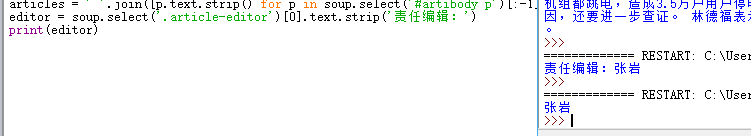
获取作者：

editor = soup.select('.article-editor')[0].text.lstrip('责任编辑：')



移除责任编辑：

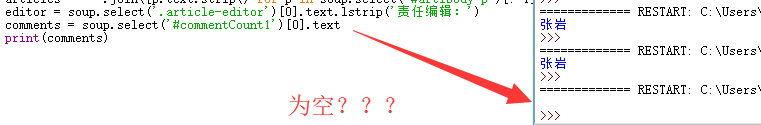
.strip(‘责任编辑：’)



或者：lstrip('责任编辑：')

获取评论数：

按照常规方法：

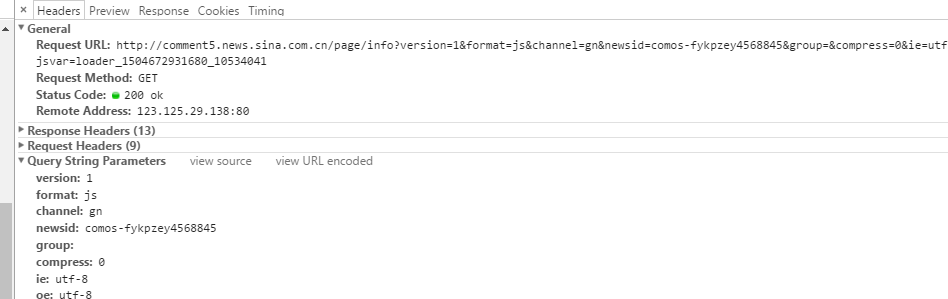


结果为空？

想到评论数可能是JavaScript动态加载的！

通过开发者工具找到：

<http://comment5.news.sina.com.cn/page/info?version=1&format=js&channel=gn&newsid=comos-fykpzey4568845&group=&compress=0&ie=utf-8&oe=utf-8&page=1&page_size=20&jsvar=loader_1504672931680_10534041>





分析URL，将改为：

<http://comment5.news.sina.com.cn/page/info?version=1&format=js&channel=gn&newsid=comos-（文章id）&group=&compress=0&ie=utf-8&oe=utf-8&page=1&page_size=20>

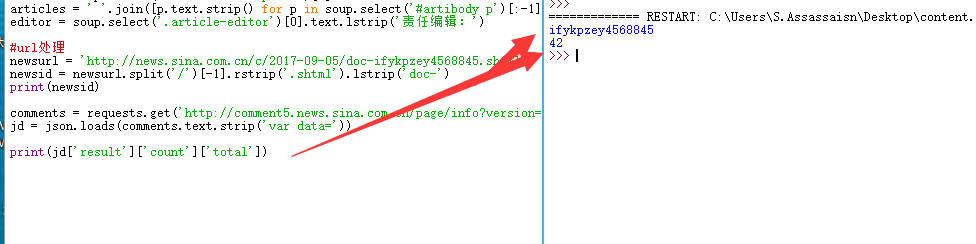
如何获取新闻id：

#url处理

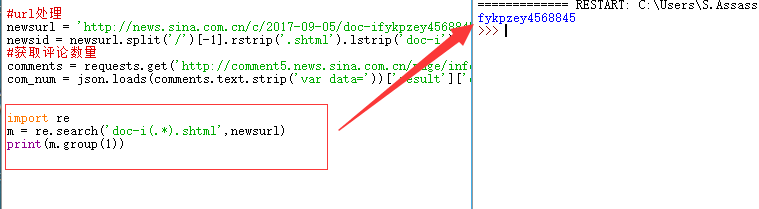
newsurl = 'http://news.sina.com.cn/c/2017-09-05/doc-ifykpzey4568845.shtml'

newsid = newsurl.split('/')[-1].rstrip('.shtml').lstrip('doc-')

print(newsid)



或者re正则表达：

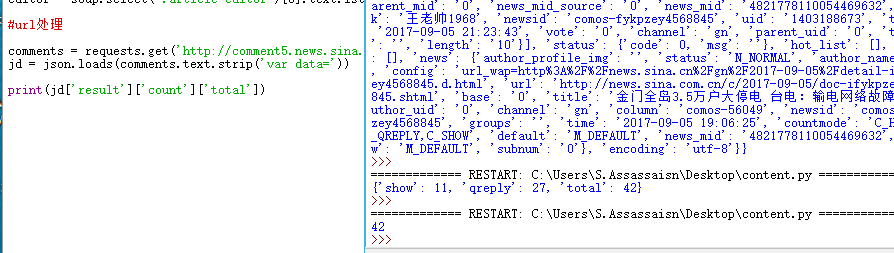


JSON数据解析获取评论数量：

comments = requests.get('http://comment5.news.sina.com.cn/page/info?version=1&format=js&channel=gn&newsid=comos-fykpzey4568845&group=&compress=0&ie=utf-8&oe=utf-8&page=1&page\_size=20')

jd = json.loads(comments.text.strip('var data='))

print(jd['result']['count']['total'])



整理如下：

