汽车之家的文章评论的爬取

这里我们以<http://www.autohome.com.cn/3586/>这个车系的文章为例

我们很容易可以得到这个车系的文章模块链接为<https://www.autohome.com.cn/3586/0/0-0-1-0/>（这里要注意链接最后有一个“/”，抓取页面的时候，没有的话会找不到这个页面），并且通过观察可知每个车系的文章模块链接的后面的数字都是一样的。



然后我们在进入文章页面后，要获得每个文章的链接，在这之前我们首先要获得文章的最大页码之后才能进行循环得到文章的全部链接



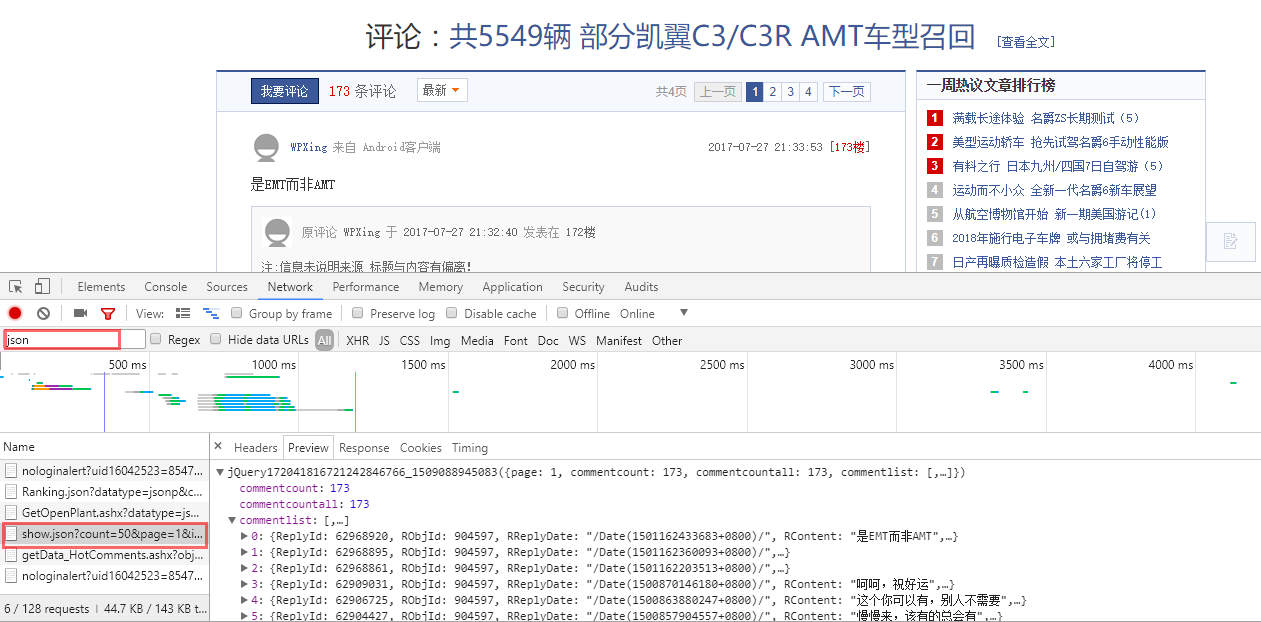
获取文章最大的页码的方法，就是先利用requests模块的get方法获取整个页面，再利用BeautifulSoup模块对页面进行解析，解析完之后利用浏览器上的开发者工具观察页码的表达形式，利用正则匹配，匹配到相应a标签的数字，得到最大页码。

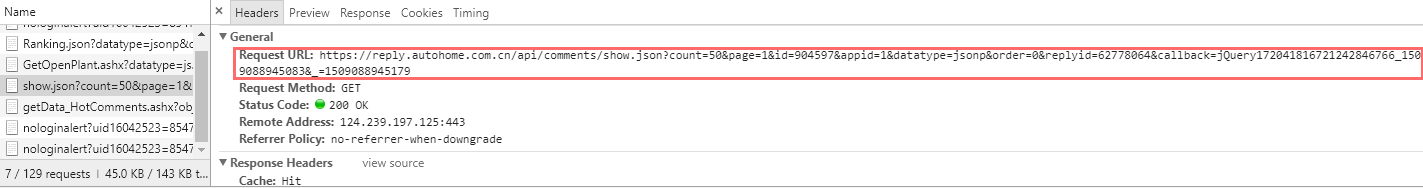


得到最大页码之后，通过观察发现文章的链接根据页码的变化只在倒数第二位相应变化，这样我们就能知道其他剩余文章的链接与页码相关。在这之上，通过对网页源代码的观察我们通过正则匹配可以得到文章的名称和链接，通过对最大页码的循环可以得到这个车系全部文章的标题与链接。



获得文章的链接之后我们进入一篇文章全部评论的页面，在进行查看源代码时我们发现评论并没有办法获取到，因为这些评论是通过动态加载过去的，之前我们能获取的只是静态的页面。这时我们通过开发者工具的network查看数据存储的方式和路径，多查看几页的评论的json数据的路径我们可以发现，评论的路径基本是相同的<https://reply.autohome.com.cn/api/comments/show.json?count=50&page=1&id=904597&appid=1>我们只需要获得页码page和文章的id就能得到全部的评论json数据。





获得文章的id可以通过在文章页面中，正则匹配到全部评论的href链接中的articleid的值来得到。评论的页码可以通过爬取到第一页中评论的json数据（json数据的解析通过json模块进行解析），获得的全部评论来计算评论的页码。最后在通过循环页码来获取评论的json数据，这样就可以爬取到你想要的评论信息啦！



最后在编程过程中注意一些编码格式和数组溢出的问题。