## 二．回顾一下上次的爬虫

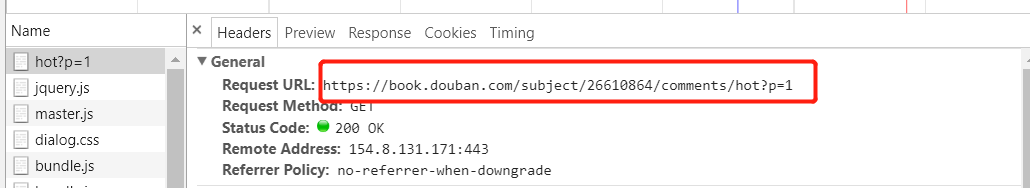
上次的爬虫呀，是获取书评链接的评论，现在呢，我们看看另外一本书的书评页面



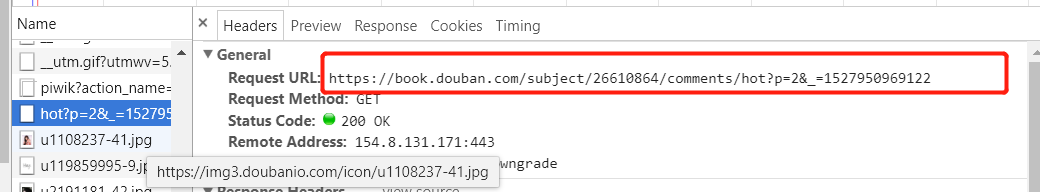
这个书评有3000多条，但是呢，网页中只显示了19（但是我看呀，每一页的条数好像不是固定的，待考证）条记录呢，那如果现在我们要把所有的评论都获取到，那要怎么办呢？我们通过网页的的翻页功能看看，先看看网页是怎么设计是怎么的



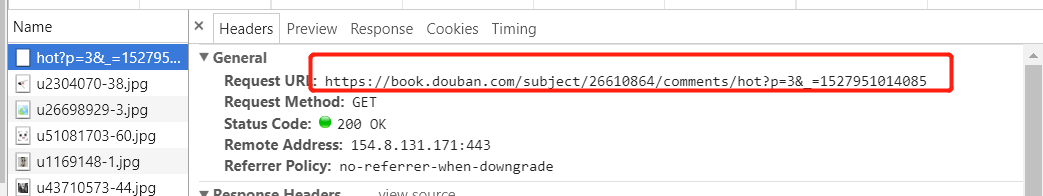
第一页数据的网页请求



第二页的数据请求



第三页的数据请求



我们发现呀，评论的数据都从<https://book.douban.com/subject/26610864/comments/hot>这个接口返回的，都是保存在html结构的里面，仔细一看，发现请求有一个参数 p ，这个参数对应着我打开的评论的页数（猜测的），验证一下之后，总共有3202条评论，每页有19条，那么总共有的页数是3202/9=169.5，确实哦



169是最后一页了，下一页是不能点击的状态

2.那现在如果我要把全部的评论获取下来，是不是要通过去这169页对应的请求获取的呢？



import requests

from bs4 import BeautifulSoup

baseUrl = 'https://book.douban.com/subject/26610864/comments/hot?p='

page = 169

while True:

url = baseUrl + str(page)

req = requests.get(url)

soup = BeautifulSoup(req.content, 'html.parser')

avatar = soup.select("#comments > ul > li > div.avatar > a > img")

for img in avatar:

print(img['src'])

# 如果当页的评论条数为0，代表已经没数据了超过了最后一页了，所以可以停止循环了

print(len(avatar))

if len(avatar) == 0:

break

# 页数加一，进入下一个循环获取下一个页面的数据逻辑

page +=1

print('job done,total page:%s' % (page))