一．项目介绍

1、使用scrapy爬取京东上所有的图书数据

2、将爬取的数据存储到MongoDB

****环境****

win7、python2、pycharm

****技术****

1、数据采集：scrapy

2、数据存储：MongoDB

****难点分析****

和其他的电商网站相比，京东的搜索类爬取主要有以下几个难点：

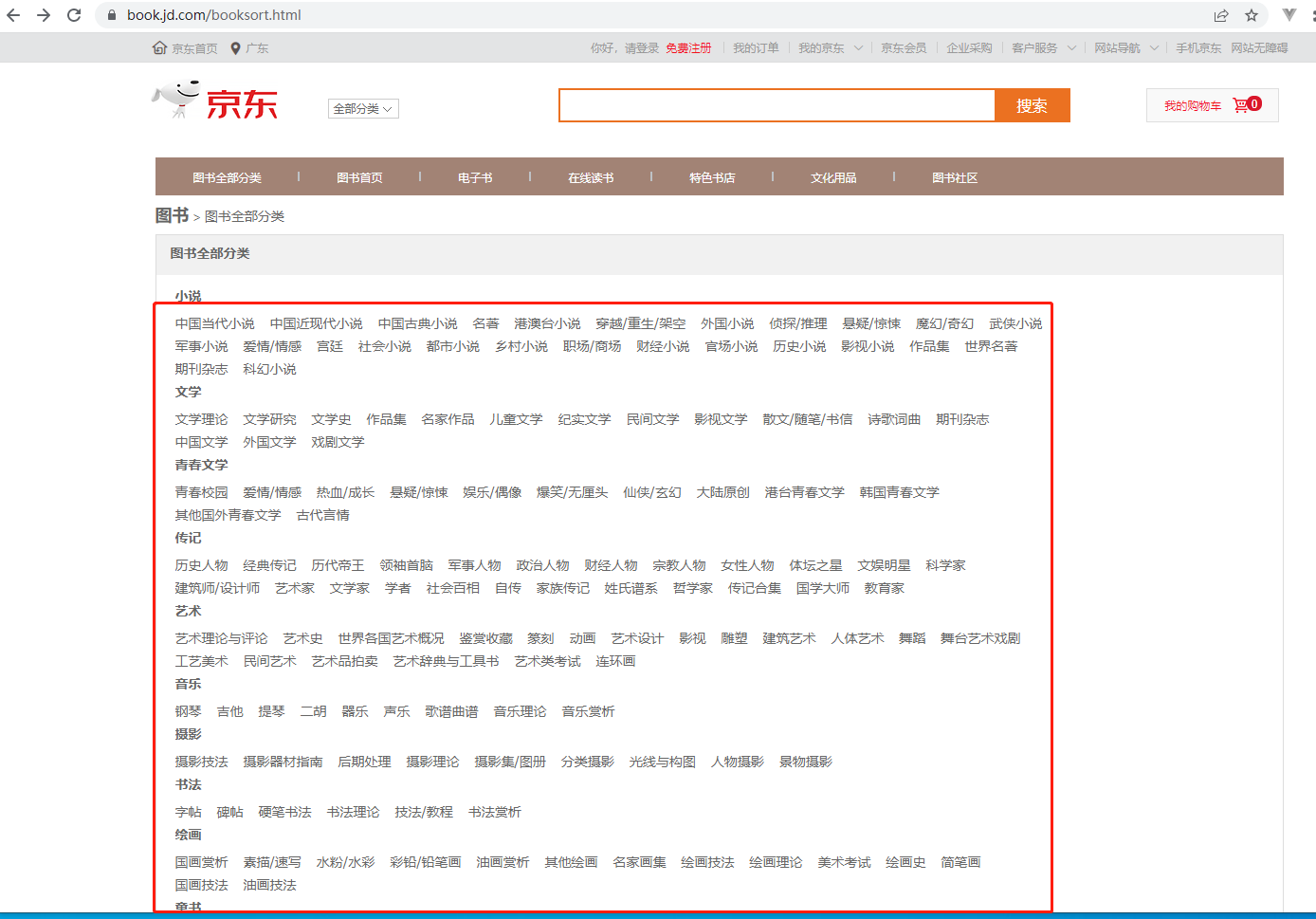
1、搜索一个商品时，一开始显示的商品数量为30个，当下拉这一页 时，又会出现30个商品，这就是60个商品了，前30个可以直接 从原网页上拿到，后30个却在另一个隐藏链接中，要访问这两个 链接，才能拿到一页的所有数据。

2、隐藏链接的构造，发现最后的那个show\_items字段其实是前30 个商品的id。

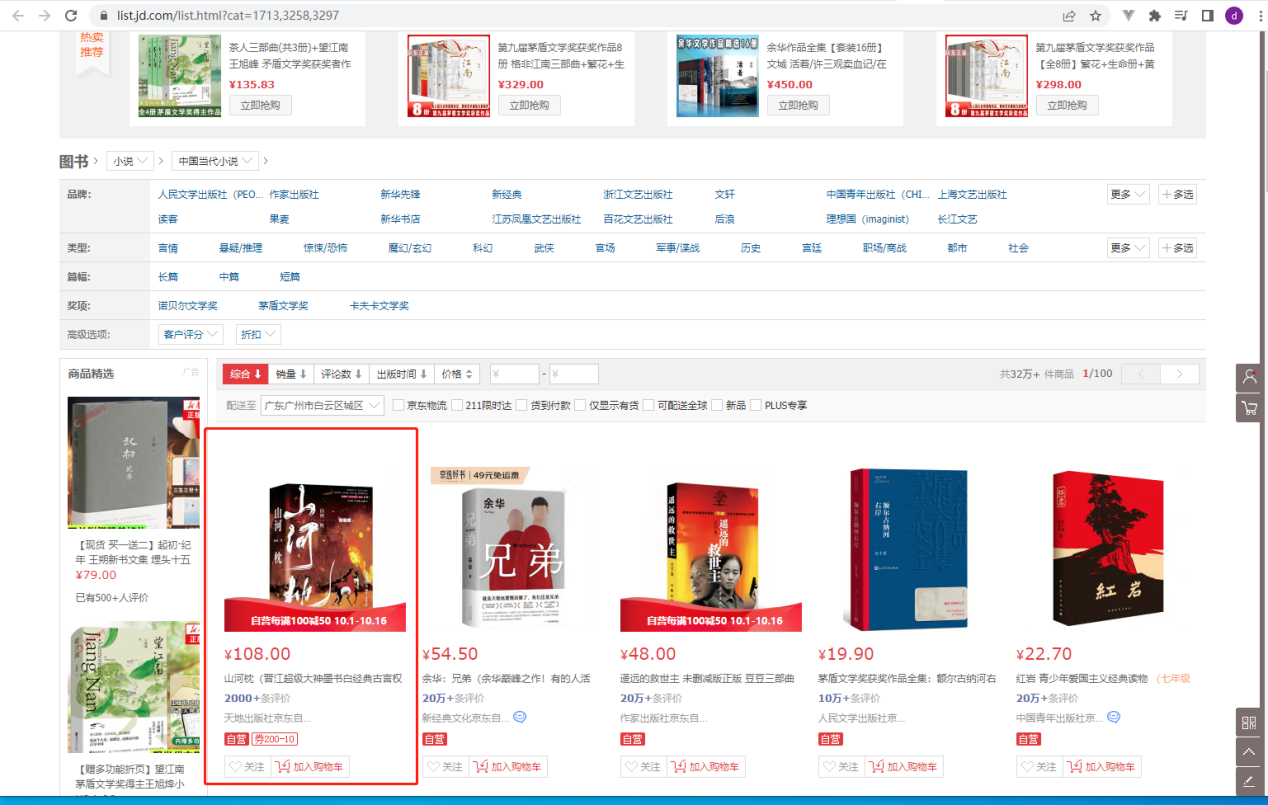
3、直接反问隐藏链接被拒绝访问，京东的服务器会检查链接的来源， 只有来自当前页的链接他才会允许访问。

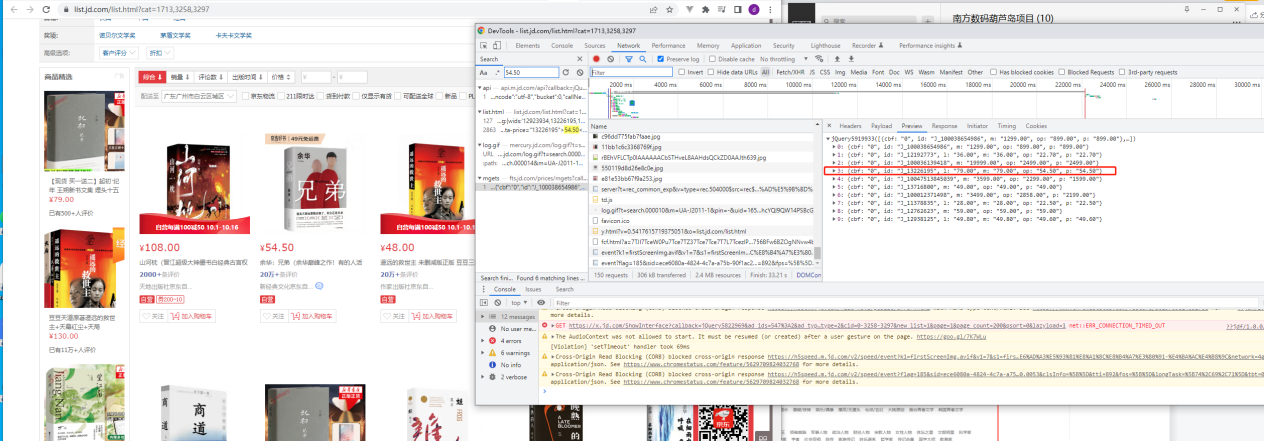
4、前30个商品的那一页的链接page字段的自增是1、3、5。。。这 样的，而后30个的自增是2、4、6。。。这样的。

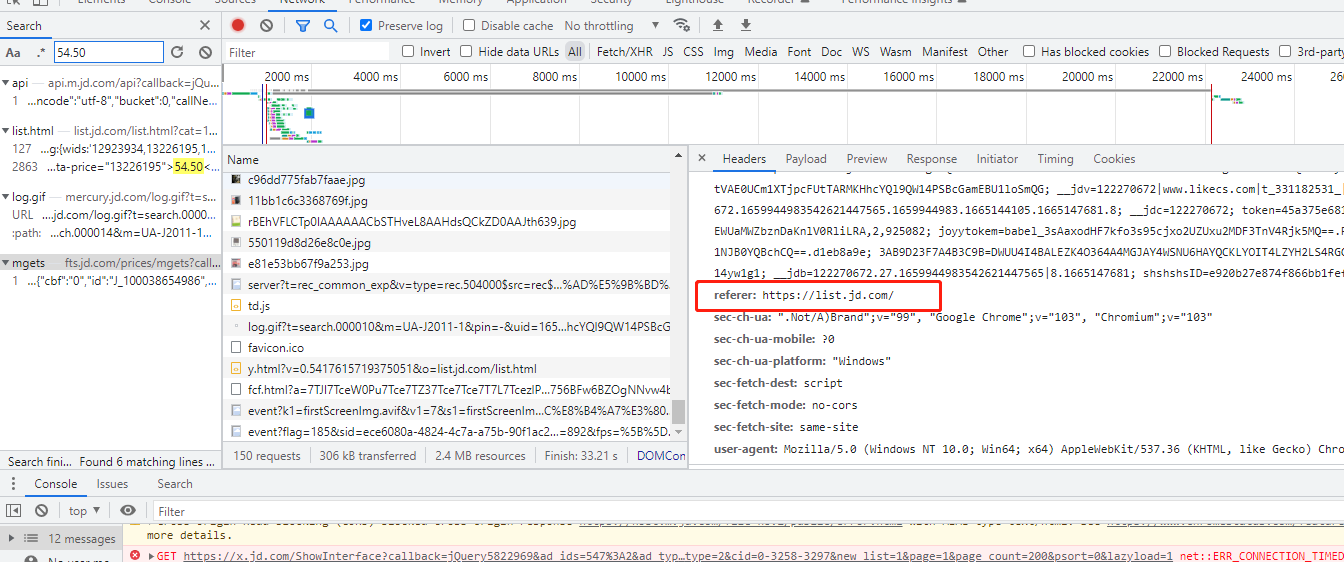
目标：[https://book.jd.com/booksort.html](https://book.jd.com/booksort.html，)



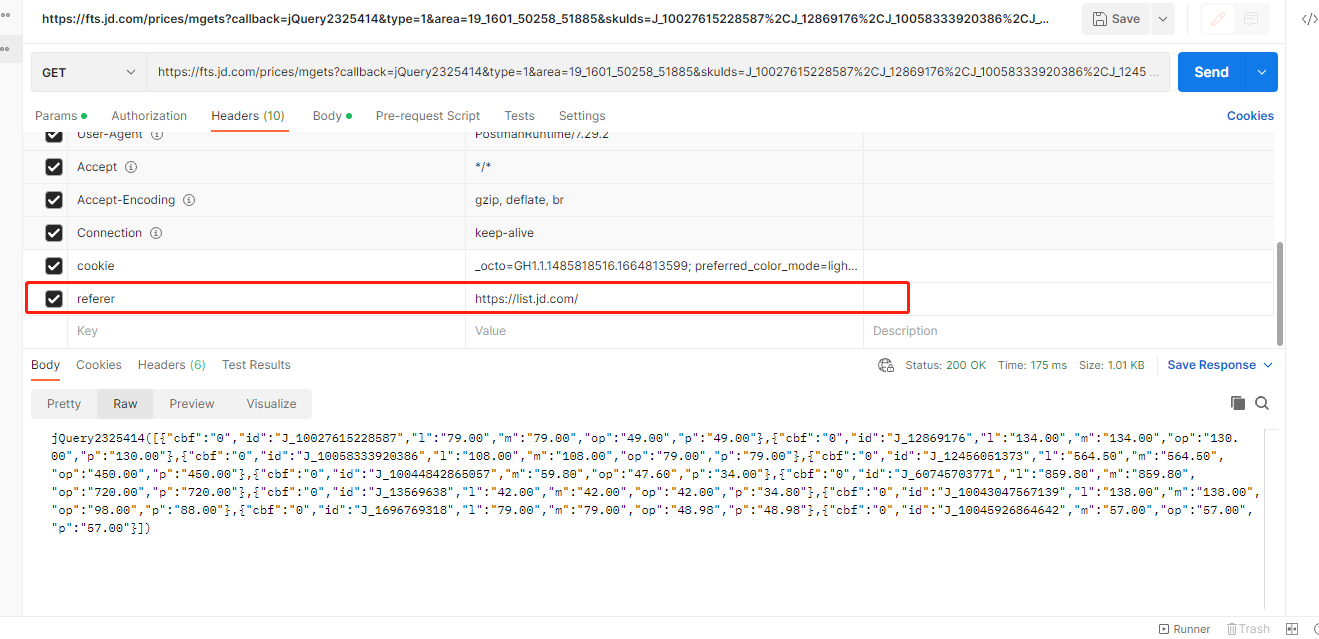
<https://list.jd.com/list.html?cat=1713,3258,3297>







在postman中模拟请求时，需带上防盗链，不然直接复制请求地址访问时得不到数据的



https://fts.jd.com/prices/mgets?callback=jQuery5919933&type=1&area=19\_1601\_50258\_51885&skuIds=J\_100038654986%2CJ\_12192773%2CJ\_100036139418%2CJ\_13226195%2CJ\_10047513845039%2CJ\_13716800%2CJ\_100012371498%2CJ\_11378835%2CJ\_12762623%2CJ\_12938125&pdbp=0&pdtk=&pdpin=&pduid=1659944983542621447565&source=search\_pc&\_=1665149787239

jQuery5919933，5919933应该为当前发起请求时的时间戳

参考：<https://www.cnblogs.com/SnowPhoenix/p/15674155.html>

使用scrapy\_redis 进行分布式爬虫

