# 一、程序

本爬虫基于python的requests库实现，选择了爬取京东的一页商品数据，具体链接为<https://list.jd.com/list.html?cat=670%2C699%2C700&page=1&s=1&click=0>

在爬取时，为防止爬虫被封或京东出现要求登录账号的情况，选择设置请求头中的User-Agent以模拟浏览器进行爬取，在访问京东的url时，用随机数选择一个User-Agent并调用requests中的get方法访问url，具体实现如图1-1所示。

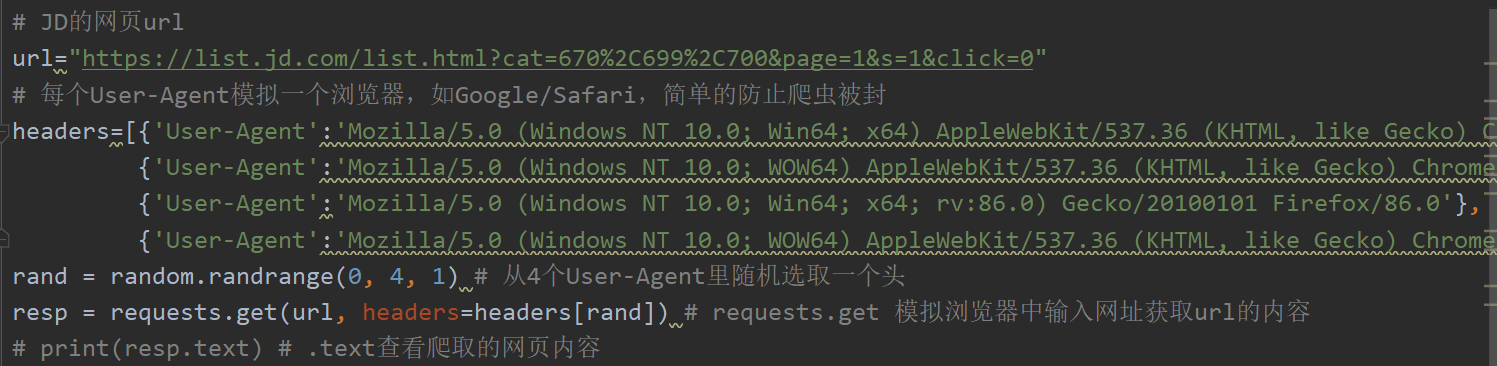


图1-1

在实现爬虫时，选择了用xpath解析爬取的网页内容，这里先用谷歌浏览器的xpath插件模拟爬虫分析，提取网页中的商品名称以及价格，如图1-2和图1-3所示。



图1-2

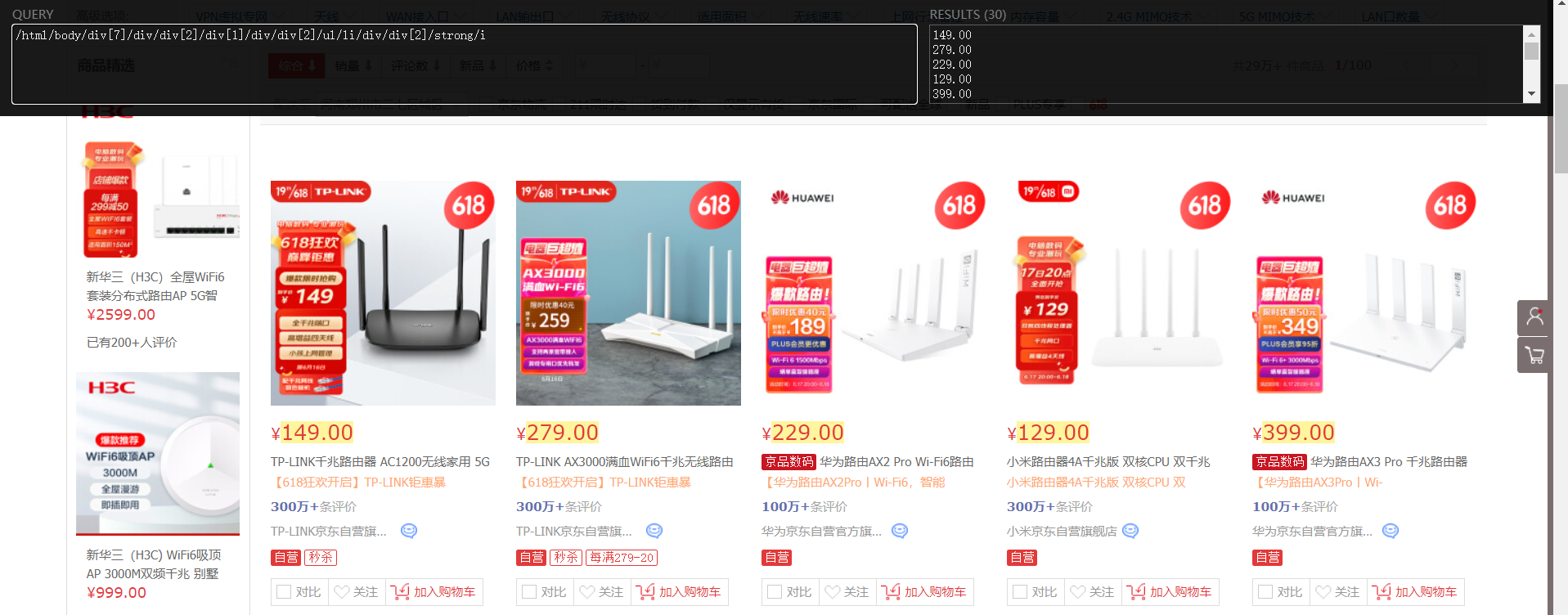


图1-3

get请求返回的页面内容以字符串文本的形式存在resp.text中，要使用xpath对页面内容进行解析，要利用lxml库中的etree，将文本形式的html处理，方面下面调用xpath进行解析。将每个爬取的商品标题和商品价格信息以文本的形式写入JDspider.txt中，具体实现如图1-4所示。



图1-4

# 二、结果截图

运行爬虫，控制台的部分输出如图2-1和图2-2所示。

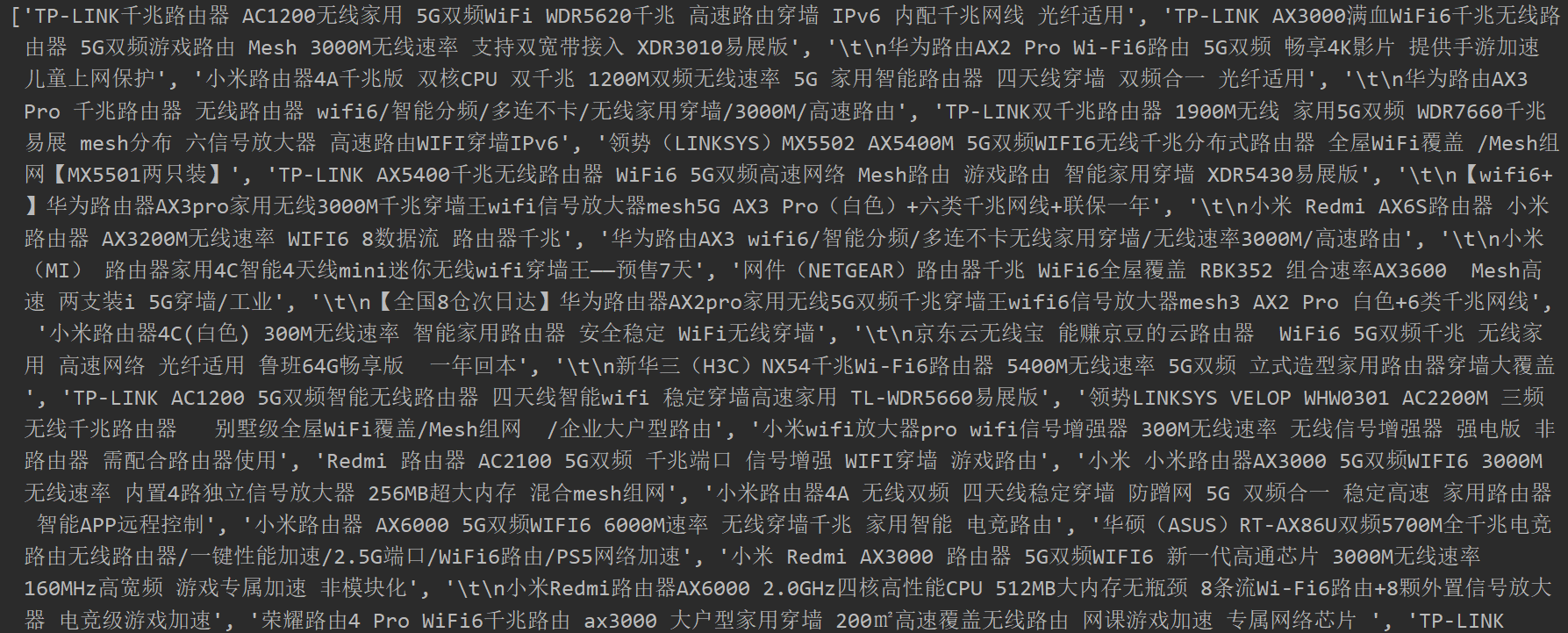


图2-1

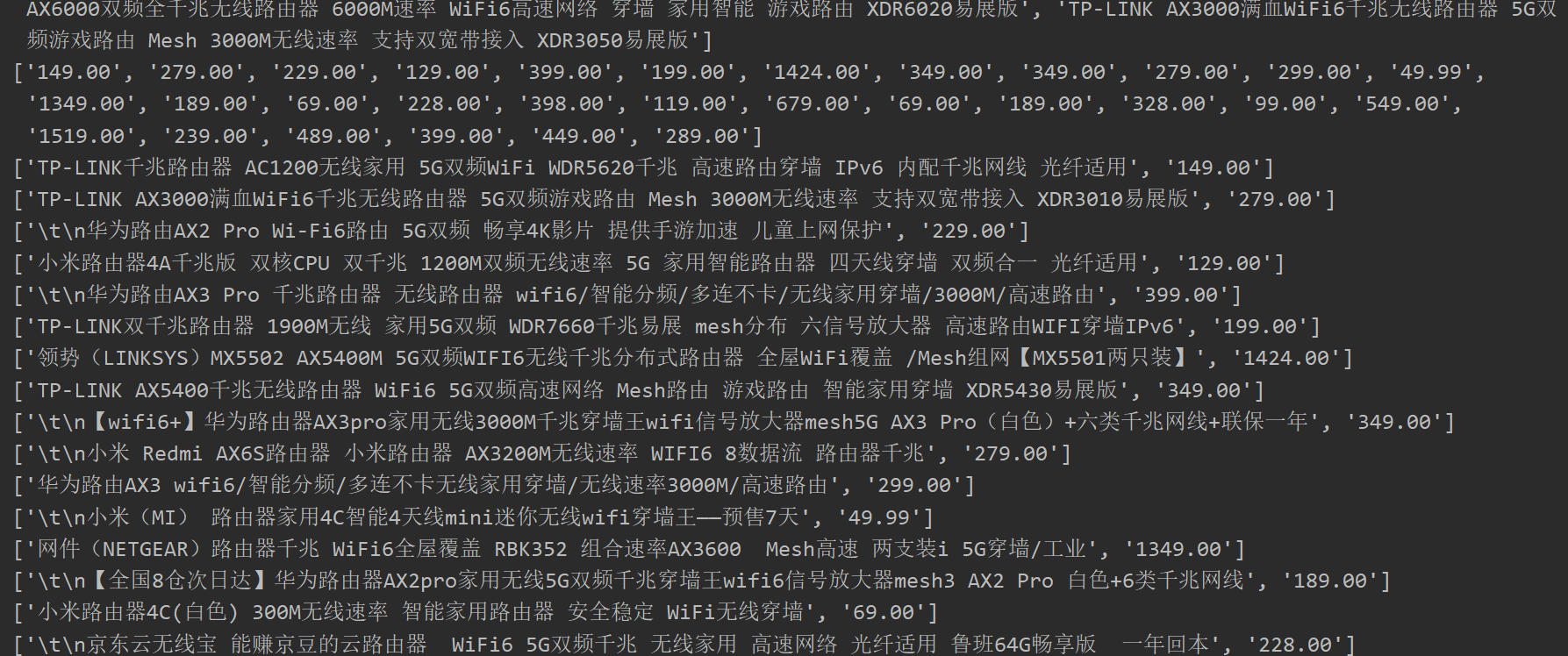


图2-2

打开爬虫写入的JDspider.txt文件，爬取的结果成功写入文件保存，如图2-3所示。

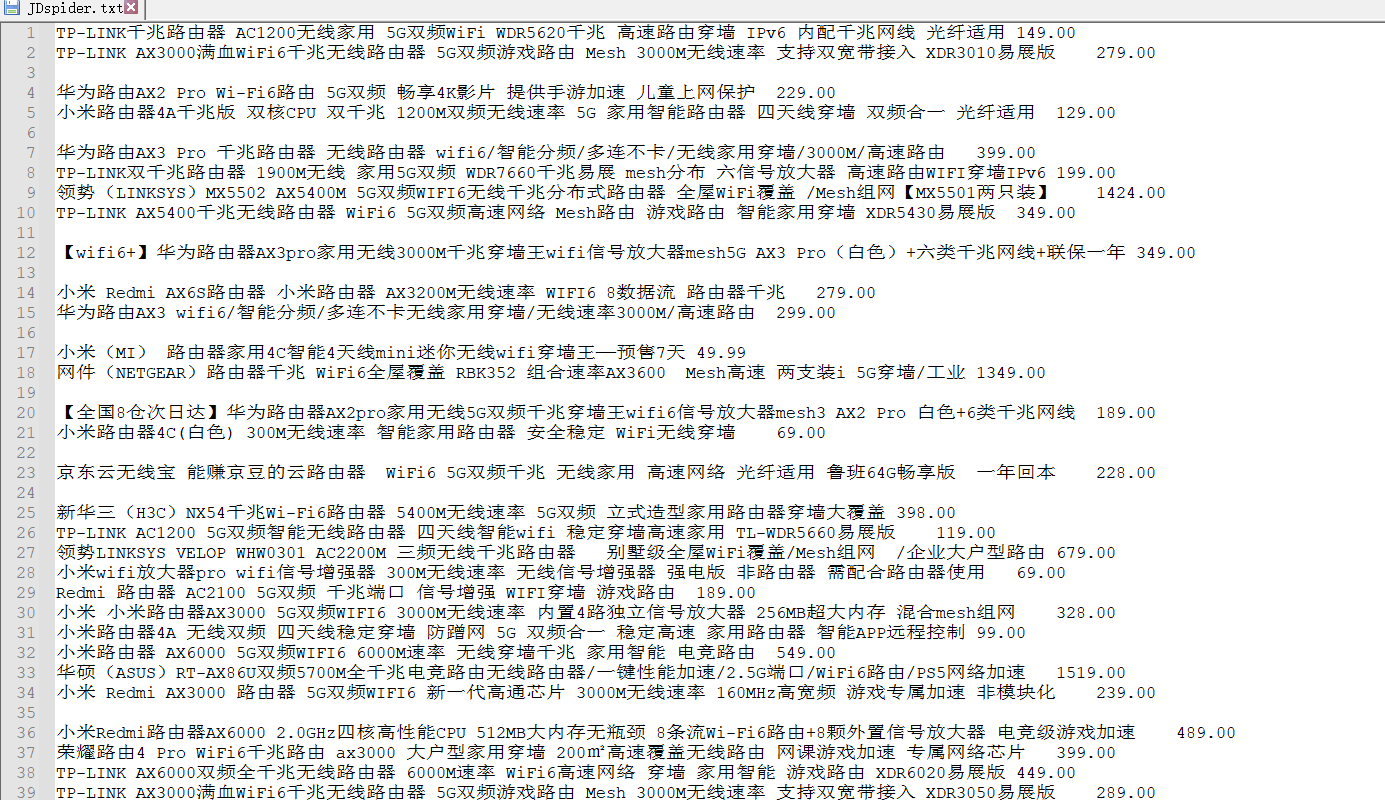


图2-3