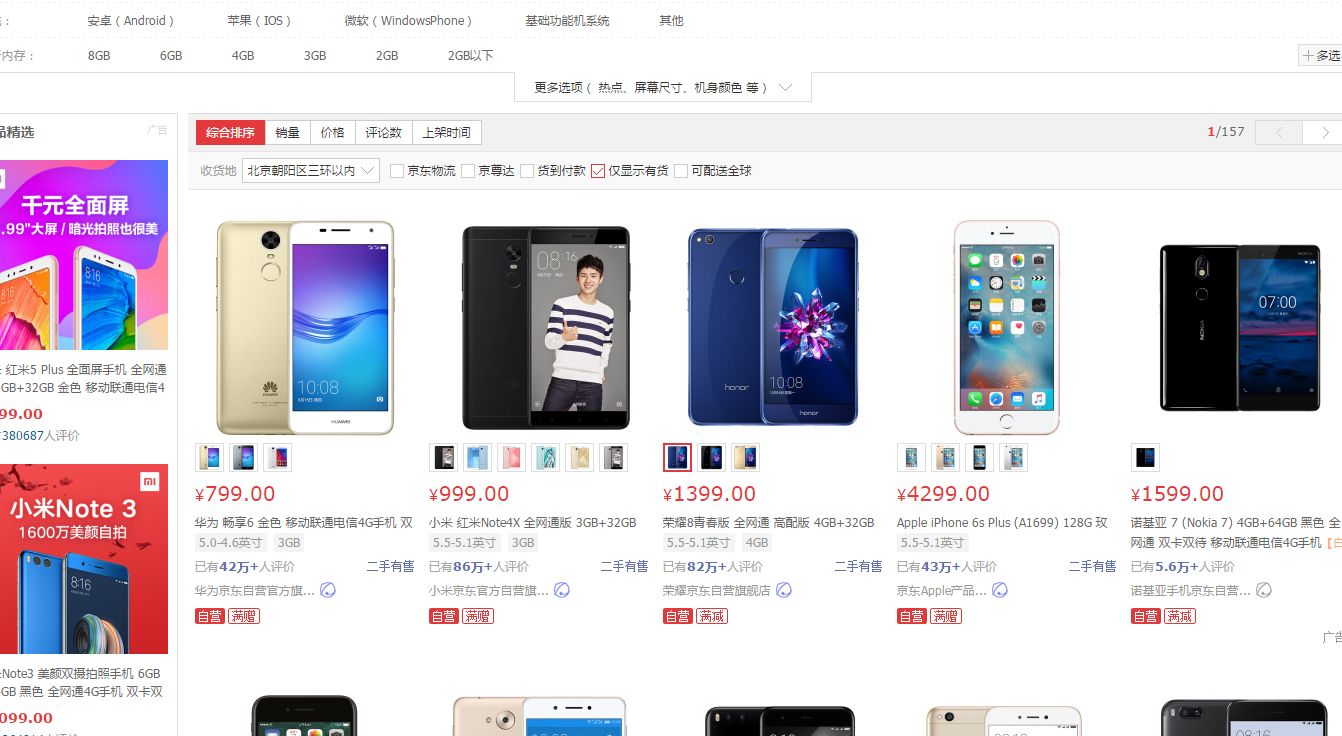
京东手机信息爬取

京东商城有很多商品信息可供爬取，这次主要爬取全部手机综合排序的第一页

一、相关爬取页面分析

页面地址：

[**https://list.jd.com/list.html?cat=9987,653,655**](https://list.jd.com/list.html?cat=9987,653,655)



对应源码：



其中价格、评论数需要动态获取

先来看价格信息，访问的URL格式为

https://p.3.cn/prices/mgets?skuIds=J\_（data-sku）

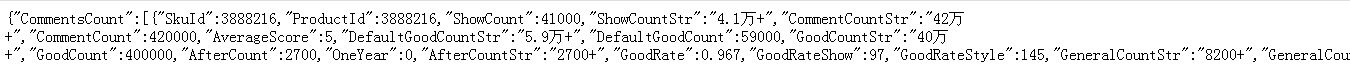
括号里的data-sku为商品id，打开url返回



再看评论信息，访问的URL格式为

https://club.jd.com/comment/productCommentSummaries.action?referenceIds=（data-sku）

括号里的data-sku为商品id，打开url返回



二、python3代码说明

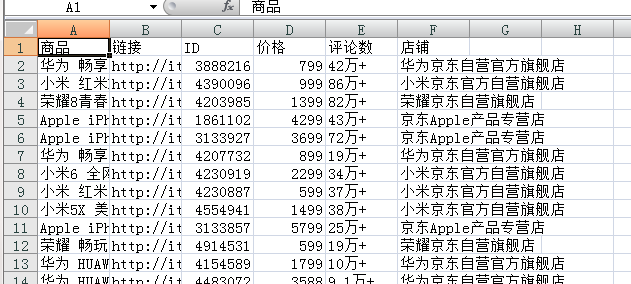
爬取思路：

1. 爬取该网页的源代码（前述分析，函数模块getHTMLText(url)）
2. Beautiful Soup提取所需内容
3. 根据网页源代码对所取内容进行精准修改以获取所需内容

  4、把内容写入到文件

主要代码（jdspider.py）

三、结果



四、小结

实现了一个爬虫的基本功能，后续可根据需要在此基础上进行扩充完善，例如多线程的并发爬取，验证码交互等。