

项目 2：大数据采集与来源

实训目标

通实操八爪鱼采集器，独立完成电商商品采集全流程，动手配置关键词循环、商品循环、分页采集规则，自主导出并人工校验采集数据，掌握零代码网页数据采集完整实操技能，树立合规采集实操规范。

实训环境

Windows10 及以上、Office/WPS 表格、浏览器、八爪鱼采集器客户端。

操作步骤

任务准备：

1、八爪鱼采集器的下载

在八爪鱼采集器官网下载最新版客户端软件（这里以 Windows 版本为例，下载得到名为 OctopusSetup8.7.7.exe 的安装文件），如图 2-1 所示。

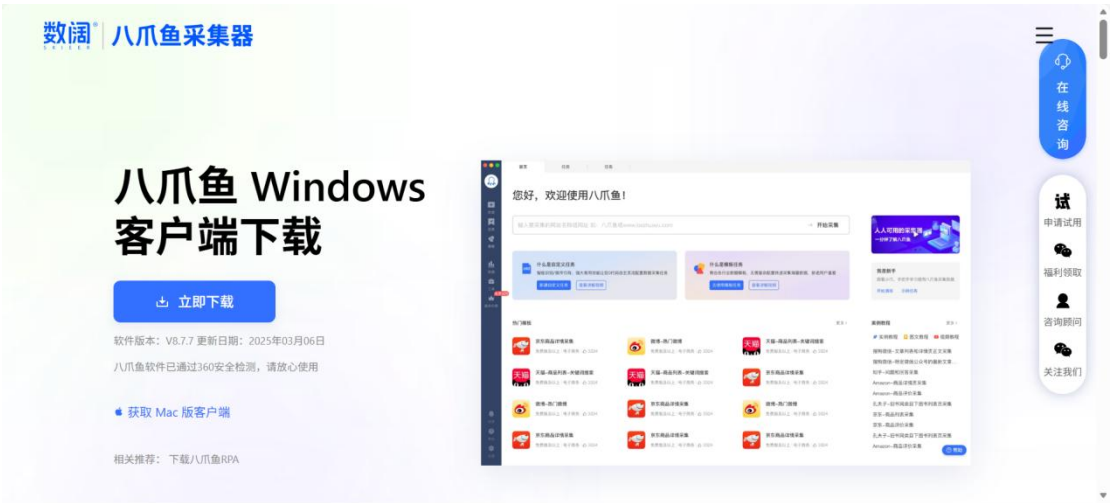


图 2-1 八爪鱼下载页面

双击可执行文件，按界面指引完成安装。安装后，在桌面双击“八爪鱼采集器”图标启动。启动后，将进入八爪鱼采集器的登录/注册页面，如图 2-2。整个安装过程简单快捷，方便用户使用。



图 2-2 八爪鱼登录页面

2、八爪鱼采集器的注册

使用八爪鱼之前，必须先注册用户。在登录页面通过手机号和验证码可以快速注册并登录八爪鱼采集器，进入八爪鱼软件首页，如图 2-3 所示。

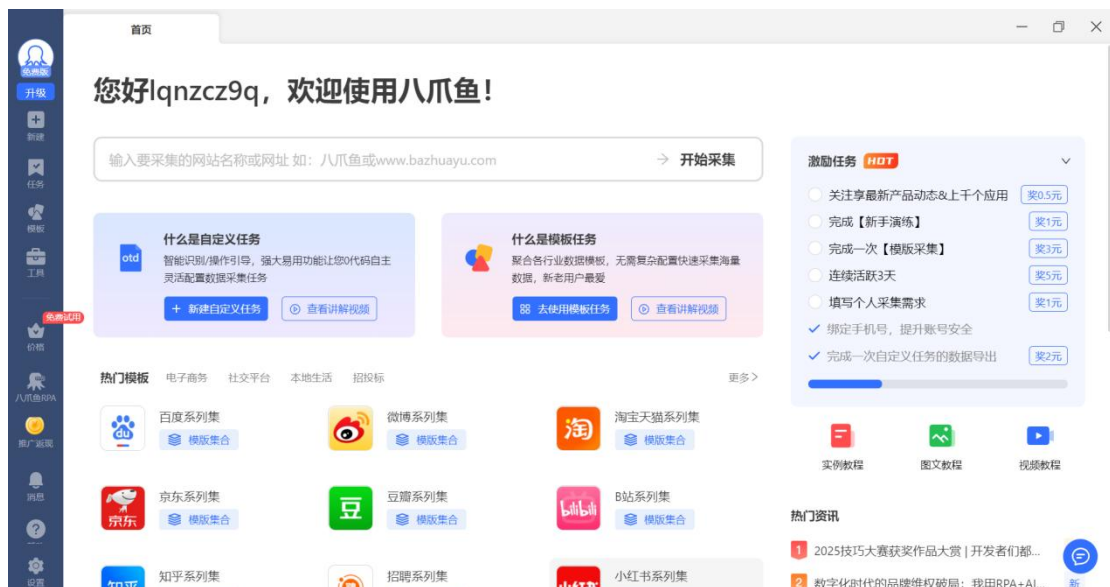


图 2-3 八爪鱼首页页面

任务实施步骤：

1、步骤 1：创建采集任务

打开八爪鱼采集器，单击【新建采集任务】，输入目标网页的 URL (如京东手机分类页)，

如图 2-4 所示，单击【开始采集】，工具会自动加载网页，如图 2-5 所示。需要注意的是，八爪鱼启动后内置浏览器会自动加载目标网页，部分电商网站可能要求登录后才能查看完整商品信息，此时需在页面右上角切换至【浏览模式】，再手动完成网站登录操作。

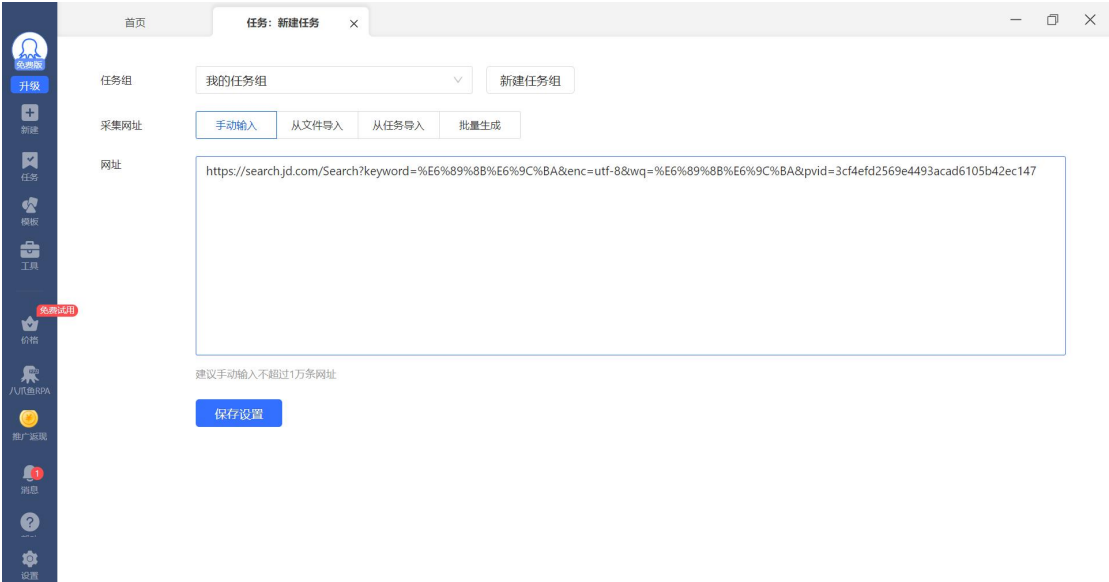


图 2-4 采集任务设置

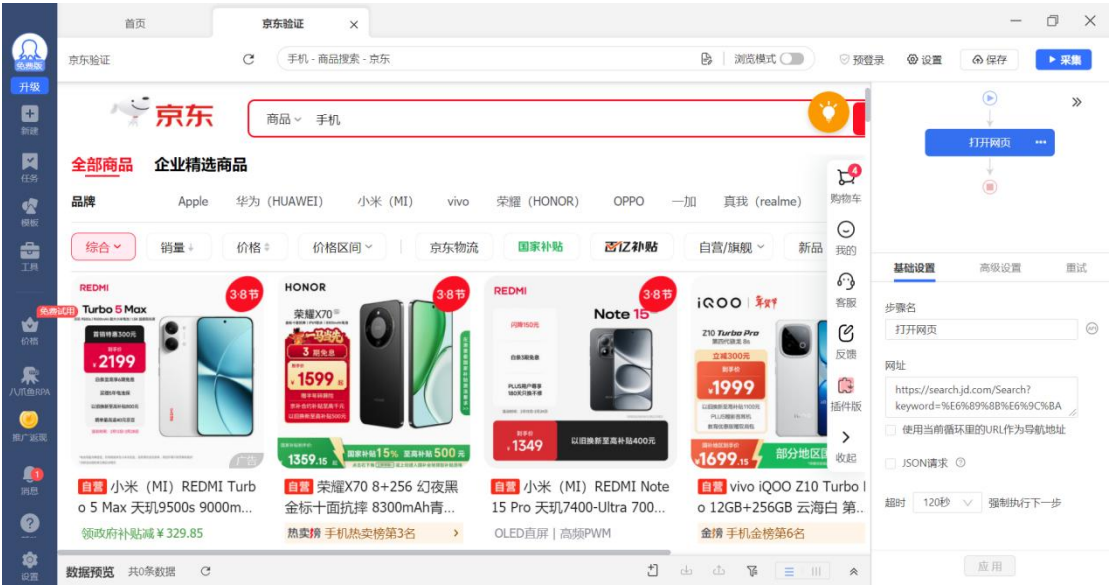


图 2-5 采集网页信息预览

2、步骤 2：批量输入、搜索关键词

网页打开后，按以下步骤批量输入多个关键词进行搜索：

①选中搜索框，在操作提示框中，单击【输入文本】。

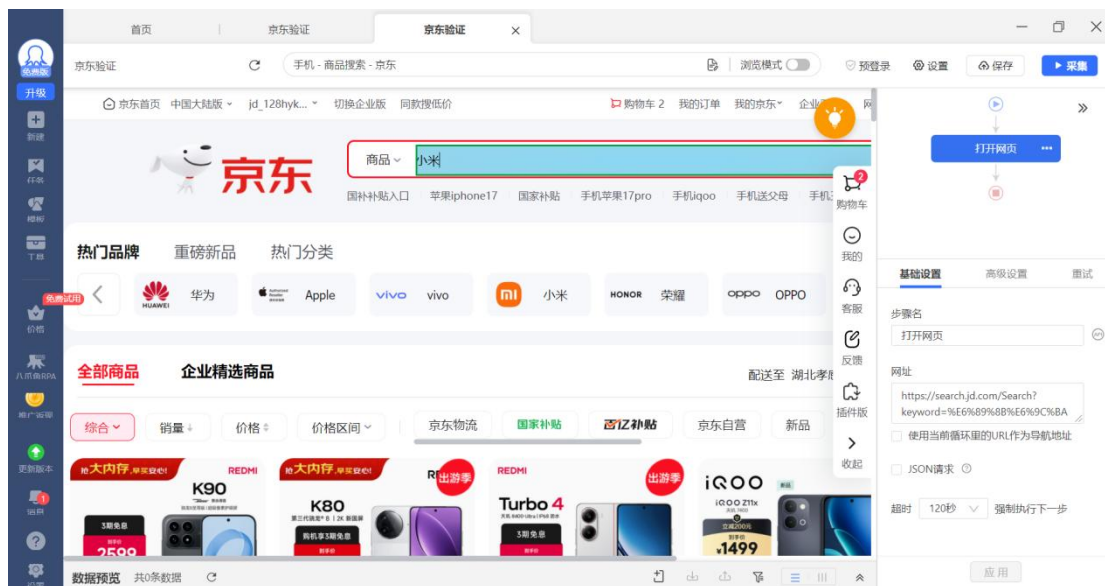


图 2-6 商品搜索页面

②单击操作框内的【批量输入文本】，将所有要搜索的关键词输入进去，单击【确定】。此时，右侧会自动生成一个循环输入文本的步骤，如图 2-7 所示。



图 2-7 批量输入文本页面

③选中【搜索】按钮，在弹出的操作提示框中单击【单击该按钮】，即可呈现关键词对应的搜索结果，如图 2-8 所示。

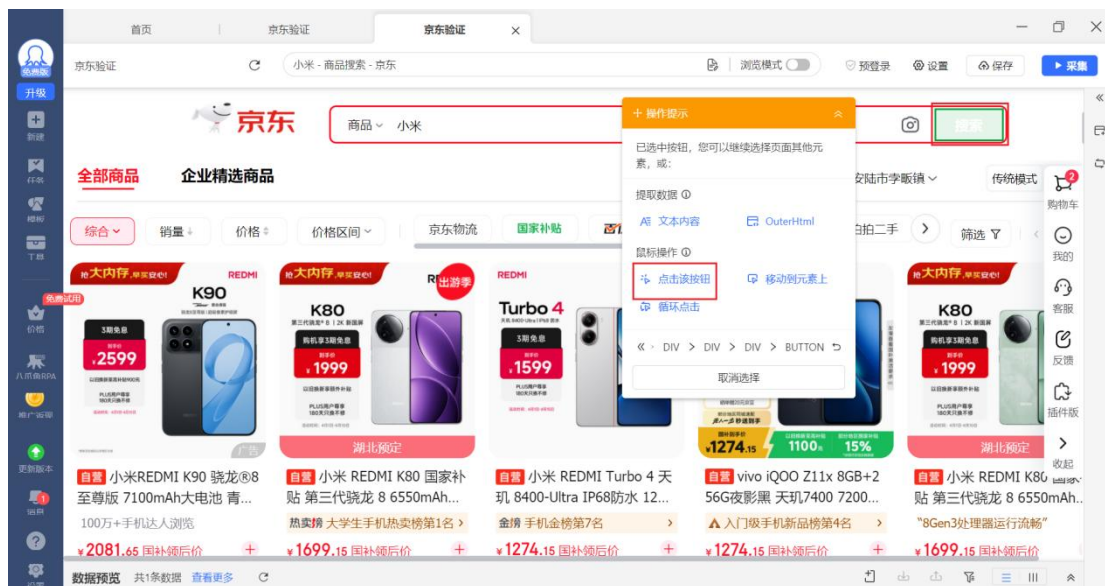


图 2-8 关键词搜索页面

3、步骤 3：创建循环列表

要采集商品列表中的全部数据，先选中包含所有所需字段的整个商品列表，接着在弹出的黄色操作提示框中依次单击【选中全部子元素】→【选择全部相似组】，最后单击【元素中数据内容】即可，如图 2-9 所示。

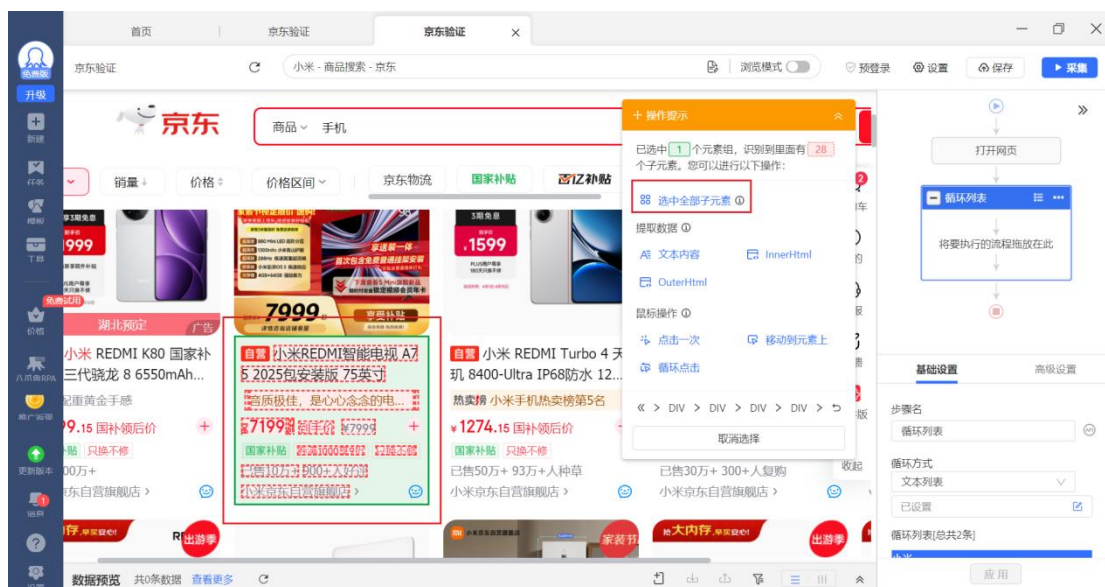


图 2-9 采集全部数据页面

4、步骤 4：设置采集字段

网页加载完成后，进入可视化采集模式，单击网页上的商品名称，工具会自动识别同类元素，在【数据预览】可删除多余字段，修改字段名，移动字段顺序。例如，将字段命名为“商品名称”，如图 2-10 所示。每个字段采集完成后，工具会在预览区显示采集结果，检查是否准确。

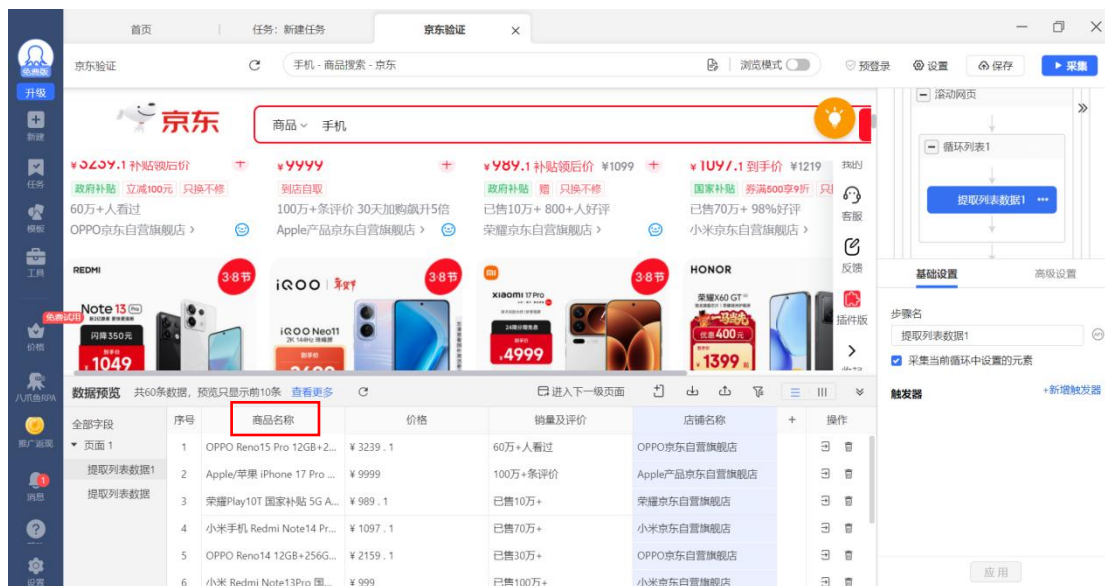


图 2-10 采集字段设置页面

5、步骤 5：设置采集规则

由于商品列表可能有多页，单击页面上的【下一页】按钮，在操作提示上单击【循环点击下一页】，创建【循环翻页】，设置分页采集规则，确保抓取所有页面的商品数据，如图 2-11 所示。



图 2-11 分页采集设置页面

6、步骤 6：执行采集任务

单击【开始采集】，工具会自动按照设置的规则抓取数据，采集过程中可实时查看进度和采集结果。其中【本地采集】是使用自己的电脑采集数据，【云采集】是使用八爪鱼提供的云采集，云采集可能需付费且涉及数据上传至第三方服务器，对敏感数据有安全风险，如图 2-12 所示。实时进度的查看如图 2-13 所示



图 2-12 信息采集页面



图 2-13 采集进度页面

数据采集完成后，单击【导出数据】，八爪鱼工具支持多种格式（如 Excel、CSV、HTML 等）的导出，根据需求选择合适格式的进行数据导出即可。此处我们选择导出为 Excel 格式，如图 2-14 所示。

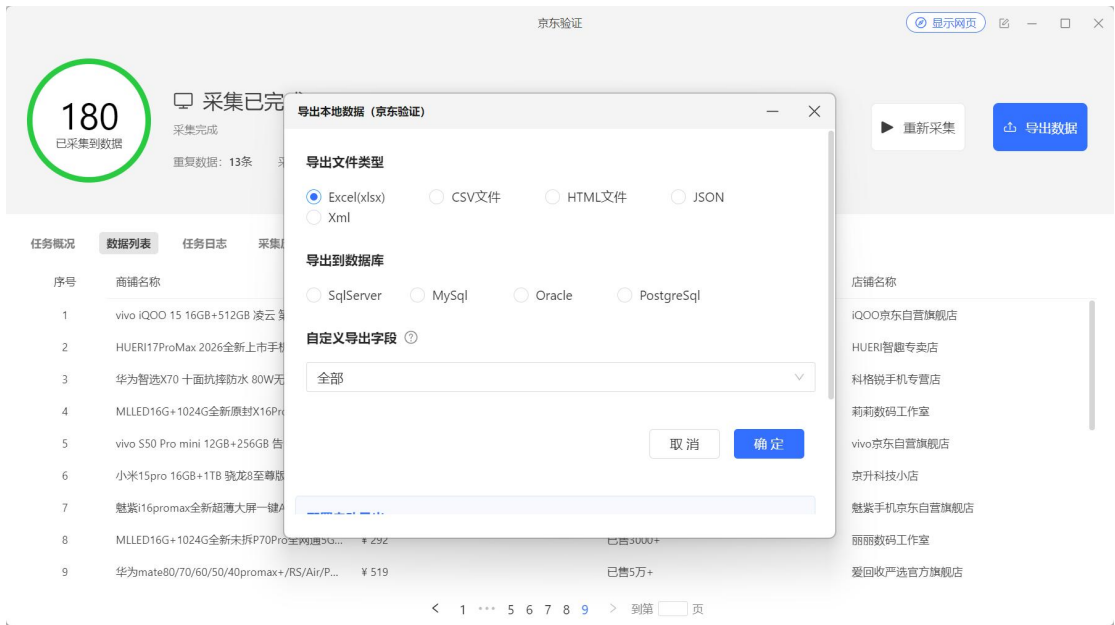


图 2-14 采集数据导出页面

7、步骤 7：实验结果分析

打开导出的 Excel 文件，检查数据的准确性和完整性：是否有缺失的字段（如部分商品没有店铺名称）、是否有错误数据（如价格格式异常），采集数据展示如图 2-15 所示。

#	A	B	C	D	E	F	G	H	I	J	K	L
1	店铺名称	价格	销量及评价	店铺名称								
2	小米 (MI) REDMI Turbo 5	¥1869.15	400万+人看过	小米京东自营旗舰店								
3	荣耀X70 8+256 幻夜黑	¥1574	已售90万+	荣耀京东自营旗舰店								
4	小米 (MI) REDMI Note15 Pro	¥1214.1	20万+条评价	小米京东自营旗舰店								
5	vivo iQOO Z10 Turbo Pro	¥1689	20万+条评价	iQOO京东自营旗舰店								
6	Apple 苹果 iPhone17/16/15	¥809	已售20万+	爱回收严选官方旗舰店								
7	荣耀X70 8+128 朱砂红	¥1374	已售90万+	荣耀京东自营旗舰店								
8	小米REDMI K80 国家补贴	¥1799	已售100万+	小米京东自营旗舰店								
9	vivo iQOO Z11 Turbo	¥2159.1	5万+条评价	iQOO京东自营旗舰店								
10	小米 (MI) REDMI Turbo 5	¥2124.15	50万+人看过	小米京东自营旗舰店								
11	荣耀X70 12+512 幻夜黑	¥1799.1	已售90万+	荣耀京东自营旗舰店								
12	Apple/苹果 iPhone 17	¥256¥5999	100万+条评价	Apple产品京东自营旗舰店								
13	小米REDMI Note14 金刚版	¥799.2	已售20万+	小米京东自营旗舰店								
14	荣耀Play10C 5G AI手机	¥719.1	已售20万+	荣耀京东自营旗舰店								
15	荣耀X70 8+128 幻夜黑	¥1374	已售90万+	荣耀京东自营旗舰店								
16	Apple/苹果 iPhone 17	¥256¥5999	100万+条评价	Apple产品京东自营旗舰店								
17	小米 (MI) Redmi 14C	¥479	已售40万+	小米京东自营旗舰店								
18	一加 Ace 6 12GB+256GB	¥2399	10万+条评价	一加手机京东自营旗舰店								
19	小米 Redmi Note13Pro	¥999	已售100万+	小米京东自营旗舰店								
20	荣耀Play10T 国家补贴	¥899.1	已售10万+	荣耀京东自营旗舰店								
21	一加 Ace 6 12GB+256GB	¥1899.2	2万+条评价	一加手机京东自营旗舰店								

图 2-15 采集数据展示

注意事项

- 1、采集必须设置访问间隔，避免触发网站反爬封禁；
- 2、业务敏感数据仅使用本地采集，不选用云采集；
- 3、不抓取手机号、住址等隐私信息，遵守网站 robots 协议；
- 4、导出后必须人工逐条校验数据，规则缺陷需重新采集。