**爬虫程序操作文档**

目录

[一、运行程序 2](#_Toc31584)

[1.1修改数据保存路径 2](#_Toc11830)

[二、淘宝操作页面 4](#_Toc22226)

[2.1步骤一： 4](#_Toc13849)

[2.1.1修改爬取页数方法： 5](#_Toc30626)

[2.1.2查看是否报错方法： 5](#_Toc5338)

[2.2步骤二： 5](#_Toc22508)

[2.3步骤三： 6](#_Toc15867)

[三、 京东操作页面 8](#_Toc16303)

[3.1扫码填价版本 8](#_Toc6620)

[3.1.1抓取商品信息 8](#_Toc15527)

[3.1.2扫码填价 8](#_Toc4801)

[3.1.2商智查销量 9](#_Toc7603)

[3.2机器填价版本 10](#_Toc13989)

[3.2.1操作流程 10](#_Toc24152)

[3.2.2多个webdriver配置 11](#_Toc6009)

[3.2.3爬取app端配置（安卓） 12](#_Toc7252)

[四、 注意事项 12](#_Toc25487)

[4.1天猫特征标签 12](#_Toc1319)

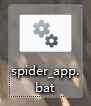
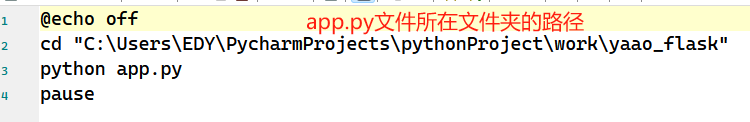
[4.2淘宝验证码问题（待优化） 13](#_Toc5016)

[4.3商智鼠标影响 13](#_Toc10953)

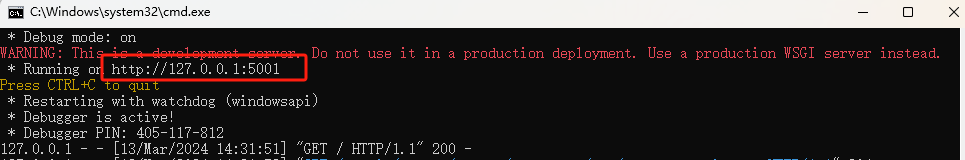
## 一、运行程序

**方法一：**

1. 先配置好bat文件（先创建txt文件，里面内容编辑为一下，然后把txt后缀改为bat）



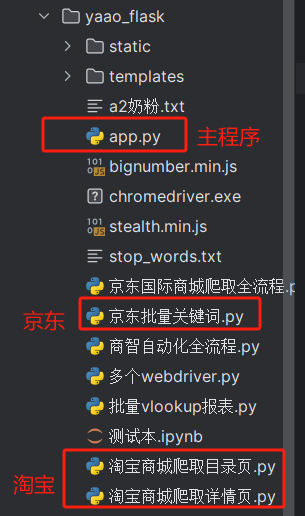
1. 双击bat文件，会弹出命令提示符窗口，和pycharm的运行窗口一致



1. 复制http://127.0.0.1:5001到浏览器打开，可以固定到浏览器顶部标签固定，就不用每次都复制

**方法二：**

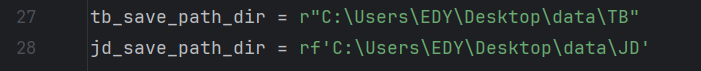
通过打开pycharm运行



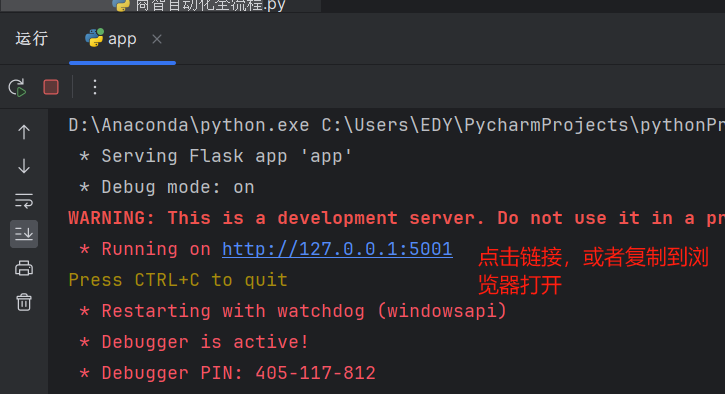
运行程序只需打开app.py文件，其余文件不用管

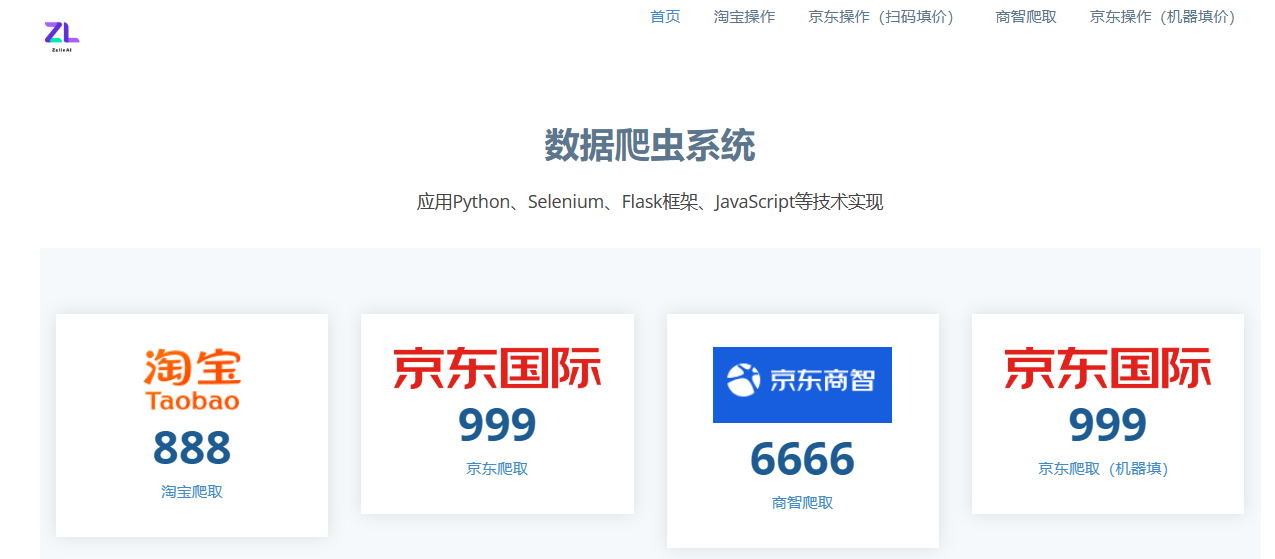
#### **1.1修改数据保存路径**

如果以下储存数据的文件夹路径还没改，需改成自己电脑的路径（自定义位置），改了以后就不用管了。代码在app.py文件的第三十行附近，将引号里的路径更换即可，一个淘宝，一个京东









## 二、淘宝操作页面

### 2.1步骤一：

爬取搜索页，只需输入关键词，点击启动（点击一次即可），每天第一次点击可能反应会慢一点，加载会久一点是正常现象，等待浏览器窗口自动打开，打开后需要扫码登录会有弹窗提示，如果一两分钟都没反应再去检查是否报错。

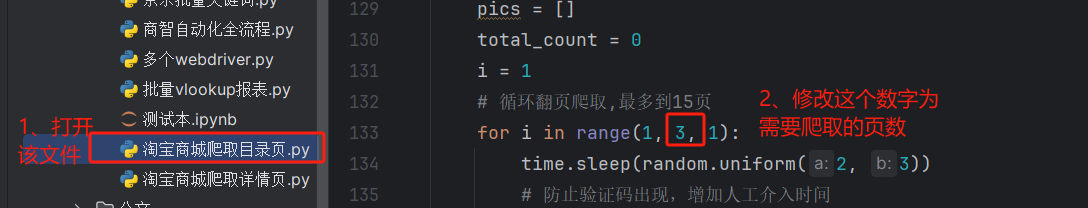


打开浏览器后，一般会先跳转到扫码登录页面，需要在80秒内扫码登录，登录后就开始自动抓取商品信息



#### 2.1.1修改爬取页数方法：

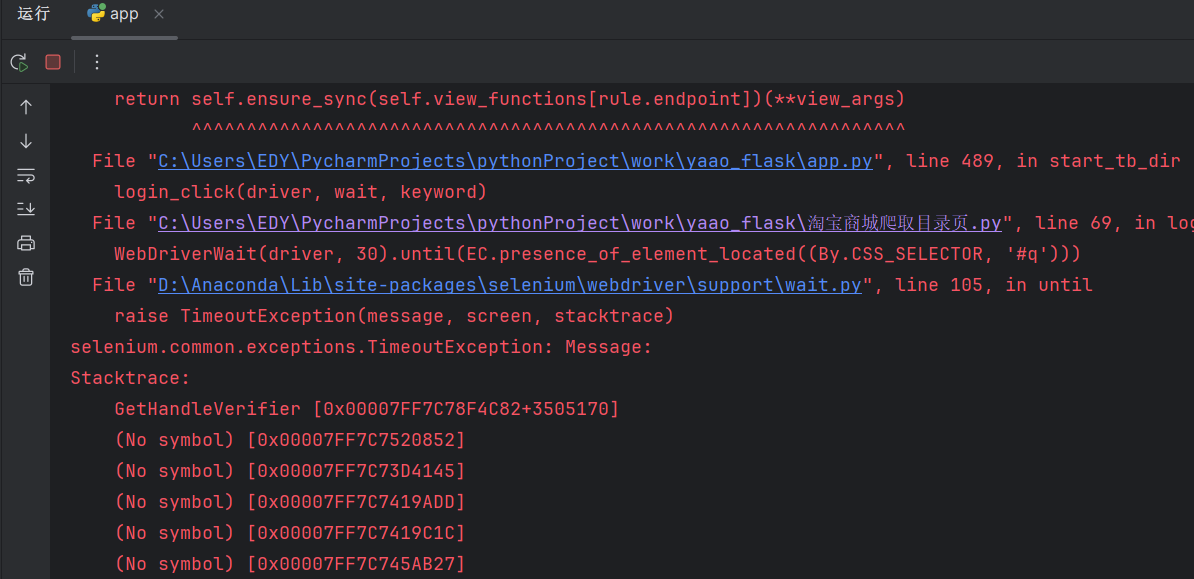
目前设置了爬取（按照销量）前三个搜索页的商品信息，如果需要更改爬取页数，打开“淘宝商城爬取目录页.py”这个文件，在大约130行左右，修改爬取页数（该数字不能超过实际商品的展示页数容量）



#### 2.1.2查看是否报错方法：

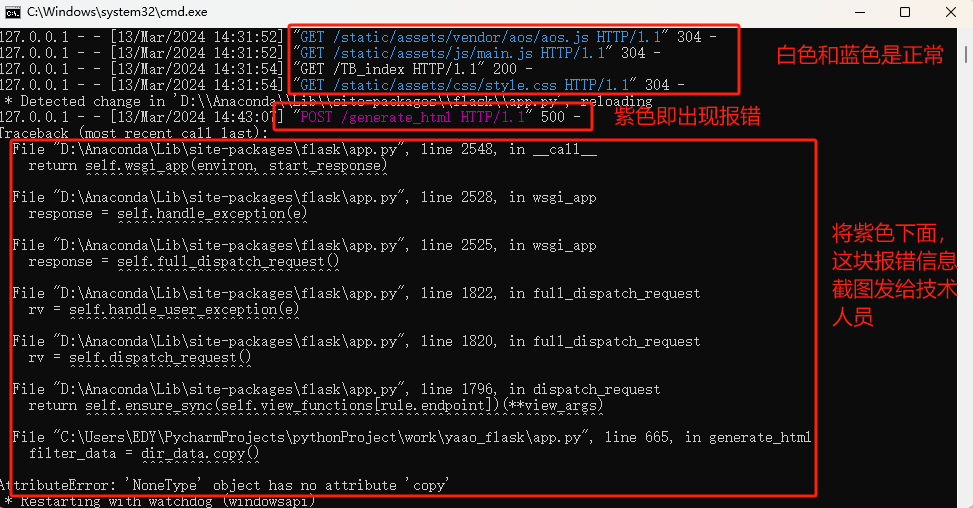
**方法一：**

如果是打开pycharm运行的，切换到pycharm软件，查看运行窗口，报错界面类似如下：（报错联系技术人员）



**方法二：**

如果是通过bat可执行文件直接运行，打开命令提示符窗口观察，

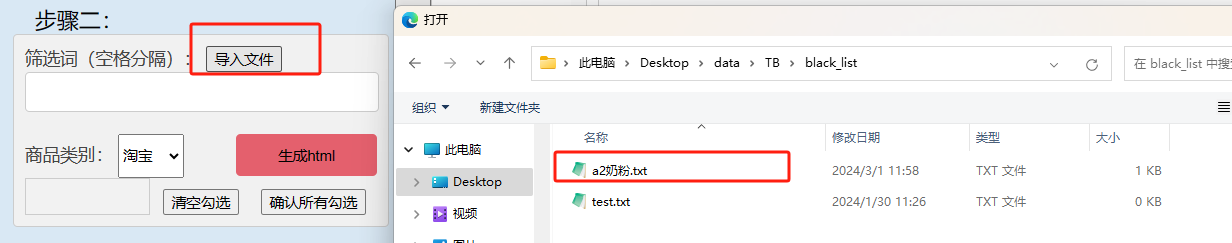


### 2.2步骤二：

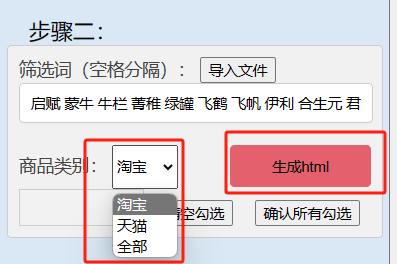
2.2.1手动输入筛选词，例如（飞鹤 飞帆 伊利 合生元 君乐宝 美赞臣 贝臻）每个词语之间一个空格分隔。目的筛掉标题中出现含以上的词语再生成html。

2.2.2点击导入文件。该文件要自己先创建对应txt文本（命名自定义，一般一个品类对应一个txt），把筛选词都添加都txt里面，格式一样是空格分隔，然后在弹窗窗口找到txt所在文件路径选择文件再点击打开，即可导入所有筛选词，该功能只是方便以后不用每次都手动输入筛选词。

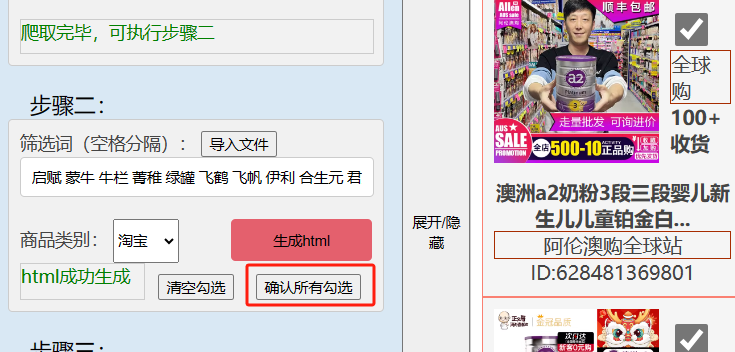
两种方法选一种即可，筛选词可根据点击展开词频帮助输入，也可以留空都不输入任何筛选词。



2.2.3接下来，在下拉菜单选择商品品类（默认是只选淘宝的），点击“生成html”。

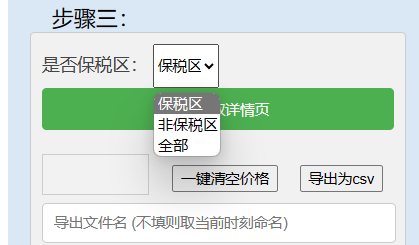


2.2.4在右侧商品信息页面根据主图勾选合适的商品，全部勾选完毕后一定要点击“确认勾选”



### 2.3步骤三：

选择是否保税区（默认是只选择保税区的），然后点击启动



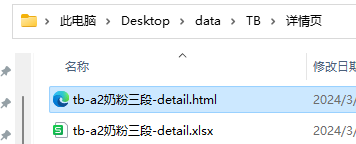
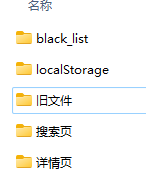
会打开浏览器，需要点击右上角一键登录，然后扫码登录，进入自动爬取



爬取完后，会在网页的右侧展示详情页所有信息，扫码填价 —》导出表格



如果该网页数据因为点击跳转到其他页面丢失原页面信息（因此，如果想要打开某个商品链接，最好不要直接点击打开，建议右键选择在新标签页打开），可以在“详情页”文件夹里面找到对应关键词的html文件，打开一样可以填价格并导出表格。



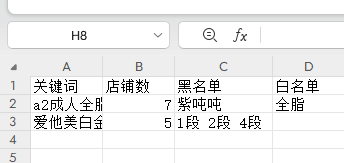
## 京东操作页面

### 3.1扫码填价版本

（抓取商品信息 ——》 扫码填价 ——》 商智查销量）

#### 3.1.1抓取商品信息

和淘宝一样，可以手动输入商品信息，也可以导入excel配置文件。配置文件中格式内容如下图，上下这两张图是等价的。填完商品配置信息，填生成表格名称（自定义），点击启动，等待浏览器打开并自动抓取。



**如果是手动输入的方法，注意以下：**

①黑名单和白名单可都不填，留空

②一个关键词如果需要多个黑名单词，多个黑名单词语间空格分隔，如左图，爱他美白金三段对应的黑名单为（1段 2段 4段）

③如果批量爬多个关键词，不同关键词间用英文逗号分隔，下面的三个框信息也要跟着逗号，即便某个关键词没有黑名单或白名单，也要逗号占位

**导入文件的方式则相对简便一点，只需建好excel表，格式和左图一致即可。**

①黑名单和白名单可都不填，留空

②一个关键词如果需要多个黑名单词，多个黑名单词语间空格分隔

③加关键词继续填多几行，一般不要超过十个，否则可能触发反爬验证码

#### 3.1.2扫码填价

1. 下图从左往右从上往下，字段依次是（二维码、主图、店铺索引、关键词、sku号、评论数、标题、店铺名、sku属性名称）
2. 其中关键词、标题、sku属性名，如果展示不全，将鼠标悬浮于其上面，会显示全部文字

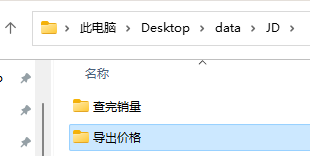
3、如果需要打开商品详情页链接，右键主图，在新标签页中打开

4、有六个价格框，分别对应（拍1件、拍2件......拍6件的价格），填多少个框都可以，大多数情况只需要填第一个价格框

5、如果遇到非想要的商品，如下图右边这个商品，则一个价格都不填，导出表格时程序会自动筛掉这个商品。



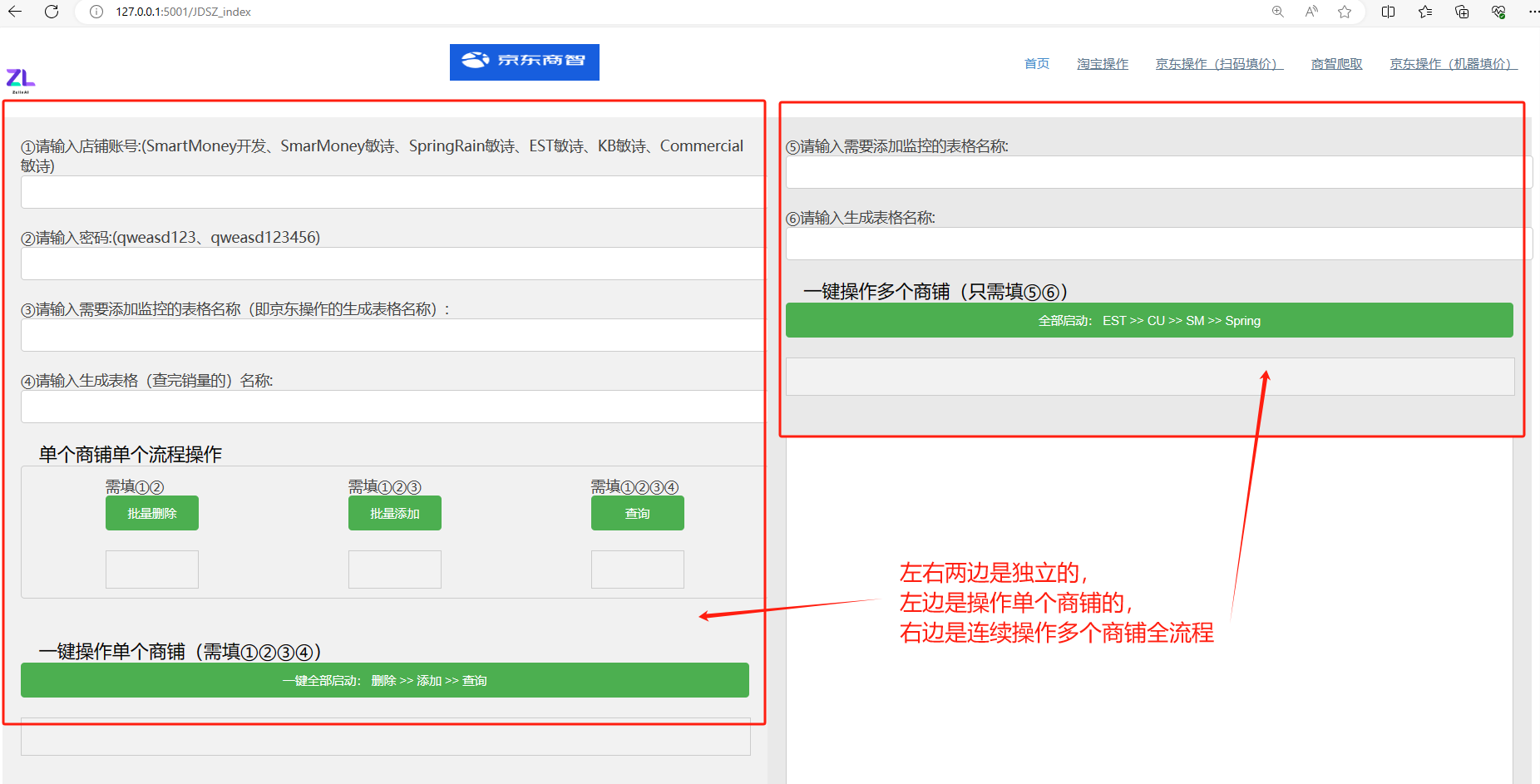
6、在填完所有需要填的价格后，点击左下角“导出excel”按钮，导出表格在“导出价格”文件夹中



#### 3.1.2商智查销量

按照网页的操作提示操作

查完销量的表格在“查完销量”文件夹中，即为最终表格



### 3.2机器填价版本

（抓取商品信息 —》 勾选符合的sku —》 商智查销量 —》 抓取app端价格数据）

#### 3.2.1操作流程



⑤查完后，再点击查价格，查完价格数据表格在“导出价格”文件夹中

④导出成功后，再点击查销量，查完销量数据表格在“查完销量”文件夹中

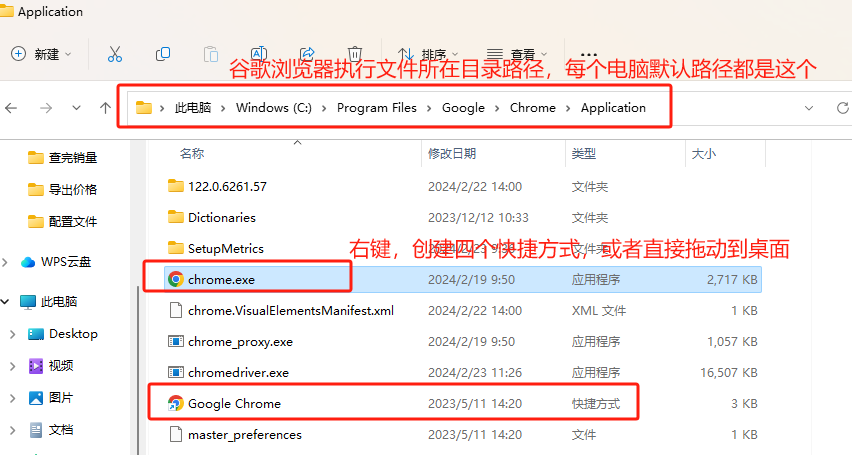
③在右侧展示面板，勾选完所有符合的商品，再点击导出excel按钮，导出文件在“已筛选完”文件夹中

②填入生成表格名称，再点击启动

①填入抓取信息或者导入文件

#### 3.2.2多个webdriver配置

针对于登录京东商智后台进行 删除SKU、增加SKU、查询销量的操作，并且目前有四个商铺，因此需要分别给每个商铺配置一个浏览器，互相独立防止登录信息相互干扰。





分别右键四个快捷方式，t1 ~ t4里面的信息都要分别配置

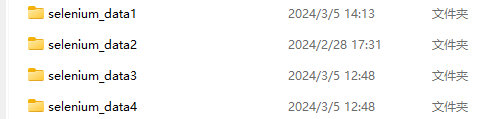
--remote-debugging-port=9222

远程调试端口，t2即对应9223，以此类推

--user-data-dir="D:\selenium\_data1"用户数据文件夹，t2即对应"D:\selenium\_data2"，以此类推

快捷键t2即对应Ctrl+Alt+2，以此类推

先在D盘新建四个空文件夹，分别作为四个商铺的user-data-dir（即用户数据文件夹）



|  |  |  |
| --- | --- | --- |
| 名称 | 目标参数信息 | 快捷键 |
| t1 | "C:\ProgramFiles\Google\Chrome\Application\chrome.exe" --remote-debugging-port=9222 --user-data-dir="D:\selenium\_data1" | Ctrl+Alt+1 |
| t2 | "C:\ProgramFiles\Google\Chrome\Application\chrome.exe" --remote-debugging-port=9223 --user-data-dir="D:\selenium\_data2" | Ctrl+Alt+2 |
| t3 | "C:\ProgramFiles\Google\Chrome\Application\chrome.exe" --remote-debugging-port=9224 --user-data-dir="D:\selenium\_data3" | Ctrl+Alt+3 |
| t4 | "C:\ProgramFiles\Google\Chrome\Application\chrome.exe" --remote-debugging-port=9225 --user-data-dir="D:\selenium\_data4" | Ctrl+Alt+4 |

#### 3.2.3爬取app端配置（安卓）

电脑中Python环境安装以下包

pip install pocoui -i <https://pypi.tuna.tsinghua.edu.cn/simple>

Airtest官方文档：[欢迎使用 - Airtest Project Docs (netease.com)](https://airtest.doc.io.netease.com/)

**总结：**

1. 在手机设置中打开开发者模式（一般在“关于手机”找到版本号位置，连续点击几次即进入开发者模式）
2. 点击打开USB调试按钮开关
3. 需要选择USB配置选择为“MTP（多媒体传输）”
4. 电脑通过数据线连上手机
5. 打开手机首页，亮屏，并且京东软件在首页
6. 点击“查价格”按钮进行爬取
7. （过程中出现是否连接的询问弹窗，点击确认即可）

## 注意事项

#### 4.1天猫特征标签

爬取淘宝网站搜索页的时候，程序需要区分淘宝格式网页和天猫格式网页，但是天猫的标签会经常改动，一般节假日淘宝官方都会更换天猫标签的图标。听不懂的话，直接看图好理解。

IMG_256

类似这种标签还有很多，传统节假日，新兴购物节都会变换，一般同时只会出现一个，但是不同节日官方会替换对应的标签

IMG_256

IMG_256



为了要程序正确区分淘宝和天猫的区别，如果遇到新的特征标签，则需要如上图操作，在淘宝网页先右键特征图标，然后点击复制图像链接，将该图像链接加入到代码中的列表里，代码在“淘宝商城爬取目录页.py”这个文件，在大约110行左右，继续按照上面的格式往下添加一行即可，每次官方更新图标，如果新的图标没有在代码列表里，都要新增进去。



#### 4.2淘宝验证码问题（待优化）

淘宝爬取详情页过程的验证码问题，目前一次性频繁的循环打开过多淘宝或者天猫的商品详情页链接，会触发出现滑动验证码，目前的解决方案是，识别弹窗网页元素，等待人工处理，滑块消失再继续执行后续操作。（但根据运营反馈，有小概率会出现网络错误的弹窗，即便人工关闭下一个商品页面依旧持续出现弹窗。

#### 4.3商智鼠标影响

商智删除和增加SKU环节，鼠标尽量不要做其余操作（移动、点击）。Chrome爬虫浏览器也不要最小化窗口。以上受鼠标影响原因是商智中的按键本身设计的问题。

#### 4.4 chromedriver版本与谷歌Chrome版本

chromedriver版本与谷歌Chrome版本需一致，后期可以设置电脑谷歌禁止自动更新，否则需保持chromedriver版本与谷歌Chrome版本一致。chromedriver.exe文件需在两个位置。一个在python.exe的同个文件夹下（一般是D:\Anaconda\下），另一个在chrome.exe的同个文件夹下（一般是C:\Program Files\Google\Chrome\Application\下）。

#### 4.5路径问题

京东的生成文件路径可能会有点杂乱，最好不要填同名excel表格文件名

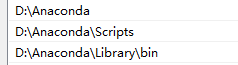
#### 4.5环境问题

安装文件路径父级目录尽量不要含中文名称文件夹

1. PC端

python环境：Python3.11（我装的是anaconda带的python.exe）

配一下电脑设置的path路径



编辑器：Pycharm（专业版和社区版都可）

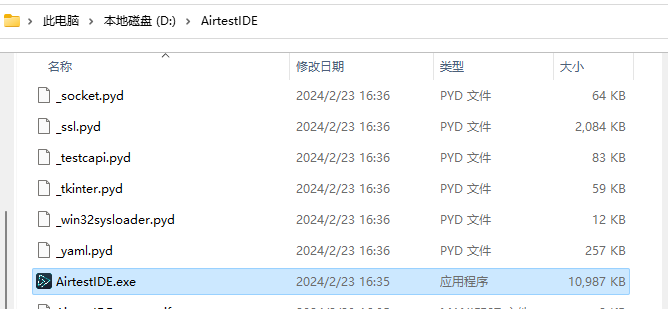
一些python包缺哪个pip哪个就行，例如selenium、jieba这些等等

pip install xxx -i <https://pypi.tuna.tsinghua.edu.cn/simple>

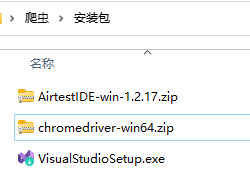
1. App端

pip install pocoui -i <https://pypi.tuna.tsinghua.edu.cn/simple>

自动化采集使用的是airtest框架，解压压缩包，点击AirtestIDE即可



需要依赖 Visual C++ 环境，安装即可



连接手机过程可能被360等软件阻止，需先关闭360