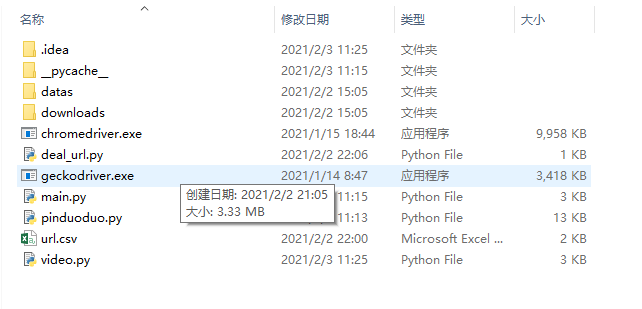
# 爬虫使用

## 文件目录结构

爬虫文件保存在桌面



目录结构如下



前两个文件为编译器自动生成文件。Geckodriver和chromedriver为浏览器驱动。

Datas：

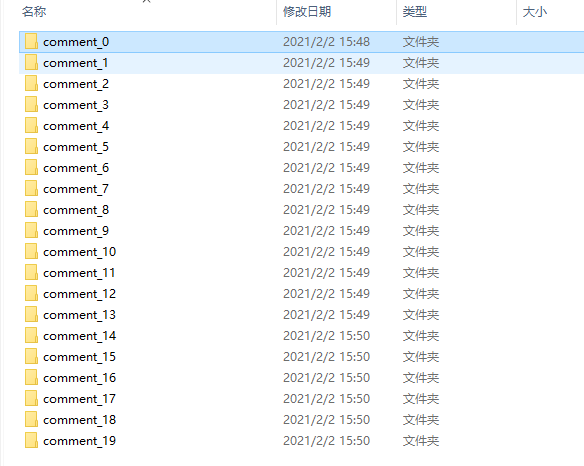
保存文本内容，包括商品信息，商品sku，评论，以及图片路径等。每个商品都保存在自己id命名的文件下。



# f91efdd7e72cb0ca844efa2ccab9568

Downloads：

保存相应的图片，分为商品图片，详情图片，sku图片和评论图片。



Pinduoduo：

爬取代码

Url.csv:

需要爬取的网址文件

Deal\_url:

对url.csv预处理代码

Video：

爬取视频网址代码

Main

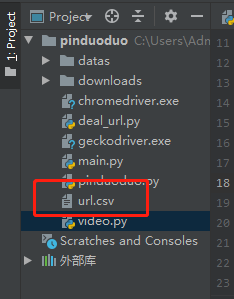
主程序

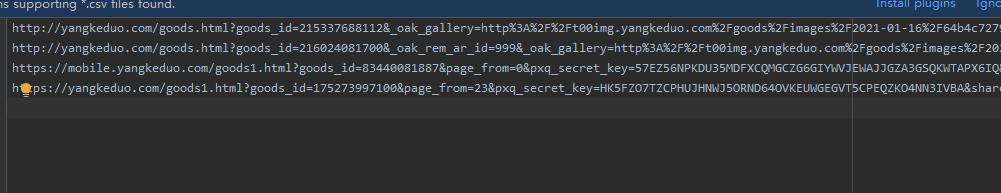
## 程序启动

右键点击桌面pinduoduo文件使用pycharm打开



可以在url.csv输入需要爬取的网址





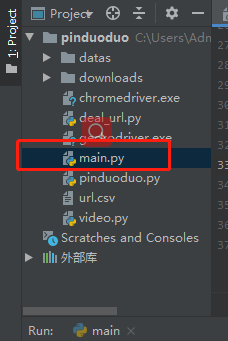
模拟浏览器

1. 查找chrome所在目录，打开cmd切换到该目录下
2. 输入“chrome.exe --remote-debugging-port=9222”命令打开浏览器，启动成功。（可以输入”127.0.0.1:9222/json”网址，如果有内容则监听成功，如果没有大概率是因为之前启动了chrome，关闭之前启动的chrome包括后台，重启）登陆账号：

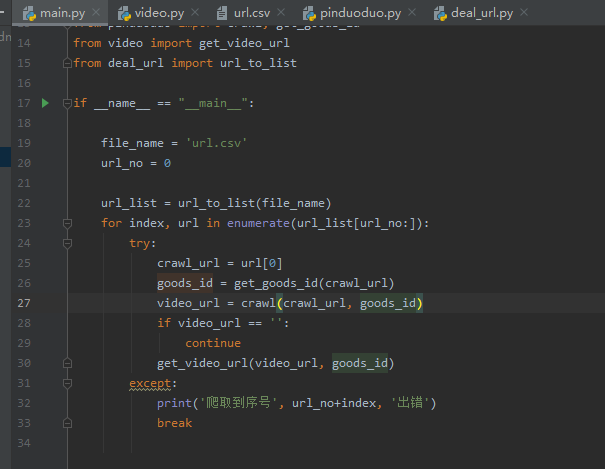
在模拟启动的浏览器输入拼多多网页版网址”http://yangkeduo.com”，进入主页个人中心登录账号（手机验证码登录）

启动程序

在pycharm中打开main.py,

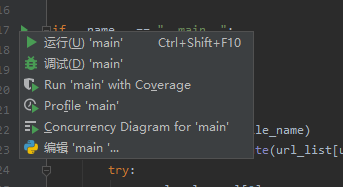


双击即可



这里第20行有一个url\_no，这个是断点续接，当爬虫出错时控制台会输出”爬取到序号？出错“，解决问题后继续爬取就可以将url\_no设置为？

在17点击绿色三角，选择运行，程序就启动了



# 环境配置

Windows环境下配置：

安装软件python3.7,pycharm2019.3,chrome,chromedriver

Python3.7

<https://mp.weixin.qq.com/s/JqU7mZUOEmVNE9SvYVClzg>

Pycharm2019.3

<https://mp.weixin.qq.com/s/JF5GL2OHUi-NsBZ3HP7q-A>

chrome官网下载即可

Chromedriver

<http://npm.taobao.org/mirrors/chromedriver/>

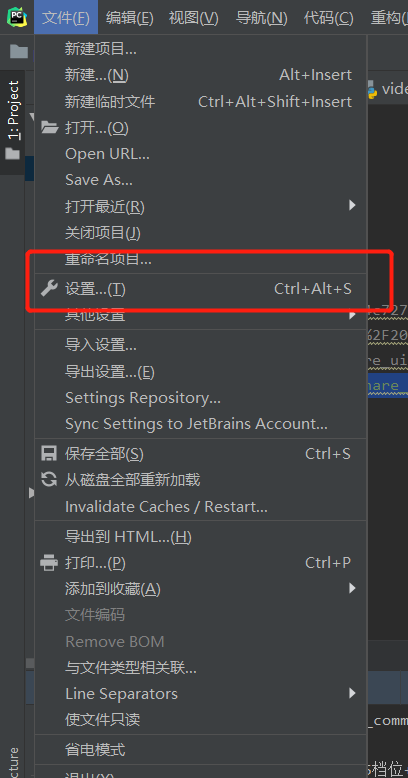
该网址有版本对应关系，和电脑的chrome版本要对应上，下载后解压，将driver文件放在项目根目录下。

安装必须的包，所需要的包都已经打包好，保存到了根目录下requirements.txt，将该文件复制粘贴到python安装后的python》python37》scripts目录下，然后cmd切换到该目录，输入以下命令等待安装包，时间较长。

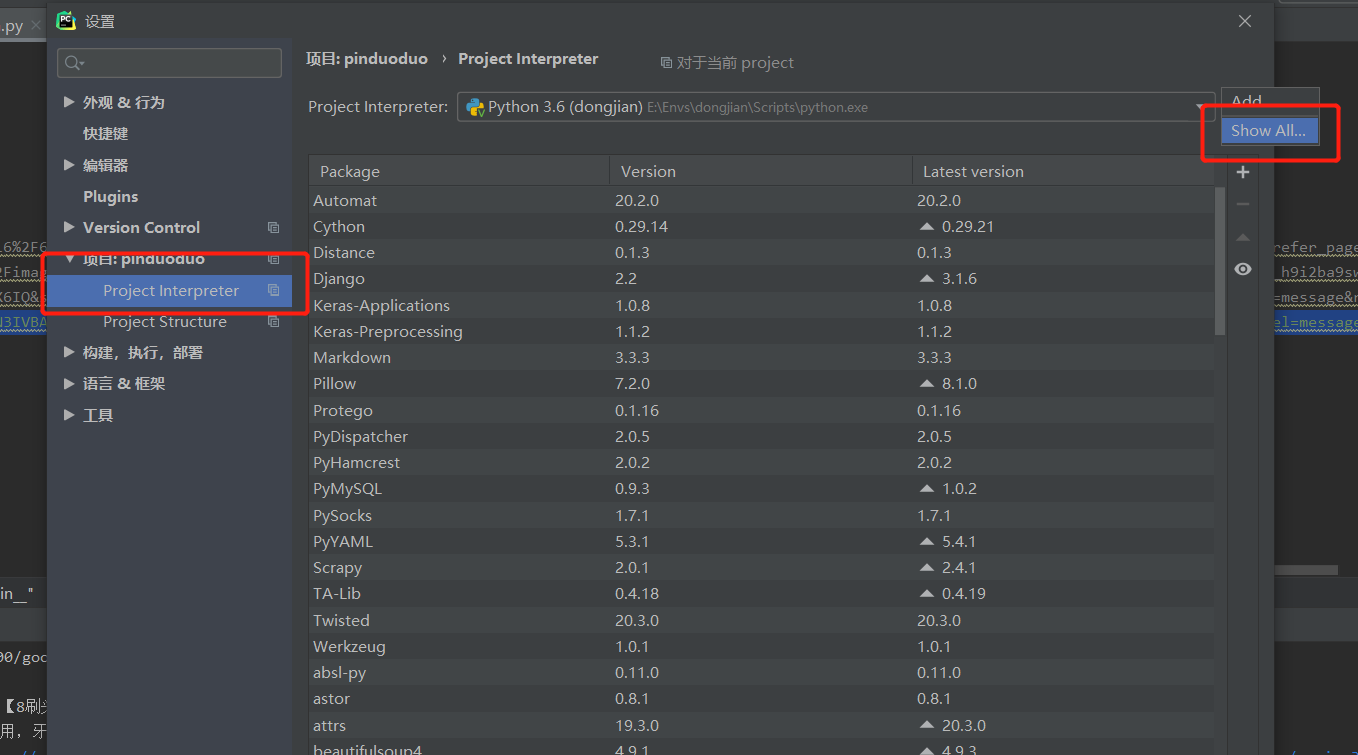
Pip3 install -r requirements.txt -i <https://pypi.tuna.tsinghua.edu.cn/simple>

右键点击pinduoduo文件用pycharm打开，打开后设置编译环境

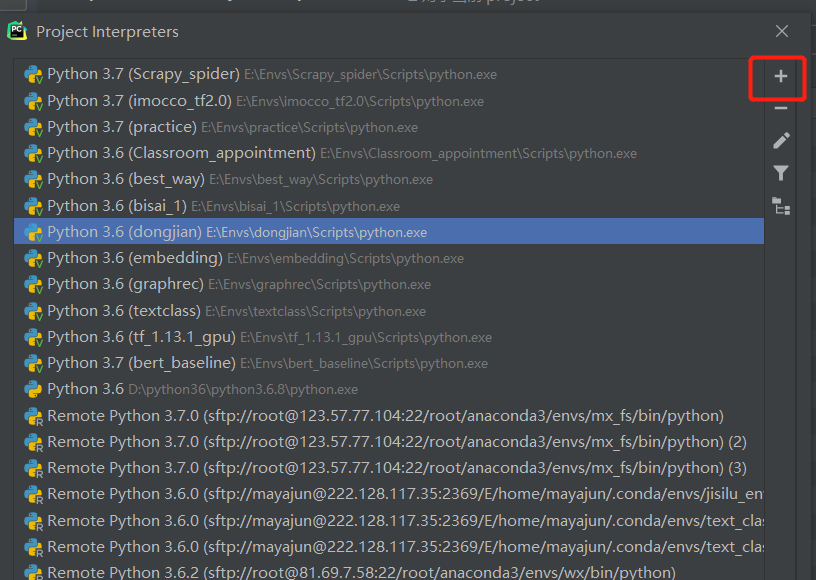
选择文件》设置



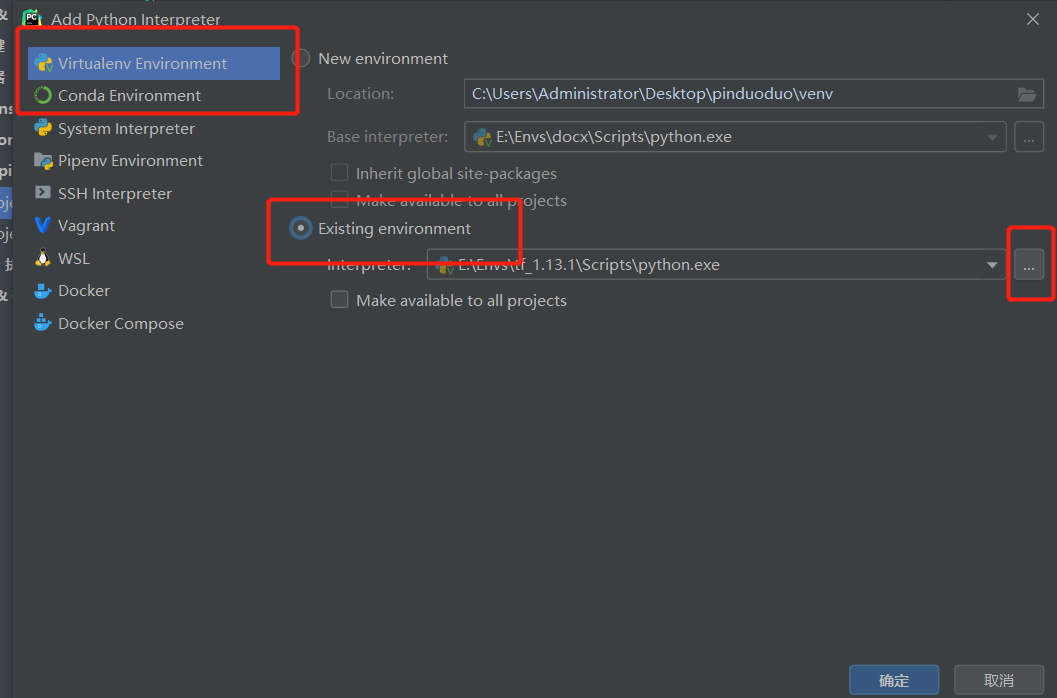
选择project interpreter，点击右侧设置选择show all，新下载的情况下框里面是空的



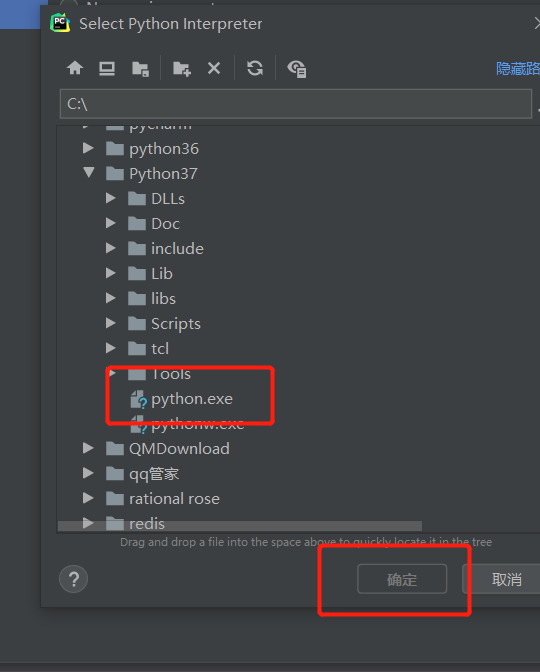
点击加号



选择已存在的环境，并点击三个点选择环境



找到安装好的python目录，选中python.exe,然后一路确定。



# 问题处理

## 浏览器弹出验证

目前见到的验证方式有两种，一种是根据箭头划手势，一种是看图选择物体。

解决：

可以尝试验证，如果通过验证，根据控制台输出的断点序号就可以继续爬取。如果不通过可以尝试用其他浏览器登录并验证，然后再在模拟浏览器刷新页面再试一次，基本可以通过。如果还不能通过，只能切换账号。

## 商品信息消失



大概情况如上图所示，这种情况并不是真的售空了，而是账号被限制了

解决：

尝试刷新页面，重新登陆等操作。在测试中，大部分情况，重新登陆即可，也有部分情况等待十到二十分钟就会解除限制，一般超过20分钟还不可以这个账号应该没有在拼多多消费过。

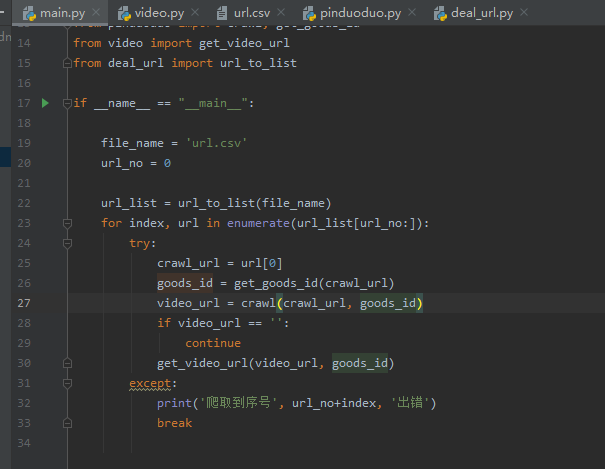
## 程序异常终止

浏览器并没有显示上述两种情况，但是代码报错并停止。这种情况大多是因为js没有完全加载

解决：

输入断点序号重新启动，注意尽量让模拟的浏览器整个浏览器不被遮挡。

## 视频网址



在代码中第30行为爬取网址的函数调用，这里不需要的话可以注释掉（ctrl+/）

如何爬取视频网址：

拼多多这个视频是采用blob加密模式加密了网址信息，直接从网页抓取的网址是不能用的，当浏览器切换成手机模式之后加密就会自动暴露成m3u8格式的视频网址。该格式的视频是不能直接下载的，需要解析出这个地址下对应的多个后缀名为ts的文件，并且下载拼接起来。

由于，浏览器不能多个端口监听，因此我才用模拟新的浏览器以手机模式拿到这个m3u8网址。（可能存在拼多多发现异常被封拿不到的情况，由于这种模拟情况无法登陆账号，所以被封就没有什么办法了）

如果想要下载该视频，使用m3u8工具输入网址，并选择输出模式，就可以实现下载（比较耗时）

