# Python期末报告

## 项目介绍

相关代码及实验课做的其他代码(如中文乱码，淘宝爬取，cookie)在相关代码文件夹中.

### 基于Scrapy的网页：爬取豆瓣top250

### 二）12306的自动登录

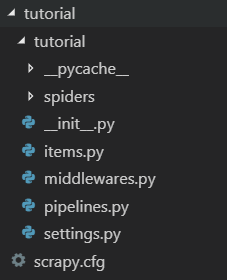
### 三）redis分布式爬虫

### 四）百度百科的深度优先及广度优先的爬虫

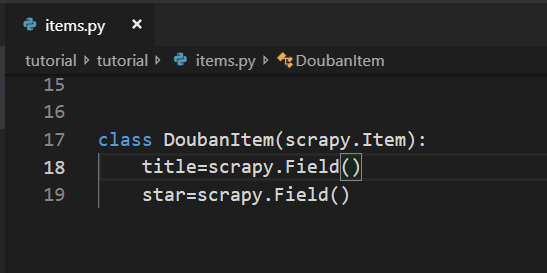
## 代码及效果截图

### 基于Scrapy的网页爬取

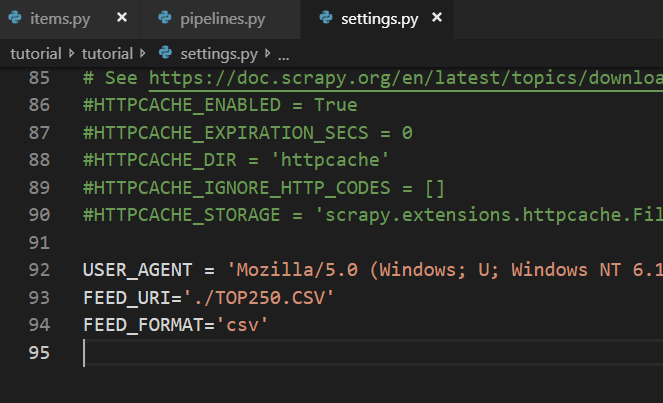
1.在cmd中输入语句创建scrapy项目，会如图生成目录



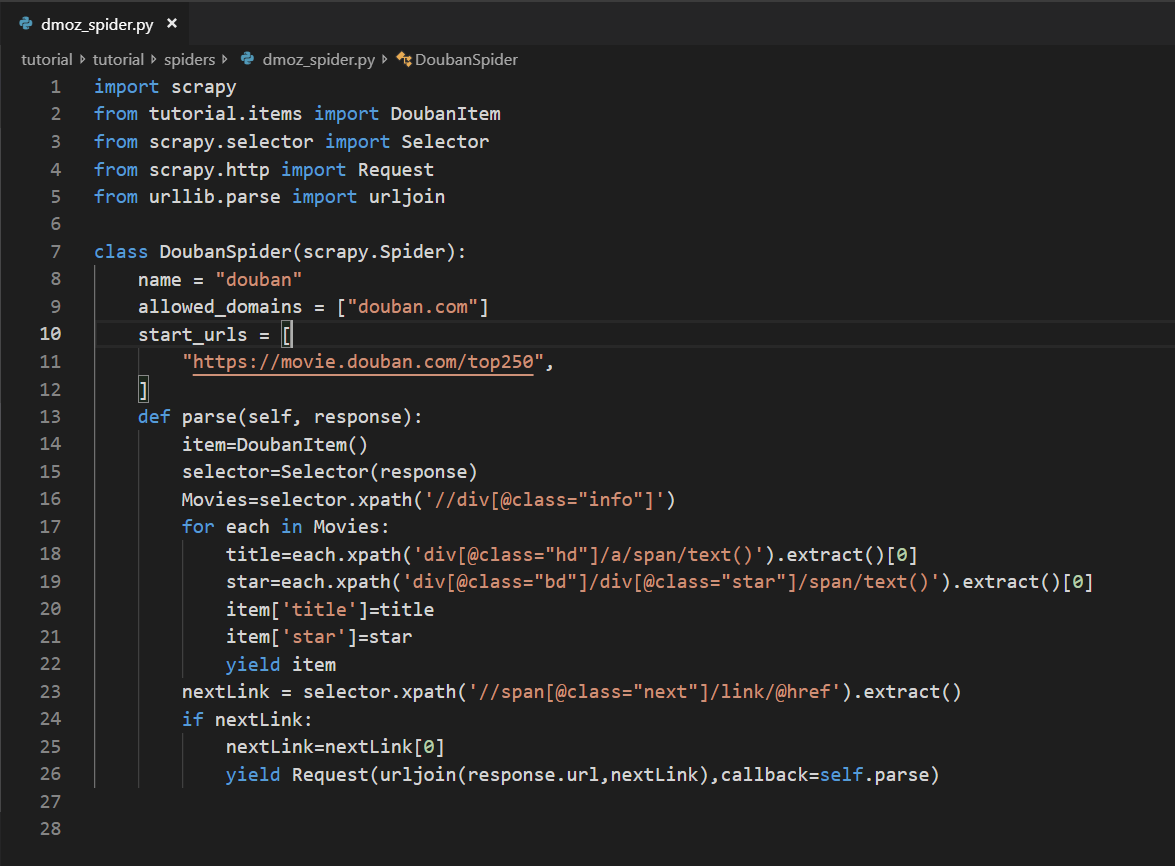
2.定义items，里面写需要存储的数据项



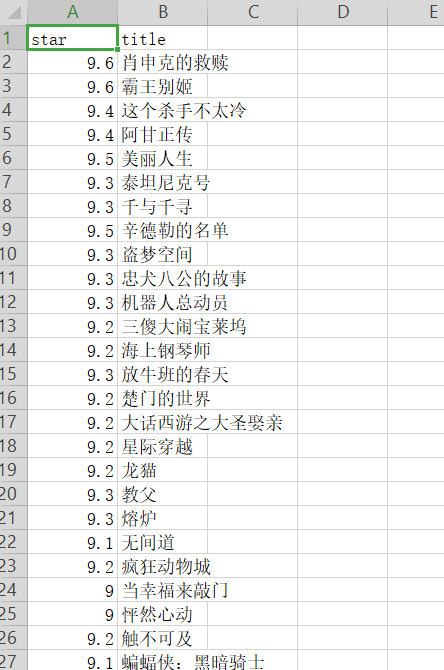
3.定义items存储地址和类型



4.创建spider去爬取数据

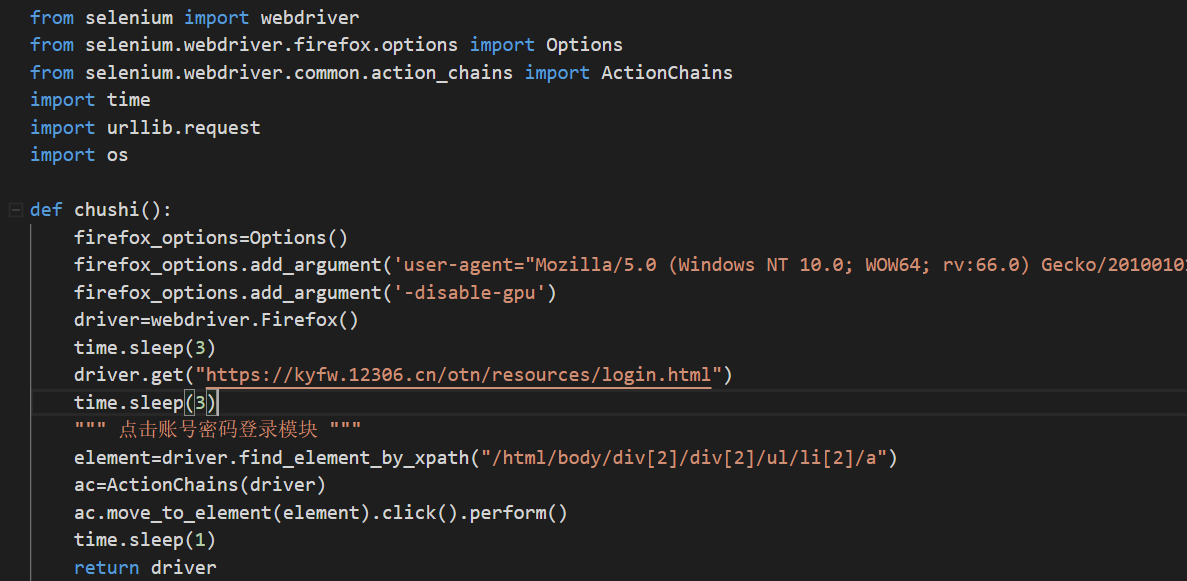


效果截图：

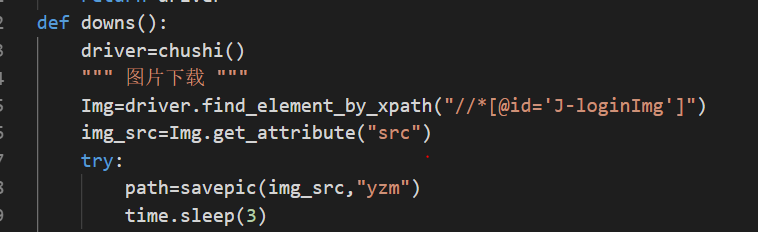


### 二）12306的自动登录

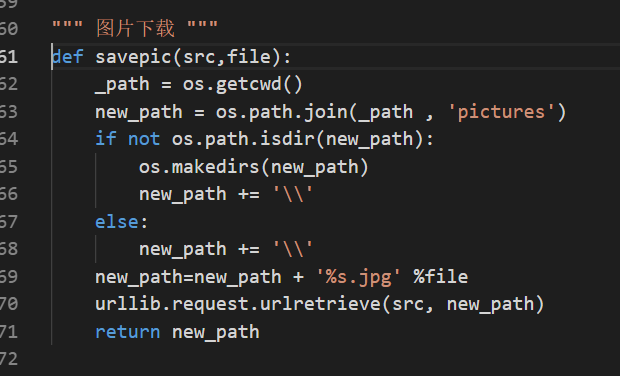
1. 自动登录利用的是selenium，首先进行初始化：
   1. 首先登录到该网站
   2. 因为登录模块基于js完成所以利用ActionChains模拟点击时间，初始化完成。



1. 完成初始化，介绍down()方法，完成其余操作
2. 图片下载
   1. 图片下载，获得该图片id,写一个方法下载它，path为该图片路径



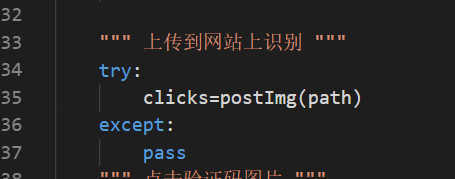
* 1. Savepic方法



* 1. 效果图：

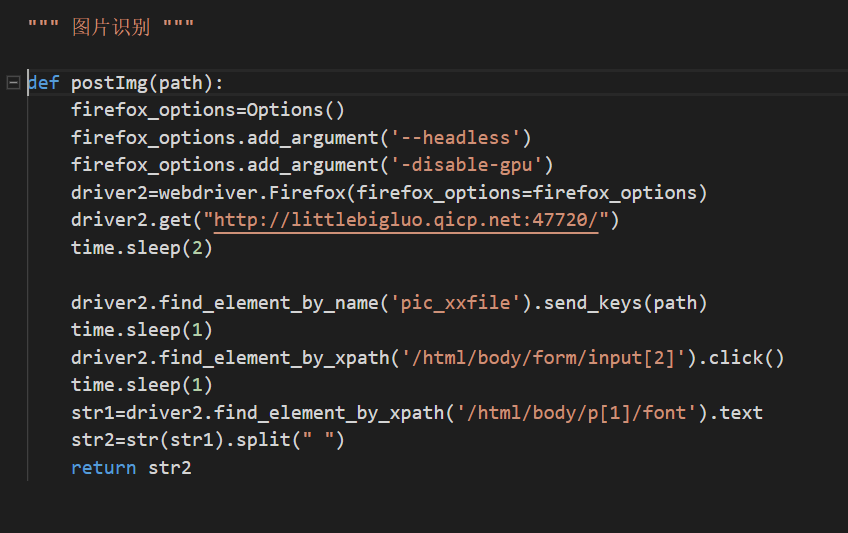


1. 将图片上传到识别网站
   1. down()中代码：path来源与图片下载部分，是图片的路径



* 1. postImg方法：

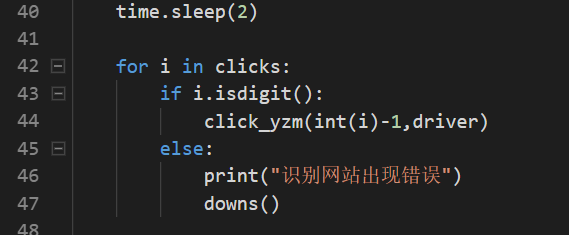
重开一个driver,打开识别网站。在上传文件的标签里键入地址，找到上传按钮模拟点击事件，然后得到一个字符串为正确答案的数字，然后分割后将字符数组传回down（）函数中。



* 1. 效果

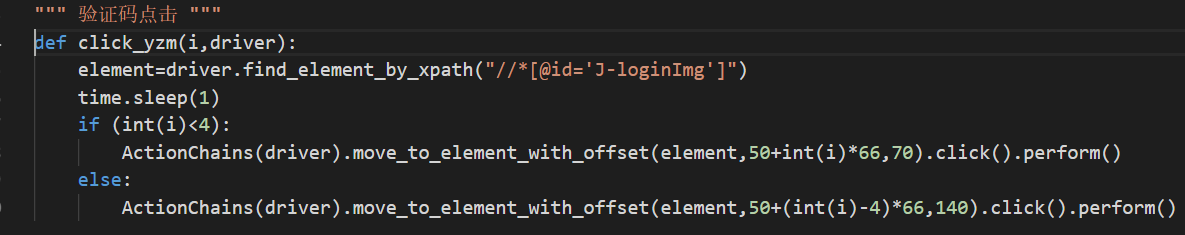


1. 点击验证码部分
   1. 判断返回的数组内是否都为数字，不是的话（就是识别网站坏了）重开downs（），是则逐个点击验证码。



* 1. click\_yzm函数：

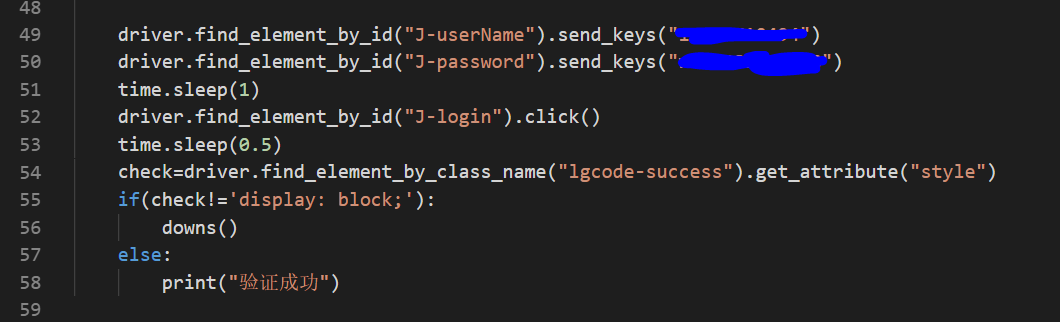
做的验证就是判断是上面那行还是下面那行。



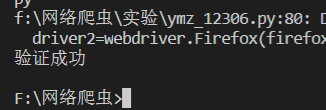
* 1. 效果



1. 最后
   1. 键入账号密码及点击登录按钮，根据登录成功模块的样式来判断是否成功



* 1. 最后截图



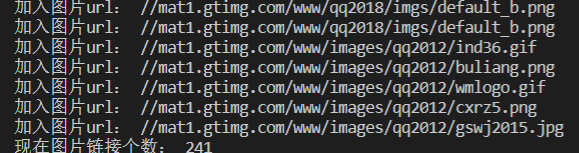


### 三）redis分布式爬虫

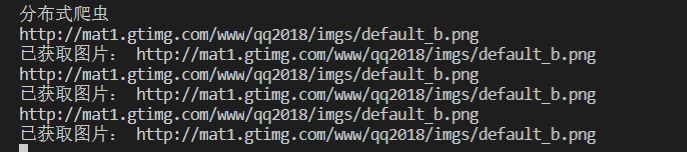
因为是书上例子就不详细说明了。

效果截图：

Mater：



Salve：



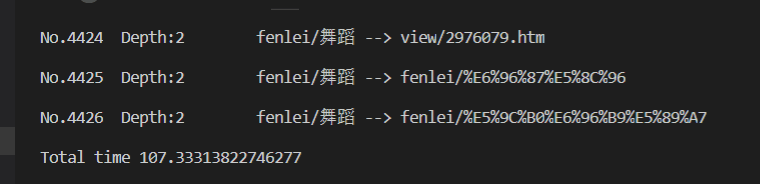


### 百度百科的深度优先及广度优先的爬虫

和书上例子差不多只是改变了a标签的正则也不详细说明了。

效果展示：

深度递归爬虫：



广度多线程爬虫：

