



《爬虫期末作业》

题 目： 网络爬虫

分 院： 数学与信息工程学院

班 级： 16计算机科学与技术一班

姓 名： 缪家亮

学 号： 16219111103

完成日期： 2019年6月14日

# 实验要求介绍

## 1、实验要求：

（1）模仿书本P36 “Request爬虫实践：TOP250电影数据”爬取相关数据，并数入据存入数据库中。（数据库软件自选）

（2）建立Django项目：

基本要求：可运行Django项目;

进阶要求：展示爬虫信息的Django项目

高阶要求：基于BeautifulSoup或其他网页模拟的爬虫的Django项目

（3）使用selenium爬取动态网站—京东手机

（4）上传代码到Github

以上均为期中作业要求，以下为期末作业要求：

（5）实现12306网站的验证码识别登录

## 2、开发工具与编程语言：

这里只是简单介绍：

使用软件：**Visual Studio code , MySQL**

Python语言中的相关模块和类库：

**Request：**Request支持HTTP连接保持和连接池，支持使用cookie保持会话，支持文件上传，支持自动响应内容的编码，支持国际化的URL和POST数据自动编码。在python内置模块的基础上进行了高度的封装，从而使得python进行网络请求时，变得人性化，使用Requests可以轻而易举的完成浏览器可有的任何操作。

**BeautifulSoup：**Beautiful Soup提供一些简单的、python式的函数用来处理导航、搜索、修改分析树等功能。它是一个工具箱，通过解析文档为用户提供需要抓取的数据，因为简单，所以不需要多少代码就可以写出一个完整的应用程序。简单来说，Beautiful Soup是python的一个库，最主要的功能是从网页抓取数据。

**Pymysql：**在python3.x中，可以使用pymysql来MySQL数据库的连接，并实现数据库的各种操作。

**Selenium：**selenium 是一套完整web应用程序测试系统，包含了测试的录制（selenium IDE）,编写及运行（Selenium Remote Control）和测试的并行处理（Selenium Grid）。Selenium的核心Selenium Core基于JsUnit，完全由JavaScript编写，因此可以用于任何支持JavaScript的浏览器上。selenium可以模拟真实浏览器，自动化测试工具，支持多种浏览器，爬虫中主要用来解决JavaScript渲染问题。

**Re**：正则表达式（Regular Expression）是字符串处理的常用工具，通常被用来检索、替换那些符合某个模式（Pattern）的文本。很多程序设计语言都支持正则表达式，像Perl、Java、C/C++。在 Python 中是通过标准库中的re 模块 提供对正则的支持。

其余的还有**time，os**两个类库就不介绍。

# 设计过程简单介绍

先前部分的效果与代码截图仍留在上次的pdf文件中，请自行查看。

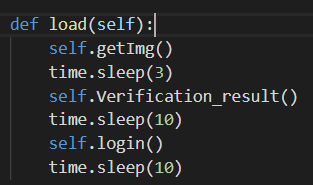
## 期末新增部分：12306验证码识别登录：

代码中设计的函数分别有：

**getImg(self)**

**Verification\_result (self)**

**login(self)**

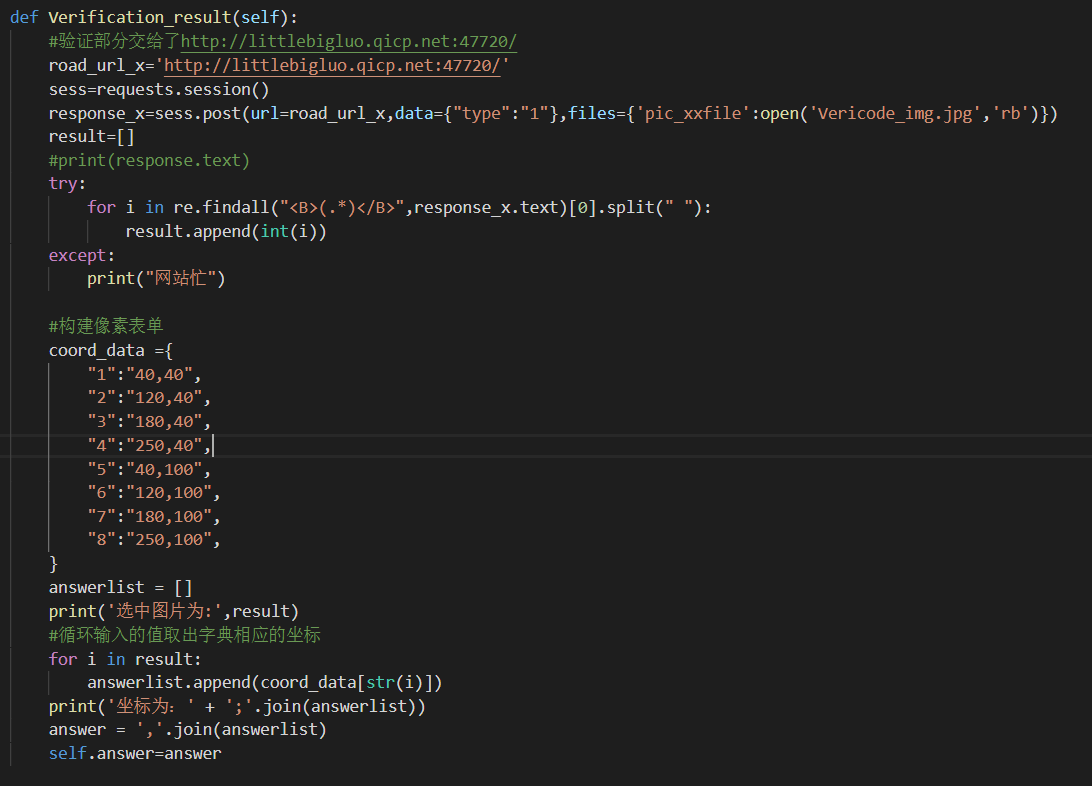


**getImg()**：从12306的登录网页获取验证码的图片并下载下来保存为**Vericode\_img.jpg**,其中需解码。



**Verification\_result()**：此处验证码识别交由#<http://littlebigluo.qicp.net:47720/>#处理，如图，上传之前getImg()函数已下载的Yz\_img.jpg，获取相应图片的位置。再构造像素表单，用获取到的图片位置，来判断验证码识别所需要的图片点击坐标，即**answer**，下个函数会讲到。

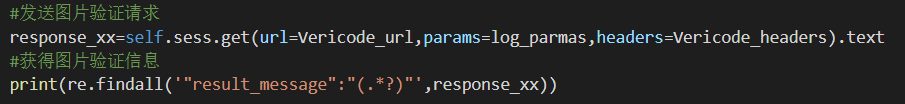


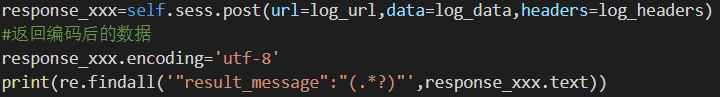


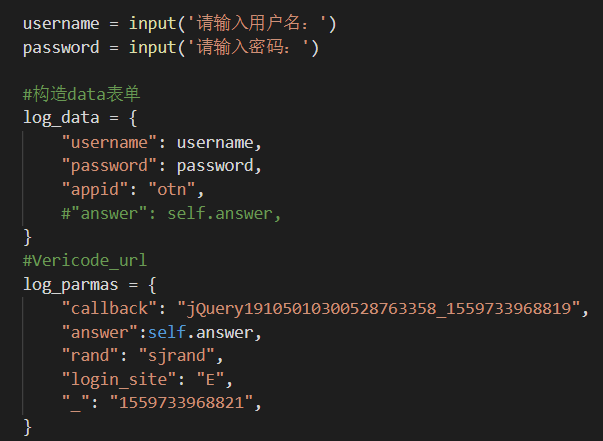
**login()**：此处涉及到了验证码识别通过和登录申请两方面。



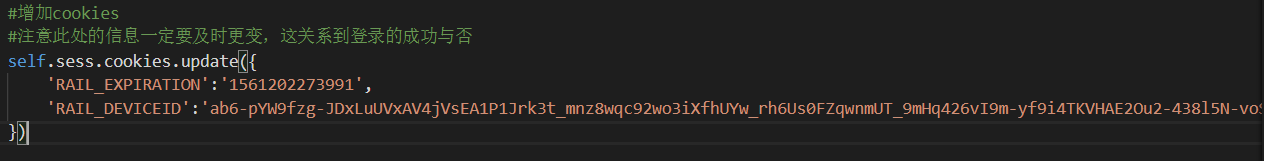
**Get()**和**post()**中的相关参数可以在<https://kyfw.12306.cn/otn/resources/login.html>网页中右击检查等操作找到参数值，如图：（之前函数里的answer就是其中一个参数）







其中登录申请时，也需要**增加cookies**（参数也可以找到）：



# 运行截图

先前部分的效果与代码截图仍留在上次的pdf文件中，请自行查看。

12306验证码识别登录：

用户名和密码是手动输的…

