温州大学瓯江学院数信分院

爬虫与数据分析 实验报告

实验名称：	爬虫与数据分析				
班　　级：	16 计算机四班	姓　　名：	姚周晖	学　　号：	16219111407
实验地点：	B7-403	日　　期：	2019.4.18		

目录

前言

  基本完成了对豆瓣 top250 的静态爬取，动态爬取京东商城的手机商品信息，完成对
12306 网站的两种方法登录，以及在实验课上的课堂实验整理，主要通过 requests 库，
BeautifulSoup 库，使用了 django，scrapy 框架。动态由 selenium 实现。

豆瓣 top250 部电影爬取

爬取 top250 的电影（静态）

```python
from bs4 import BeautifulSoup
import pymysql
import requests
import re
import os


def connect_db():
    connect = pymysql.connect(
        user="root",
        password="admin",
        host="localhost",
        db="scraping",
        port=3306,
        charset=("utf8"),
        use_unicode=True,
    )
    return connect


def get_html(web_url):
    header = {
        "User-Agent": "Mozilla/5.0 (Windows NT 10.0; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/47.0.2526.108
Safari/537.36 2345Explorer/8.5.1.15355"}
    html = requests.get(url=web_url, headers=header).text
    Soup = BeautifulSoup(html, "lxml")
    data = Soup.find("ol").find_all("li")
    return data


def get_info(all_move, connect, cursor):
    for info in all_move:

        nums = re.findall(r'<em class="">\d+</em>', str(info), re.S | re.M)
        nums = re.findall(r'\d+', str(nums), re.S | re.M)
        num = nums[0]



        names = info.find("span")
        name = names.get_text()
```

```python
        charactors = info.find("p")

        charactor = charactors.get_text().replace(" ", "").replace("\n", "")

        charactor = charactor.replace("\xa0", "").replace("\xee", "").replace("\xf6", "").replace("\u0161",
"").replace("\xf4", "").replace("\xfb", "").replace("\u2027", "")


        data = {'num':num, 'name':name, 'charactor':charactor, 'remark':remark, 'score':score}

        print(data)

        cursor.execute("insert into doubantop(num,name,charactor)values(%s,%s,%s)",
                        [data['num'], data['name'], data['charactor']])

        connect.commit()

    return


if __name__ == "__main__":

    connect = connect_db()

    cursor = connect.cursor()

    page = 0

    while page <= 225:

        web_url = "https://movie.douban.com/top250?start=%s&filter=" % page

        all_move = get_html(web_url)

        data = get_info(all_move, connect, cursor)

        page += 25


    connect.close()
```
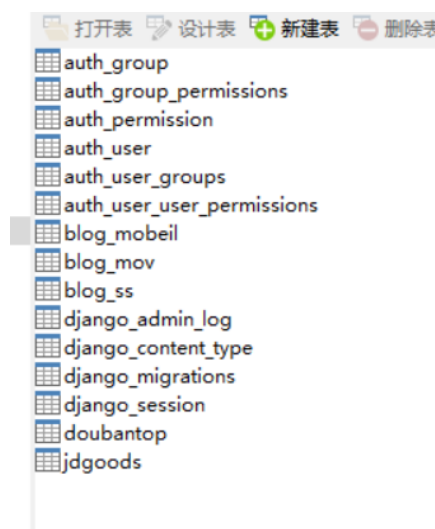
数据库文件：


```
打开表  设计表  新建表  删除表
auth_group
auth_group_permissions
auth_permission
auth_user
auth_user_groups
auth_user_user_permissions
blog_mobeil
blog_mov
blog_ss
django_admin_log
django_content_type
django_migrations
django_session
doubantop
jdgoods
```

Doubantop：

django 显示数据：

Setting:

```
INSTALLED_APPS = [
    'django.contrib.admin',
    'django.contrib.auth',
    'django.contrib.contenttypes',
    'django.contrib.sessions',
    'django.contrib.messages',
    'django.contrib.staticfiles',
    'blog',
]
```

```
TEMPLATES = [
    {
        'BACKEND': 'django.template.backends.django.DjangoTemplates',
        'DIRS': [os.path.join(BASE_DIR,'templates')],
        'APP_DIRS': True,
        'OPTIONS': {
            'context_processors': [
                'django.template.context_processors.debug',
                'django.template.context_processors.request',
                'django.contrib.auth.context_processors.auth',
                'django.contrib.messages.context_processors.messages',
            ],
        },
    },
]
```

```
DATABASES = {
    'default': {
        'ENGINE': 'django.db.backends.mysql',
        'NAME': 'ojmovie',  #数据库名称
        'USER':'root',
        'PASSWORD':'admin',
        'Host':'localhost', #用户名
        'PORT': '3306',  # 数据库使用的端口
    }
}
```

Models.py:

```
class MOV(models.Model):
    id=models.BigIntegerField
    name = models.CharField(max_length=255)
    charactor = models.CharField(max_length=255, blank=True, null=True)
    def __str__(self):
        return self.id
```

Views.py:

```
# Create your Views he
def index1(request):

  # movies=movie.objects.all()
  movies=MOV.objects.all()
  return render(request,'blog/index.html',
  {
      'movies':movies,
  })
```

```
        return self.id
class SS(models.Model):
    id=models.BigIntegerField
    title = models.CharField(max_length=255)
    price = models.CharField(max_length=255)
    comment = models.CharField(max_length=255, blank=True, null=True)
    def __str__(self):
        return self.id
```

```
def ccc(request):
    cursor=connection.cursor()
    a = request.GET.get('tex')  #30

    test = a+'%'

    #tex=request.get("tex")
    cursor.execute(("select * from doubantop where name like %s"),test)
    rows=cursor.fetchall()
    return render(request,'blog/chaxun.html',
    {'content':rows
    })
```

Urls:

```
from django.contrib import admin
from django.conf.urls import include,url
from blog  import views
from blog import search
urlpatterns = [
    url(r'^admin/', admin.site.urls),
    url(r'^blog/',include('blog.urls')),
]
```

看网上视频将 urls 分写到 blog 项目中：（blog.urls.py）

```
1    from django.contrib import admin
2    from django.urls import path
3    from django.conf.urls import url
4    from blog  import views
5    from blog import search
6
7    urlpatterns = [
8        url(r'^index/', views.index1 ),
9        url(r'^good/', views.index),
10       url(r'^chaxun/', views.ccc),
11       url(r'^chaxun/',views.search),
12       url(r'^blog/chaxun/$',views.chaxun,name='chaxun')
13
14   ]
```

页面设计（主页和搜索页）：
Index.html:

```
<!DOCTYPE html>
{% load staticfiles %}
<html>


<head>
    <title>MyBlog</title>
    <link rel="stylesheet" href="{% static 'css/main.css' %}"/>
    <style>
        body{
            background-color: yellow;
        }
```

```html
            b{

                float: right;;


            }
    </style>
</head>


<body >
    <div id="main"><img src="{% static "img/5.jpg" %}" width="100%" height="200px"></div>
    <form action="http://127.0.0.1:8000/blog/chaxun/" method="GET"> <!-- 将 http 这修改为后台地址   method get/post  与后台
的要一致 -->
    <input type="text" name='tex'> <!-- 这里就收参数 name 要与后台一致 -->


    <input type="submit" value="搜索">
    </form>
            <br>
            <br>
        <center>
            <table border="1px">


                <tr>


                    <th>排名</th>
                    <th>电影名</th>
                    <th>导演</th>
                    <th>前往观看</th>



                </tr>


            {% for new  in  movies %}
                <tr>
                    <td><center>{{new.id}}</center></td>
                    <td><center>{{new.name}}</center></td>
                    <td><center>{{new.charactor}}</center></td>
                    <th><center><a href="https://movie.douban.com/">点击观看</a></center></th>
                </tr>


            {% endfor%}
             </table>
            </center>
```

```
</body>



</html>
```

Chaxun.html:

```html
<!DOCTYPE html>

{% load staticfiles %}

<html>


<head>

    <title>MyBlog</title>

    <link rel="stylesheet" href="{% static 'css/main.css' %}"/>

    <style>

            body{

                background-color: yellow;

            }

            label{


                float: right;


            }

    </style>

</head>


<body >



    <div id="main"><img src="{% static "img/5.jpg" %}" width="100%" height="200px"></div>

    <br>

    <br>

    <laebl>

            <a href="http://127.0.0.1:8000/blog/index/">

                    <input type="button" value="返回首页">

            </a>

    </label>

    <br>

        <center> <font size="50px">    <tr>搜索结果</tr></font></center>

            <br>

            <br>

                <tr>

                        <center> {{content}} <a href="https://movie.douban.com/">点击观看</a></center>

                </tr>
```

```
</body>

</html>
```

效果展示：

搜索功能：

动态爬取京东手机产品(装了谷歌的驱动)：

代码如下：

```python
from selenium.webdriver import Chrome
from config import *
from selenium.webdriver.support.ui import WebDriverWait
from selenium.webdriver.support import expected_conditions as EC
from selenium.webdriver.common.by import By
from selenium.webdriver.common.keys import Keys
from time import sleep
from selenium.common.exceptions import NoSuchElementException
import pymysql


def next_page(client, wait, page_num):
    END_PAGE = 1
    while len(client.find_elements_by_class_name('gl-item')) < 60:
        client.execute_script('window.scrollTo(0, document.body.scrollHeight)')
        sleep(1)
    print(" 第{}加载完成".format(page_num))



    parse_page(page_num, client)



    page_num += 1
    if page_num > END_PAGE:
        print('前{}页爬取成功'.format(END_PAGE))
        return



    wait.until(
        EC.presence_of_element_located(
            (By.CSS_SELECTOR, '#J_bottomPage > span.p-skip > input')
        )
    )


    wait.until(
        EC.element_to_be_clickable(
            (By.CSS_SELECTOR, '#J_bottomPage > span.p-skip > a')
        )
    )
```

```python
    input_ = client.find_element_by_css_selector('#J_bottomPage > span.p-skip > input')

    input_.clear()

    input_.send_keys(page_num)


    input_.send_keys(Keys.ENTER)


    wait.until(

        EC.text_to_be_present_in_element(

            (By.CSS_SELECTOR, '#J_bottomPage > span.p-num > a.curr'),

            str(page_num)

        )

    )


    next_page(client, wait, page_num)


def connect_db():

    connect = pymysql.connect(

        user="root",

        password="admin",

        host="localhost",

        db="jddog",

        port=3306,

        charset=("utf8"),

        use_unicode=True,

    )

    return connect

def parse_page(page_num, client):

    print("[+] 开始解析第{}页数据".format(page_num))

    items = client.find_elements_by_class_name('gl-item')

    index = 1


    for item in items:

        print("[{}] ".format(index), end="")

        try:

            title = item.find_element_by_css_selector("div.p-name > a > em").text

        except NoSuchElementException:

            title = None

        try:

            price = item.find_element_by_css_selector("div.p-price > strong > i").text

        except NoSuchElementException:

            price = None



        try:
```

```python
            comment = item.find_element_by_css_selector(".p-commit a").text

        except NoSuchElementException:

            comment = None


        print("{} >>> {}   >>> {}".format(title, price, comment))


        connect=connect_db()

        cursor = connect.cursor()

        cursor.execute("insert into JDGOODS(id,TITLE,PRICE,COMMENT)values (%s,%s,%s,%s)",(index,title, price, comment))

        connect.commit()

        index += 1
    print(" 解析第{}页数据完成".format(page_num))

    connect.close()


def search(client, url, keyword,wait):

    client.get(url)

    wait.until(

        EC.presence_of_element_located(

            (By.ID, 'key')

        )

    )


    wait.until(

        EC.element_to_be_clickable(

            (By.CSS_SELECTOR, '#search > div > div.form > button > i')

        )

    )




    input_ = client.find_element_by_id('key')

    input_.send_keys(keyword)




    botton = client.find_element_by_css_selector('#search > div > div.form > button > i')

    botton.click()




    page_num = 1

    next_page(client, wait, page_num)


def main():

    client = Chrome()
```

```
    url = "http://www.jd.com"

    KEYWORD = '手机'

    wait = WebDriverWait(client, 10)

    search(client, url, KEYWORD, wait)


if __name__ == '__main__':

    main()
```

数据库：

Jdgoods:

id	TITLE	PRICE	COMMENT
1	荣耀10青春版 幻彩渐变 240	1299	二手有售
2	荣耀8X 千元屏霸 91%屏占比	1299	二手有售
3	Apple iPhone XR (A2108)	5899	二手有售
4	【KPL官方比赛用机】vivo i	3298	二手有售
5	荣耀V20 胡歌同款 麒麟980	2799	二手有售
6	vivo U1 水滴全面屏 AI智慧	799	12万+
7	vivo X27 8GB+256GB大内	3598	二手有售
8	OPPO Reno 全面屏拍照手机	2999	200+
9	小米 红米Redmi Note7 幻	1199	53万+
10	荣耀畅玩8C两天一充 莱茵护	899	39万+
11	黑鲨游戏手机2 8GB+128GI	3499	3.6万+
12	小米 红米6 4GB+64GB 流沙	799	77万+
13	小米8SE 全面屏智能游戏拍	1599	二手有售
14	华为 HUAWEI Mate 20 麒	3989	二手有售
15	小米 红米Redmi 7 幻彩渐变	799	4.1万+
16	Apple iPhone X (A1865) 6	6349	二手有售
17	Apple iPhone XS Max (A2	9699	二手有售
18	荣耀10 GT游戏加速 AIS手持	2198	二手有售
19	vivo Z3 6GB+64GB 极光蓝	1598	二手有售

页面设计：

Models.py:

```
class Mobeil(models.Model):
    id=models.BigIntegerField
    title = models.CharField(max_length=255)
    price = models.CharField(max_length=255)
    comment = models.CharField(max_length=255, blank=True, null=True)
    def __str__(self):
        return self.id
class SS(models.Model):
    id=models.BigIntegerField
    title = models.CharField(max_length=255)
    price = models.CharField(max_length=255)
    comment = models.CharField(max_length=255, blank=True, null=True)
    def __str__(self):
        return self.id
```

Views,py:

```python
def index(request):

    # movies=movie.objects.all()
    mobeils=Mobeil.objects.all()
    return render(request,'blog/good.html',
    {
        'mobeils':mobeils,
    })
```

Urls.py:

```python
from django.contrib import admin
from django.urls import path
from django.conf.urls import url
from blog  import views
from blog import search

urlpatterns = [
    url(r'^index/', views.index1 ),
    url(r'^good/', views.index),
    url(r'^chaxun/', views.ccc),
    url(r'^chaxun/',views.search),
    url(r'^blog/chaxun/$',views.chaxun,name='chaxun')

]
```

页面设计（good.html）：

```html
<!DOCTYPE html>

{% load staticfiles %}

<html>

<head>

    <title>MyBlog</title>

    <style>

        body{

            background-color: wheat;




        }

        table{

            margin-left: 20px;

            float: left;




        }

        label
```

```
                {    color: yellow;

                     float:right;

                     margin-right: 10%;


                }


    </style>
</head>


<body>
        <div id="a">
            <img src="{% static "img/4.jpg" %}" width="100%" height="200px">
            <br/>
            <br/>


        </div>



            <table  border="1px"  >


                    <tr>


                        <th>排名</th>
                        <th>手机名</th>
                        <th>价格</th>
                        <th>         </th>
                        <th>其他信息</th>
                        <th>         </th>
                        <th>更多...</th>


                    </tr>


                    {% for a  in  mobeils %}
                      <tr>
                          <td><center>{{a.id}}</center></td>
                          <td><center>{{a.title}}</center></td>
                          <td><center>{{a.price}}</center></td>
                          <td>         </td>
                          <td><center>{{a.comment}}</center></td>
                          <td>         </td>
                          <td><a
href="https://search.jd.com/Search?keyword=%E6%89%8B%E6%9C%BA&enc=utf-8&wq=%E6%89%8B%E6%9C%BA&pvid=d8c7439106bc4b409
39fc102b20246df">点解了解</td>
```

```
            </tr>
        {% endfor%}


            </table>
    <label>
                <font size="100px" color:yellow> 著名手机品牌</font>
        <br/>
        <a href="https://www.vmall.com/?cid=91895">
        <img src="{% static "img/huawei.jpeg" %}"  width="400px" height="100px">
        </a>
        <br/><br/><br/><br/><br/>
        <a href="https://www.apple.com/cn/?afid=p238%7C11DPVC7K_mtid_18707vxu38484&cid=aos-cn-kwsg-brand"></a>
        <img src="{% static "img/iphone.jpg" %}" width="400px" height="100px">
        </a>
        <br/><br/><br/><br/><br/>
        <a href="https://www.mi.com/"></a>
        <img src="{% static "img/xiaomi.png" %}" width="400px" height="100px">
        </a>
        <br/><br/><br/><br/><br/>
        <a href="https://www.nokia.com/zh_int/"></a>
        <img src="{% static "img/rjy.jpg" %}" width="400px" height="100px">
        </a>
    </label>



</body>


</html>
```

效果展示：

12306 的爬取

使用 requests 爬取：

代码：

```
        import requests
from json import loads
from PIL import Image
import config
session = requests.Session()
url = "https://kyfw.12306.cn/otn/login/init"
response = session.get(url)
captcha_url="https://kyfw.12306.cn/passport/captcha/captcha-image?login_site=E&module=login&rand=sjrand&0.5604
448691516117"
cap_response = session.get(captcha_url)
with open('1.jpg', 'wb') as f:
    f.write(cap_response.content)
img = Image.open('1.jpg')
img.show()
ver = input("请输入验证码>>>")
soList = ver.split(',')
yanSol = ['35,35', '105,35', '175,35', '245,35', '35,105', '105,105', '175,105', '245,105']
yanList = []
for item in soList:
```

```python
    print(item)
    yanList.append(yanSol[int(item)])
yanStr = ','.join(yanList)
f.close()
check_url = "https://kyfw.12306.cn/passport/captcha/captcha-check"
data = {
    "answer": yanStr,
    "login_site": "E",
    "rand": "sjrand"
    }
check_response = session.post(check_url, data=data)
print(check_response.text)
check_res = check_response.json()
if not check_res['result_code'] == '4':
    exit('验证码错误')
login_url = "https://kyfw.12306.cn/passport/web/login"
login_data = {
    "username": config.username,
    "password": config.password,
    "appid": "otn",
    }
login_response = session.post(login_url, data=login_data)
print(login_response.text)
```
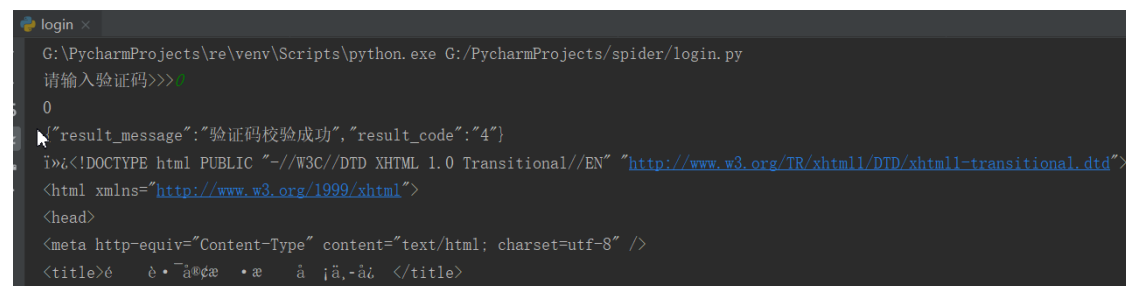
结果：

```
login ×
 G:\PycharmProjects\re\venv\Scripts\python.exe G:/PycharmProjects/spider/login.py
 请输入验证码>>>0
 0
 {"result_message":"验证码校验成功","result_code":"4"}
 <!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN" "http://www.w3.org/TR/xhtml1/DTD/xhtml1-transitional.dtd">
 <html xmlns="http://www.w3.org/1999/xhtml">
 <head>
 <meta http-equiv="Content-Type" content="text/html; charset=utf-8" />
 <title>é    è·¯ å¢æ¥ ·æ    å ¡ä¸-ä¿ </title>
```

使用 selenium 登录 12306：

基本步骤如下：

（1）先加载驱动：

```python
def login(self):
    login_url = "https://kyfw.12306.cn/otn/resources/login.html1"
    driver = webdriver.Chrome()
    driver.set_window_size(1200, 900)
    driver.get(login_url)
    account = driver.find_element_by_class_name("login-hd-account")
    account.click()
    userName = driver.find_element_by_id("J-userName")
    userName.send_keys("531218020@qq.com")
    password = driver.find_element_by_id("J-password")
    password.send_keys("**********")
    self.driver = driver
```

（2）：将 12306 的图片爬取打开：

```python
def getimage(self):
    try:
        img_element = WebDriverWait(self.driver, 100).until(
            EC.presence_of_element_located((By.ID, "J-loginImg"))
        )
    except Exception as e:

        print(u"网络开小差,请稍后尝试")
    base64_str = img_element.get_attribute("src").split(",")[-1]
    imgdata = base64.b64decode(base64_str)
    with open('1.jpg', 'wb') as file:
        file.write(imgdata)
    self.img_element = img_element

    img = Image.open('1.jpg')
```

（3）:将图片展示给用户，得到验证码（0-7）：

```python
def getnum(self):
    ver = input("请输入验证码（如1,2）:")
    yanSol = ['35,35', '105,35', '175,35', '245,35', '35,105', '105,105', '175,105', '245,105']
    yanList = []
    soList = ver.split(',')
    for i in soList:
        print(i)
    print(soList)
    self.result = soList
```

（4）模拟点击验证码的：

```python
def moveAndClick(self):
    num=[]
    for j in self.result:
        num.append(int(j))
    print(num)
    element = self.driver.find_element_by_class_name('touclick-wrapper')

    for i in num:
        if i <= 4:
            ActionChains(self.driver).move_to_element_with_offset(element, 50 + i * 65, 70).click().perform()
            print(i)
        else:
            i = i - 4
            ActionChains(self.driver).move_to_element_with_offset(element, 50 + i * 65, 140).click().perform()
            print(i)
```

（5）调用函数：

```python
def submit(self):
    btn = self.driver.find_element_by_id("J-login")
    btn.click()

def __call__(self):
    self.login()

    time.sleep(3)
    self.getimage()
    time.sleep(3)
    self.getnum()
    time.sleep(3)
    self.moveAndClick()
    time.sleep(1)
    self.submit()
    time.sleep(10000)


Demo()()
```

代码如下：

```python
from selenium import webdriver
from selenium.webdriver.common.keys import Keys
from selenium.webdriver.common.by import By
from selenium.webdriver.support.ui import WebDriverWait
from selenium.webdriver.support import expected_conditions as EC
from selenium.webdriver.common.action_chains import ActionChains
import requests
from PIL import Image
import base64
import re
import time


class Demo():
```

```python
    def login(self):
        login_url = "https://kyfw.12306.cn/otn/resources/login.html"
        driver = webdriver.Chrome()
        driver.get(login_url)
        account = driver.find_element_by_class_name("login-hd-account")
        account.click()
        userName = driver.find_element_by_id("J-userName")
        userName.send_keys("15157757688")
        password = driver.find_element_by_id("J-password")
        password.send_keys("yzh980513")
        self.driver = driver

    def getimage(self):
        try:
            img_element = WebDriverWait(self.driver, 100).until(
                EC.presence_of_element_located((By.ID, "J-loginImg"))
            )
        except Exception as e:

            print(u"网络开小差,请稍后尝试")
        base64_str = img_element.get_attribute("src").split(",")[-1]
        imgdata = base64.b64decode(base64_str)
        with open('1.jpg', 'wb') as file:
            file.write(imgdata)
        self.img_element = img_element

        img = Image.open('1.jpg')
        img.show()

    def getnum(self):
        ver = input("请输入验证码（如 1,2）:")
        yanSol = ['35,35', '105,35', '175,35', '245,35', '35,105',
'105,105', '175,105', '245,105']
        yanList = []
        soList = ver.split(',')
        for i in soList:
            print(i)
        print(soList)
        self.result = soList

    def moveAndClick(self):
        num=[]
        for j in self.result:
            num.append(int(j))
```

```python
        print(num)
        element = self.driver.find_element_by_class_name('touclick-wrapper')

        for i in num:
            if i <= 4:
                ActionChains(self.driver).move_to_element_with_offset(element, 50 + i * 65, 70).click().perform()
                print(i)
            else:
                i = i - 4
                ActionChains(self.driver).move_to_element_with_offset(element, 50 + i * 65, 140).click().perform()
                print(i)


    def submit(self):
        btn = self.driver.find_element_by_id("J-login")
        btn.click()
        print()

    def __call__(self):
        self.login()
        time.sleep(3)
        self.getimage()
        time.sleep(3)
        self.getnum()
        time.sleep(3)
        self.moveAndClick()
        time.sleep(1)
        self.submit()
        time.sleep(10000)


Demo()()
```
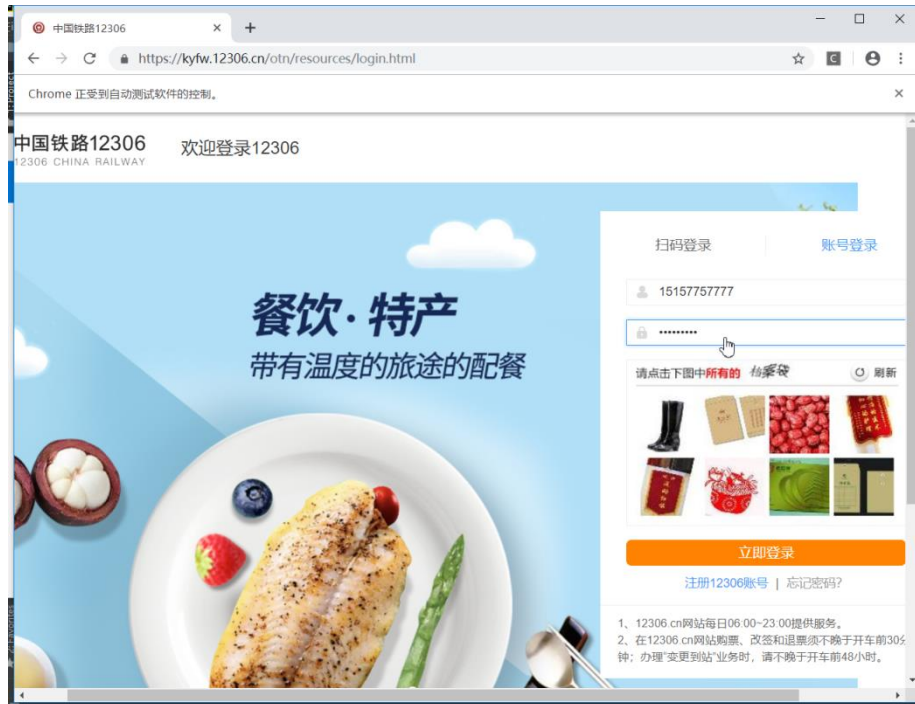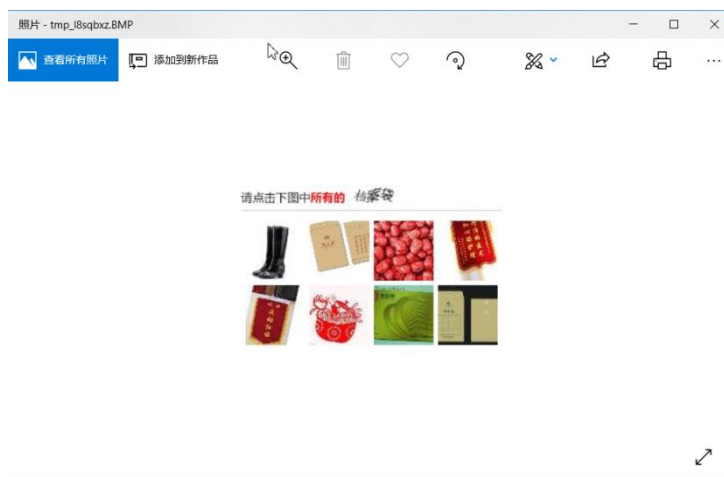
效果展示：

验证码：

扫码登录　　　　账号登录

👤　15157757777

🔒　••••••••

请点击下图中所有的 档案袋　　○ 刷新

立即登录

注册12306账号 ｜ 忘记密码?

1、12306.cn网站每日06:00~23:00提供服务。
2、在12306.cn网站购票、改签和退票须不晚于开车前30分钟；办理"变更到站"业务时，请不晚于开车前48小时。

验证成功，跳转中...　　　○ 刷新

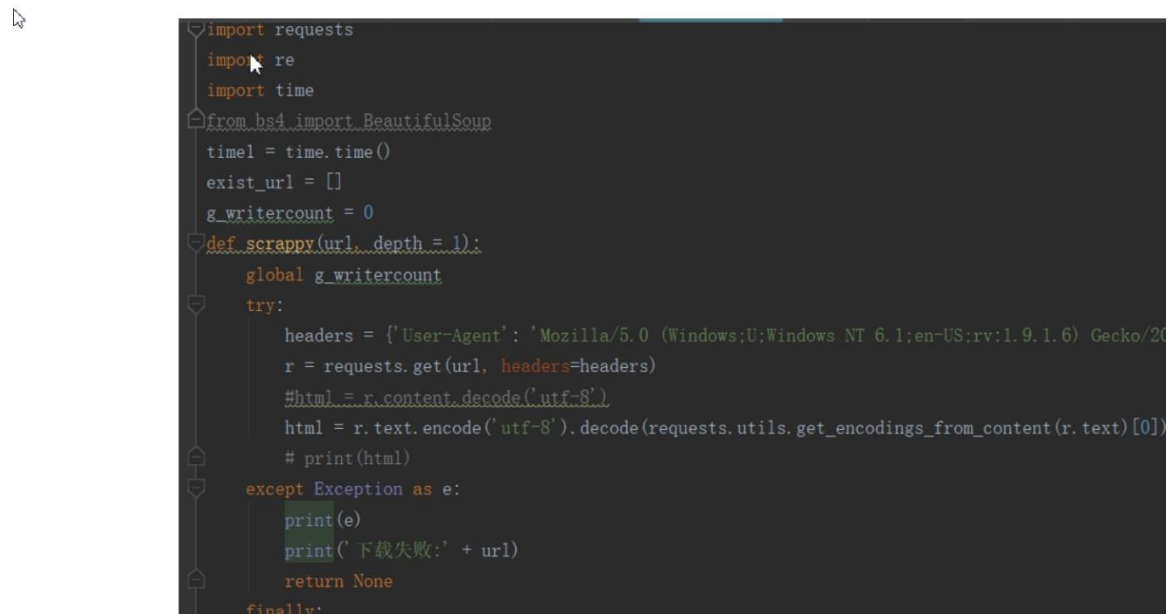恭喜！完成验证。

立即登录

注册12306账号 ｜ 忘记密码?

实验课各项作业展示：

点击后跳转到代码和效果展示：
例如：
深度优先：

深度优先代码:

结果: