温州大学瓯江学院

爬虫与数据分析　　实验报告

实验名称：	python 爬虫期末作业				
班　　级：	16 计算机三班	姓　　名：	王宜家	学　　号：	16219112208
实验地点：		日　　期：	6.16.2019		

一、项目介绍：

该项目在前端设计运用了 bootstrap 框架，css 和 jq，实现响应式布局。

运用 python 分别进行静态爬虫和动态爬虫，并把数据存储到 mysql 数据库中。

再运用 django 技术把内容显示到网站上。

并且实现了 12306 爬虫自动登录

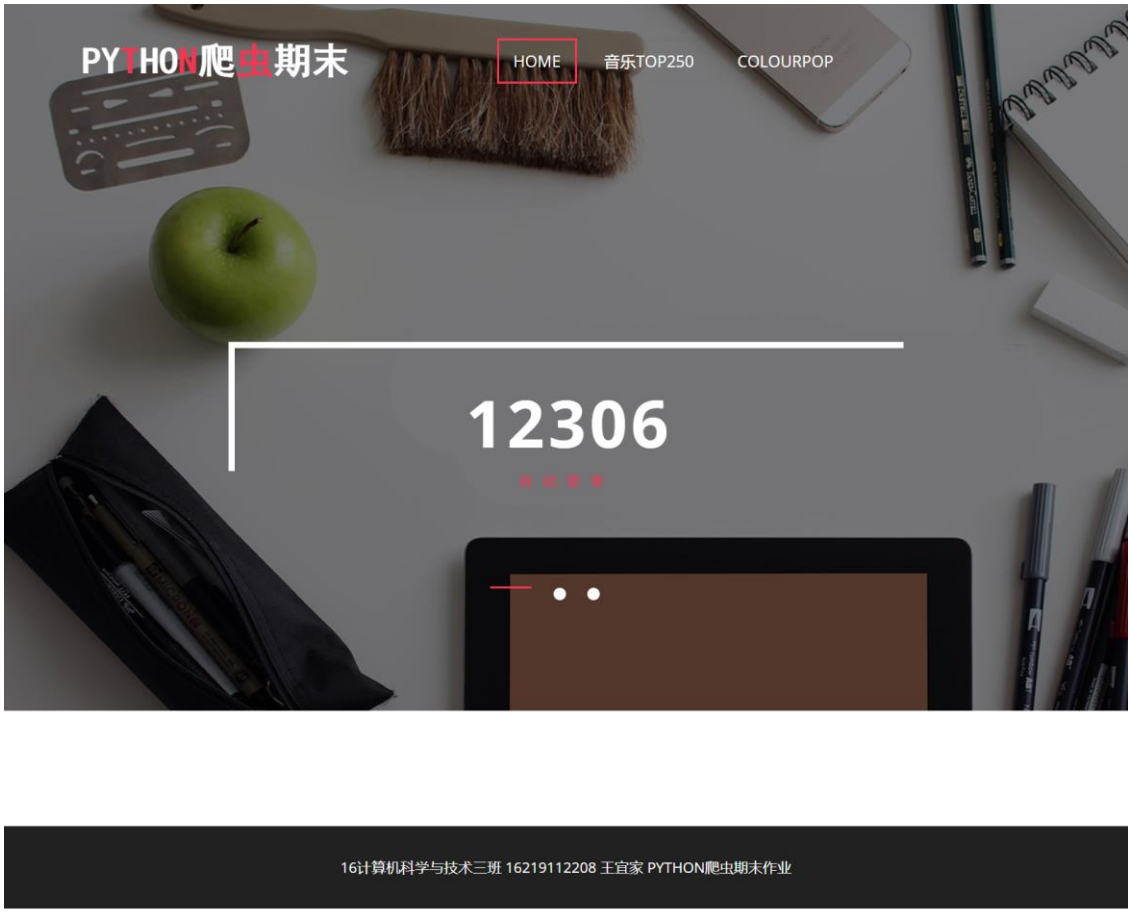
二、设计思路：

网站页界面大致可分为四个部分，分别是：导航条、title、爬虫内容和 ending。
导航条可以实现页面切换，并且标出现在所在的页面；title 说明了爬虫的来源；爬虫内容通过 django 和前端来显示。
页面布局以表格的形式呈现，且运用了 django 分页技术；ending 介绍了我的基本信息。
还设计了主页，在主页加入了图片轮播。

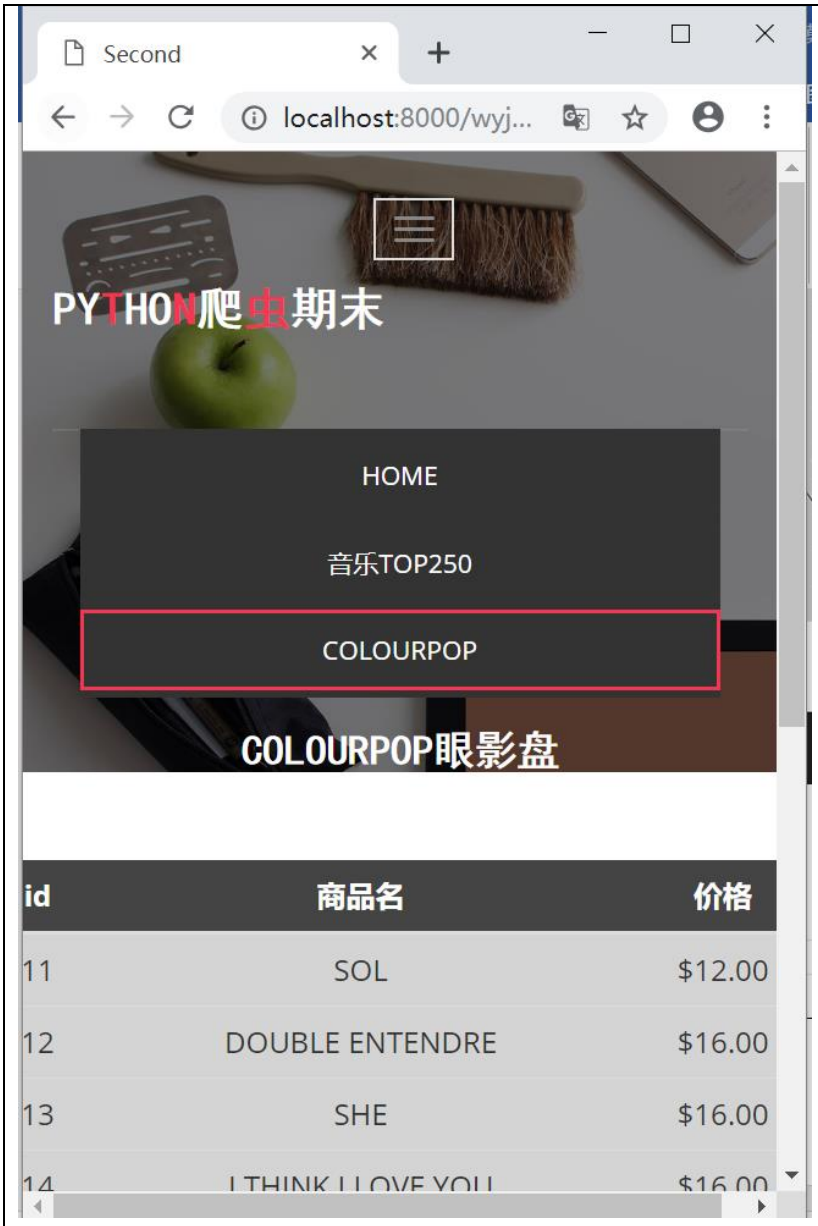
三、效果截图：

12306 自动登录

首页：


PYTHON爬虫期末
HOME
音乐TOP250
COLOURPOP
12306
16计算机科学与技术三班 16219112208 王宜家 PYTHON爬虫期末作业

响应式布局：


Second
localhost:8000/wyj...
PYTHON爬虫期末
HOME
音乐TOP250
COLOURPOP
COLOURPOP眼影盘
id
商品名
价格
11
SOL
$12.00
12
DOUBLE ENTENDRE
$16.00
13
SHE
$16.00
14
I THINK I LOVE YOU
$16.00

静态网页爬取：

排名	歌名	详细内容	评分
1	We Sing. We Dance. We Steal Things.	Jason Mraz / 2008-05-13 / Import / Audio CD / 民谣	9.1
2	Viva La Vida	Coldplay / 2008-06-17 / 专辑 / CD / 摇滚	8.7
3	华丽的冒险	陈绮贞 / 2005-09-23 / 专辑 / CD / 流行	8.9
4	范特西	周杰伦 / 2001-09-14 / 专辑 / CD / 流行	9.2
5	後。青春期的詩	五月天 / 2008-10-23 / 专辑 / CD / 摇滚	8.8
6	是时候	孙燕姿 / 2011-03-08 / 专辑 / CD / 流行	8.6
7	Lenka	Lenka / 2008-09-23 / 专辑 / Audio CD / 流行	8.5
8	Start from Here	王若琳 / 2008-01-11 / 专辑 / CD / 爵士	8.7
9	旅行的意义	陈绮贞 / 2004-02-02 / 单曲 / CD / 流行	9.2
10	太阳	陈绮贞 / 2009-01-22 / 专辑 / CD / 流行	8.6
11	Once (Soundtrack)	Glen Hansard,Marketa Irglova / 2007-05-22 / Soundtrack / CD / 原声	9.1
12	Not Going Anywhere	Keren Ann / 2004-08-24 / Import / Audio CD / 民谣	8.9
13	American Idiot	Green Day / 2004-09-21 / Explicit Lyrics / Audio CD / 摇滚	8.9
14	思念是一种病	张震岳 Csun Yuk / 2007-07-06 / 专辑 / CD / 流行	8.8
15	無與倫比的美麗	苏打绿 / 2007-11-02 / 专辑 / CD / 流行	8.6

动态网页爬取：

id	商品名	价格
11	SOL	$12.00
12	DOUBLE ENTENDRE	$16.00
13	SHE	$16.00
14	I THINK I LOVE YOU	$16.00
15	GOOD SPORT	$16.00
16	MISUNDERSTOOD.	$22.00
17	PERCEPTION	$24.00
18	YOU HAD ME AT HELLO	$18.00
19	THROUGH MY EYES PALETTE	$24.00
20	SALVAJE PALETTE	$18.00

« 1 2 3 4 »

16计算机科学与技术三班 16219112208 王宜家 PYTHON爬虫期末作业

分页美化：

1 2 3 4 5 »

« 1 2 3 4 5 6 »

« 13 14 15 16 17

12306 自动登录代码

```python
import requests
from selenium.webdriver import Chrome
from selenium import webdriver
from selenium.webdriver import ChromeOptions
import time
from selenium.webdriver.common.keys import Keys
from selenium.webdriver.common.action_chains import ActionChains
from PIL import Image
import win32api
import win32con
import os
import re

def main():
    username='18767305246'
    userpassword='www980124'
    start_station='温州南'
    end_station='嘉兴南'
    date='2019-06-15'
    loginandget(username,userpassword,start_station,end_station,date)

def loginandget(username,userpassword,start_station,end_station,date):
    pic_name=None
    options = webdriver.ChromeOptions()
    options.add_argument('user-agent="Mozilla/5.0 (Windows NT 10.0; WOW64) AppleWebKit/
    driver = webdriver.Chrome(chrome_options=options)
    driver.maximize_window()
    driver.get('https://kyfw.12306.cn/otn/login/init')

    time.sleep(2)

    driver.find_element_by_xpath('//*[@id="username"]').send_keys(username) #发送帐号名
    driver.find_element_by_xpath('//*[@id="password"]').send_keys(userpassword) #发送帐
    time.sleep(2)
    pic_name=get_a_verify_pic(driver)            #截取12306验证码图片
    time.sleep(5)
    body_list=ana_pic(pic_name)  #破解12306验证码

    click_pic(driver,body_list)
    time.sleep(2)

    driver.find_element_by_xpath('//*[@class="btn200s"]').click()

    time.sleep(10) # 等待cookie加载完成
    cookies = driver.get_cookies()
    print(cookies)
```

```
    url = 'https://kyfw.12306.cn/otn/resources/js/framework/station_name.js?station_ver
    response = requests.get(url, verify=False)
    stations = re.findall(r'([\u4e00-\u9fa5]+)\|([A-Z]+)', response.text)
    station_codes=dict(stations)

    driver.get('https://kyfw.12306.cn/otn/leftTicket/init?linktypeid=dc&fs='
               +start_station+','+station_codes[start_station]+'&ts='+end_station
               +','+station_codes[end_station]+'&date='+date+'&flag=N,N,Y')

    time.sleep(5)
    lt = driver.find_elements_by_xpath('//div[@id="t-list"]//tr[@class]')#.get_attribut
    for i in lt:
        try:
            number=i.find_element_by_xpath('.//div[@class="train"]//a').text
            start=i.find_element_by_xpath('.//div[@class="cdz"]//strong[1]').text
            end=i.find_element_by_xpath('.//div[@class="cdz"]//strong[2]').text
            starttime=i.find_element_by_xpath('.//div[@class="cds"]//strong[1]').text
            endtime=i.find_element_by_xpath('.//div[@class="cds"]//strong[2]').text
            times=i.find_element_by_xpath('.//div[@class="ls"]//strong[1]').text
            print("number:"+number+",start:"+start+",end:"+end+",starttime:"+
                  starttime+",endtime:"+endtime+",time:"+times)
        except:
            None

VK_CODE ={'enter':0x0D, 'down_arrow':0x28}
#键盘键按下
def keyDown(keyName):
    win32api.keybd_event(VK_CODE[keyName], 0, 0, 0)
#键盘键抬起
def keyUp(keyName):
    win32api.keybd_event(VK_CODE[keyName], 0, win32con.KEYEVENTF_KEYUP, 0)
#截取一张验证码图片，保存
def get_a_verify_pic(b):
    if(os.path.exists("C:/Users/44876/Downloads/captcha-image.jpeg")):
        os.remove("C:/Users/44876/Downloads/captcha-image.jpeg")

    element=b.find_element_by_xpath("//*[@class='touclick-image']")
    action = ActionChains(b).move_to_element(element)#移动到该元素
    action.context_click(element).perform()#右键点击该元素
    time.sleep(1)
    #按v
    win32api.keybd_event(86, 0, 0, 0)
    win32api.keybd_event(86, 0, win32con.KEYEVENTF_KEYUP, 0)
    time.sleep(2)
    #按enter
    keyDown("enter")
```

```
    keyUp("enter")
    time.sleep(1)

    return "captcha-image.jpeg"

#破解图片验证码
def ana_pic(pic_name):
    body_list=[]
    url='http://littlebigluo.qicp.net:47720/'
    files={'pic_xxfile':(pic_name,open("C:/Users/44876/Downloads/"+pic_name,'rb'),'imag
    res=requests.post(url,files=files)#post  pic

    if res.status_code == 200:#return  ok
        if u"图片貌似选" in res.text:#识别验证码成功
            body_str_1=res.text.split(u'<B>')
            body_str=body_str_1[1].split(u'<')[0].split()
            for index in body_str:
                body_list.append(int(index))
            return body_list


#按输入的下标，点击一张验证码图片
def click_one_pic(b,i):
    try:
        imgelement=b.find_element_by_xpath("//*[@class='touclick-image']")
        if i<=4:
            ActionChains(b).move_to_element_with_offset(imgelement,40+72*(i-1),73).clic
        else:
            i -= 4
            ActionChains(b).move_to_element_with_offset(imgelement,40+72*(i-1),145).cli
    except:
        print("Wa  --  click  one  pic  except!!!")

#按bodylist  指示，点击指定验证图片
def click_pic(b,body_list):
    for i in range(len(body_list)):
        click_one_pic(b,body_list[i])
        time.sleep(1)

if __name__ == "__main__":
    main()
```

豆瓣音乐 TOP250 静态爬虫代码

```
import MySQLdb
import requests
from lxml import etree
import re

def get_page(num):
    conn=MySQLdb.connect(host='localhost',user='root',passwd='123456',db='wyjqz',charset='utf8')
    cur=conn.cursor()

    url='https://music.douban.com/top250?start=%s' % num
    headers={'User-Agent':'Mozilla/5.0 (Windows NT 10.0; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/69.0.3497.100 Safari/537.36'}
    response=requests.get(url,headers=headers)
    tree=etree.HTML(response.text)
    name=tree.xpath('//tr[@class="item"]//div[@class="pl2"]//a[1]/text()')
    detail=tree.xpath('//div[@class="pl2"]//p/text()')
    stars=tree.xpath('//div[@class="star clearfix"]//span[2]/text()')

    for i in name:
        if(i.strip()==''):
            name.remove(i)

    for i in range(len(name)):
        sql='insert into testmodel_test (name,detail1,detail2) values ("%s","%s","%s")' % (name[i],detail[i],stars[i])
        cur.execute(sql)
    cur.close()
    conn.commit()
    conn.close()

def get_all():
    for i in range(0,250,25):
        get_page(i)

get_all()
```

COLOURPOP 眼影盘动态爬虫代码

```
from selenium.webdriver.chrome.options import Options
import MySQLdb
import requests
from lxml import etree
from selenium import webdriver
from selenium.webdriver.common.keys import Keys
from selenium.webdriver import ActionChains
import time

def getPhone():
    chorme_options=Options()
    chorme_options.add_argument("--headless")
    chorme_options.add_argument("--disable-gpu")
    driver=webdriver.Chrome(chrome_options=chorme_options)
    #driver=webdriver.Chrome()
    #driver.maximize_window()
    #driver.get("https://www.sina.com.cn")
    driver.get("https://colourpop.com/collections/eye-palettes")

    time.sleep(3)
    js="window.scrollTo(0,document.body.scrollHeight)"
    driver.execute_script(js)

    conn=MySQLdb.connect(host='localhost',user='root',passwd='123456',db='wyjqz',charset='utf8')
    cur=conn.cursor()
    time.sleep(1)
    lt=driver.find_elements_by_xpath("//div[@class='collectioncontainer']//div[@class='product-info-inner']")

    for i in lt:
        try:
            name=i.find_element_by_xpath('.//div[@id="prod-title-price"]').text
            price=i.find_element_by_xpath('.//div[@id="price-html"]').text
            sql="insert into testmodel_test2 (name,price) values ('%s','%s')"%(name,price)
            cur.execute(sql)
        except:
            price=""
            name=""

    cur.close()
    conn.commit()
    conn.close()

getPhone()
```