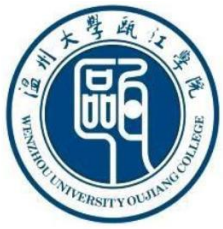


温州大學瓯江學院

WENZHOU UNIVERSITY OUJIANG COLLEGE

《爬虫期中作业》

题　　目：爬虫期中作业

分　　院：数学与信息工程学院

班　　级：16 计算机科学与技术三班

姓　　名：余依伦

学　　号：16219111319

完成日期：2019 年 4 月 25 日

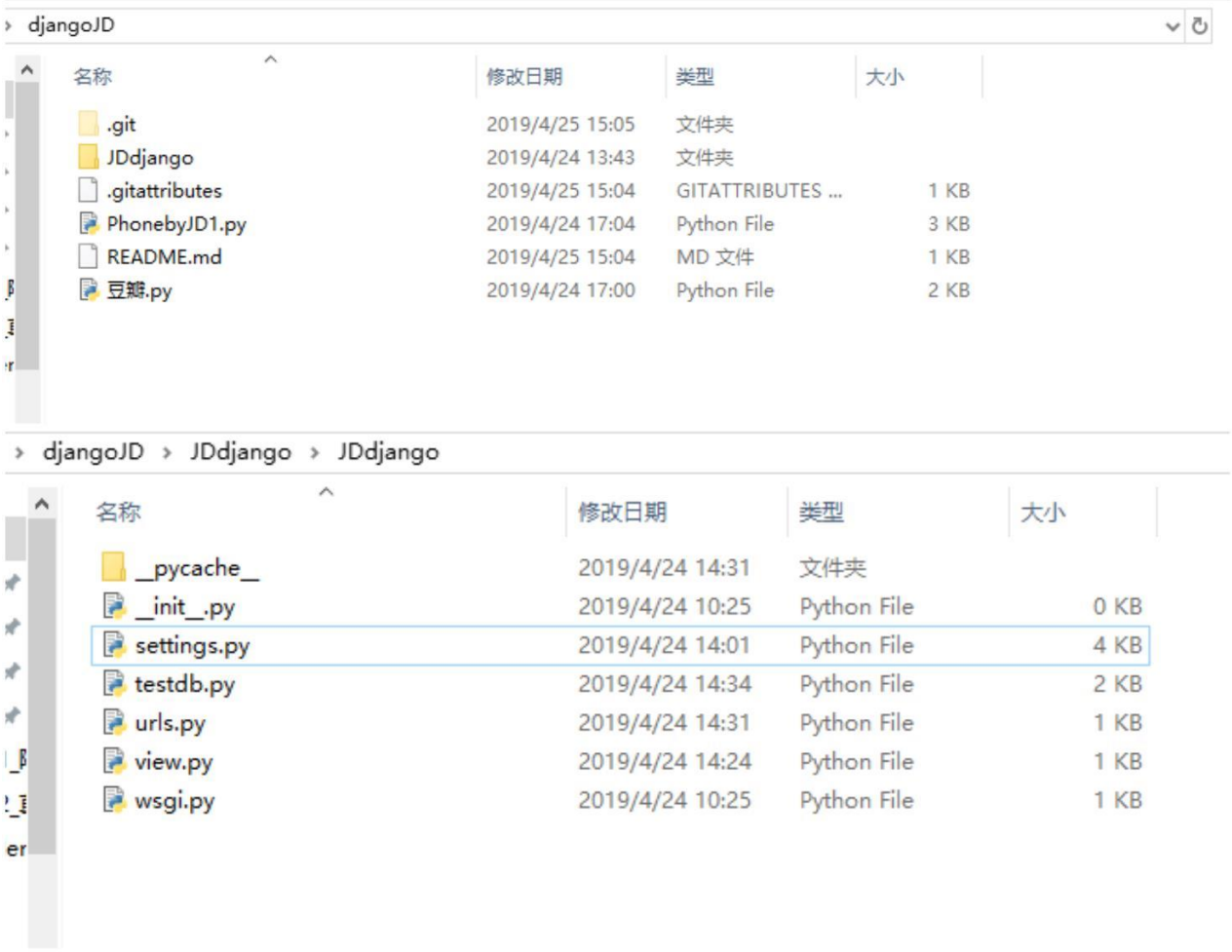
温州大学瓯江学院教务部二〇一九年四月制

实验环境

环境：VS code 编辑器,Django,Python3.5,Mysql,bootstrap

Python 需要安装 django,lxml,selenium,bs4,requests,mysqlclient 等第三方库如:pip install requests

项目结构



Dejango 配置

创建好 django 项目后打开 setting.py 文件,编辑数据库配置

```
DATABASES = {
    'default': {
        'ENGINE': 'django.db.backends.mysql',
        'NAME': 'python_crawler',
        'USER': 'root',
        'PASSWORD': 'yyl123',
        'HOST':'localhost',
        'PORT':'3306',
    }
}
```

使用 cmd 命令，cd 到项目根目录
运行命令 python manage.py migrate 创建相关数据表输入 python manage.py runserver +本机 ip+ --insecure 即可启动 django 项目

页面效果展示

电影页

手机页

点击删除，插入，更新，可以对数据库数据进行相关操作

Django 源代码

爬虫以爬取豆瓣 TOP250 电影为例 movies.py

```
import requests
from bs4 import BeautifulSoup
import pymysql


def get_movies():

conn=pymysql.connect(host='localhost',user='root',password='wangjiayue',db='mypac
hou',charset="utf8")
    cur=conn.cursor()
    headers={
        'user-agent':'Mozilla/5.0(windows
NT6.1;win64;x64)AppleWebKit/537.36(KHTML,like Gecko) Chrome/52.0.2743.82
Safari/537.36',
        'Host':'movie.douban.com'
        }
    movielist=[]
```

```
    for i in range(0,10):
        link='http://movie.douban.com/top250?start='+str(i*25)
        r=requests.get(link,headers=headers,timeout=10)
        print(str(i+1),"页面相应码：",r.status_code)
        soup=BeautifulSoup(r.text,'lxml')
        div_list=soup.find_all('div',class_='hd')
        for each in div_list:
            movie=each.a.span.text.strip()
            movielist.append(movie)
            cur.execute("INSERT INTO testmodel_movie (name)
VALUES(\"%s\")"%(movie))
    cur.close()
    conn.commit()
    conn.close()
    return movielist


movies=get_movies()
```

PhonebyJD1.py

```
import time

from selenium import webdriver

from selenium.webdriver.support import expected_conditions as EC

from selenium.webdriver.common.by import By

from selenium.webdriver.support.ui import WebDriverWait



from lxml import etree

import pymysql



browser = webdriver.Chrome()

browser.get("https://www.baidu.com")



wait = WebDriverWait(browser,50)

def search():

    browser.get('https://www.jd.com/')

    try:
```

```
            input = wait.until(
                EC.presence_of_all_elements_located((By.CSS_SELECTOR,"#key"))
            )#llist
            submit = wait.until(
                EC.element_to_be_clickable((By.CSS_SELECTOR,"#search > div > div.form > button"))
            )
            #input = browser.find_element_by_id('key')
            input[0].send_keys('phone')
            submit.click()

            total = wait.until(
                EC.presence_of_all_elements_located(
                    (By.CSS_SELECTOR,'#J_bottomPage > span.p-skip > em:nth-child(1) > b')
                )
            )
            html = browser.page_source
            prase_html(html)
            return total[0].text
        except TimeoutError:
            search()

def next_page(page_number):
        try:
            #滑动到底部，加载出后三十个货物信息
            browser.execute_script("window.scrollTo(0, document.body.scrollHeight);")
```

```
            time.sleep(10)
        #翻页动作
            button = wait.until(
                EC.element_to_be_clickable((By.CSS_SELECTOR,'#J_bottomPage > span.p-num > a.pn-next > em'))
            )
            button.click()
            wait.until(
                EC.presence_of_all_elements_located((By.CSS_SELECTOR,"#J_goodsList > ul > li:nth-child(20)"))
            )
        #判断翻页成功
            wait.until(
                EC.text_to_be_present_in_element((By.CSS_SELECTOR,"#J_bottomPage > span.p-num > a.curr"),str(page_number))
            )
            html = browser.page_source
            prase_html(html)
        except TimeoutError:
            return next_page(page_number)

def prase_html(html):
        conn=pymysql.connect(host='localhost',user='root',password='wangjiayue',db='mypachou',charset="utf8")
        cur=conn.cursor()
        html = etree.HTML(html)
        items = html.xpath('//li[@class="gl-item"]')
        for i in range(len(items)):
```

```
            if html.xpath('//div[@class="p-img"]//img')[i].get('data-lazy-img') != "done":
                img=html.xpath('//div[@class="p-img"]//img')[i].get('data-lazy-img')
                print("img:", img)
            else :
                img=html.xpath('//div[@class="p-img"]//img')[i].get('src')
                print("img:",img)
            name=html.xpath('//div[@class="p-name p-name-type-2"]//em')[i].xpath('string(.)')
            print("title:", name)
            price=html.xpath('//div[@class="p-price"]//i')[i].text
            print("price:",price)
            commit=html.xpath('//div[@class="p-commit"]//a')[i].text
            print("commit", commit)
            cur.execute("INSERT INTO testmodel_test (name,price,img,commit)
VALUES('%s','%s','%s','%s')"%(name,price,img,commit))
            print("++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++")
        cur.close()
        conn.commit()
        conn.close()


def main():
    print("第",1,"页：")
    total=int(search())
    for i in range(2,total+1):
        time.sleep(3)
        print("第",i,"页：")
        next_page(i)
```

```
if __name__=="__main__":

    main()
```

爬取天气

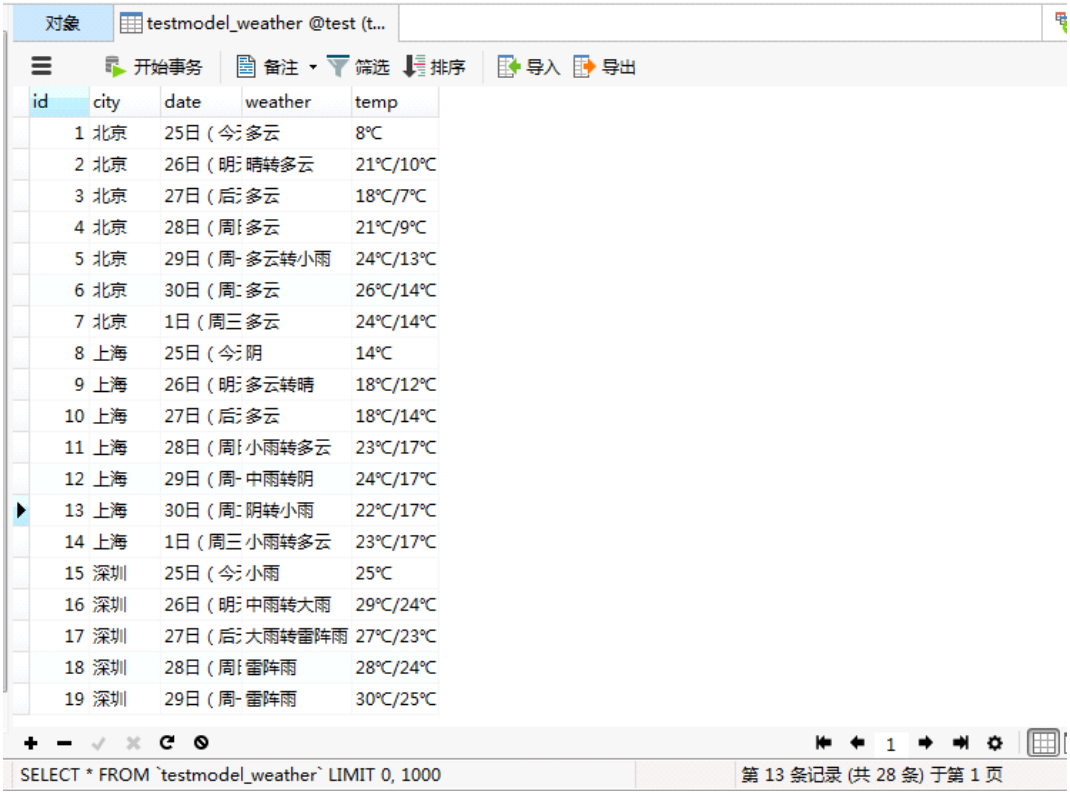
代码：

```
from bs4 import BeautifulSoup
from bs4 import UnicodeDammit
import urllib.request
import pymysql

conn=pymysql.connect(host='localhost',user='root',passwd='1234',db='test',charset="utf8")
cursor=conn.cursor()

headers={'user-agent':'Mozilla/5.0(Windows;U;Windows NT 6.0 x64;en-us;rv:1.9pre)Gecko/2008072421
MineField/3.0.2pre'}
citycode={"北京":"101010100","上海":"101020100","广州":"101280101","深圳":"101280601"}
for city in citycode:
  url="http://www.weather.com.cn/weather/"+citycode[city]+".shtml"
  try:
    req=urllib.request.Request(url,headers=headers)
    data=urllib.request.urlopen(req)
    data=data.read()
    dammint=UnicodeDammit(data,["utf-8","gbk"])
    data=dammint.unicode_markup
    soup=BeautifulSoup(data,"lxml")
    lis=soup.select("ul[class='t clearfix'] li")
    n=0
    for li in lis:
      try:
        date=li.select('h1')[0].text
        print(date)
        weather=li.select("p[class='wea']")[0].text
        if n>0:
          temp=li.select("p[class='tem'] span")[0].text+"/"+li.select("p[class='tem'] i")[0].text
        else:
          temp=li.select("p[class='tem'] i")[0].text
        cursor.execute("insert into testmodel_weather(city,date,weather,temp)
values(%s,%s,%s,%s)",(city,date,weather,temp))
        n=n+1
      except Exception as err:
        print(err)
  except Exception as err:
    print(err)
cursor.close()
conn.commit()
conn.close()
```

数据库截图



对象 testmodel_weather @test (t...

开始事务 备注 筛选 排序 导入 导出

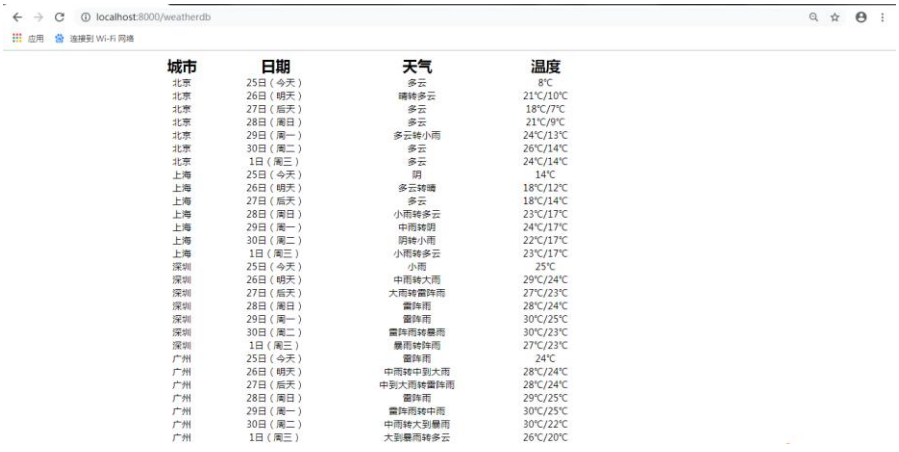
id	city	date	weather	temp
1	北京	25日（今)	多云	8℃
2	北京	26日（明)	晴转多云	21℃/10℃
3	北京	27日（后)	多云	18℃/7℃
4	北京	28日（周[	多云	21℃/9℃
5	北京	29日（周-	多云转小雨	24℃/13℃
6	北京	30日（周:	多云	26℃/14℃
7	北京	1日（周三	多云	24℃/14℃
8	上海	25日（今)	阴	14℃
9	上海	26日（明)	多云转晴	18℃/12℃
10	上海	27日（后)	多云	18℃/14℃
11	上海	28日（周[	小雨转多云	23℃/17℃
12	上海	29日（周-	中雨转阴	24℃/17℃
13	上海	30日（周:	阴转小雨	22℃/17℃
14	上海	1日（周三	小雨转多云	23℃/17℃
15	深圳	25日（今)	小雨	25℃
16	深圳	26日（明)	中雨转大雨	29℃/24℃
17	深圳	27日（后)	大雨转雷阵雨	27℃/23℃
18	深圳	28日（周[	雷阵雨	28℃/24℃
19	深圳	29日（周-	雷阵雨	30℃/25℃

1

SELECT * FROM `testmodel_weather` LIMIT 0, 1000

第 13 条记录 (共 28 条) 于第 1 页

Django截图：



localhost:8000/weatherdb

城市	日期	天气	温度
北京	25日（今天）	多云	8℃
北京	26日（明天）	晴转多云	21℃/10℃
北京	27日（后天）	多云	18℃/7℃
北京	28日（周日）	多云	21℃/9℃
北京	29日（周一）	多云转小雨	24℃/13℃
北京	30日（周二）	多云	26℃/14℃
北京	1日（周三）	多云	24℃/14℃
上海	25日（今天）	阴	14℃
上海	26日（明天）	多云转晴	18℃/12℃
上海	27日（后天）	多云	18℃/14℃
上海	28日（周日）	小雨转多云	23℃/17℃
上海	29日（周一）	中雨转阴	24℃/17℃
上海	30日（周二）	阴转小雨	22℃/17℃
上海	1日（周三）	小雨转多云	23℃/17℃
深圳	25日（今天）	小雨	25℃
深圳	26日（明天）	中雨转大雨	29℃/24℃
深圳	27日（后天）	大雨转雷阵雨	27℃/23℃
深圳	28日（周日）	雷阵雨	28℃/24℃
深圳	29日（周一）	雷阵雨	30℃/25℃
深圳	30日（周二）	雷阵雨转暴雨	30℃/23℃
深圳	1日（周三）	暴雨转阵雨	27℃/23℃
广州	25日（今天）	雷阵雨	24℃
广州	26日（明天）	中雨转中到大雨	28℃/24℃
广州	27日（后天）	中到大雨转雷阵雨	28℃/24℃
广州	28日（周日）	雷阵雨	29℃/25℃
广州	29日（周一）	雷阵雨转中雨	30℃/25℃
广州	30日（周二）	中雨转大到暴雨	30℃/22℃
广州	1日（周三）	大到暴雨转多云	26℃/20℃

12306

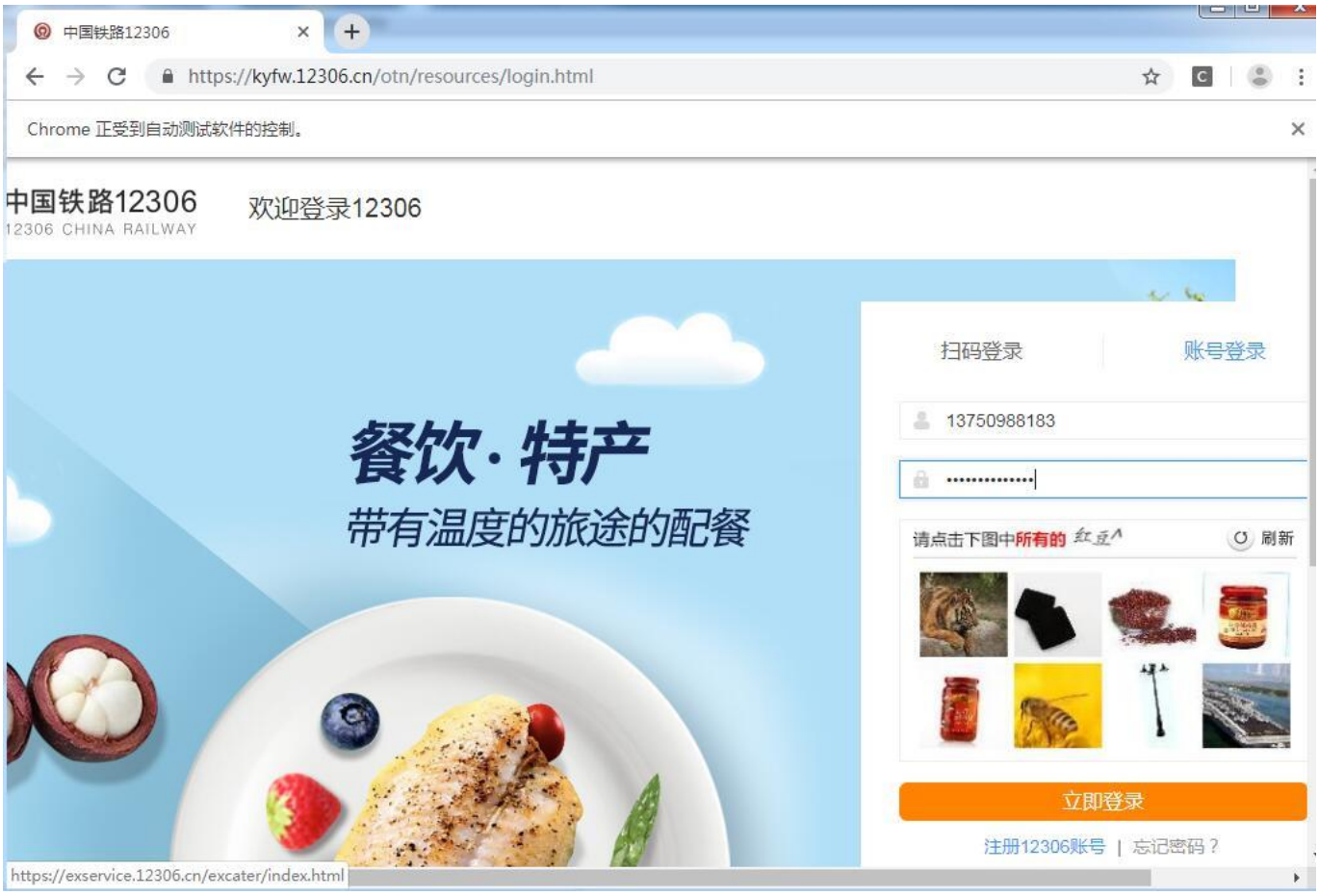
```
import re
import time
import base64
import requests
import sys
from selenium import webdriver
from selenium.webdriver.support.ui import WebDriverWait
import urllib.request
```

```
class Login(object):
    #初始化函数
    def __init__(self):
        self.login_url = "https://kyfw.12306.cn/otn/resources/login.html"
        self.totalFlush = 0
        self.startTime = time.time()
        # self.driver = '' #驱动chrome浏览器进行操作
        driver = webdriver.Chrome()
        self.driver = driver #驱动chrome浏览器进行操作


    def login_input(self):
        self.driver.get(self.login_url)
        time.sleep(0.2)
        account = self.driver.find_element_by_class_name("login-hd-account")
        account.click()
        userName = self.driver.find_element_by_id("J-userName")
        userName.send_keys("13750988183")  # 12306账号
        password = self.driver.find_element_by_id("J-password")
        password.send_keys("yuyilun123")  # 12306密码




if __name__ == '__main__':
    spider = Login()
    spider.login_input()
```


中国铁路12306
https://kyfw.12306.cn/otn/resources/login.html
Chrome 正受到自动测试软件的控制。
中国铁路12306
12306 CHINA RAILWAY
欢迎登录12306
餐饮·特产
带有温度的旅途的配餐
扫码登录
账号登录
13750988183
请点击下图中所有的 红豆
刷新
立即登录
注册12306账号 | 忘记密码？
https://exservice.12306.cn/excater/index.html


中国铁路12306
https://kyfw.12306.cn/otn/resources/login.html
Chrome 正受到自动测试软件的控制。
中国铁路12306
12306 CHINA RAILWAY
欢迎登录12306
餐饮·特产
带有温度的旅途的配餐
扫码登录
账号登录
用户名 / 邮箱 / 手机号
密码
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？