



# 温州大学瓯江学院

WENZHOU UNIVERSITY OUJIANG COLLEGE

## 《爬虫期末作业》

题    目： 爬虫期末作业

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： 陈一波

学    号： 16219111321

完成日期： 2019 年 4 月 25 日

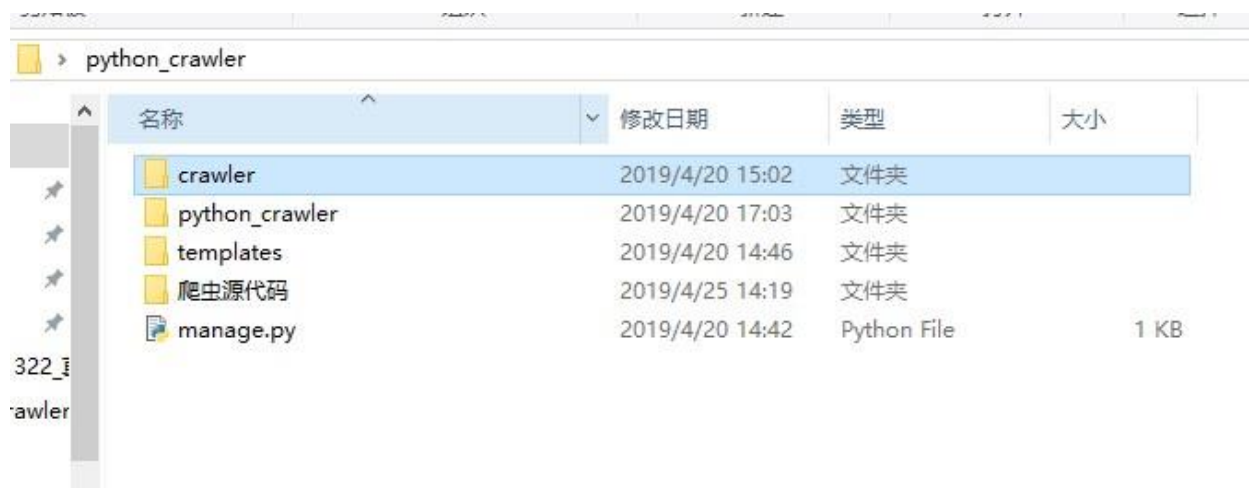
温州大学瓯江学院教务部二〇一九年四月制

## 实验环境

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

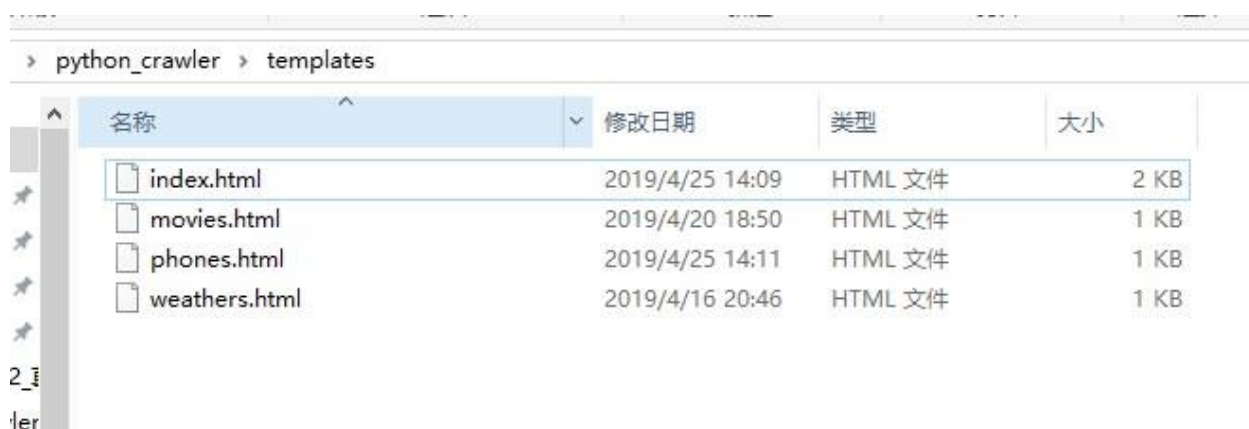
Python 需要安装 django, lxml, selenium, bs4, requests, mysqlclient 等第三方库如: pip install requests

## 项目结构



The screenshot shows a file explorer window with the path 'python\_crawler' selected. The table below represents the contents of this directory.

名称	修改日期	类型	大小
crawler	2019/4/20 15:02	文件夹	
python_crawler	2019/4/20 17:03	文件夹	
templates	2019/4/20 14:46	文件夹	
爬虫源代码	2019/4/25 14:19	文件夹	
manage.py	2019/4/20 14:42	Python File	1 KB



The screenshot shows a file explorer window with the path 'python\_crawler > templates' selected. The table below represents the contents of this directory.

名称	修改日期	类型	大小
index.html	2019/4/25 14:09	HTML 文件	2 KB
movies.html	2019/4/20 18:50	HTML 文件	1 KB
phones.html	2019/4/25 14:11	HTML 文件	1 KB
weathers.html	2019/4/16 20:46	HTML 文件	1 KB

python_crawler > python_crawler				
名称	修改日期	类型	大小	
_pycache_	2019/4/25 14:10	文件夹		
_init_.py	2019/4/20 14:42	Python File	0 KB	
chromedriver.exe	2019/4/20 15:09	应用程序	8,348 KB	
control.py	2019/4/25 14:09	Python File	2 KB	
getmovies.py	2019/4/25 14:06	Python File	2 KB	
getphones.py	2019/4/25 14:06	Python File	3 KB	
getweathers.py	2019/4/25 14:06	Python File	2 KB	
settings.py	2019/4/25 14:05	Python File	4 KB	
urls.py	2019/4/20 18:47	Python File	2 KB	
view.py	2019/4/20 17:41	Python File	1 KB	
wsgi.py	2019/4/20 14:42	Python File	1 KB	

三个 getmovies.py, getphones.py, getweathers.py 文件是爬虫文件，chromedriver.exe 是谷歌浏览器驱动，urls.py 绑定 url 与后台函数，view.py 处理前台页面内容, control.py 负责数据库爬虫操作

## 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': 'cyb123',
        '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

首页

欢迎来到爬虫网

[首页](#) | [豆瓣前250部电影](#) | [各地天气情况](#) | [京东手机](#) | [删除](#) | [插入](#) | [更新](#) |

Hello

## 电影页

欢迎来到爬虫网

[首页](#) | [豆瓣前250部电影](#) | [各地天气情况](#) | [京东手机](#) | [删除](#) | [插入](#) | [更新](#) |

## 豆瓣前250部电影

1. 肖申克的救赎 导演: 弗兰克·德拉邦特 Frank Darabont 主演: 蒂姆·罗宾斯 Tim Robbins / ... 1994 / 美国 / 犯罪 剧情 2019-04-25 14:11:25
2. 霸王别姬 导演: 陈凯歌 Kaige Chen 主演: 张国荣 Leslie Cheung / 张丰毅 Fengyi Zha... 1993 / 中国大陆 香港 / 剧情 爱情 同性 2019-04-25 14:11:26
3. 这个杀手不太冷 导演: 吕克·贝松 Luc Besson 主演: 让·雷诺 Jean Reno / 娜塔莉·波特曼 ... 1994 / 法国 / 剧情 动作 犯罪 2019-04-25 14:11:26
4. 阿甘正传 导演: 罗伯特·泽米吉斯 Robert Zemeckis 主演: 汤姆·汉克斯 Tom Hanks / ... 1994 / 美国 / 剧情 爱情 2019-04-25 14:11:26
5. 美丽人生 导演: 罗伯托·贝尼尼 Roberto Benigni 主演: 罗伯托·贝尼尼 Roberto Beni... 1997 / 意大利 / 剧情 喜剧 爱情 战争 2019-04-25 14:11:26
6. 泰坦尼克号 导演: 詹姆斯·卡梅隆 James Cameron 主演: 莱昂纳多·迪卡普里奥 Leonardo... 1997 / 美国 / 剧情 爱情 灾难 2019-04-25 14:11:26
7. 千与千寻 导演: 宫崎骏 Hayao Miyazaki 主演: 柊瑠美 Rumi Hiragi / 入野自由 Miy... 2001 / 日本 / 剧情 动画 奇幻 2019-04-25 14:11:26
8. 辛德勒的名单 导演: 史蒂文·斯皮尔伯格 Steven Spielberg 主演: 连姆·尼森 Liam Neeson... 1993 / 美国 / 剧情 历史 战争 2019-04-25 14:11:26
9. 盗梦空间 导演: 克里斯托弗·诺兰 Christopher Nolan 主演: 莱昂纳多·迪卡普里奥 Le... 2010 / 美国 英国 / 剧情 科幻 悬疑 冒险 2019-04-25 14:11:26
10. 忠犬八公的故事 导演: 莱塞·霍尔斯特姆 Lasse Hallström 主演: 理查·基尔 Richard Ger... 2009 / 美国 英国 / 剧情 2019-04-25 14:11:26
11. 机器人总动员 导演: 安德鲁·斯坦顿 Andrew Stanton 主演: 本·贝尔特 Ben Burt / 艾丽... 2008 / 美国 / 爱情 科幻 动画 冒险 2019-04-25 14:11:26
12. 三傻大闹宝莱坞 导演: 拉库马·希拉尼 Rajkumar Hirani 主演: 阿米尔·汗 Aamir Khan / 卡... 2009 / 印度 / 剧情 喜剧 爱情 歌舞 2019-04-25 14:11:26
13. 海上钢琴师 导演: 朱塞佩·托纳多雷 Giuseppe Tornatore 主演: 蒂姆·罗斯 Tim Roth / ... 1998 / 意大利 / 剧情 音乐 2019-04-25 14:11:26
14. 放牛班的春天 导演: 克里斯托夫·巴拉蒂 Christophe Barratier 主演: 热拉尔·朱尼奥 Gè... 2004 / 法国 瑞士 德国 / 剧情 音乐 2019-04-25 14:11:26
15. 楚门的世界 导演: 彼得·威尔 Peter Weir 主演: 金·凯瑞 Jim Carrey / 劳拉·琳妮 Lau... 1998 / 美国 / 剧情 科幻 2019-04-25 14:11:26
16. 大话西游之大圣娶亲 导演: 刘镇伟 Jeffrey Lau 主演: 周星驰 Stephen Chow / 吴孟达 Man Tat Ng... 1995 / 香港 中国大陆 / 喜剧 爱情 奇幻 古装 2019-04-25 14:11:26
17. 星际穿越 导演: 克里斯托弗·诺兰 Christopher Nolan 主演: 马修·麦康纳 Matthew Mc... 2014 / 美国 英国 加拿大 冰岛 / 剧情 科幻 冒险 2019-04-25 14:11:26
18. 龙猫 导演: 宫崎骏 Hayao Miyazaki 主演: 日高法子 Noriko Hidaka / 坂本千夏 Ch... 1988 / 日本 / 动画 奇幻 冒险 2019-04-25 14:11:26
19. 教父 导演: 弗朗西斯·福特·科波拉 Francis Ford Coppola 主演: 马龙·白兰度 M... 1972 / 美国 / 剧情 犯罪 2019-04-25 14:11:26
20. 熔炉 导演: 黄东赫 Dong-hyuk Hwang 主演: 孔侑 Yoo Gong / 郑有美 Yu-mi Jeong ... 2011 / 韩国 / 剧情 2019-04-25 14:11:26
21. 无间道 导演: 刘伟强 / 麦兆辉 主演: 刘德华 / 梁朝伟 / 黄秋生 2002 / 香港 / 剧情 犯罪 悬疑 2019-04-25 14:11:26
22. 疯狂动物城 导演: 拜伦·霍华德 Byron Howard / 瑞奇·摩尔 Rich Moore 主演: 金妮弗·... 2016 / 美国 / 喜剧 动画 冒险 2019-04-25 14:11:26
23. 当幸福来敲门 导演: 加布里尔·穆奇诺 Gabriele Muccino 主演: 威尔·史密斯 Will Smith ... 2006 / 美国 / 剧情 传记 家庭 2019-04-25 14:11:26

## 天气页

## 各地天气

所在位置	日期	风级	最低温度	最高温度	天气
黑龙江>哈尔滨> 城区	25日 (今天)	<3级	1°C	13	多云
吉林>长春> 城区	25日 (今天)	3-4级	3°C	14	多云
辽宁>沈阳> 城区	25日 (今天)	5-6级转3-4级	3°C	15	阵雨转多云
内蒙古>呼和浩特> 城区	25日 (今天)	3-4级	2°C	15	晴
山西>太原> 城区	25日 (今天)	3-4级转<3级	6°C	20	多云
陕西>西安> 城区	25日 (今天)	3-4级	15°C	32	晴
山东>济南> 城区	25日 (今天)	4-5级转3-4级	7°C	15	阴转多云

手机页

## 京东手机

价格: ¥ 3198.00 【预售】魅族 16s 骁龙855全面屏拍照游戏手机 6GB+128GB 碳纤维黑 全网通移动联通电信4G手机 双卡双待 2019-04-25 14:15:12

价格: ¥ 5698.00 Apple iPhone XR (A2108) 128GB 黑色 移动联通电信4G手机 双卡双待 2019-04-25 14:15:12

价格: ¥ 3298.00 【KPL官方比赛用机】vivo iQOO 44W超快闪充 8GB+128GB电光蓝 全面屏拍照手机 骁龙855电竞游戏 全网通4G手机 2019-04-25 14:15:12

价格: ¥ 3988.00 华为 HUAWEI P30 超感光徕卡三摄麒麟980AI智能芯片全面屏屏内指纹版手机8GB+64GB亮黑色全网通双4G手机双 2019-04-25 14:15:12

价格: ¥ 1299.00 荣耀8X 千元屏霸 91%屏占比 2000万AI双摄 4GB+64GB 幻夜黑 移动联通电信4G全面屏手机 双卡双待 2019-04-25 14:15:12

价格: ¥ 1199.00 小米 红米Redmi Note7 幻彩渐变AI双摄 4GB+64GB 梦幻蓝 全网通4G 双卡双待 水滴全面屏拍照游戏智能手机 2019-04-25 14:15:12

价格: ¥ 1299.00 荣耀10青春版 幻彩渐变 2400万AI自拍 全网通版4GB+64GB 渐变蓝 移动联通电信4G全面屏手机 双卡双待 2019-04-25 14:15:12

价格: ¥ 799.00 vivo U1 水滴全面屏 AI智慧拍照手机 3GB+32GB 极光色 移动联通电信全网通4G手机 2019-04-25 14:15:12

价格: ¥ 2999.00 联想Z6 Pro 8GB+128GB 黑色 骁龙855 4800万AI四摄 4000mAh大电池 PC级液冷散热 游戏手机 全网通4G 双卡双待 2019-04-25 14:15:12

价格: ¥ 799.00 小米 红米6 4GB+64GB 铂银灰 全网通4G手机 双卡双待 2019-04-25 14:15:12

价格: ¥ 2799.00 荣耀V20 胡歌同款 麒麟980芯片 魅眼全视屏 4800万深感相机 6GB+128GB 幻夜黑 移动联通电信4G全面屏手机 2019-04-25 14:15:12

价格: ¥ 899.00 荣耀畅玩8C两天一充 莱茵护眼 刘海屏 全网通版4GB+32GB 幻夜黑 移动联通电信4G全面屏手机 双卡双待 2019-04-25 14:15:12

价格: ¥ 1399.00 小米8SE 全面屏智能游戏拍照手机 6GB+64GB 灰色 骁龙710处理器 全网通4G 双卡双待 2019-04-25 14:15:12

价格: ¥ 3299.00 小米9 4800万超广角三摄 8GB+128GB全息幻彩蓝 骁龙855 全网通4G 双卡双待 水滴全面屏拍照游戏智能手机 2019-04-25 14:15:12

价格: ¥ 1499.00 小米8青春版 镜面渐变AI双摄 6GB+64GB 梦幻蓝 骁龙 全网通4G 双卡双待 全面屏拍照游戏智能手机 2019-04-25 14:15:12

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 getmovies.py

```
import requests from bs4 import
BeautifulSoup import MySQLdb
import time from crawler.models
import Movies def get_movies():

#conn=MySQLdb.connect(host="localhost",user="gjj",passwd="gjj8897",db="python_cra
wler",charset="utf8") #cur=conn.cursor()
headers={'user-agent':'Mozilla/5.0 (Windows NT 6.1;
Win64;x64) AppleWebKit/537.36 (KHTML,like Gecko) Chrome/52..02743.82
Safari/537.36','Host':'movie.douban.com'} for i in range(0,10):

link='https://movie.douban.com/top250?start='+str(i*25)
r=requests.get(link,headers=headers,timeout=10)
soup=BeautifulSoup(r.text,"lxml")
div_list=soup.find_all('div',class_='info') for each
in div_list:
    url=each.div.a['href']
title=each.div.a.span.text.strip()
synopsis=each.contents[3].p.get_text().strip()
now=time.strftime('%Y-%m-%d %H:%M:%S',time.localtime(time.time()))
record=Movies(name=title,url=url,synopsis=synopsis,time=now)
record.save()

# cur.execute("insert into crawler_movies(name,url,synopsis,time)
values(%s,%s,%s,%s)",(title,url,synopsis,now)) print("电影爬取成功! ")
# cur.close()
# conn.commit()
# conn.close()
```

引入了 django 的模型，所以无需配置数据库连接，直接在 setting.py 修改即可，但也因此无法本地运行，若要直接 python 运行要删除模型导入 from crawler.models import Movies，并把 conn 和 cur 的注释取消，删除 record

### view.py

```
from django.shortcuts import render from
django.http import HttpResponse from
crawler.models import Movies from crawler.models
import Weathers from crawler.models import Phones
def index(request): context = {}
context['hello']='Hello' return
render(request, 'index.html', context)
```

```
def
movies(request):
    movies=Movies.objects.all()      return render(request,
'movies.html', {'movies': movies})
def
weathers(request):
    weathers=Weathers.objects.all()    return render(request,
'weathers.html', {'weathers': weathers})
def
phones(request):
    phones=Phones.objects.all()
    return render(request, 'phones.html', {'phones': phones})
```

[illegible]

---

```
</body>
</html>
```

其余 html 继承 index，只需编辑<body>内容例如 movie.html

```
{% extends "index.html" %}
{% block mainbody %}
<h1>豆瓣前 250 部电影</h1>
<div style="text-align:left">
    <ol>
        {% for movie in movies %}
            <li><a
href="{{ movie.url }}">{{ movie.name }}</a>&nbsp;{{ movie.synopsis }}&nbsp;{{ movie.time }}</li>
        {% endfor %}
    </ol>
</div>
{% endblock %}
```

**control.py** 调用爬虫，实现数据更新

```
from . import getmovies,getphones,getweathers,view import
time
from crawler.models import Movies from
crawler.models import Weathers from
crawler.models import Phones from
django.shortcuts import render
from django.http import HttpResponse
```

```
def deletelall(request):
context = {}    try:
    Movies.objects.all().delete()
    Weathers.objects.all().delete()
    Phones.objects.all().delete()
context['hello']='删除成功！'    except:
    context['hello']='删除失败，请先写入数据！'
    return render(request, 'index.html', context)
```

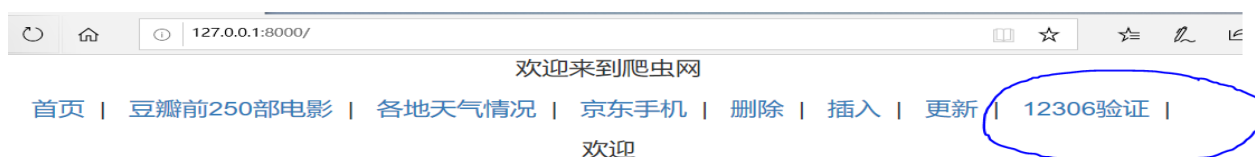
```
def insertdata(request):
context = {}    try:
    getmovies.get_movies()    time.sleep(2)
getweathers.get_weather()    time.sleep(2)
getphones.get_phones('手机')
    time.sleep(2)    context['hello']='插入成
功！'    except:
    context['hello']='插入失败！'
    return render(request, 'index.html', context)
```



```
def updatabase(request):
deletelall(request)
insertdata(request) context = {}
context['hello']='更新成功！'
return render(request, 'index.html', context)
```

## 12306 登录

点击 12306 验证,selenium 打开 chrome 跳转到 12306 登陆界面,自动将账号密码填入,并进行图片验证,自动识别并点击正确图片,提示验证成功,但由于无法跨域, 所以不能跳转登录



扫码登录

账号登录

1111111111

.....

请点击下图中所有的 **路灯**

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。

请点击下图中**所有的** 路灯

刷新



立即登录

[注册12306账号](#) | [忘记密码?](#)

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?](#)

1 12306.cn网站每日06:00~23:00提供服务