



# 温州大学瓯江学院

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

djangoJD				
名称	修改日期	类型	大小	
.git	2019/4/25 15:05	文件夹		
JDdjango	2019/4/24 13:43	文件夹		
.gitattributes	2019/4/25 15:04	GITATTRIBUTES ...	1 KB	
PhonebyJD1.py	2019/4/24 17:04	Python File	3 KB	
README.md	2019/4/25 15:04	MD 文件	1 KB	
豆瓣.py	2019/4/24 17:00	Python File	2 KB	

  

djangoJD > JDdjango > JDdjango				
名称	修改日期	类型	大小	
__pycache__	2019/4/24 14:31	文件夹		
__init__.py	2019/4/24 10:25	Python File	0 KB	
settings.py	2019/4/24 14:01	Python File	4 KB	
testdb.py	2019/4/24 14:34	Python File	2 KB	
urls.py	2019/4/24 14:31	Python File	1 KB	
view.py	2019/4/24 14:24	Python File	1 KB	
wsgi.py	2019/4/24 10:25	Python File	1 KB	

## 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': 'yy1123',
        '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

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 Burtt / 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 Gé... 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: 加布里埃尔·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

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 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 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"))
        )#list
        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()
```