



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： 数信分院

班 级： 16 计算机科学与技术三

姓 名： 王霜霜

学 号： 16219111324

完成日期： 2019 年 4 月 24 日

温州大学瓯江学院教务部

二〇一二年十一月制

目录

一、 豆瓣 Top-250 电影的爬取	1
1.1 内容介绍	1
1.1.1 Python 获取电影名称	1
1.1.2 使用 django 模板继承提高代码的复用性	1
1.1.3 Mysql 数据库的添加与内容的显示	1
1.2 截图	2
二、 天气预报的爬取	3
2.1 主要内容	3
2.2 截图	3
三、 京东手机的爬取	5
3.1 主要内容	5
3.2 截图	5
四、 淘宝书包的定向爬取	7
4.1 主要内容	7
4.2 截图	7
五、 Github 的上传及下载地址	8
5.1 下载地址	8
六、 个人信息	9

一、豆瓣 Top-250 电影的爬取

1.1 内容介绍

1.1.1 Python 获取电影名称

1. 观察每页地址得出规律通过 requests 获得相关页面信息
2. 使用 etree.HTML 将字符串格式的 html 片段解析成 html 文档
3. etree 和 xpath 结合使用: `tree.xpath('//span[@class="title"] [1]/text()')` 获取电影名称
4. 通过循环将电影名称放入空列表中

1.1.2 使用 django 模板继承提高代码的复用性

Templates/nav.html: 采用 bootstrap 样式, js 制作导航

Templates/base.html: 把公用的 HTML 部分提取出来, 放到 base.html 文件中 `{% include 'nav.html' %}` 包含 nav.html 模板; `{% block mainbody %}{% endblock %}` 用于子模板重载

movie.html, weather.html, phone.html 具体页面继承 base.html

1.1.3 Mysql 数据库的添加与内容的显示

创建 App myApp 设置 settings.py, models.py 等相关信息, 进行同步数据库操作, 系统自动创建表

通过循环使用游标 cur 操作 execute() 方法将获得的数据插入 Mysql 数据库中

showDB 方法中获取 Movies 数据表中的所有信息, 并用 render 传递给 movie.html

movie.html 中通过循环将电影编号, 电影名称放入表格每行每列 (表格, 图片轮播使用 bootstrap 样式), 调用 urls.py 中设置的地址, 查看相关信息的显示

1.2 截图

添加数据:

特工队', '三块广告牌', '无敌破坏王', '雨中曲', '冰川时代', '你的名字。', '燃情岁月', '我是山姆', '爆裂鼓手', '人工智能']
['未麻的部屋', '穿越时空的少女', '魂断蓝桥', '一个叫欧维的男人决定去死', '模仿游戏', '猜火车', '房间', '忠犬八公物语', '恐怖游轮', '罗生门', '完美陌生人', '魔女宅急便', '阿飞正传', '香水', '哪吒闹海', '浪潮', '黑客帝国3: 矩阵革命', '海街日记', '朗读者', '可可西里', '谍影重重2', '谍影重重', '战争之王', '牯岭街少年杀人事件', '地球上的星星']
['惊魂记', '青蛇', '疯狂的石头', '一次别离', '追随', '天书奇谭', '终结者2: 审判日', '源代码', '初恋这件小事', '步履不停', '小萝莉的猴神大叔', '新龙门客栈', '再次出发之纽约遇见你', '撞车', '爱在午夜降临前', '梦之安魂曲', '海蒂和爷爷', '无耻混蛋', '东京物语', '城市之光', '绿里奇迹', '彗星来的那一夜', '血钻', '2001 太空漫游', '这个男人来自地球']
['E.T. 外星人', '末路狂花', '聚焦', '功夫', '勇闯夺命岛', '变脸', '发条橙', '黄金三镖客', '黑鹰坠落', '秒速5厘米', '非常嫌疑犯', '我爱你', '卡萨布兰卡', '国王的演讲', '千钧一发', '奇迹男孩', '疯狂的麦克斯4: 狂暴之路', '遗愿清单', '美国丽人', '驴得水', '荒岛余生', '碧海蓝天', '枪火', '四个春天', '新世界']
数据添加成功!

对象 myapp_movies @scraping (...)			
开始事务 备注 筛选			
	id	mId	mName
	1	1	肖申克的救赎
	2	2	霸王别姬
	3	3	这个杀手不太冷
	4	4	阿甘正传
	5	5	美丽人生
+ - ✓ ✕ ↺ ⌂ ⏮ ⚙			
SE		第 1 条记录 (共 250 条) 于第 1 页	

显示数据:

TOP250-MOVIES

127.0.0.1:8000/m_showDB/

爬虫 豆瓣电影TOP-250 天气预报 京东手机 淘宝书包 关于

TOP1 肖申克的救赎

AZKABAN

SHAWSHANK

TOP250-Movies

排名	电影名称
1	肖申克的救赎
2	霸王别姬
3	这个杀手不太冷
4	阿甘正传
5	美丽人生
6	泰坦尼克号
7	千与千寻

辛德勒的名单

二、 天气预报的爬取

2.1 主要内容

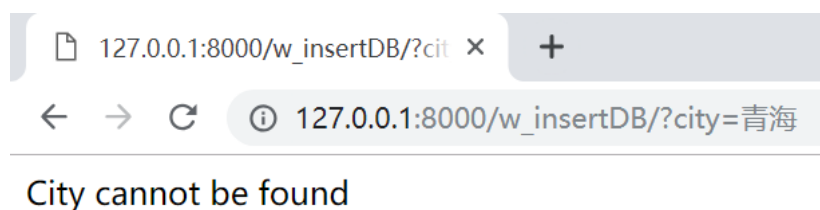
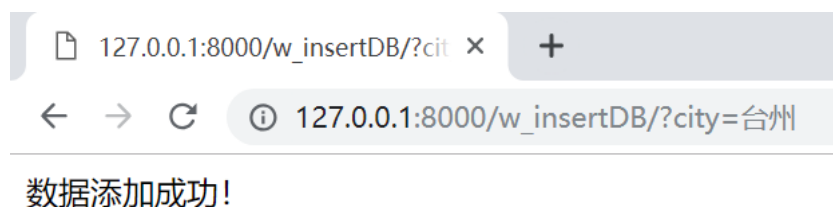
获取数据：通过使用 `urllib.request` 模块模拟浏览器的一个请求发起过程获取网页内容；`BeautifulSoup` 和 `lxml` 解析网页查找网页中的元素获得日期，天气，温度等信息

添加数据：通过使用 Django 自带的 ORM 添加数据方法循环将数据存储到 Mysql 数据库中（`weatherDB=Weathers(wCity=city,wDate=date[i],wWeather=weather[i],wTemp=temp[i])` `weatherDB.save()`）

显示数据：通过 dropdown 下拉菜单分城市显示数据；通过循环将之前保存的电影信息保存到数据库中，设置 `urls.py`，通过开启网页调用添加数据方法，完成后通过 `HttpResponse` 向浏览器发送数据添加成功的字符串

2.2 截图

添加数据：



显示数据：

爬虫

豆瓣电影TOP-250

天气预报 ▾

京东手机

城市

台州

台州

温州

宁波

杭州

绍兴

13日 (今天)

7-WEATHERS

127.0.0.1:8000/w_showDB/?city=台州

爬虫 豆瓣电影TOP-250 天气预报 ▾ 京东手机 淘宝书包 关于

台州

温州

宁波

杭州

绍兴

7-Weathers			
	日期	天气	温度
台州	13日 (今天)	阴转小雨	20/12°C
台州	14日 (明天)	小雨转多云	22/11°C
台州	15日 (后天)	多云转小雨	19/11°C
台州	16日 (周二)	多云	19/12°C
台州	17日 (周三)	晴	24/13°C
台州	18日 (周四)	晴	25/16°C
台州	19日 (周五)	晴转阴	26/19°C

127.0.0.1:8000/w_showDB/?city=台州

7-WEATHERS

127.0.0.1:8000/w_showDB/?city=温州

爬虫 豆瓣电影TOP-250 天气预报 ▾ 京东手机 淘宝书包 关于

台州

温州

宁波

杭州

绍兴

7-Weathers			
	日期	天气	温度
温州	13日 (今天)	小雨转多云	19/13°C
温州	14日 (明天)	小雨转阴	22/12°C
温州	15日 (后天)	多云转小雨	20/11°C
温州	16日 (周二)	小雨转晴	19/12°C
温州	17日 (周三)	晴转多云	26/14°C
温州	18日 (周四)	晴转多云	24/16°C
温州	19日 (周五)	晴转阴	27/18°C

127.0.0.1:8000/w_showDB/?city=温州

三、 京东手机的爬取

3.1 主要内容

通过使用 selenium(无头模式 headless) 模拟浏览器完成抓取, 将动态网页爬取变成静态网页爬取

`driver.find_element_by_xpath` 查找页面元素获取数据

`driver.execute_script("window.scrollTo(0, 7000)", '1000')` 控制浏览器滚动条已获取更多未加载出的数据

`driver.find_element_by_xpath("//span[@class='p-num']/a[@class='pn-next']")` 获取翻页, 实现翻页爬取

利用 bootstrap 实现页面的美化, 实现翻页功能









3.2 截图

```
C:\Windows\System32\cmd.exe - python manage.py runserver
005967 华为 (HUAWEI) 华为 3099.00 005967. jpg
005968 华为 (HUAWEI) 488.00 005968. jpg
005969 华为 (HUAWEI) 599.00 005969. jpg
005970 华为 (HUAWEI) 1798.00 005970. jpg
005971 OPPO 2599.00 005971. jpg
005972 华为 (HUAWEI) 1588.00 005972. jpg
005973 21KE 499.00 005973. jpg
005974 魅族 (MEIZU) 836.00 005974. jpg
005975 华为 (HUAWEI) 2178.00 005975. jpg
005976 魅族 (MEIZU) 838.00 005976. jpg
005977 易百年 459.00 005977. jpg
005978 华为 (HUAWEI) 843.00 005978. png
005979 迪美 (DIM) 397.00 005979. jpg
005980 华为 (HUAWEI) 1299.00 005980. jpg
005981 华为 (HUAWEI) 3699.00 005981. jpg
Message: Unable to locate element: //span[@class='p-num']/a[@class='pn-next']
Spider completed.....
Total 3609 seconds elapsed
Spider completed.....
```

JD-PHONES

127.0.0.1:8000/p_showDB/





爬虫 豆瓣电影TOP-250 天气预报 京东手机 淘宝书包 关于

 <p>¥ 4899.00</p> <p>华为 HUAWEI Mate 20 Pro 麒麟980芯片 全面屏超微距影像超大广角徕卡三摄 6GB+128GB亮黑色全网通版双4G</p>	 <p>¥ 5899.00</p> <p>Apple iPhone XR (A2108) 128GB 黑色 移动联通电信4G手机 双卡双待</p>	 <p>¥ 3298.00</p> <p>【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</p>	 <p>¥ 1299.00</p> <p>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 双卡双待</p>
 <p>¥ 4499.00</p> <p>华为 HUAWEI Mate20X 麒麟980芯片 全面屏超微距影像超大广角徕卡三摄 6GB+128GB宝石蓝全网通版双4G游戏</p>	 <p>¥ 2999.00</p> <p>OPPO R17 2500万美颜拍照 6.4英寸水滴屏 光感屏幕指纹 8G+128G 流光蓝 全网通 移动联通电信4G 双卡双待</p>	 <p>¥ 3198.00</p> <p>vivo 【新品上市】X27 4800万广角夜景三摄 零界全面屏拍照手机 移动联通电信全网通4G 雀羽蓝 8GB+128GB</p>	 <p>¥ 949.00</p> <p>魅族 Note8 全面屏手机 4GB+64GB 曜黑 全网通移动联通电信4G手机 双卡双待</p>

JD-PHONES

127.0.0.1:8000/p_showDB/









1 2 3 4 5 »

 <p>¥ 3099.00</p> <p>Apple iPhone 6s Plus (A1699) 128G 玫瑰金色 移动联通电信4G手机</p>	 <p>¥ 599.00</p> <p>酷派 (Coolpad) 酷玩8 Lite 6"高清全面屏 1300万双摄 私密双系统 梦幻紫 3GB+32GB 双卡双待全网通</p>	 <p>¥ 3298.00</p> <p>vivo iQOO 水滴全面屏 超广角AI三摄拍照 高通骁龙855 4G全网通 电竞游戏智能手机 焰岩橙 8GB 128GB</p>	 <p>¥ 1499.00</p> <p>华为 HUAWEI 畅享MAX 4GB+64GB 幻夜黑 全网通版 珍珠屏占比 全景声大电池 移动联通电信4G 双卡双待</p>
---	---	---	--

JD-PHONES

127.0.0.1:8000/p_showDB/?page=2

« 1 2 3 4 5 »

 <p>¥ 749.00</p> <p>魅族 V8 全面屏手机 4GB+64GB 曜黑 全网通移动联通电信4G手机 双卡双待</p>	 <p>¥ 899.00</p> <p>Meitu 美图M8s 芭比粉 4GB+64GB 自拍 云美化 夜景美化 智能 正品 手机 电影人像 4G全网通 移动版</p>	 <p>¥ 138.00</p> <p>天语 (K-TOUCH) Q31 超长待机 直板按键 三防老人手机 双卡双待 移动/联通2G 黑色</p>	 <p>¥ 139.00</p> <p>飞利浦 (PHILIPS) E163K 陨石黑 移动联通2G直板按键老人手机 双卡双待 超长待机 老年 学生备用功能机</p>
 <p>¥ 1199.80</p> <p>华为 (HUAWEI) 荣耀9i手机 (荣耀直供 限时抢购) 幻夜黑 全网通4+64GB标配版</p>	 <p>¥ 1399.00</p> <p>OPPO 【限时下单立减50+3期免息】K1 首款千元屏下指纹水滴屏手机 4G+64GB版 梵星蓝</p>	 <p>¥ 1899.00</p> <p>小米Max3 大屏游戏智能手机 6GB+128GB 黑色 骁龙处理器 全网通4G 双卡双待</p>	 <p>¥ 2299.00</p> <p>小米9 SE 4800万超广角三摄 骁龙712 水滴全面屏 游戏智能拍照手机 6GB+128GB 全息幻影蓝 全网通4G双卡双待</p>

四、 淘宝书包的定向爬取

4.1 主要内容

使用正则表达式和 requests 库对淘宝进行定向爬取

核心代码：plt=re.findall(r'"view_price"\:\.[\d\.]*"',html)

tit=re.findall(r'"raw_title"\:\.[^?]*"',html)

（使用正则表达式获取书包的价格及标题）

插入数据库方法如豆瓣电影

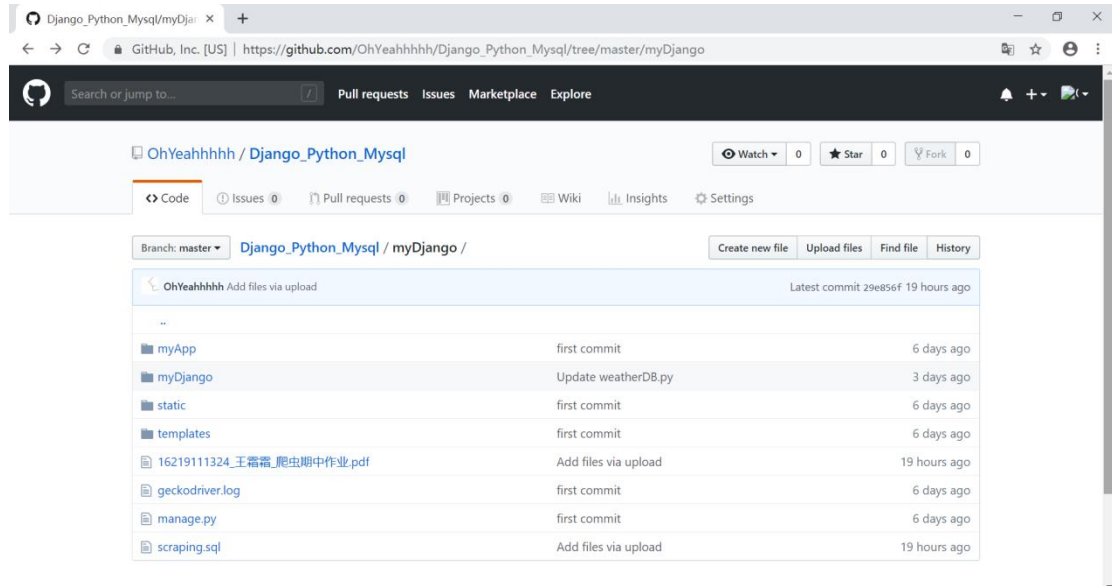
4.2 截图

8	199.00	小米 米兔儿童书包 6-12岁男女小学生潮双肩背包幼儿园大容量背包
9	79.00	双肩包男士背包大容量旅行包电脑休闲女时尚潮流高中初中学生书包
10	109.00	七匹狼商务双肩包男书包中学生女电脑包旅行包休闲男士背包大容量
11	148.00	佑一良品男士背包双肩包男韩版大学生书包男时尚潮流大容量旅行包
12	69.00	巴布豆旗舰店书包1-3年级护脊减负儿童书包男4-6小学生书包轻便
13	299.00	BOPAI博牌电脑背包男户外旅行休闲双肩包商务书包出差多功能男包
14	49.00	小学生书包6-12周岁 女童双肩包 3-5年级女童背包 1-3年级女孩
15	45.80	儿童书包小学生男童1-3年级6-12周岁4-6年级男孩双肩背包轻便减负
16	59.80	商务背包男士双肩包韩版潮流旅行包休闲女学生书包简约时尚电脑包
17	168.00	双肩包男书包男士时尚潮流青年休闲简约潮牌旅行背包大学生电脑包

SELECT * FROM `myapp_packages` LIMIT 1 第 1 条记录 (共 92 条) 于第 1 页

TAOPBAO-Packages	
价格	标题
¥ 45.80	小学生书包男生1-3-4-6年级6-12周岁儿童
¥ 39.90	迪卡侬双肩包运动背包男女健身书包儿童学生户外旅行包KIPSTA
¥ 119.00	kk树书包小学生女孩6-12周岁儿童1-3-6年级女童双肩背包护脊减负
¥ 499.00	Fjallraven/北极狐双肩包kanken classic书包女户外旅行背包23510
¥ 129.00	小米双肩包简约休闲多功能书包男女笔记本电脑包时尚潮流旅行背包
¥ 258.00	电视剧款JanSport旗舰店官网杰斯伯双肩包时尚女书包背包男大容量
¥ 348.00	爆款anello官方旗舰店日本ins潮风双肩女背包男离家出走包包
¥ 199.00	小米 米兔儿童书包 6-12岁男女小学生潮双肩背包幼儿园大容量背包
¥ 79.00	双肩包男士背包大容量旅行包电脑休闲女时尚潮流高中初中学生书包
¥ 109.00	七匹狼商务双肩包男书包中学生女电脑包旅行包休闲男士背包大容量
¥ 148.00	佑一良品男士背包双肩包男韩版大学生书包男时尚潮流大容量旅行包
¥ 69.00	巴布豆旗舰店书包1-3年级护脊减负儿童书包男4-6小学生书包轻便

五、Github 的上传及下载地址



5.1 下载地址

<https://github.com/OhYeahhhhh/own.git>

六、 个人信息

