

# 爬虫期中作业报告

目录

爬虫期中作业报告 .....1

    一、项目需求 .....3

    二、技术简介 .....3

    三、项目实施： .....4

        1、静态页面爬虫程序（豆瓣电影 top250） .....4

        2、动态页面爬虫程序（京东手机） .....6

        3、新建 Django 项目来展示数据： .....8

        4、完整项目目录 ..... 12

## 一、项目需求

编写爬虫程序爬取网页数据并存入数据库，再将数据利用 Django 展示在网页上。

## 二、技术简介

爬虫：请求网站并提取数据的自动化程序。

Mysql：最流行的关系型数据库管理系统。

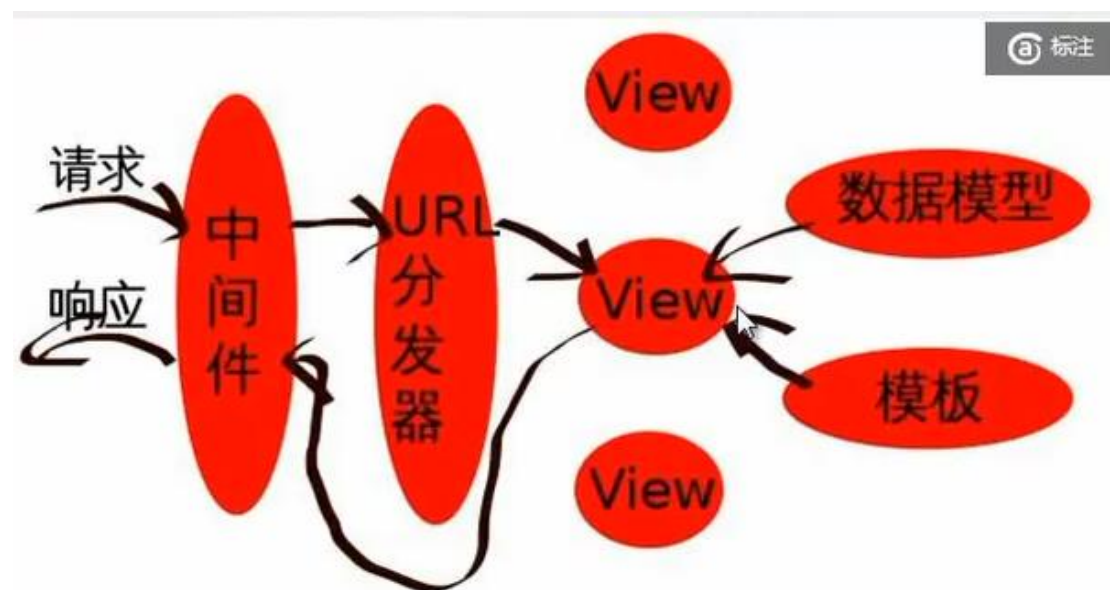
Django：一个开放源代码的 Web 应用框架，由 Python 写成。采用了 MVC 框架。适用于快速高效的上手开发一个网站。

Django MTV 模式：

m-models：用于业务对象和数据库之间的映射

t-templates：用于展示业务的模板页面

v-views：负责业务逻辑。



### 三、项目实施:

## 1、静态页面爬虫程序（豆瓣电影 top250）

### 1.1、实现思路:

用 request 库请求豆瓣页面，得到第一页的 html 代码；用正则表达式 re 库分析每条电影的正则，得到电影名、主演等信息放入字典。这样就得到第一页的 25 条电影信息；再点击下一页观察页面 url 的变化 <https://movie.douban.com/top250?start=2>，发现只有 start 的值改变，因此只需改变每次传入的 start 的值并循环调用 main (start) 方法 10 次即可。最后用多进程来提高爬取速度。

### 1.2、主要代码:

```
#请求网页
def get_one_page(url):
    try:
        response=requests.get(url)
        if response.status_code==200:
            return response.text
        return None
    except RequestException:
        return None


#解析单个页面
def parse_one_page(html):
    pattern = re.compile('<li>. *<em.*?>(\\d+)</em>. *src="(.*?)".*"title">(.*?)</span>' +'. *?<p class="">(.*?) &nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&(.*)<br>(.*?)&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&/' + '(.*?)&nbsp;&nbsp;&nbsp;&nbsp;&~(.*?)</p>. *"v:average">(.*?)</span>. *?</li>', re.S)
    items = re.findall(pattern, html)
    for item in items:
        data = {
            'index': item[0],
            'title': item[2],
            'director': item[3].strip()[3:],
            'actor': item[4][3:],
```

```

        'time': item[5].strip(),
        'region': item[6],
        'type': item[7].strip(),
        'score': item[8],
        'images': item[1]
    }
    write_to_mysql(data)
    return items

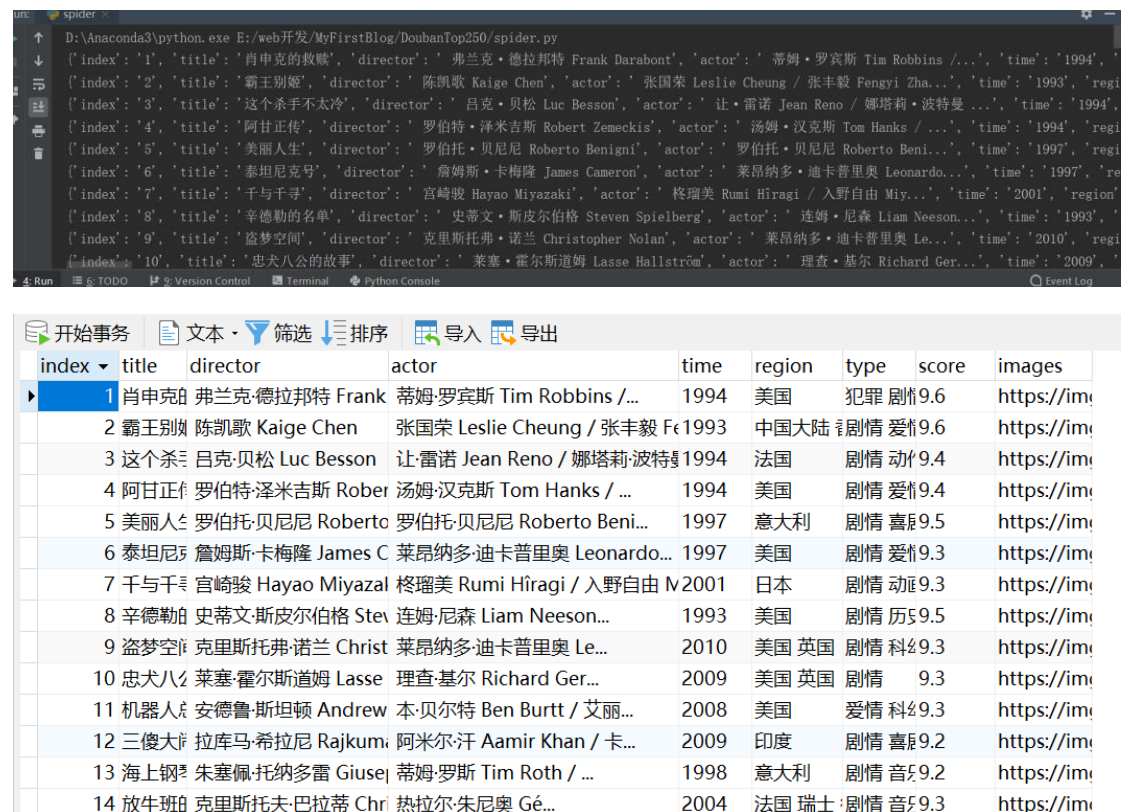
#存入mysql
def write_to_mysql(data):
    db = pymysql.connect(host='localhost', user='root',
password='123456', db='myblog', charset='utf8')
    cur = db.cursor()
    sqlc = '''
        insert into mymovie_movies
        values(null,%s,%s,%s,%s,%s,%s,%s,%s)
    '''
    try:
        if cur.execute(sqlc, (data["title"], data["director"],
data["actor"], data["time"], data["region"], data["type"],
data["score"], data["images"])):
            print('Successful')
            db.commit()
    except Exception as e:
        print(e)
        print('Failed')
        db.rollback()
    cur.close()
    db.close()

def main(start):
    url = 'https://movie.douban.com/top250?start=' + str(start)
    html = get_one_page(url)
    for item in parse_one_page(html):
        print(item)

if __name__ == '__main__':
    pool = Pool()
    pool.map(main, [i*25 for i in range(10)])

```

### 1.3、效果截图;



The top part of the image shows a terminal window with a Python script outputting a list of movie data. The bottom part shows a web browser displaying a table with the same data.

index	title	director	actor	time	region	type	score	images
1	肖申克的救赎	弗兰克·德拉邦特 Frank Darabont	蒂姆·罗宾斯 Tim Robbins / ...	1994	美国	犯罪 剧情	9.6	https://img...
2	霸王别姬	陈凯歌 Kaige Chen	张国荣 Leslie Cheung / 张丰毅 Fengyi Zha...	1993	中国大陆	剧情 爱情	9.6	https://img...
3	这个杀手不太冷	吕克·贝松 Luc Besson	让·雷诺 Jean Reno / 娜塔莉·波特曼 ...	1994	法国	剧情 动作	9.4	https://img...
4	阿甘正传	罗伯特·泽米吉斯 Robert Zemeckis	汤姆·汉克斯 Tom Hanks / ...	1994	美国	剧情 爱情	9.4	https://img...
5	美丽人生	罗伯托·贝尼尼 Roberto Benigni	罗伯托·贝尼尼 Roberto Beni...	1997	意大利	剧情 喜剧	9.5	https://img...
6	泰坦尼克号	詹姆斯·卡梅隆 James Cameron	莱昂纳多·迪卡普里奥 Leonardo...	1997	美国	剧情 爱情	9.3	https://img...
7	千与千寻	宫崎骏 Hayao Miyazaki	柊瑠美 Rumi Hiragi / 入野自由 Miy...	2001	日本	剧情 动画	9.3	https://img...
8	辛德勒的名单	史蒂文·斯皮尔伯格 Steven Spielberg	连姆·尼森 Liam Neeson...	1993	美国	剧情 历史	9.5	https://img...
9	盗梦空间	克里斯托弗·诺兰 Christopher Nolan	莱昂纳多·迪卡普里奥 Le...	2010	美国 英国	剧情 科幻	9.3	https://img...
10	忠犬八公	莱塞·霍尔斯道姆 Lasse Hallström	理查·基尔 Richard Ger...	2009	美国 英国	剧情 科幻	9.3	https://img...
11	机器人总动员	安德鲁·斯坦顿 Andrew Stanton	本·贝尔特 Ben Burtt / 艾丽...	2008	美国	爱情 科幻	9.3	https://img...
12	三傻大闹宝莱坞	拉库马·希拉尼 Rajkumar Hirani	阿米尔·汗 Aamir Khan / 卡...	2009	印度	剧情 喜剧	9.2	https://img...
13	海上钢琴师	朱塞佩·托纳多雷 Giuseppe Tornatore	蒂姆·罗斯 Tim Roth / ...	1998	意大利	剧情 音乐	9.2	https://img...
14	放牛班的春天	克里斯托夫·巴拉蒂 Christophe Barratier	热拉尔·朱尼奥 Gérald Lamoignon	2004	法国 瑞士	剧情 音乐	9.3	https://img...

## 2、动态页面爬虫程序（京东手机）

### 2.1、实现思路：

用 selenium 库模拟浏览器来爬取京东手机商品信息。首先下载对应的 chromedriver，并设置为无头浏览；通过 css 选择器来找到商品的输入框并输入“手机”且提交；同样的模拟翻页的情况，总共有 100 页；引入 pyquery 库，通过分析节点来获取到商品的不同信息；最后存入到数据库。

### 2.2、主要代码：

```
def search():
    try:
        browser.get("https://www.jd.com/")
        input = wait.until(EC.presence_of_element_located((By.ID,
'key'))))
        submit =
```

```

wait.until(EC.presence_of_element_located((By.CSS_SELECTOR,
'#search > div > div.form > button'))))
    input.send_keys(keyword)
    submit.click()
    total =
wait.until(EC.presence_of_element_located((By.CSS_SELECTOR,
'#J_bottomPage > span.p-skip > em:nth-child(1) > b'))))
    get_products()
    return total.text
except TimeoutException:
    return search()

def next_page(page_number):
    print('正在翻页', 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(5)

        input =
wait.until(EC.presence_of_element_located((By.CSS_SELECTOR,
'#J_bottomPage > span.p-skip > input'))))
        submit =
wait.until(EC.element_to_be_clickable((By.CSS_SELECTOR,
'#J_bottomPage > span.p-skip > a'))))
        input.clear()
        input.send_keys(page_number)
        submit.click()
        wait.until(EC.text_to_be_present_in_element((By.CSS_SELECTOR,
'#J_bottomPage > span.p-num > a.curr'), str(page_number)))
        get_products()
    except TimeoutException:
        next_page(page_number)

def get_products():
    wait.until(EC.presence_of_element_located((By.CSS_SELECTOR,
'#J_goodsList > ul'))))
    html = browser.page_source
    doc = pq(html)
    items = doc('#J_goodsList .gl-warp .gl-item .gl-i-wrap').items()
    images = browser.find_element_by_xpath('//div[@class="gl-i-
wrap"]/div[1]/a/img')
    #获取商品信息列表

```

```

for item in items:
    product = {
        'name': re.search('.*?\n', item.find('.p-
name').text()).group(0)[: -1],
        'price': item.find('.p-price').text()[2:],
        'commit': re.sub('\n', '', item.find('.p-
commit').text()),
        'shop': item.find('.p-shop').text(),
        'icons': re.sub('\n', '', item.find('.p-icons').text()),
        'image': images.get_attribute('src')
    }
    print(product)
    write_to_mysql(product)

```

## 2.3、效果截图：

<div> <span>开始事务</span> <span>文本</span> <span>筛选</span> <span>排序</span> <span>导入</span> <span>导出</span> </div>						
id	name	price	commit	shop	icons	image
1	OPPO Re	3599.00	1300+条评	OPPO京	自营 赠	https://img
2	Apple iPl	5899.00	二手有售91	Apple产	自营 券5	https://img
3	【KPL官	73298.00	二手有售8.4	vivo京东	自营 新品	https://img
4	荣耀8X 千	1298.00	二手有售14	荣耀京东	自营 放心	https://img
5	荣耀10青	1298.00	二手有售47	荣耀京东	自营 放心	https://img
6	vivo U1 z	799.00	13万+条评	vivo京东	自营 放心	https://img
7	小米 红米	1199.00	57万+条评	小米京东	自营 放心	https://img
8	荣耀V20	2798.00	二手有售23	荣耀京东	自营 满赠	https://img
9	OPPO Re	2999.00	0条评价	牧申手机	赠 险	https://img
10	荣耀畅玩	898.00	40万+条评	荣耀京东	自营 放心	https://img
11	小米 红米	799.00	二手有售70	小米京东	自营 放心	https://img

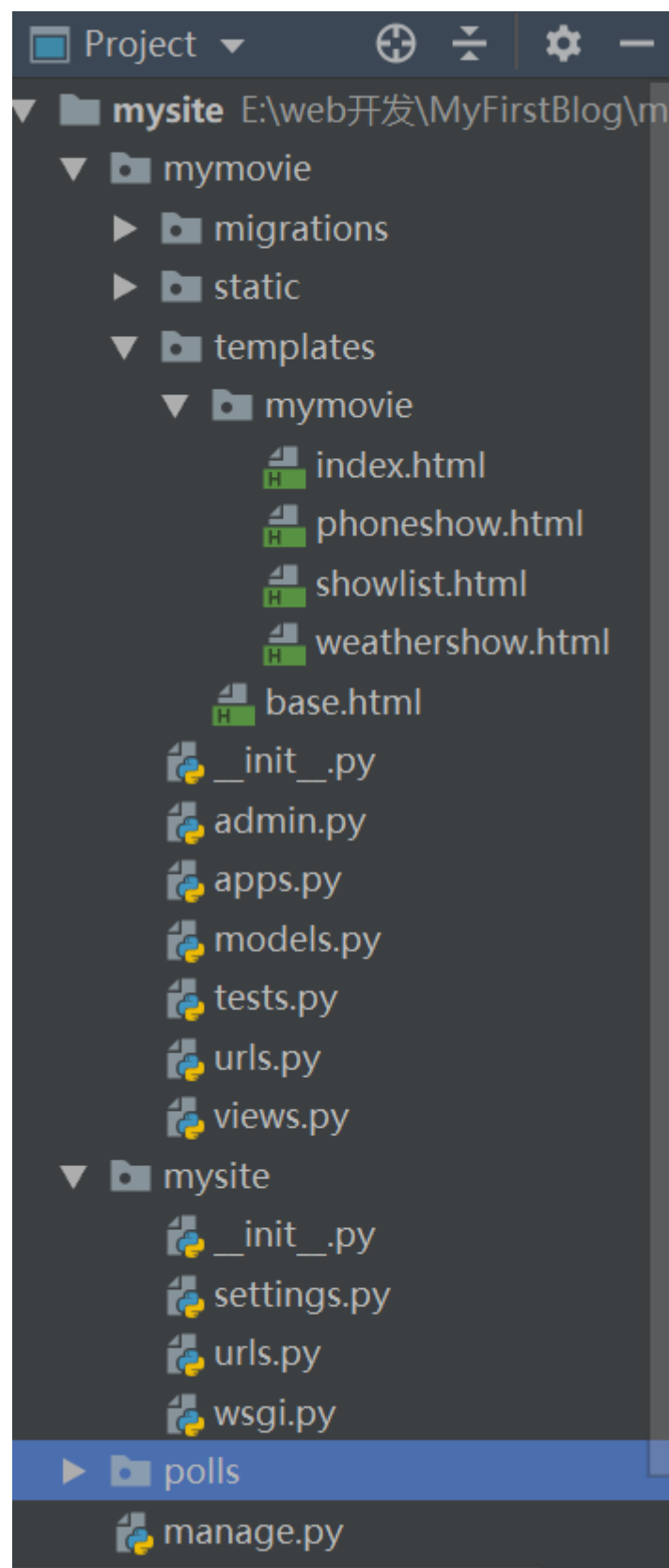
## 3、新建 Django 项目来展示数据：

### 3.1、实现步骤：

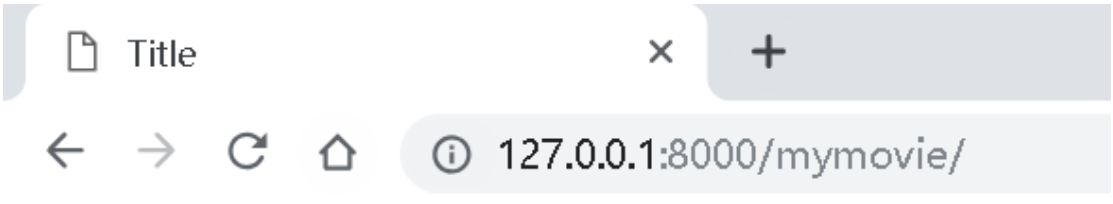
首先创建 models，映射到 mysql；创建 templates，编写要展示的 html 页面；在 views 里编写功能并调用 templates；编写 urls。



### 3.2、项目目录结构：



3.3、效果截图：



[豆瓣电影top250](#) [京东手机商品展示](#) [天气预报](#)



豆瓣电影Top250							
排名	标题	导演	主演	发行时间	地区	类型	评分
1	肖申克的救赎	弗兰克·德拉邦特 Frank Darabont	蒂姆·罗宾斯 Tim Robbins / ...	1994	美国	犯罪 剧情	9.6
2	霸王别姬	陈凯歌 Kaige Chen	张国荣 Leslie Cheung / 张丰毅 Fengyi Zha...	1993	中国大陆 香港	剧情 爱情 同性	9.6
3	这个杀手不太冷	吕克·贝松 Luc Besson	让·雷诺 Jean Reno / 娜塔莉·波特曼 ...	1994	法国	剧情 动作 犯罪	9.4
4	阿甘正传	罗伯特·泽米吉斯 Robert Zemeckis	汤姆·汉克斯 Tom Hanks / ...	1994	美国	剧情 爱情	9.4
5	美丽人生	罗伯托·贝尼尼 Roberto Benigni	罗伯托·贝尼尼 Roberto Beni...	1997	意大利	剧情 喜剧 爱情 战争	9.5
6	泰坦尼克号	詹姆斯·卡梅隆 James Cameron	莱昂纳多·迪卡普里奥 Leonardo...	1997	美国	剧情 爱情 灾难	9.3
7	千与千寻	宫崎骏 Hayao Miyazaki	柊瑠美 Rumi Hiragi / 入野自由 Miy...	2001	日本	剧情 动画 奇幻	9.3
8	辛德勒的名单	史蒂文·斯皮尔伯格 Steven Spielberg	连姆·尼森 Liam Neeson...	1993	美国	剧情 历史 战争	9.5
9	盗梦空间	克里斯托弗·诺兰 Christopher Nolan	莱昂纳多·迪卡普里奥 Le...	2010	美国 英国	剧情 科幻 悬疑 冒险	9.3
10	忠犬八公的故事	莱塞·霍尔斯道姆 Lasse Hallström	理查·基尔 Richard Ger...	2009	美国 英国	剧情	9.3
11	机器人总动员	安德鲁·斯坦顿 Andrew Stanton	本·贝尔特 Ben Burtt / 艾丽...	2008	美国	爱情 科幻 动画 冒险	9.3
12	三傻大闹宝莱坞	拉库马·希拉尼 Rajkumar Hirani	阿米尔·汗 Aamir Khan / 卡...	2009	印度	剧情 喜剧 爱情 歌舞	9.2
13	海上钢琴师	朱塞佩·托纳多雷 Giuseppe Tornatore	蒂姆·罗斯 Tim Roth / ...	1998	意大利	剧情 音乐	9.2
14	放牛班的春天	克里斯托夫·巴拉蒂 Christophe Barratier	热拉尔·朱尼奥 Gé...	2004	法国 瑞士 德国	剧情 音乐	9.3
15	楚门的世界	彼得·威尔 Peter Weir	金·凯瑞 Jim Carrey / 劳拉·琳妮 Lau...	1998	美国	剧情 科幻	9.2



京东手机商品展示

127.0.0.1:8000/mymovie/showphone/

京东手机商品展示

编号	手机	价格	评价人数	店铺	标签	图片
1	OPPO Reno 全面屏拍照	3599.00	1300+条评价	OPPO京东自营官方旗舰店	自营 赠	
2	Apple iPhone XR (A2108) 128GB 黑色 移动联通电信4G	5899.00	二手有售91万+条评价	Apple产品京东自营旗舰店	旗舰店 /qi jian dian/ flagship store	

Windows taskbar with icons for File Explorer, Chrome, Firefox, etc.

15:18 2019/4/26

WeatherForecast

127.0.0.1:8000/mymovie/showweather/







未来8天天气预报

日期	天气	温度	风向	风力
周四 (25日)	雨转阴	28°C/19°C	东南风转西北风	<3级
周五 (26日)	阴	29°C/19°C	南风	<3级
周六 (27日)	雨	25°C/18°C	南风转西南风	<3级
周日 (28日)	雨	25°C/17°C	东风转东北风	<3级
周一 (29日)	雨	24°C/15°C	东风转北风	<3级
周二 (30日)	雨	23°C/16°C	东风转南风	<3级
周三 (1日)	雨	25°C/15°C	西南风转东北风	<3级
周四 (2日)	雨	20°C/14°C	北风	<3级

Windows taskbar with icons for File Explorer, Chrome, Firefox, etc.

15:18 2019/4/26

## 4、完整项目目录

 DoubanTop250
 JDPhones
 WeatherForecast
 mysite
 README.md
 django_admin.txt

DouanTop250 为静态页面爬虫程序，JDPhones 为动态页面爬虫程序，mysite 为 Django 项目。