爬取豆瓣电影 TOP250，django 显示

douban.py 爬取文件

1. url 分析

豆瓣的 url 基本格式

```
https://movie.douban.com/top250?start={}
```

2. 数据分析

```
<li>
    <div class="item">
        <div class="pic">
            <em class="">1</em>
            <a href="https://movie.douban.com/subject/1292052/">
                <img width="100" alt="肖申克的救赎" src="https://img3.doubanio.com/view/photo/s_ratio_poster/public/p480747492.webp
            </a>
        </div>
        <div class="info">
            <div class="hd">
                <a href="https://movie.douban.com/subject/1292052/" class="">
                    <span class="title">肖申克的救赎</span>
                            <span class="title"> / The Shawshank Redemption</span>
                        <span class="other"> / 月黑高飞(港)  /  刺激1995(台)</span>
                </a>


                    <span class="playable">[可播放]</span>
            </div>
            <div class="bd">
                <p class="">
                    导演: 弗兰克·德拉邦特 Frank Darabont   主演: 蒂姆·罗宾斯 Tim Robbins /...<br>
                    1994 / 美国 / 犯罪 剧情
                </p>


                <div class="star">
                        <span class="rating5-t"></span>
                        <span class="rating_num" property="v:average">9.6</span>
                        <span property="v:best" content="10.0"></span>
                        <span>1400939人评价</span>
                </div>

                    <p class="quote">
                        <span class="inq">希望让人自由。</span>
                    </p>
            </div>
        </div>
    </div>
</li>
```

需要的就是标签里的数据，于是调用 lxml 库用 Xpath 解析

xpath语法

表达式	说明
article	选取所有article元素的所有子节点
/article	选取根元素article
article/a	选取所有属于article的子元素的a元素
//div	选取所有div子元素(不论出现在文档任何地方)
article//div	选取所有属于article元素的后代的div元素，不管它出现在article之下的任何位置
//@class	选取所有名为class的属性

xpath语法-谓语

表达式	说明
/article/div[1]	选取属于article子元素的第一个div元素
/article/div[last()]	选取属于article子元素的最后一个div元素
/article/div[last()-1]	选取属于article子元素的倒数第二个div元素
//div[@lang]	选取所有拥有lang属性的div元素
//div[@lang='eng']	选取所有lang属性为eng的div元素

xpath语法

表达式	说明
/div/*	选取属于div元素的所有子节点
//*	选取所有元素
//div[@*]	选取所有带属性的 div 元素
/div/a | //div/p	选取所有div元素的a和p元素
//span | //ul	选取文档中的span和ul元素
article/div/p| //span	选取所有属于article元素的div元素的p元素 以及文档中所有的span元素

这里还用到 `requests` 库传输 `url`

3. 爬取数据

代码解析

链接数据库

```
conn =
pymysql.connect(host='localhost',user='root',password='123456',db='qzpy
',port=3306,charset='utf8')
cursor=conn.cursor()    #连接数据库
```

页面的 url

```
if __name__=='__main__':     #主程序入口
    urls=['https://movie.douban.com/top250?start={}'.format(i*25) for i
in range(0,10)]     #页面

    for url in urls:
        get_info(url)
        time.sleep(random.random()*2)
    conn.commit()
```

访问头 headers，`res.status_code`**判断是否响应，执行爬取代码。Xpath 用法如上，**

`cursor.execute`**将相应字段与数据库的字段对应，写入数据库**

```
headers={
     'User-Agent':'Mozilla/5.0 (Windows NT 10.0; WOW64) AppleWebKit/537.36
(KHTML, like Gecko) Chrome/63.0.3239.26 Safari/537.36 Core/1.63.5478.400
QQBrowser/10.1.1550.400'
}



def get_info(url):
    res=requests.get(url,headers=headers)
    if res.status_code==200:
        selector=etree.HTML(res.text)
        infos=selector.xpath('//div[@class="item"]')
        for info in infos:

name=info.xpath('div[2]/div[@class="hd"]/a/span[1]/text()')[0]
            url=info.xpath('div[1]/a/@href')[0]

movies_infos=info.xpath('div[2]/div[@class="bd"]/p[1]/text()')[0].strip
('\n').strip('\xa0')

#movies_infos=info.xpath('div[2]/div[@class="bd"]/p[1]/text()')

            director = movies_infos[:movies_infos.find('主演')].strip()
```

```
            actor = movies_infos[movies_infos.find('主演'):].strip()


ping_infos=info.xpath('div[2]/div[@class="bd"]/p[1]/text()')[1].strip('
\n').strip('\xa0')
            year =ping_infos.split('/')[0].strip()
            region =ping_infos.split('/')[1].strip()
            type =ping_infos.split('/')[2].strip()


rate=info.xpath('div[2]/div[2]/div[@class="star"]/span[2]/text()')[0]

comments=info.xpath('div[2]/div[2]/p[@class="quote"]/span[1]/text()')
            if len(comments)!=0:
                comment=comments[0]
            else:
                comment='空'        #防止无评论

            #print(comment)
            cursor.execute("insert into
doubanmovie(name,url,director,actor,year,region,type,rate,comment)
values(%s,%s,%s,%s,%s,%s,%s,%s,%s)",

(str(name),str(url),str(director),str(actor),str(year),str(region),str(
type),str(rate),str(comment)))     #按对应字段写入数据库


    else:
        print('failed')
```

运行结果

id	name	url	director	actor	year	region	type	rate	comment
1	肖申克的救赎	https://movie.dou	导演: 弗兰克·德拉邦	主演: 蒂姆·罗宾斯 Tim Rob	1994	美国	犯罪 剧情	9.6	希望让人自由。
2	霸王别姬	https://movie.dou	导演: 陈凯歌 Kaige	主演: 张国荣 Leslie Cheung	1993	中国大陆 香港	剧情 爱情 同性	9.6	风华绝代。
3	这个杀手不太冷	https://movie.dou	导演: 吕克·贝松 Luc	主演: 让·雷诺 Jean Reno /	1994	法国	剧情 动作 犯罪	9.4	怪蜀黍和小萝莉不得不说的
4	阿甘正传	https://movie.dou	导演: 罗伯特·泽米吉	主演: 汤姆·汉克斯 Tom Har	1994	美国	剧情 爱情	9.4	一部美国近现代史。
5	美丽人生	https://movie.dou	导演: 罗伯托·贝尼尼	主演: 罗伯托·贝尼尼 Robert	1997	意大利	剧情 喜剧 爱情 战争	9.5	最美的谎言。
6	泰坦尼克号	https://movie.dou	导演: 詹姆斯·卡梅隆	主演: 莱昂纳多·迪卡普里奥	1997	美国	剧情 爱情 灾难	9.3	失去的才是永恒的。
7	千与千寻	https://movie.dou	导演: 宫崎骏 Haya	主演: 柊瑠美 Rumi Hiragi /	2001	日本	剧情 动画 奇幻	9.3	最好的宫崎骏，最好的久石
8	辛德勒的名单	https://movie.dou	导演: 史蒂文·斯皮尔	主演: 连姆·尼森 Liam Nees	1993	美国	剧情 历史 战争	9.5	拯救一个人，就是拯救整个
9	盗梦空间	https://movie.dou	导演: 克里斯托弗·诺	主演: 莱昂纳多·迪卡普里奥	2010	美国 英国	剧情 科幻 悬疑 冒险	9.3	诺兰给了我们一场无法盗取
10	忠犬八公的故事	https://movie.dou	导演: 莱塞·霍尔斯道	主演: 理查·基尔 Richard Ge	2009	美国 英国	剧情	9.3	永远都不能忘记你所爱的人
11	机器人总动员	https://movie.dou	导演: 安德鲁·斯坦顿	主演: 本·贝尔特 Ben Burtt /	2008	美国	爱情 科幻 动画 冒险	9.3	小瓦力，大人生。
12	三傻大闹宝莱坞	https://movie.dou	导演: 拉库马·希拉尼	主演: 阿米尔·汗 Aamir Khan	2009	印度	剧情 喜剧 爱情 歌舞	9.2	英俊版憨豆，高情商版谢耳
13	海上钢琴师	https://movie.dou	导演: 朱塞佩·托纳多	主演: 蒂姆·罗斯 Tim Roth /	1998	意大利	剧情 音乐	9.2	每个人都要走一条自己坚定
14	放牛班的春天	https://movie.dou	导演: 克里斯托夫·巴	主演: 热拉尔·朱尼奥 Gé...	2004	法国 瑞士 德国	剧情 音乐	9.3	天籁一般的童声，是最接近
15	楚门的世界	https://movie.dou	导演: 彼得·威尔 Pet	主演: 金·凯瑞 Jim Carrey /	1998	美国	剧情 科幻	9.2	如果再也不能见到你，祝你
16	大话西游之大圣	https://movie.dou	导演: 刘镇伟 Jeffre	主演: 周星驰 Stephen Cho	1995	香港 中国大陆	喜剧 爱情 奇幻 古装	9.2	一生所爱。
17	星际穿越	https://movie.dou	导演: 克里斯托弗·诺	主演: 马修·麦康纳 Matthew	2014	美国 英国 加拿大	剧情 科幻 冒险	9.2	爱是一种力量，让我们超越
18	龙猫	https://movie.dou	导演: 宫崎骏 Haya	主演: 日高法子 Noriko Hid	1988	日本	动画 奇幻 冒险	9.2	人人心中都有个龙猫，童年
19	教父	https://movie.dou	导演: 弗朗西斯·福特	主演: 马龙·白兰度 M...	1972	美国	剧情 犯罪	9.3	千万不要记恨你的对手，这
20	熔炉	https://movie.dou	导演: 黄东赫 Dong	主演: 孔侑 Yoo Gong / 郑	2011	韩国	剧情	9.3	我们一路奋战不是为了改变

这个 id 列是 django 生成的表默认的列

接下来实现 Django 显示

Helloworld 文件夹

安装使用可以借鉴：http://www.runoob.com/django/django-first-app.html

代码实现

Helloword 目录

Settings.py

加入 app 名称

```
INSTALLED_APPS = [
    'django.contrib.admin',
    'django.contrib.auth',
    'django.contrib.contenttypes',
    'django.contrib.sessions',
    'django.contrib.messages',
    'django.contrib.staticfiles',
    'ysyapp',
]
```

数据库配置连接

```
DATABASES = {
    'default': {
        'ENGINE': 'django.db.backends.mysql',
        'NAME':'qzpy',
        'USER':'root',
        'PASSWORD':'123456',
        'HOST':'localhost',
        'PORT':'3306',
    }
}
```

urls.py:加入路径导向，由于不会直接调用网站命令

在 app 文件里加入了 urls.py 做指引

```
path('ysy/', include('ysyapp.urls')),
```

app 目录下

Models.py:字段名与数据库相对应

```python
class Doubanmovie(models.Model):
    name = models.CharField(max_length=255, blank=True, null=True)
    url = models.CharField(max_length=255, blank=True, null=True)
    director = models.CharField(max_length=255, blank=True, null=True)
    actor = models.CharField(max_length=255, blank=True, null=True)
    year = models.CharField(max_length=255, blank=True, null=True)
    region = models.CharField(max_length=255, blank=True, null=True)
    type = models.CharField(max_length=255, blank=True, null=True)
    rate = models.CharField(max_length=255, blank=True, null=True)
    comment = models.CharField(max_length=255, blank=True, null=True)

    class Meta:
        managed = True
        db_table = 'doubanmovie'
```

不过他生成是有 id 列的新的 doubanmovie 表，不想再爬一次可以参考

https://www.cnblogs.com/smiling-crying/p/9237452.html

迁移映射命令：

```
D:\codefile\Helloword>python manage.py makemigrations ysyapp
Migrations for 'ysyapp':
  ysyapp\migrations\0001_initial.py
    - Create model Doubanmovie

D:\codefile\Helloword>python manage.py migrate
Operations to perform:
  Apply all migrations: admin, auth, contenttypes, sessions, ysyapp
Running migrations:
  Applying contenttypes.0001_initial... OK
  Applying auth.0001_initial... OK
  Applying admin.0001_initial... OK
  Applying admin.0002_logentry_remove_auto_add... OK
  Applying admin.0003_logentry_add_action_flag_choices... OK
  Applying contenttypes.0002_remove_content_type_name... OK
  Applying auth.0002_alter_permission_name_max_length... OK
  Applying auth.0003_alter_user_email_max_length... OK
  Applying auth.0004_alter_user_username_opts... OK
  Applying auth.0005_alter_user_last_login_null... OK
  Applying auth.0006_require_contenttypes_0002... OK
  Applying auth.0007_alter_validators_add_error_messages... OK
  Applying auth.0008_alter_user_username_max_length... OK
  Applying auth.0009_alter_user_last_name_max_length... OK
  Applying sessions.0001_initial... OK
```

Views.py:主要连接数据库

```
import MySQLdb
def get_data(sql):#获取数据库的数据
    conn = MySQLdb.connect('127.0.0.1','root','123456','qzpy',port=3306)
#test 为数据库名
    cur = conn.cursor()
    cur.execute(sql)
    results = cur.fetchall() # 搜取所有结果
    cur.close()
    conn.close()
    return results
def order(request):# 向页面输出订单
    sql = "select * from doubanmovie"
    m_data = get_data(sql)
    return render(request,'index.html',{'order':m_data})
```

urls.py

```
urlpatterns = [

    path('', views.order, name='order'),

]
```

网页展示，原本想试试 django-tables2 仍是新手，遇到部分不兼容啥的

写了个很粗糙的网页

App.templates.index.html

```
<html>
    <head>
    <title>豆瓣电影 top250</title>
    </head>
    <body>
       <font color="#FF0000" size="20px"><center> 豆瓣电影</center></font>
    <br>
    <table border="2" align="center">
        <tr>
            <td width="100" id="title">排名</td>
            <td width="100" name="title">电影</td>
            <td width="100" url="title">链接</td>
            <td width="100" author="title">导演</td>
            <td width="100" publisher="title">主演</td>
            <td width="100" year="title">年份</td>
```

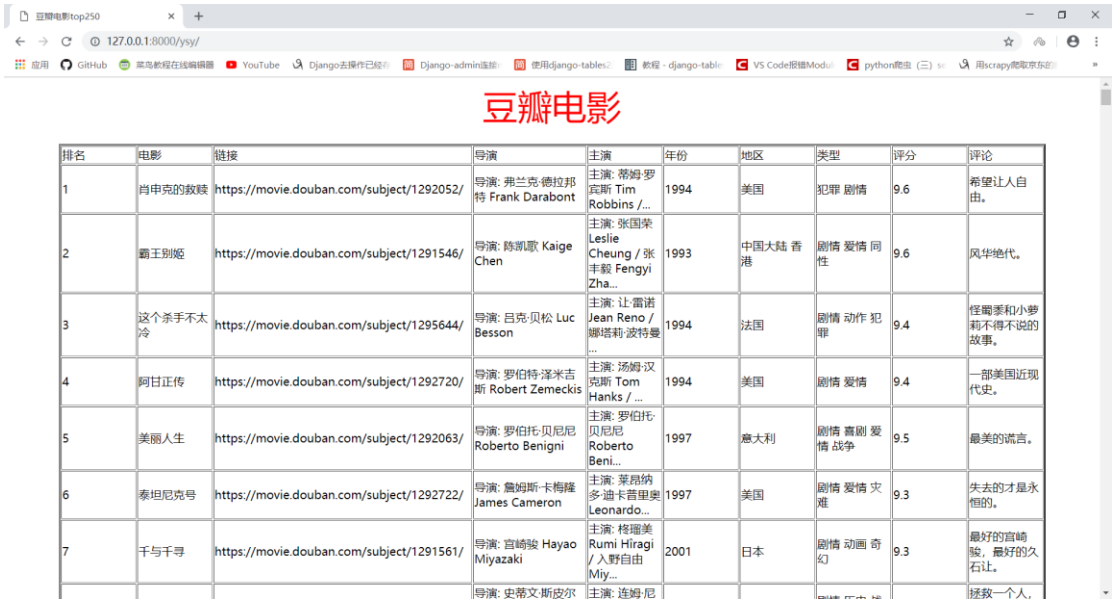
```
            <td width="100" price="title">地区</td>
            <td width="100" rate="title">类型</td>
            <td width="100" name="title">评分</td>
            <td width="100" comments="title">评论</td>
        </tr>
        {% for i in order %}
        <tr>
            <td width="100">{{i.0}}</td>
            <td width="100">{{i.1}}</td>
            <td width="100">{{i.2}}</td>
            <td width="100">{{i.3}}</td>
            <td width="100">{{i.4}}</td>
            <td width="100">{{i.5}}</td>
            <td width="100">{{i.6}}</td>
            <td width="100">{{i.7}}</td>
            <td width="100">{{i.8}}</td>
            <td width="100">{{i.9}}</td>
        </tr>
        {% endfor %}


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

显示结果：

```
D:\codefile\Helloword>python manage.py runserver
Performing system checks...

System check identified no issues (0 silenced).
April 26, 2019 - 11:27:42
Django version 2.1.7, using settings 'Helloword.settings'
Starting development server at http://127.0.0.1:8000/
Quit the server with CTRL-BREAK.
[26/Apr/2019 11:28:00] "GET /ysy/ HTTP/1.1" 200 154274
Not Found: /favicon.ico
[26/Apr/2019 11:28:02] "GET /favicon.ico HTTP/1.1" 404 2079
```



排名	电影	链接	导演	主演	年份	地区	类型	评分	评论
1	肖申克的救赎	https://movie.douban.com/subject/1292052/	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	9.6	希望让人自由。
2	霸王别姬	https://movie.douban.com/subject/1291546/	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	9.6	风华绝代。
3	这个杀手不太冷	https://movie.douban.com/subject/1295644/	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	9.4	怪蜀黍和小萝莉不得不说的故事。
4	阿甘正传	https://movie.douban.com/subject/1292720/	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	9.4	一部美国近现代史。
5	美丽人生	https://movie.douban.com/subject/1292063/	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	9.5	最美的谎言。
6	泰坦尼克号	https://movie.douban.com/subject/1292722/	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	9.3	失去的才是永恒的。
7	千与千寻	https://movie.douban.com/subject/1291561/	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	9.3	最好的宫崎骏，最好的久石让。
			导演: 史蒂文·斯皮尔	主演: 连姆·尼					拯救一个人，