

# 温州大学瓯江学院理工分院

## 《爬虫》实验报告

实验名称:	实验 2 移动、云及数字化工具				
班 级:	16 计算机科学与技术 2	姓 名:	汤嘉伟	学 号:	16219111211
实验地点:		日 期:	2019 年 4 月 26 日		

### 一、实验目的:

[实验目的和要求]

1. 静态与动态爬虫
2. DJANGO

### 二、实验环境:

1、python, pycharm, mysql

### 三、实验内容和要求:

1.静态爬虫

2.动态爬虫

3.制作 DJANGO 网站

### 四、实验步骤与结果:

(对实验步骤的说明应该能够保证根据该说明即可重复完整的实验内容, 得到正确结果。)

爬虫

步骤:

- 1) 确认要爬取的网站为豆瓣读书, 确认 URL
- 2) 使用 requests 与 beautifulsoup 库获取网页源代码
- 3) 使用浏览器查看源码, 分析需要爬取的图片, 标题等的特征
- 4) 使用 findall 函数获取数据存入列表中
- 5) 分析 url 爬取更多页数编写循环
- 6) 将得到的数据存入数据库

代码:

```
import requests

from bs4 import BeautifulSoup

def get_movies():

    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'

    }

    movie_list = []

    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_='hd')

        for each in div_list:

            movie = each.a.span.text.strip()

            movie_list.append(movie)

    return movie_list
```

结果:

```
D:\tang\venv\Scripts\python.exe D:/作业/app1/h.py
['自申克的救赎', '霸王别姬', '这个杀手不太冷', '阿甘正传', '美丽人生', '泰坦尼克号', '千与千寻', '辛德勒的名单', '盗梦空间', '忠犬八公的故事', '机器人总动员', '三傻大闹宝莱坞', '海上钢琴师', '放

Process finished with exit code 0
```

Django

步骤:

- 1) 使用 pycharm 创建项目，创建一个应用
- 2) 把我的静态爬虫文档导入到应用文档下，在 view 中应用

```

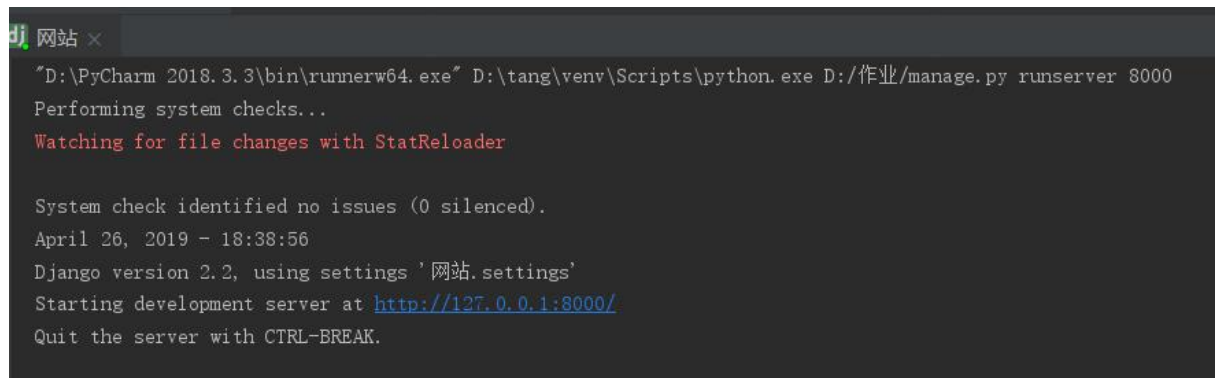
from django.shortcuts import render
from appl.h import get_movies
from appl.models import appl
# Create your views here.
def gettitle(request):
    if appl.objects.count()==0:
        l=get_movies()
        for i in l:
            temp = appl(title=i)
            temp.save()
        ll=appl.objects.all()
    return render(request,'index.html',{'li':ll})

```

使用系统自带的 SQLLIST3

显示数据

截图：



```

网站 ×
"D:\PyCharm 2018.3.3\bin\runnerw64.exe" D:\tang\venv\Scripts\python.exe D:/作业/manage.py runserver 8000
Performing system checks...
Watching for file changes with StatReloader

System check identified no issues (0 silenced).
April 26, 2019 - 18:38:56
Django version 2.2, using settings '网站.settings'
Starting development server at http://127.0.0.1:8000/
Quit the server with CTRL-BREAK.

```

网站：

追风筝的人

解忧杂货店

小王子

白夜行

围城

三体

嫌疑人X的献身

活着

挪威的森林

百年孤独

红楼梦

看见

平凡的世界（全三部）

三体Ⅱ

1▷ 完成

---