# <DoubanTVSpider>开发文档

版本修订日志

|  |  |  |
| --- | --- | --- |
| 修订时间 | 修订人 | 修订内容 |
| 2016年10月20日 | 杨恬 | 创建 |
|  |  |  |
|  |  |  |
|  |  |  |

## 1．1 doubanTVSpider 豆瓣电视剧爬取程序

1. 程序编号：doubanTVSpider
2. 程序名称：豆瓣电视剧爬取程序
3. 程序说明：本程序为豆瓣电视剧信息爬取程序，目前是为搜集各电视剧信息功能研发，主要研发思路：访问豆瓣电视剧主页，根据不同的标签搜集所有的电视剧ID，根据ID访问该电视剧的豆瓣主页，抓取该电视剧的基本信息，包含的程序模块如下：

模块A：电视剧ID抓取功能

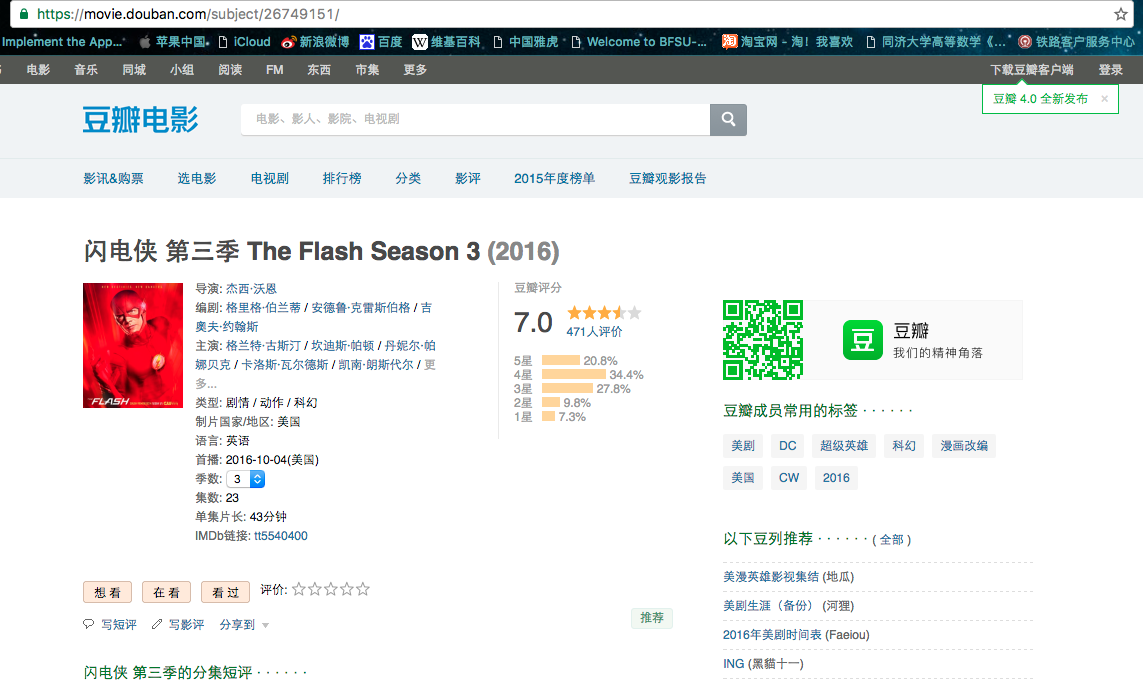
模块B： 电视剧基本信息抓取功能

1. 程序功能可视化说明：

在没有输入文件的情况下，程序会访问豆瓣电视剧的主页搜索电视剧ID：



访问所有标签尽可能搜索更多的电视剧ID，得到所有ID的中间文件。然后抽取其中的ID，根据ID得到该电视剧的主页：



抓取该页面的电视剧信息并保存到一个json文件中，同时记录抓取过的ID以避免重复抓取。

1. 程序输入与输出说明：

输入: 无需输入或需要抓取的电视剧ID列表

输出: 包含电视剧信息的json字符串，如：

{"rate": "7.0", "rate\_num": "None", "movie\_id": "26749151", "name": {"0": "None", "1": "闪电侠 第三季 The Flash Season 3"}, "basic\_info": {"director": "杰西·沃恩", "type": "剧情", "init\_time": "2016-10-04(美国)", "actors": "格兰特·古斯汀", "area": " 美国"}}

{"rate": "9.3", "rate\_num": "None", "movie\_id": "26387813", "name": {"0": "嫌疑人", "1": "意中人", "2": "疑犯追踪 第五季 Person of Interest Season 5"}, "basic\_info": {"director": "克里斯·费舍", "type": "剧情", "init\_time": "2016-05-03(美国)", "actors": "吉姆·卡维泽", "area": " 美国"}}

1. 处理过程流程图：

../../../豆瓣.pdf

1. 开放接口（若被调用，需要提供开放的接口说明）：

无

1. 引入的包及接口：

retring: 错误重试模块

scrapy：运用其中的Selector类进行html解析和元素定位

1. 程序源码目录结构说明：

.

├── \_\_configure\_\_

│   ├── ConfigureParser.py

│   ├── \_\_init\_\_.py

│   └── configure.xml

├── \_\_doc\_\_

│   └── DoubanTVSpider文档.docx

├── \_\_file\_\_

│   ├── douban\_finished.json

│   ├── douban\_tv\_id.txt

│   └── finished\_id.txt

├── \_\_spider\_\_

│   ├── \_\_init\_\_.py

│   └── douban.py

├── \_\_util\_\_

│   ├── \_\_init\_\_.py

│   └── library.py

└── main.py

│

└── run.sh

\_\_configure\_\_: 包含配置文件和配置文件解析类

\_\_doc\_\_：包含程序文档

\_\_file\_\_：包含输入输出以及临时文件

\_\_spider\_\_: 包含爬虫代码

\_\_util\_\_：包含一些常用的函数

main.py: 启动程序

run.sh：执行文件

1. 程序启动方式说明

修改run.sh执行权限，执行run.sh

1. 程序配置文件及说明

配置文件: configure.xml

pattern: 匹配模式，其中的class属性包括regex和xpath，分别表示正则表达式模式和xpath路径。

files: 程序文件

input\_file: 电视剧ID列表

output\_file: 电视剧信息json文件

temp\_file：已执行的电视剧ID列表

(三个文件如无必要都不用重新设置)

1. 部署程序打包方式：

上传文件夹即可

1. 程序部署说明：

暂无

1. 程序潜在的问题，推荐解决方案及其他特殊说明

豆瓣容易被封，建议停顿3-5秒，或者半夜运行（貌似不需要停顿）

1. 版本号和CHANGES

doubanTVSpider 1.0