# 专栏：009：高评分电影抓取

E – 爬虫技术　爬虫知识

> **用理工科思维看待这个世界**

系列爬虫专栏

崇尚的学习思维是：输入，输出平衡，且平衡点不断攀升。

曾经有大神告诫说：没事别瞎写文章；所以，很认真的写的是能力范围内的，看客要是看不懂，不是你的问题，问题在我，得持续输入，再输出。

今天的主题是：实战爬取电影，并存储至MySQL数据库

---

# 1：框架

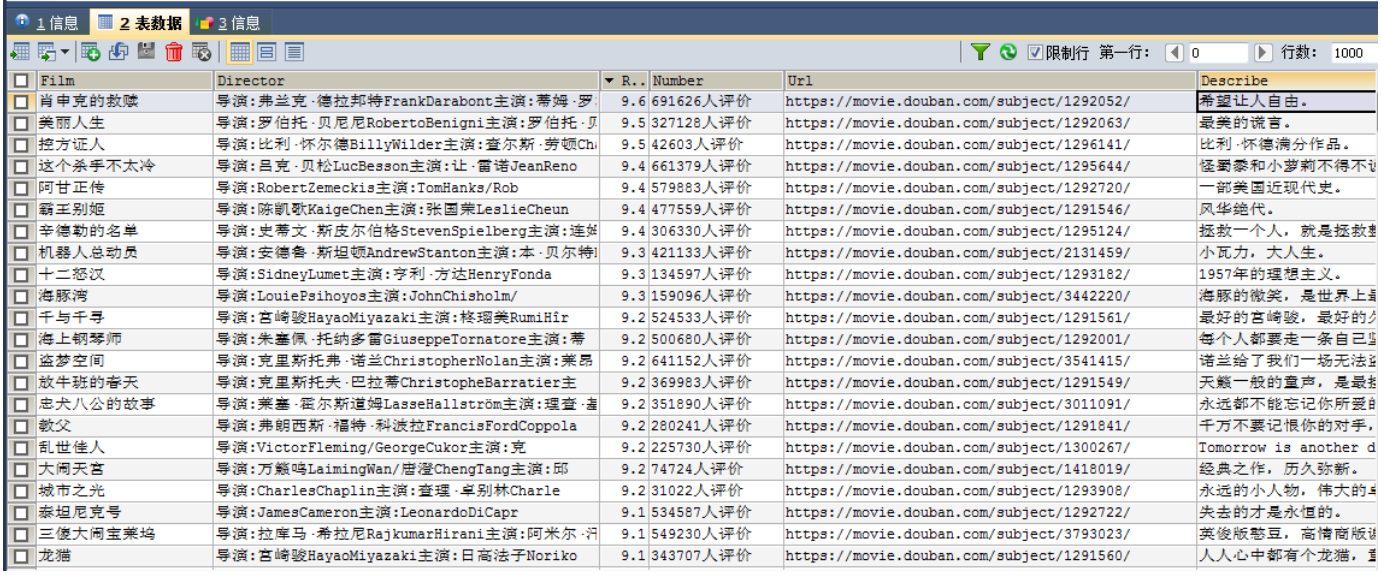
| 序号 | 目标 | 说明 |
| --- | --- | --- |
| 01 | 抓取目标分析 | -目标是什么 |
| 02 | 分解任务 | – |
| 03 | MySQL建表操作 | 本地建表 |
| 04 | 实战抓取 | – |
| 05 | 参考及总结 | – |

---

# 2：目标

任务是：抓取网站数据，存放至MySQL数据库中。

效果显示：抓取的目标存放至本地MySQL数据库



| Film | Director | R.. | Number | Url | Describe |
|---|---|---|---|---|---|
| 肖申克的救赎 | 导演:弗兰克·德拉邦特FrankDarabont主演:蒂姆·罗 | 9.6 | 691626人评价 | https://movie.douban.com/subject/1292052/ | 希望让人自由。 |
| 美丽人生 | 导演:罗伯托·贝尼尼RobertoBenigni主演:罗伯托·贝 | 9.5 | 327128人评价 | https://movie.douban.com/subject/1292063/ | 最美的谎言。 |
| 控方证人 | 导演:比利·怀尔德BillyWilder主演:查尔斯·劳顿Cha | 9.5 | 42603人评价 | https://movie.douban.com/subject/1296141/ | 比利·怀德满分作品。 |
| 这个杀手不太冷 | 导演:吕克·贝松LucBesson主演:让·雷诺JeanReno | 9.4 | 661379人评价 | https://movie.douban.com/subject/1295644/ | 怪蜀黍和小萝莉不得不让 |
| 阿甘正传 | 导演:RobertZemeckis主演:TomHanks/Rob | 9.4 | 579883人评价 | https://movie.douban.com/subject/1292720/ | 一部美国近现代史。 |
| 霸王别姬 | 导演:陈凯歌KaigeChen主演:张国荣LeslieCheun | 9.4 | 477559人评价 | https://movie.douban.com/subject/1291546/ | 风华绝代。 |
| 辛德勒的名单 | 导演:史蒂文·斯皮尔伯格StevenSpielberg主演:连姆 | 9.4 | 306330人评价 | https://movie.douban.com/subject/1295124/ | 拯救一个人，就是拯救整 |
| 机器人总动员 | 导演:安德鲁·斯坦顿AndrewStanton主演:本·贝尔特 | 9.3 | 421133人评价 | https://movie.douban.com/subject/2131459/ | 小瓦力，大人生。 |
| 十二怒汉 | 导演:SidneyLumet主演:亨利·方达HenryFonda | 9.3 | 134597人评价 | https://movie.douban.com/subject/1293182/ | 1957年的理想主义。 |
| 海豚湾 | 导演:LouiePsihoyos主演:JohnChisholm/ | 9.3 | 159096人评价 | https://movie.douban.com/subject/3442220/ | 海豚的微笑，是世界上最 |
| 千与千寻 | 导演:宫崎骏HayaoMiyazaki主演:柊瑠美RumiHir | 9.2 | 524533人评价 | https://movie.douban.com/subject/1291561/ | 最好的宫崎骏，最好的少 |
| 海上钢琴师 | 导演:朱塞佩·托纳多雷GiuseppeTornatore主演:蒂 | 9.2 | 500680人评价 | https://movie.douban.com/subject/1292001/ | 每个人都要走一条自己坚 |
| 盗梦空间 | 导演:克里斯托弗·诺兰ChristopherNolan主演:莱昂 | 9.2 | 641152人评价 | https://movie.douban.com/subject/3541415/ | 诺兰给了我们一场无法盗 |
| 放牛班的春天 | 导演:克里斯托夫·巴拉蒂ChristopheBarratier主 | 9.2 | 369983人评价 | https://movie.douban.com/subject/1291549/ | 天籁一般的童声，是最接 |
| 忠犬八公的故事 | 导演:莱塞·霍尔斯道姆LasseHallström主演:理查·基 | 9.2 | 351890人评价 | https://movie.douban.com/subject/3011091/ | 永远都不能忘记你所爱的 |
| 教父 | 导演:弗朗西斯·福特·科波拉FrancisFordCoppola | 9.2 | 280241人评价 | https://movie.douban.com/subject/1291841/ | 千万不要记恨你的对手， |
| 乱世佳人 | 导演:VictorFleming/GeorgeCukor主演:克 | 9.2 | 225730人评价 | https://movie.douban.com/subject/1300267/ | Tomorrow is another d |
| 大闹天宫 | 导演:万籁鸣LaimingWan/唐澄ChengTang主演:邱 | 9.2 | 74724人评价 | https://movie.douban.com/subject/1418019/ | 经典之作，历久弥新。 |
| 城市之光 | 导演:CharlesChaplin主演:查理·卓别林Charle | 9.2 | 31022人评价 | https://movie.douban.com/subject/1293908/ | 永远的小人物，伟大的卓 |
| 泰坦尼克号 | 导演:JamesCameron主演:LeonardoDiCapr | 9.1 | 534587人评价 | https://movie.douban.com/subject/1292722/ | 失去的才是永恒的。 |
| 三傻大闹宝莱坞 | 导演:拉库马·希拉尼RajkumarHirani主演:阿米尔·汗 | 9.1 | 549230人评价 | https://movie.douban.com/subject/3793023/ | 英俊版憨豆，高情商版谢 |
| 龙猫 | 导演:宫崎骏HayaoMiyazaki主演:日高法子Noriko | 9.1 | 343707人评价 | https://movie.douban.com/subject/1291560/ | 人人心中都有个龙猫，重 |

初始URL

```
url = https://movie.douban.com/top250
```

字段：

Film : 电影名称

Director: 电影导演

Rates : 评分数

Numbers : 评分人数

Url: 电影链接

Describe: 电影介绍 (网站的一句话，经典台词之类的)

---

# 3：任务分解

具体点击网页审查元素：链接

- 字段的正则表达式

电影名称：

```
Film_pattern = r'<span class="title">(.*?)</span>'
```

电影导演：

```
Director_pattern = r'<p class="">(.*?)</p>'
```

评分数：

```
Rates_pattern = r'<span class="rating_num" property="v:average">(.*?)</span>'
```

评分人数：先抓大，再在大的里面匹配所需的文本信息

```
Number_pattern_large = r'<div class="star">(.*?)</div>'
Number_pattern_small = r'<span>(.*?)</span>'
```

电影链接：先抓大，再在大的里面匹配所需的文本信息

```
Urlfilm_pattern_large = r'<div class="hd">(.*?)</div>'
Urlfilm_pattern_small = r'<a href="(.*?)"'
```

电影介绍：

```
Describe_pattern = r'<span class="inq">(.*?)</span>'
```

- URL的分析：

由翻页，自己匹配出网址：(也可以网址抓取)。10页。

`urls = ["https://movie.douban.com/top250?start={}&filter=".format(i) for i in range(0,250,25)]`

网址分析完成，正则分析完成。任务完成了大半。

---

# 4：数据库建表操作

在本地：数据库名为：exercise 下创建一个表名为：douban_film

建表的SQL语法：参照w3school

```
CREATE TABLE `douban_film` (
  `Film` CHAR(32) DEFAULT NULL COMMENT '电影名称',
  `Director` CHAR(32) DEFAULT NULL COMMENT '电影导演',
  `Rates` FLOAT DEFAULT NULL COMMENT '评分数',
  `Number` CHAR(16) DEFAULT NULL COMMENT '评分人数',
  `Url` CHAR(128) DEFAULT NULL COMMENT '评分人数',
  `Describe` CHAR(128) DEFAULT NULL COMMENT '电影介绍'
) ENGINE=INNODB DEFAULT CHARSET=utf8 COMMENT='豆瓣电影250介绍'
`douban_film`
# 执行sql语句就可以创建一个表，各字段及其属性如上示
```

---

# 5：实战抓取

单独使用正则，会出现很多难以匹配(可能没有尝试其他匹配规则)。需要对网页进行不断的分析。

抓取核心代码：(大神轻拍代码...)

```
# 网页抓取字段示例
 def content_json(self, content):
     Film_all = re.findall(self.Film_pattern, content, re.S)
     Film = []
     for one_film in Film_all:
         if " " not in one_film:
             Film.append(one_film)
     Director_all = re.findall(self.Director_pattern, content, re.S)
     Director = []
     for one_Director in Director_all:
         one = self.str_replace(one_Director)
         Director.append(one)
     Rates = re.findall(self.Rates_pattern, content, re.S)
     Number_large = re.findall(self.Number_pattern_large, content, re.S)
     Number = []
     for one_number in Number_large:
         Number_one = re.findall(self.Number_pattern_small, one_number, re.
S)[0]
         Number.append(Number_one)
     Describe = re.findall(self.Describe_pattern, content, re.S)
     Url_largre = re.findall(self.Urlfilm_pattern_large, content, re.S)
     Url =[]
     for one_url in Url_largre:
         Url_one = re.findall(self.Urlfilm_pattern_small, one_url, re.S)[0]
         Url.append(Url_one)
     Film_collection = []
     for Film, Director, Rates, Number, Describe, Url in zip(Film, Directo
r, Rates, Number, Describe, Url):
         data = {
             "Film": Film,
             "Director": Director,
             "Rates": Rates,
             "Number": Number,
             "Describe": Describe,
             "Url": Url
         }
         Film_collection.append(data)
     return Film_collection
```

```
# 文本匹配会有很多不需要的字段，如下函数实现数据清洗
def str_replace(self, str_one):
    str_one_1 = str_one.replace("\n", '')
    str_one_2 = str_one_1.replace("<br>", '')
    str_one_3 = str_one_2.replace(" ", '')
    str_one_4 = str_one_3.replace("\t", "")
    str_one_5 = str_one_4.replace(" ", '').strip()
    return str_one_5
```

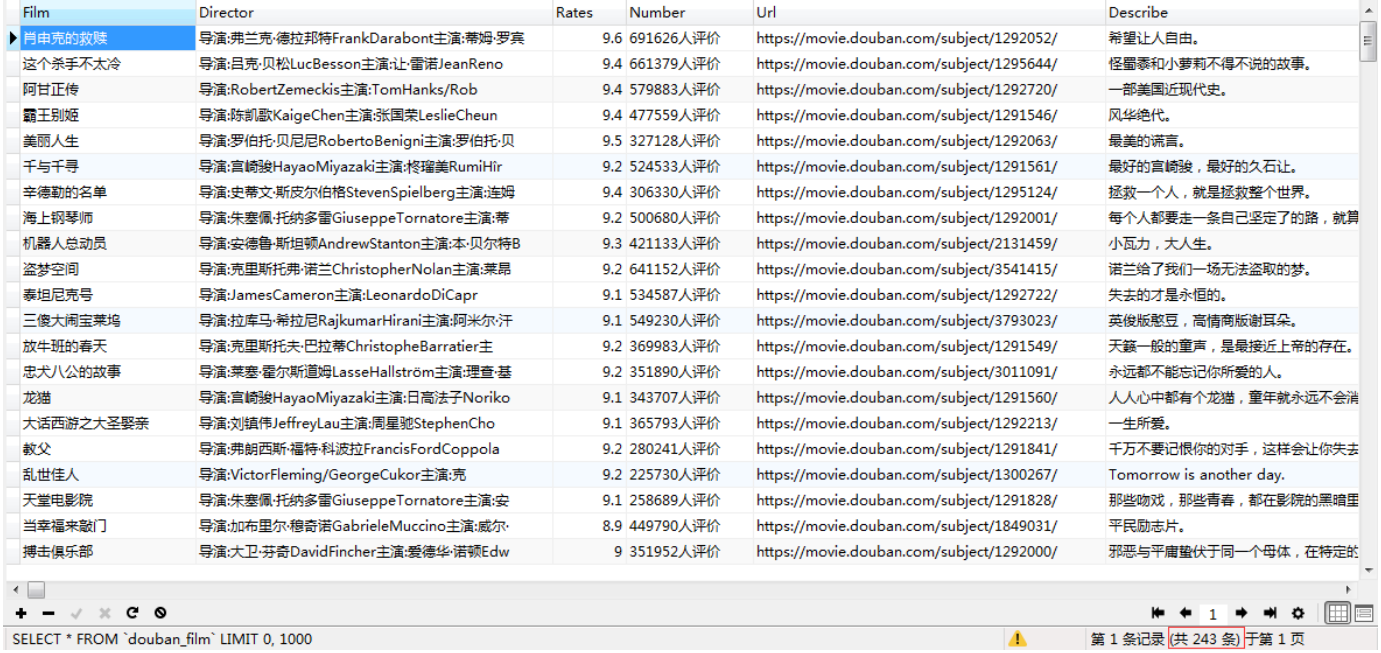
```
# 单独写sql语句比较繁琐，如下函数实现JSON格式数据转换成SQL语句
import copy
def json_to_mysql(json_obj, table, sql_type="insert"):
    local_copy = copy.deepcopy(json_obj)
    if sql_type == "insert":
        sql_part1 = "insert into " + table
        keys = local_copy.keys()

        sql_part2 = "("
        for key in keys:
            sql_part2 += "`%s`"%(key)
            sql_part2 += ","
        sql_part2 = sql_part2.rstrip(",")
        sql_part2 += ")"

        sql_part3 = "("
        for key in keys:
            sql_part3 += "'" + (local_copy[key]) + "'"
            sql_part3 += ","
        sql_part3 = sql_part3.rstrip(",")
        sql_part3 += ")"

        sql = sql_part1 + " " + sql_part2 + " values " + sql_part3
    return sql
```

核心代码已经完成：
整体思路如下：

- 分析首页文本信息的正则表达式
- 抓取首页的字段
- 对字段进行数据的清洗，去掉不需要的信息
- 将数据结构化
- 循环操作
- 获取的全部信息执行sql语句，存入已经建表的MySQL数据库中

完整版代码：完整版代码

另一款数据库可视化工具显示效果：



| Film | Director | Rates | Number | Url | Describe |
|---|---|---|---|---|---|
| 肖申克的救赎 | 导演:弗兰克·德拉邦特FrankDarabont主演:蒂姆·罗宾 | 9.6 | 691626人评价 | https://movie.douban.com/subject/1292052/ | 希望让人自由。 |
| 这个杀手不太冷 | 导演:吕克·贝松LucBesson主演:让·雷诺JeanReno | 9.4 | 661379人评价 | https://movie.douban.com/subject/1295644/ | 怪蜀黍和小萝莉不得不说的故事。 |
| 阿甘正传 | 导演:RobertZemeckis主演:TomHanks/Rob | 9.4 | 579883人评价 | https://movie.douban.com/subject/1292720/ | 一部美国近现代史。 |
| 霸王别姬 | 导演:陈凯歌KaigeChen主演:张国荣LeslieCheun | 9.4 | 477559人评价 | https://movie.douban.com/subject/1291546/ | 风华绝代。 |
| 美丽人生 | 导演:罗伯托·贝尼尼RobertoBenigni主演:罗伯托·贝 | 9.5 | 327128人评价 | https://movie.douban.com/subject/1292063/ | 最美的谎言。 |
| 千与千寻 | 导演:宫崎骏HayaoMiyazaki主演:柊瑠美RumiHir | 9.2 | 524533人评价 | https://movie.douban.com/subject/1291561/ | 最好的宫崎骏，最好的久石让。 |
| 辛德勒的名单 | 导演:史蒂文·斯皮尔伯格StevenSpielberg主演:连姆 | 9.4 | 306330人评价 | https://movie.douban.com/subject/1295124/ | 拯救一个人，就是拯救整个世界。 |
| 海上钢琴师 | 导演:朱塞佩·托纳多雷GiuseppeTornatore主演:蒂 | 9.2 | 500680人评价 | https://movie.douban.com/subject/1292001/ | 每个人都要走一条自己坚定了的路，就算 |
| 机器人总动员 | 导演:安德鲁·斯坦顿AndrewStanton主演:本·贝尔特B | 9.3 | 421133人评价 | https://movie.douban.com/subject/2131459/ | 小瓦力，大人生。 |
| 盗梦空间 | 导演:克里斯托弗·诺兰ChristopherNolan主演:莱昂 | 9.2 | 641152人评价 | https://movie.douban.com/subject/3541415/ | 诺兰给了我们一场无法盗取的梦。 |
| 泰坦尼克号 | 导演:JamesCameron主演:LeonardoDiCapr | 9.1 | 534587人评价 | https://movie.douban.com/subject/1292722/ | 失去的才是永恒的。 |
| 三傻大闹宝莱坞 | 导演:拉库马·希拉尼RajkumarHirani主演:阿米尔·汗 | 9.1 | 549230人评价 | https://movie.douban.com/subject/3793023/ | 英俊版憨豆，高情商版谢耳朵。 |
| 放牛班的春天 | 导演:克里斯托夫·巴拉蒂ChristopheBarratier主 | 9.2 | 369983人评价 | https://movie.douban.com/subject/1291549/ | 天籁一般的童声，是最接近上帝的存在。 |
| 忠犬八公的故事 | 导演:莱塞·霍尔斯道姆LasseHallström主演:理查·基 | 9.2 | 351890人评价 | https://movie.douban.com/subject/3011091/ | 永远都不能忘记你所爱的人。 |
| 龙猫 | 导演:宫崎骏HayaoMiyazaki主演:日高法子Noriko | 9.1 | 343707人评价 | https://movie.douban.com/subject/1291560/ | 人人心中都有个龙猫，童年就永远不会消 |
| 大话西游之大圣娶亲 | 导演:刘镇伟JeffreyLau主演:周星驰StephenCho | 9.1 | 365793人评价 | https://movie.douban.com/subject/1292213/ | 一生所爱。 |
| 教父 | 导演:弗朗西斯·福特·科波拉FrancisFordCoppola | 9.2 | 280241人评价 | https://movie.douban.com/subject/1291841/ | 千万不要记恨你的对手，这样会让你失去 |
| 乱世佳人 | 导演:VictorFleming/GeorgeCukor主演:克 | 9.2 | 225730人评价 | https://movie.douban.com/subject/1300267/ | Tomorrow is another day. |
| 天堂电影院 | 导演:朱塞佩·托纳多雷GiuseppeTornatore主演:安 | 9.1 | 258689人评价 | https://movie.douban.com/subject/1291828/ | 那些吻戏，那些青春，都在影院的黑暗里 |
| 当幸福来敲门 | 导演:加布里尔·穆奇诺GabrieleMuccino主演:威尔· | 8.9 | 449790人评价 | https://movie.douban.com/subject/1849031/ | 平民励志片。 |
| 搏击俱乐部 | 导演:大卫·芬奇DavidFincher主演:爱德华·诺顿Edw | 9 | 351952人评价 | https://movie.douban.com/subject/1292000/ | 邪恶与平庸蛰伏于同一个母体，在特定的 |

SELECT * FROM `douban_film` LIMIT 0, 1000

第 1 条记录 (共 243 条) 于第 1 页

可知：抓取了243条信息。全站存在250条数据。代码存在Bug…

你懂的。可以继续重构。

---

# 6：参考及总结

**在通往牛逼的路上，别人一定存在许多值得借鉴的地方。**

**学习学习在学习**

**实践实践再实践**

**感谢伟大的互联网时代**

Github

> 关于本人：
>
> > 国内小硕，半路出家的IT学习者。
> > 兴趣领域：爬虫 ， 数据科学
> > 本人正在构建一个共同成长爬虫小型社群。QQ群:539065787
> > 持续精进。如果理念相似，欢迎加入。
> > 文档及代码托管在Github上。