



大作业三 影视爬虫与检索系统

第三周大作业





实验目的



- ◆ 掌握**Python**的简单爬虫功能
- ◆ 掌握用正则表达式提取关键信息
- ◆ 掌握**Django**的**MVC**框架思想
- ◆ 编写一个信息整合与检索系统



实验内容及要求



◆ (1) 数据爬取

⊕ 从中文影评网站中任选，爬取其中的影视信息

- ◆ 豆瓣电影 <https://movie.douban.com/> 时光网 <http://www.mtime.com/> 等。
- ◆ 爬取的信息至少包括名称、演员、简介、图片1张和网站提供的其他信息。
- ◆ 总数量不少于1000。

⊕ 爬取演员信息

- ◆ 同时爬取与已爬到的影视相关的演员信息。
- ◆ 至少包括姓名、简介、个人照片1张和网站提供的其他信息。

⊕ 爬取相关影评信息

- ◆ 同时爬取与影视相关的影评信息。
- ◆ 每部影视至少爬取5条影评信息。



实验内容及要求



◆ 相关事项

- ⊕ 可选取[电影/电视剧/电视节目/.....]中的某一类进行爬取。
- ⊕ 爬取影视时只爬取演员信息即可，导演、编剧等其他人员信息不做要求。
如果其中演员数量太多，爬取前**10**个即可。
- ⊕ 影评爬取长评或者短评均可，其中的图片等媒体数据不作要求。
- ⊕ 爬取后的数据要使用某种规范格式组织存储（例如：数据库、文件系统、**json**文件、**XML**文件等）。



根节点：影视列表



豆瓣电影

搜索电影、电视剧、综艺、影人

影讯&购票 选电影 电视剧 排行榜 分类 影评 2019年度榜单 2019书影音报告

选电影

热门

最新

经典

可播放

豆瓣高分

冷门佳片

华语

欧美

韩国

日本

动作

喜剧

爱情

科幻

悬疑

恐怖

文艺

☒按热度排序

☐按时间排序

☐按评价排序

☐我没看过的

☐可在线播放



釜山行2: 半岛 5.1



铁雨2：首脑峰会 5.8



新 独一无二的伊万 6.8



1917 8.5



超能计划 5.6



猎谎者 6.9



乔乔的异想世界 8.4



误杀 7.7



妙先生 6.2



多力特的奇幻冒险 5.8



影里 7.4



#活着 5.6



灰猎犬号 8.2



进击的巨人：编年



少年的你 8.3

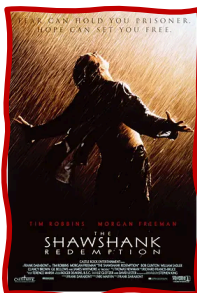


一个影视信息页面



No.1 豆瓣电影Top250

肖申克的救赎 The Shawshank Redemption (1994)



导演: 弗兰克·德拉邦特

编剧: 弗兰克·德拉邦特 / 斯蒂芬·金

主演: 蒂姆·罗宾斯 / 摩根·弗里曼 / 鲍勃·冈顿 / 威廉姆·赛德勒 / 克兰西·布朗 / 吉尔·贝罗斯 / 马克·罗斯顿 / 詹姆斯·惠特摩 / 杰弗里·德曼 / 拉里·布兰登伯格 / 尼尔·吉恩托利 / 布赖恩·利比 / 大卫·普罗瓦尔 / 约瑟夫·劳格诺 / 祖德·塞克利拉 / 保罗·麦克兰尼 / 芮妮·布莱恩 / 阿方索·弗里曼 / V·J·福斯特 / 弗兰克·梅德诺 / 马克·迈克尔斯 / 尼尔·萨默斯 / 耐德·巴拉米 / 布赖恩·戴拉特 / 唐·麦克马纳斯

类型: 剧情 / 犯罪

制片国家/地区: 美国

语言: 英语

上映日期: 1994-09-10(多伦多电影节) / 1994-10-14(美国)

片长: 142分钟

又名: 月黑高飞(港) / 刺激1995(台) / 地狱诺言 / 铁窗岁月 / 肖申克的救赎

IMDb链接: tt0111161

豆瓣评分

9.7  2129995人评价

5星  85.2%
4星  13.3%
3星  1.3%
2星  0.1%
1星  0.1%

好于 99% 剧情片

好于 99% 犯罪片

想看

看过

评价: ☆☆☆☆☆

 写短评  写影评  分享到

推荐

肖申克的救赎的剧情简介 · · · · ·

20世纪40年代末，小有成就的青年银行家安迪（蒂姆·罗宾斯 Tim Robbins 饰）因涉嫌杀害妻子及她的情人而锒铛入狱。在这座名为鲨堡的监狱内，希望似乎虚无缥缈，终身监禁的惩罚无疑注定了安迪接下来灰暗绝望的人生。未过多久，安迪尝试接近囚犯中颇有声望的瑞德（摩根·弗里曼 Morgan Freeman 饰），请求对方帮自己搞来小锤子。以此为契机，二人逐渐熟稔，安迪也仿佛在鱼龙混杂、罪恶横生、黑白混淆的牢狱中找到属于自己的求生之道。他利用自身的专业知识，帮助监狱管理层逃税、洗黑钱，同时凭借与瑞德的交往在犯人中间也渐渐受到礼遇。表面看来，他已如瑞德那样对那堵高墙从憎恨转变为处之泰然，但是对自由的渴望仍促使他朝着心中的希望和目标前进。而关于其罪行的真相，似乎更使这一切朝前推进了一步……

本片根据著名作家斯蒂芬·金（Stephen Edwin King）的... [\(展开全部\)](#) ©豆瓣

清华大学

https://movie.douban.com/subject/1292052/?tag=%E8%B1%86%E7%93%A3%E9%AB%98%E5%88%86&from=gaia_video



影评信息



肖申克的救赎的短评 ····· (全部 388257 条)

我要写短评

热门 / 最新 / 好友

如小果 看过 ★★★★★ 2008-02-27

17498 有用

恐惧让你沦为囚犯，希望你重获自由。——《肖申克的救赎》

犀牛 看过 ★★★★★ 2005-10-28

16218 有用

当年的奥斯卡颁奖礼上，被如日中天的《阿甘正传》掩盖了它的光彩，而随着时间的推移，这部电影在越来越多的人心中的地位已超越了《阿甘》。每当现实令我疲惫得产生无力感，翻出这张碟，就重获力量。毫无疑问，本片位列男人必看的电影前三名！回顾那一段经典台词：“有的人的羽翼是如此光辉，即使世界上最黑暗的牢狱，也无法长久地将他困住！”

赫克托尔 看过 ★★★★★ 2007-06-06

9096 有用

有种鸟是关不住的。

EvelClassified 看过 ★★★★★ 2008-05-09

8037 有用

“这是一部男人必看的电影。”人人都这么说。但单纯从性别区分，就会让这电影变狭隘。《肖申克的救赎》突破了男人电影的局限，通篇几乎充满令人难以置信的温馨基调，而电影里最伟大的主题是“希望”。当我们无奈地遇到了如同肖申克一般囚禁了心灵自由的那种图图，我们是无奈的老布鲁克，灰心的瑞德，还是智慧的安迪？运用智慧，信任希望，并且勇敢面对恐惧心理，去打败它？经典的电影之所以经典，因为他们都在做同一件事—— (展开)

游牧人·芳汀 看过 ★★★★★ 2010-10-06

2369 有用

墙里的人老是喊自由，可是墙外的我们真正得到了吗？

> 更多短评 388257条

肖申克的救赎的影评 ····· (全部 10429 条)

我要写影评

热门 / 最新 / 好友

隱居雲上 ★★★★★ 2005-07-12 11:23:39

终于找到了郁闷人生的原因——观《肖申克的救赎》有感

周末看了一部美国影片《肖申克的救赎》（《The Shawshank Redemption》）讲的是一位因冤案入狱的年轻银行家在牢中如何追寻自由的故事。不同的人看同样的影片可能都有不同的感受。对于目前无力改变现状的我，看完这部影片后最深的感受就是：才华、毅力两样，是任何人在任何境... (展开)

△ 10937 ▽ 765 801回应

大头绿豆 ★★★★★ 2005-05-12 20:44:13

十年·肖申克的救赎

距斯蒂芬·金（Stephen King）和德拉邦特（Frank Darabont）们缔造这部伟大的作品已经有十年了。我知道美好的东西想必大家都能感受，但是很抱歉，我的聒噪仍将一如既往。在我眼里，肖申克的救赎与信念、自由和友谊有关。 [1] 信念 瑞德（Red）说，希望是危险的东西，是精... (展开)

△ 16194 ▽ 697 936回应

[已注销] ★★★★★ 2006-01-07 20:56:26

Andy Dufresne

Andy Dufresne，一个永垂电影史册的名字。1 关于《The Shawshank Redemption》的评论，太多，该说的差不多都已说了千万遍。对于这样一个热门的话题，再想要抒发一些个人的喜爱之情，不免有拾人牙慧之嫌。为了避免这样没新意的事情发生，许多单词我就不再提了，譬如“希望”... (展开)

△ 3470 ▽ 108 248回应

方枪枪 ★★★★★ 2018-03-30 16:46:05

#为何《肖申克的救赎》在IMDb和豆瓣都能排第一？

为何《肖申克的救赎》在IMDb和豆瓣都能排第一？

时间会证明经典的价值，虽然在某些时刻会被误判和忽视。用上面这句话来描述电影《肖申克的救赎》丝毫不会让人觉得会有任何夸张和吹捧的意味，因为这部电影在后来多年牢牢占据IMDb和豆瓣电影榜单第一名的位置足以说明一切。但就是这么一部经典的电影，却在其锋芒初露的时候被人... (展开)



演员信息页面



莱昂纳多·迪卡普里奥 Leonardo DiCaprio



增改描述、换头像

性别: 男

星座: 天蝎座

出生日期: 1974-11-11

出生地: 美国,加利福尼亚,洛杉矶

职业: 演员 / 制片人 / 配音 / 编剧

更多外文名: Leonardo Wilhelm DiCaprio (本名) / Lenny D (昵称) / Leo (昵称)

更多中文名: 李奥纳多·迪卡普里奥 / 里安纳度·迪卡比奥 / 小李(昵称) / 小李子(昵称)

家庭成员: Gisele Bündchen(前女友) / Bar Refaeli(前女友) / Erin

Heatherton(前女友) / Kelly Rohrbach(前女友)

imdb编号: [nm0000138](#)

官方网站: [leonardodicaprio.com](#)

收藏

分享到 ▼

推荐

影人简介 ·····

莱昂纳多·迪卡普里奥 (Leonardo DiCaprio), 1974年11月11日出生于美国加利福尼亚州洛杉矶, 好莱坞电影男演员。1974年, 一位叫艾莫琳的德国孕妇站在莱昂纳多·达·芬奇的画展上久久不能自己, 当她沉浸在达芬奇的艺术魅力之中时, 腹中的宝贝突然有了类似的悸动, 她即刻领会了这股神奇力量, 于是将后来诞下的这名男婴取名为莱昂纳多·威尔海姆·迪卡普里奥, 仿佛她已经知道这孩子将来注定要从事艺术工作。莱昂纳多的母亲艾莫琳是个有俄罗斯血统的德国人, 他的父亲是...(展开全部)



实验内容及要求



◆ (2) Web系统设计

- ⊕ 使用爬取到的信息设计一个网站，用来浏览和查询爬取的影视信息。
- ⊕ 网站需要包含以下页面（功能）：
 - ◆ 1. 列表页
 - ◆ 2. 影视信息页
 - ◆ 3. 演员信息页
 - ◆ 4. 搜索和搜索结果页



实验内容及要求



◆ (2.1) Web系统设计：列表页

- ⊕ 需要实现两个页面，即**影视列表页**和**演员列表页**。
- ⊕ **系统主页为影视列表页**。
- ⊕ 两个列表页通过**分页**方式列出系统中的所有影视和演员，并显示影视或演员的总数量。（首页尾页、页码按钮跳转、输入页码跳转）



- ⊕ 列表页中的每个条目需要展示名字和照片（可参考豆瓣的影视列表页）。
- ⊕ 点击列表页中的条目可以跳转到对应的影视信息页或演员信息页。





实验内容及要求



◆ (2.2) Web系统设计：影视信息页

- ⊕ 系统为每个爬取到的影视创建一个影视信息页，影视信息页包含两部分内容：基本信息与影评。
- ⊕ 基本信息：显示名称、演员、照片、简介等信息；
 - ◆ 基本信息中的演员列表应以类似列表页中的条目形式展示，显示演员姓名和照片，点击姓名或照片可跳转至演员信息页。



- ⊕ 影评：展示该影视的影评信息。



实验内容及要求



◆ (2.3) Web系统设计：演员信息页

- ⊕ **演员信息页**包含三部分内容：**基本信息**、**参演影视**、**合作演员**。
- ⊕ **基本信息**：显示姓名、简介、照片等信息；
- ⊕ **参演影视**：展示系统中该演员参演的所有影视作品。
 - ◆ 参演影视列表应以类似列表页中的条目形式展示，显示影视名称和图片，点击名称或图片可跳转至影视信息页。
- ⊕ **合作演员**：展示与该演员合作次数最高的**10**名其他演员。
 - ◆ 合作演员列表应以类似列表页中的条目形式展示，显示演员名、图片和**合作作品数量**，点击演员姓名或图片可跳转至其信息页。
 - ◆ 只考虑本系统中的影视作品的合作。



马丁·斯科塞斯
合作作品(24)



罗伯特·德尼罗
合作作品(14)



汤姆·汉克斯 合
作作品(12)



布拉德·皮特 合
作作品(10)



凯特·布兰切特
合作作品(9)



昆汀·塔伦蒂诺
合作作品(9)



实验内容及要求



◆ (2.4) Web系统设计：搜索

- ⊕ 搜索功能可单独设计为一个搜索页，或者在前述各页面的顶部设计搜索栏的方式能进行搜索也可。
- ⊕ 搜索功能包括一个搜索栏、一个单选组（影视、演员、影评），和一个搜索按钮。
- ⊕ 搜索栏输入一个关键词或一段文本（**20**字以内），根据单选组选项，搜索系统中存在的影视、演员和影评，搜索后跳转到搜索结果页。
- ⊕ 搜索考虑的字段包括：
 - ◆ 影视：名称、演员名称。
 - ◆ 演员：姓名、参演影视名称。
 - ◆ 影评：影评文本



实验内容及要求



◆ (2.4) Web系统设计：搜索举例

- ⊕ 1. 搜索【演员】with “莱昂纳多”，搜索结果应该包括“莱昂纳多·迪卡普里奥”。
 - ⊕ 2. 搜索【影视】with “泰坦尼克”，搜索结果应该包括“泰坦尼克号”。
 - ⊕ 3. 搜索【影视】with “莱昂纳多”，搜索结果应该包括“泰坦尼克号”等电影。
 - ⊕ 4. 搜索【演员】with “泰坦尼克”，搜索结果应该包括“莱昂纳多·迪卡普里奥”等人。
 - ⊕ 5. 搜索【影评】with “好看的电影”，搜索结果应该包括所有包含“好看的电影”这个字符串的影评。
- ⊕ 搜索只考虑单关键词精确搜索即可，即关键字为原字段字符串的子串。



实验内容及要求



◆ (2.4) Web系统设计：搜索结果页

- ⊕ 搜索结果页包含所有的（影视/演员/影评）搜索结果列表，以**分页**形式展示。
- ⊕ 搜索结果页中的每个条目设计为超链接格式，点击可跳转到相应信息页
- ⊕ 点击影评的搜索结果应该跳转到该影评所在的影视信息页。
- ⊕ 搜索结果页需要显示搜索条目数量和搜索花费的时间（后端计时）。



实验内容及要求



◆ (3) 注意事项

- ⊕ 建议对网页使用**CSS**和**JS**进行美化，网页设计效果会有少量加分。
- ⊕ 建议对查询功能通过各种技术（建立索引等）进行优化，查询时间过长会有少量扣分。
- ⊕ 网站各个页面之间应通过点击控件来跳转，不允许输入**URL**跳转。
- ⊕ 注意网站可能存在的反爬机制，并调整爬虫策略
 - ◆ 不要野蛮抓取、两次爬取中间建议停顿一定时间
 - ◆ 设**Headers**，动态**User-Agent**头
 - ◆ 换**Cookie**
 - ◆ （有条件）代理池
- ⊕ 不限制爬虫所使用的**Python**第三方库(**bs4**、**Scrapy**等) 【考试限制】



提交内容



◆ 源代码

- ⊕ 爬虫代码
- ⊕ **Web**工程代码
- ⊕ （爬虫数据无需提交）

◆ 文档

- ⊕ 对每个功能进行介绍展示
- ⊕ 爬虫数据总量及查询时间等性能统计信息
- ⊕ 各个功能涉及的技术和实现方式



Thanks and Have Fun

如对该作业有任何疑问，请e-mail到
zhaoty17@mails.tsinghua.edu.cn

清华大学