

实验 1 信息获取与检索分析

实验背景

豆瓣 (www.douban.com) 是一个中国知名的社区网站，以书影音起家，用户可以在豆瓣上查看感兴趣的电影、书籍、音乐等内容，还可以关注自己感兴趣的豆友。

本实验要求各位同学面向豆瓣平台，爬取指定的电影、书籍的主页，并解析其相关信息 (Stage 1)；结合给定的标签信息，分别实现电影、书籍的搜索引擎并评估其效果 (Stage 2)；在此基础上，结合用户的评价信息及其他边信息，进行个性化电影、书籍推荐 (Stage 3)。

实验要求

本次实验要求分组完成，每组最多 3 人（可以少于 3 人，但无优惠政策）。

实验将分为爬虫、检索、个性化检索（推荐）三个阶段，持续时间约为 3-10 教学周，实验报告的具体提交时间将于第二阶段公布。

第一阶段任务如下：

- 1. **爬虫：**针对给定的电影、书籍 ID，爬取其豆瓣主页，并解析其基本信息。以下图电影数据为例，其主页包含导演编剧等基本信息、剧情简介、演职员表、相关视频图片、获奖情况等。

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

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: 肖申克飞越 / 刺激1995(台) / 监狱岁月 / 铁窗岁月 / 肖申克的救赎

IMDb: tt0111161

豆瓣评分

9.7

2696881人评价

5星 85.8%

4星 12.6%

3星 1.2%

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 ·····

一场谋杀案使银行家安迪（蒂姆·罗宾斯 Tim Robbins 饰）蒙冤入狱，谋杀妻子及其情人的指控将囚禁他终生。在肖申克监狱的首次现身就让监狱“大哥”瑞德（摩根·弗里曼 Morgan Freeman 饰）对他另眼相看。瑞德帮助他搞到一把石锤和一块女明星海报，两人渐成患难之交。很快，安迪在监狱里大显其才，担任监狱图书馆管理员，并利用自己的金融知识帮助监狱官避税，引起了典狱长的注意，被招致麾下帮助典狱长洗黑钱。偶然一次，他得知一名新入狱的小偷能够作证帮他洗脱谋杀罪。燃起一丝希望的安迪找到了典狱长，希望他能帮自己翻案。阴险伪善的狱长假装答应，随后却派人杀害了这个小偷。他唯一的希望破灭，渺茫的进望没有绝望，在一个电闪雷鸣的风雨夜，一场酝酿几十年的逃脱计划让他自我救赎，重获自由！老朋友瑞德在他的鼓舞和帮助下，也勇敢地奔向自由。

本片获得1995年度... (展开全部) ©豆瓣

肖申克的救赎的演职员表 ····· (全部 43)

 弗兰克·德拉邦特 导演

 蒂姆·罗宾斯 饰 安迪·杜佛兰

 摩根·弗里曼 饰 艾利斯·波德

 鲍勃·冈顿 饰 监狱长山姆·诺顿

 威廉姆·赛德勒 饰 海伍德·威布勒

 克兰西·布朗 饰 上尉哈德利

肖申克的救赎的视频和图片 ····· (预告片2 | 视频评论3 · 添加 | 图片1040 · 添加)







肖申克的救赎的获奖情况 ····· (全部)

第67届奥斯卡金像奖 最佳影片(提名) 斯蒂芬·金

第52届金球奖 电影类 剧情片最佳男主角(提名) 摩根·弗里曼

第19届日本电影学院奖 最佳外语片

喜欢这部电影的人也喜欢 ·····











在哪儿看这部电影 ·····

优酷视频

VIP免费观看

腾讯视频

VIP免费观看

爱奇艺视频

VIP免费观看

哔哩哔哩

VIP免费观看

本片原声正在播放 ·····

去豆瓣音乐收听

以下片单推荐 ····· (全部)

★豆瓣高分电影榜★ (上) 9.7-8.6分 (影志)

【励志电影】 (影志)

他们是离智商玩家 (中间元素)

那些和谐的腐剧，腐动画，腐电影们。 (墨尘)

经典商业与法律电影 (passionfly)

谁在看这部电影 ·····

 Orion

刚刚看过 ★★★★★

 艾利奥典拉

1分钟前看过 ★★★★★

 lijay1_jiay

2分钟前看过 ★★★★★

4335540人看过 / 426458人想看

订阅肖申克的救赎的评论:

feed: rss 2.0

任务要求如下：

- a) 对于电影数据，至少爬取其基本信息、剧情简介、演职员表；
- b) 对于书籍数据，至少爬取其基本信息、内容简介、作者简介；
- c) 爬虫方式不限，网页爬取和 API 爬取两种方式都可，介绍使用的爬虫方式工具；
- d) 针对所选取的爬虫方式，发现并分析平台的反爬措施，并介绍采用的应对策略；
- e) 针对所选取的爬虫方式，使用不同的内容解析方法，并提交所获取的数据。
- f) 该阶段无评测指标要求，在实验报告中说明爬虫（反爬）策略和解析方法即可。

数据集介绍

本次提供的数据来源于豆瓣电影、豆瓣读书，包含电影、书籍 ID、标签信息、用户评价和用户间相互关注的社交数据，助教组会每周更新下一阶段的实验数据。

1. **爬虫：**给定了需要爬取电影、书籍 ID 数据各 1000 条。基本信息如下：

Movie_id.txt & Book_id.txt

以电影数据为例，如第 0 行 ID 1292052 对应电影《肖申克的救赎》

<https://movie.douban.com/subject/1292052/>

类似的，书籍数据的第 0 行 ID 1046265 对应书籍《挪威的森林》

<https://book.douban.com/subject/1046265/>

你可以在这里下载到本次实验的数据集：

链接：<https://rec.ustc.edu.cn/share/821dd0b0-3762-11ed-a74b-47619140a7ac> 密码：b2ok

提交说明

请于截止日期（待定）以前提交到课程邮箱 ustcweb2022@163.com，具体要求如下：

1. 邮件标题以及压缩包命名为"学号 1-姓名 1-学号 2-姓名 2-学号 3-姓名 3-实验 1"格式。
2. 因未署名造成统计遗漏责任自行承担，你可以将邮件抄送你的队友。
3. 实验报告请务必独立完成，如果发现抄袭按 0 分处理。
4. 迟交实验将不被接收。
5. 后续版本会进一步更新具体实验报告要求和截止日期时间。