倒排索引

1. 选题介绍

本实验旨在基于豆瓣电影排行Top250，设计并实现一个简单的倒排索引系统。通过Python构建，实现对电影信息的快速检索，并验证其在多维度关键词查询中的有效性。

2. 自己在小组中承担的工作、以及小组分工

独立完成

3. 项目实现过程，以及遇到的困难和收获

3.1 数据获取与处理

3.1.1 实验数据

通过https://blog.51cto.com/wangnp/12015519?articleABtest=0爬取得到豆瓣电影TOP250评分排行数据的csv文件，筛选部分关键词后得到所要用的数据

	A	B	C	D	E	F	G
1	电影名字	评分	导演	主演	年份	国家	类型
2	肖申克的救赎	9.7	导演：弗兰	主演：蒂姆	1994	美国	犯罪剧情
3	霸王别姬/再见	9.6	导演：陈凯	主演：张国	1993	中国大陆中	剧情爱情同性
4	阿甘正传/For	9.5	导演：罗伯	主演：汤姆	1994	美国	剧情爱情
5	泰坦尼克号/T	9.5	导演：詹姆	主演：莱昂	1997	美国墨西哥	剧情爱情灾难
6	千与千寻/千と	9.4	导演：宫崎	主演：柊瑠	2001	日本	剧情动画奇幻
7	这个杀手不太	9.4	导演：吕克	主演：让·雷	1994	法国美国	剧情动作犯罪
8	美丽人生/Lav	9.5	导演：罗伯	主演：罗伯	1997	意大利	剧情喜剧爱情战争
9	星际穿越/Int	9.4	导演：克里	主演：马修	2014	美国英国加	剧情科幻冒险
10	盗梦空间/Inc	9.4	导演：克里	主演：莱昂	2010	美国英国	剧情科幻悬疑冒险
11	楚门的世界/T	9.4	导演：彼得	主演：金·凯	1998	美国	剧情科幻
12	辛德勒的名单	9.5	导演：史蒂	主演：连姆	1993	美国	剧情历史战争
13	忠犬八公的故	9.4	导演：莱塞	主演：理查	2009	美国英国	剧情
14	海上钢琴师/L	9.3	导演：朱塞	主演：蒂姆	1998	意大利	剧情音乐
15	三傻大闹宝莱	9.2	导演：拉库	主演：阿米	2009	印度	剧情喜剧爱情歌舞
16	放牛班的春天	9.3	导演：克里	主演：让-巴	2004	法国瑞士德	剧情音乐
17	机器人总动员	9.3	导演：安德	主演：本·伯	2008	美国	科幻动画冒险
18	疯狂动物城/Z	9.2	导演：拜伦	主演：金妮	2016	美国	喜剧动画冒险
19	无间道/無間道	9.3	导演：刘伟	主演：刘德	2002	中国香港	剧情犯罪惊悚

关键词提取依据为

字段	索引处理方式

电影名字	仅索引中文主标题
评分	
导演	
主演	仅索引第一主演姓名
年份	
国家	多值拆分
类型	多值拆分

3.1.2 数据预处理

运行 `data_preprocessing.py` 文件

- 数据清洗：清理导演和主演中的前缀和一些多余信息
- 多值拆分：将国家和类型按分隔符拆分成独立的关键词列表，确保每个国家和类型都能被独立索引
- 评分处理：将评分转换为字符串

3.2 倒排索引构建与实现

参考文献

倒排索引的定义结构与工作原理-智能开放搜索 OpenSearch-阿里云

https://zhuanlan.zhihu.com/p/324378430

Step-by-Step Guide to Building an Inverted Index in Python

3.2.1 索引构建

从预处理阶段得到的 `movie_documents.json` 文件中加载数据，每条电影的记录包含如图

```
[
    {
        "id": 1,
        "title": "肖申克的救赎",
        "rating": "9.7",
        "director": "弗兰克·德拉邦特 Frank Darabont",
        "actor": "蒂姆·罗宾斯 Tim Robbins /...",
        "year": "1994",
        "country": [
            "美国"
        ],
        "genre": [
            "剧情",
            "犯罪"
        ],
        "raw_title": "肖申克的救赎/TheShawshankRedemption/月黑高飞(港)/刺激1995(台)"
    },
    {
        "id": 2
```


文档数据通过 `id` 存储到 `self.documents` 字典中

遍历每一部电影根据索引字段列表 `['title', 'rating', 'director', 'actor', 'year', 'country', 'genre']` 抽取关键词

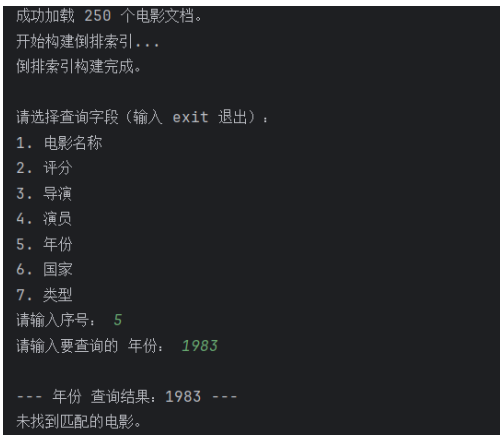
例如，更新时

```
index["director"]["诺兰"] -> [3, 15, 22]
index["genre"]["科幻"] -> [1, 4, 6, 10, 31]
index["year"]["1994"] -> [1, 3, 30, 59]
```

3.2.2 验证

运行 `inverted_index.py`

对于最初只有一个对应值的，以年份为例



```
请输入序号： 5
请输入要查询的 年份： 1993

--- 年份 查询结果：1993 ---
[2] 霸王别姬 | 评分 9.6 | 陈凯歌 Kaige Chen | 1993 | 剧情, 爱情
[11] 辛德勒的名单 | 评分 9.5 | 史蒂文·斯皮尔伯格 Steven Spielberg | 1993 | 剧情, 历史, 战争
[99] 唐伯虎点秋香 | 评分 8.7 | 李力持 Lik-Chi Lee | 1993 | 喜剧, 爱情, 古装
[123] 完美的世界 | 评分 9.1 | 克林特·伊斯特伍德 Clint Eastwood | 1993 | 剧情, 犯罪
[149] 喜宴 | 评分 9.0 | 李安 Ang Lee | 1993 | 剧情, 喜剧, 爱情, 家庭
[155] 射雕英雄传之东成西就 | 评分 8.7 | 刘镇伟 Jeffrey Lau | 1993 | 喜剧, 奇幻, 武侠, 古装
[214] 青蛇 | 评分 8.6 | 徐克 Hark Tsui | 1993 | 剧情, 爱情, 奇幻, 古装
```

有多个值的，以类型为例

```
7. 类型
请输入序号： 7

可选类型：
剧情 犯罪 爱情 灾难 奇幻 动画 动作 喜剧 战争 科幻 冒险 音乐 历史 传记 家庭 惊悚 悬疑 歌舞 古装 西部 运动 儿童 情色
请输入要查询的类型： 爱情

--- 类型查询结果：爱情 ---
[2] 霸王别姬 | 评分 9.6 | 陈凯歌 Kaige Chen | 1993 | 剧情, 爱情
[3] 阿甘正传 | 评分 9.5 | 罗伯特·泽米吉斯 Robert Zemeckis | 1994 | 剧情, 爱情
[4] 泰坦尼克号 | 评分 9.5 | 詹姆斯·卡梅隆 James Cameron | 1997 | 剧情, 爱情, 灾难
[7] 美丽人生 | 评分 9.5 | 罗伯托·贝尼尼 Roberto Benigni | 1997 | 剧情, 喜剧, 爱情, 战争
[14] 三傻大闹宝莱坞 | 评分 9.2 | 拉库马·希拉尼 Rajkumar Hirani | 2009 | 剧情, 喜剧, 爱情, 歌舞
[20] 大话西游之大圣娶亲 | 评分 9.2 | 刘镇伟 Jeffrey Lau | 1995 | 喜剧, 爱情, 奇幻, 古装
[28] 怦然心动 | 评分 9.1 | 罗伯·莱纳 Rob Reiner | 2010 | 剧情, 喜剧, 爱情
[34] 乱世佳人 | 评分 9.3 | 维克多·弗莱明 Victor Fleming | 1939 | 剧情, 爱情, 历史, 战争
[37] 哈尔的移动城堡 | 评分 9.1 | 宫崎骏 Hayao Miyazaki | 2004 | 爱情, 冒险, 奇幻, 动画
[49] 大话西游之月光宝盒 | 评分 9.0 | 刘镇伟 Jeffrey Lau | 1995 | 喜剧, 爱情, 奇幻, 古装
[54] 罗马假日 | 评分 9.1 | 威廉·惠勒 William Wyler | 1953 | 剧情, 喜剧, 爱情
[57] 天堂电影院 | 评分 9.2 | 朱塞佩·托纳多雷 Giuseppe Tornatore | 1988 | 剧情, 爱情
```

4. 个人在本项目的收获

理解了倒排索引的数据结构与原理，掌握了如何处理多类型字段，学会设计更合理的索引结构，提高了对检索系统查询流程的理解