

# 实验 1 豆瓣数据的爬取、检索与推荐

## 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爬取指定的电影、书籍的主页，并解析其基本信息 (Stage1)；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 (Stage2)；在此基础上，结合用户的评价信息及用户间社交关系，进行个性化电影、书籍推荐 (Stage3)。

## 实验要求

实验将分为爬虫、检索、个性化检索（推荐）三个阶段，本周发布实验第二个阶段的任务要求，要求基于第一阶段爬取的豆瓣 Movie/Book 信息，实现电影或书籍的搜索引擎（如果你想，也可以合在一起或者分别做一遍，不一定有加分）。对于给定的查询，能够以精确查询或模糊语义匹配的方法返回最相关的书籍或者电影集。

### 第二阶段任务---布尔检索

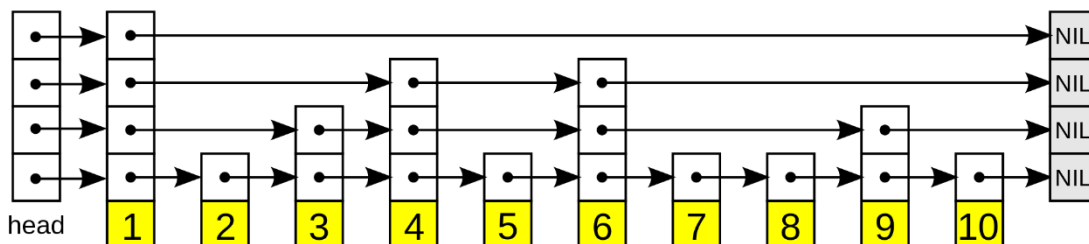
实现电影、书籍的 bool 检索。首先基于阶段一爬取的电影和书籍简介等数据，自行选择并提取需要使用的字段信息。以电影数据为例，对于剧情简介字段，将剧情简介视作一个文档，对其进行分词、去停用词处理，将剧情简介表征为一系列关键词集合；同时对于电影类型字段，如“剧情”、“犯罪”，可直接将其加入电影表征后的关键词集。具体而言，检索流程大致如下[0]：

#### 1. 对一阶段中爬取的电影和书籍数据进行预处理，将文本表征为关键词集合

- 中文没有显示分隔符，分词过程中存在歧义与新词识别的难题。你需要选择一个合适的分词算法来解决这些问题，提升分词效果；或者选择已有的分词工具直接处理文本。你可以适当地结合数据解释你使用的算法原理和采用原因。
- 在上一步中可能分出来含有许多意思相近但表达不同的词语，这些词会影响索引大小和检索准确度。你可以利用现有的 [Word2Vec 数据集](#)或手工标注的[同/近义词表](#)来合并这些词语（这里助教给出了一些例子，你也可以使用自己找到的工具），提升索引效果。类似地，去除停用词和进行根据编辑距离的纠错[1]也对提升搜索效果有不小的帮助。

#### 2. 在经过预处理的数据集上建立倒排索引表 $S$ ，并以合适的方式存储生成的倒排索引文件

- 倒排索引本身的编码和压缩方式有许多种，你可以选择一个自己认为比较合适（方便）的算法进行。或者干脆像助教一样用现成的高速压缩算法。
- 跳表指针可以有多层，感兴趣的同学可以查看 Skip List 对应的论文，很有意思。



3. 对于给定的 bool 查询 $Q_{bool}$  (例如 **动作 and 剧情**), 根据你生成的倒排索引表 $S$ , 返回符合查询规则 $Q_{bool}$ 的电影或/和书籍集合 $A_{bool} = \{A_1^{bool}, A_2^{bool}, \dots\}$ , 并以合适的方式展现给用户 (例如给出电影名称和分类或显示部分简介等)。

- 可以回想一下课上讲过的优化方法。在这里上一步设计的数据结构会体现出效率的差别。

## 数据集说明

本次实验除了采取一阶段中自行爬取的数据, 我们还提供了电影和书籍的用户标记 tag。你也可以将这些 tag 一起加入到对应电影/书籍的描述中, 也可以使用自己爬取的新数据。

电影和书籍的 tag 下载地址: <https://rec.ustc.edu.cn/share/fb282fa0-465a-11ed-8a8a-038094755c62> 密码: aoow

## 脚注

[0]:电影和书籍数据你可以分别建立索引进行搜索, 也可以放在一起建立一个索引。此外, 每一项任务下提示的主要作用是帮同学减少踩坑的次数, 同时给出一些思考方向, **学有余力的同学可以适当扩展**, 虽然做得好有可能可以拿到加分, *但请平衡好实验和生活, 不要卷坏*。

[1]:中文与英文不同, 根据编辑距离纠错通常需要基于拼音或五笔输入的相似度进行, 而非直接计算词距离, 这里有一个供参考的[例子](#)。

本阶段提交的实验报告中应包含实验方法、关键代码说明, 并对检索结果进行分析及展示; 代码请和实验报告一起包含在压缩包内提交。在结果展示中, 你需要在[豆瓣电影 Top 250](#) 中编号与你学号最后两位一致的电影/书籍中找出一些合适的关键词, 并展示这些词的检索结果 (如果一组有多个成员则需要展示对应的多份)。

## 提交说明

请于截止日期 (**11 月 13 日**) 以前提交到课程邮箱 [ustcweb2022@163.com](mailto:ustcweb2022@163.com), 具体要求如下:

1. 邮件标题以及压缩包命名为"组长学号-组长姓名-实验 1"格式。邮件正文中请列出小组所有成员的姓名、学号。
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。