实验一第 1 阶段 豆瓣数据的爬取与检索

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，然后结合给定的标签信息，实现电影和书籍的检索并评估其效果（1.1）；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（1.2）。

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）。

本周发布实验一第 1 阶段（1.1）的任务要求：首先爬取豆瓣 Movie&Book 信息，并实现电影和书籍的检索（可以合在一起做或者分别做一遍）。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。

实验内容

1. 爬虫

（1）爬虫要求

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。

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；鼓励抓取更多有用信息（可能有益于第 2 阶段的分析）

b) 对于书籍数据，至少爬取其基本信息、内容简介、作者简介；鼓励抓取更多有用信息（可能有益于第 2 阶段的分析）

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。

（2）爬虫数据集介绍

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，助教组后续会更新下一阶段的实验数据。

在提供的数据集中，我们给定了需要爬取电影、书籍 ID 数据各 1200 条。基本信息如下：Movie_id.csv & Book_id.csv

以电影数据为例，如第 1 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 1 行 ID 1046265 对应书籍《挪威的森林》

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

本次实验提供的数据集下载地址如下：

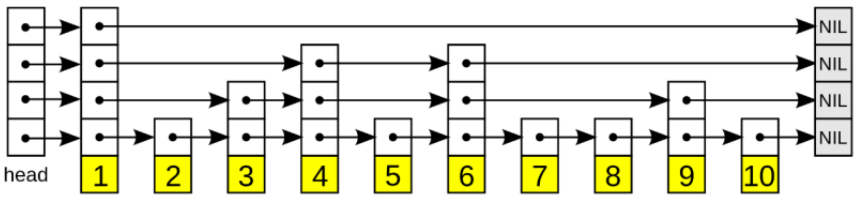
链接：https://rec.ustc.edu.cn/share/eb631310-547a-11ee-bdef-659e1b5709dc

2. 检索

（1）检索要求

实现电影、书籍的 bool 检索。首先基于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（如果采用现成的分词工具，请至少使用两种或以上工具，并比较其效果的差异性）。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（也可以使用其他自己找到的工具），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**
   - 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[2]，很有意思（可选，不一定要做成多层）。



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} = $\{A^{bool}, A^{bool}, \dots\}$，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。
4. **任选一种课程中介绍过的索引压缩加以实现，如按块存储、前端编码等，并比较压缩后的索引在存储空间和检索效率上与原索引的区别**

此部分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）。

（2）检索数据集说明

本次实验除了采取自行爬取的数据，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：

链接：https://rec.ustc.edu.cn/share/baa71590-547a-11ee-95b6-1d30d3692b76

（3）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。

[2]: Pugh W. Skip lists: a probabilistic alternative to balanced trees[J]. Communications of the ACM, 1990, 33(6): 668-676.

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"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。
6. 整个实验一只需提交一份实验报告，请等待 1.2 发布，并在全部完成实验一后再统一提交