实验一 第 1 阶段 豆瓣数据的检索

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结合给定的电影、书籍的标签信息，实现电影和书籍的检索并评估其效果。

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）。

豆瓣数据的检索：根据给定的豆瓣 Movie&Book 的 tag 信息，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，通过分词、构建索引、索引优化等流程，实现面向给定查询条件的精确索引。

实验内容

1.检索

（1）数据集说明

① 原始数据集：

“Movie_score.csv” 与 **“Book_score.csv”** 为用户的评分信息，具体内容格式如下：

User ID, Item (Movie/Book) ID, Rating (0-5), Timestamp[, Tag 1, Tag 2, …]

例如： 1000001, 1293510, 3, 2005-06-26T20:41:22+08:00, black humor 表明，ID 为 1000001 的用户给电影 1293510 打了 3 分，时间为 2005-06-26T20:41:22+08:00，同时留下了 **black humor** 的标签。

② 标签数据集：

“selected_book_top_1200_data_tag.csv” 、 **“selected_movie_top_1200_data_tag.csv”**

为我们从原始数据集中提取的标签信息，具体内容格式如下：

Movie/Book ID, Tags

例如：1046265,"{'大学读的', 'tonylist', '传说中的村上'}" 表明书籍 1046265 的标签是“大学读的、tonylist、传说中的村上”

在进行本次实验时，可以选择从**原始数据集中获取 tag 信息，也可以直接使用我们筛选过的标签数据集。**

具体数据集地址如下：

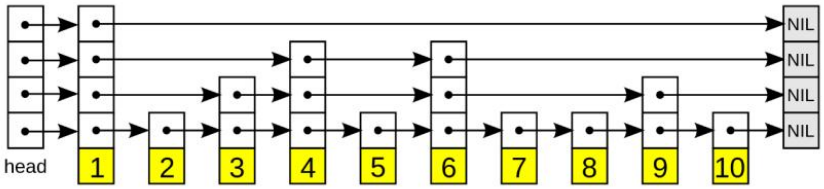
链接：https://rec.ustc.edu.cn/share/46ede9b0-763c-11ef-9ef2-35ffc59e01a0 密码：1234

（2）任务说明

基于给定的电影、书籍的标签信息，实现电影、书籍的 bool 检索。具体而言，检索流程大致如下[0]：

1. 对一阶段中给定的电影和书籍数据进行预处理，将文本表征为关键词集合：
   - 由于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。因此，你需要选择一个合适的分词算法来解决这些问题（如课上讲过的HMM等模型），或者选择已有的分词工具直接处理文本（注意：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。为此，你可以利用现有的 Word2Vec 数据集、手工标注的同/近义词表或其他合适的工具，通过词语的语义相似度来合并这些词语。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。请设计合理的方案，并基于数据分析处理后的效果。
2. 基于前一阶段形成的分词结果，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. 优化你生成的倒排索引表S，对于给定的包含任意组合（如括号）的布尔查询Q_{bool}（例如 (动作 and 剧情) or (科幻 and not 恐怖)），使其能够支持复杂的布尔查询操作。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（例如输出tag等）。

- 查询条件由自己设计，不作统一要求，但请设计多个不同的查询条件，并且具有一定的复杂度，以便于进行查询效率的分析。
- 请比较不同词项处理顺序对于最终耗时的影响。
- 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. 任选两种课程中介绍过的索引压缩方法加以实现，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。

（3）相关说明

[0] 对于电影和书籍数据，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，但请平衡好实验和生活。

[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。邮件正文和实验报告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，请等待实验一-第 2 阶段发布，并在全部完成实验一后再统一提交