实验一 信息检索部分

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）。

根据本课程**信息检索部分**内容（对应第 2-9 周课程），完成以下实验：

- 设计网络爬虫，通过网页端或 API 爬取数据【选做】
- 对指定文档进行文字处理，提取规范化词项
- 构建与优化倒排表和索引
- 基于倒排表进行文档查询
- 面向特定用户的个性化查询/推荐【选做】

数据背景

数据集来自著名在线活动组织网站 Meetup，可通过科大云盘下载：

链接：https://pan.ustc.edu.cn/share/index/a7510487e45d4e7e9330

密码：xlIf

数据使用方式：解压后，将“All.pak”文件拖曳至“FilePackager.exe”文件上，将自动进行解压缩操作。因文件数量巨大，请耐心等待。

该数据集原始文件从 Meetup 旧版官方 API 获得，以.xml 格式进行存储，一共包含四类文件，分别对应社团信息（Group）、事件（Event）、用户参与（RSVP）及用户信息（Member）。一共包含 437 个社团、82770 个用户及 93512 个事件。

其中，每个社团由若干名用户组成，用户可以随时加入或退出。用户可以在社团内发起事件/活动，所有社团成员（也仅有社团成员）会收到事件邀请。社团成员可以选择是否参加事件（Yes/No/Maybe），但不是所有人都会回应。

以事件（Event）信息为例，其 XML 文件格式如下：

实验内容

前排友情提示：

（1） 选做内容不影响分数，请根据时间及兴趣量力而行，如无精力可以不做。

（2） 可根据课程中介绍的方法进行实验，也可根据自己掌握的知识修改方案，但请注意不要内卷。

1. 【选做】网络爬虫实验

如果认为课程组提供的数据量不足以进行实验，或希望锻炼自己的爬虫技能，可以通过以下 1-2 种方式选做该部分实验：

（1） 直接爬取并解析网页

通过浏览 Meetup 的官方主页，可以获得有关社团、活动、成员等各类信息。

以 New York Tech & Beer 社团为例（页面为 https://www.meetup.com/nyctnb/），可以在该社团页面上获取该社团组织过的所有活动及成员信息。注意，获取部分信息可能需要注册账号。

（2） 通过 Meetup 所提供的 API 获取信息

Meetup 目前通过其官方 API（https://www.meetup.com/api/guide/）提供数据

接口。可通过阅读该文档尝试获取数据。但需注意，相较于旧版文件，目前的返回文件格式为 json，且其格式和内容存在较大差异。

另外，通过搜索引擎可获取其他 Meetup 有关 API，但普遍版本较陈旧，可用性存疑，请注意甄别。

2. 【必做】文档解析与规范化处理

在获取实验所需的文档后，需要从 xml（或额外抓取的 json）文档中，提取出所需的内容，并进行规范化处理。具体流程包括：

（1） 结合自选编程语言中的工具包，解析文档中所需的部分内容。文件范围和内容自定（但至少应包括 Event 类文件及其中的 Description 部分内容），并将从一个文件中解析出的内容合并为一篇待检索的文档。

（2） 对文档中的文本进行分词处理，即将成段的文字拆分为单字词和短语。短语可简单根据连字符等制定规则进行拆分，也可引入外部词库协助拆分。当然，如果想偷懒的话，也可以直接只保留单字词（例如，将 To be or not to be 拆成 6 个基本单词）。

（3） 根据第 3 节课内容，对分词后的所有单词进行规范化处理，从而形成规范化的词项（Token），包括并不限于去除停用词、数字、标点符号和其他特殊字符，对单词进行归一化处理（词干提取、词形还原）等等。该部分可通过手动编写规则进行，也可以通过寻找工具包来进行。

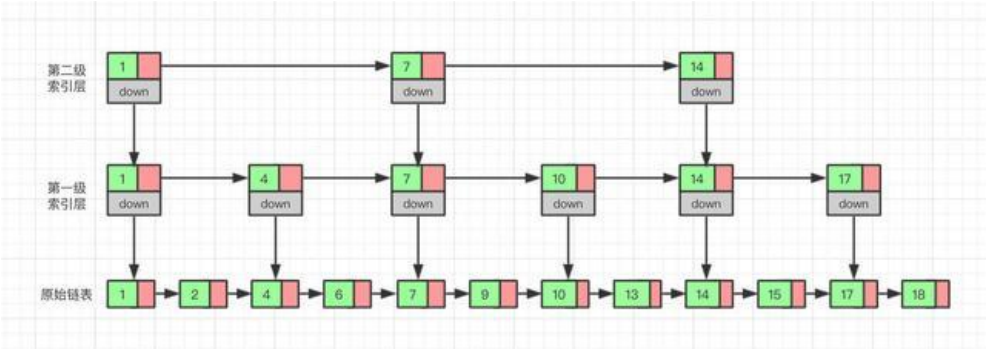
3. 【必做】倒排表的构建

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，并尝试形成相应的跳表指针。具体流程包括：

（1） 根据第 4 节课内容，对前一部分所获得的所有文档中的所有规范化词项构建倒排表。

（2） 为实现面向倒排表的快速检索，设计合适的跳表指针。

（3） **【选做】**感兴趣的同学可以查看 Skip List[1] 对应的论文，设计多层跳表指针，并比较与单层跳表指针的性能和开支差异。



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

4. 【必做】倒排表的扩展与优化

根据第 4 节课后半部分的内容，尝试对前一阶段所构建的倒排表进行扩展与优化。具体流程包括：

（1） 在倒排表中加入词项的位置信息，以应对短语检索需求。

（2） 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上与原索引的区别。

5. 【必做】多种形式的信息检索

根据第 4、6、7 节课程的内容，尝试对先前生成的文档库进行检索实践。注意：本环节所涉及的查询条件（Query）自行设计。具体流程包括：

A. 布尔检索

（1） 自行设计不少于 3 种复杂查询条件，以布尔表达式的形式呈现。并通过实验分析同一个布尔表达式的不同处理顺序对时间开支的影响。

（2） 根据倒排表进行检索，并比较索引压缩前后在检索效率上的差异。

（3） 至少设计 1 次面向短语的检索，并分析加入词项位置信息的扩展倒排表在应对短语检索任务时的效果。

（4） 选择不同的跳表指针步长，并分析其对存储性能/检索效率的影响。

B. **向量空间模型：**自行设计查询条件，计算查询条件与文档的 TF-IDF 值，通过向量空间模型进行文档检索实践。

C. **【选做】**基于文档表征的检索：自行选择合适的表征模型，基于文档/查

询条件的表征与相似性计算实现文档检索，并分析比较其与向量空间模型的效果差异。

6. 【选做】基于用户画像的个性化检索/推荐

根据第 8、9 节课程的内容，利用提供数据中的用户信息及活动记录（RSVP）信息，预测用户对活动的参与兴趣。具体流程包括：

（1） 从原始文档中解析出与个性化检索任务相关的信息。

（2） 采用协同过滤方式，仅利用用户-活动的反馈矩阵（即 Yes/No/Maybe）进行评分预测。可以自行在基本协同过滤的基础上自行添加部分策略，如考虑评分的情境或周期性等因素，并讨论这些因素会对排序精度产生何种影响。

（3） 充分利用更丰富的信息，如活动介绍、个人介绍、活动的时间/地点/开支等，构建更复杂的模型进行预测。

注意：请结合预测结果分析方法的有效性。在本次实验中，训练/测试数据集的划分及评价指标均自行选择，不做统一要求。（坚决反卷！

提交说明

请于截止日期（初步定于 **2025 年 11 月 1 日晚 23:59**，可能根据实际情况进行调整）以前提交到课程邮箱 **ustcweb2025@163.com**，具体要求如下：

1. 请在邮件中提交实验报告，并将实验代码、索引文件和其他你认为有必要提交的附件上传至科大云盘（https://pan.ustc.edu.cn）以备检查。请将云盘的链接填入实验报告中。
2. 邮件标题以及压缩包以"组长学号-组长姓名-Web 信息处理实验 1"格式命名。邮件正文及实验报告正文中请列出小组所有成员的姓名、学号。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。