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，本周发布实验第三个阶段的任务要求，要求基于第一阶段爬取的豆瓣 Movie/Book 信息、我们提供的豆瓣电影与书籍的评分记录以及用户间的社交关系，判断用户的偏好。

第三阶段任务

在这个阶段中，你们需要对用户交互过的 item（电影/书籍）进行（基于得分预测的）排序。

数据说明

这次我们在阶段二的基础上提供了社交网络信息和用户评分信息。你可以在这里下载你需要使用的数据：

链接：https://rec.ustc.edu.cn/share/1f91c220-5337-11ed-a3b2-49e51d1bec2f 密码：i97l

“contacts.txt”为社交网络信息。

例如，一条记录为：A: B, C, D，则意味着 **A** 与 **B、C、D** 三位用户之间存在社交关系，这里的社交关系是双向的（或无向的）。

因为实验数据进行了筛选，而社交网络数据没有做筛选，所以其中可能包含若干未在评分记录中出现的用户 **ID**。**是否需要利用社交网络信息**，如何利用这部分数据请同学自定。

“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。

实验任务说明

在这次实验中，你们需要自行划分训练集与测试集，在测试集上为用户对书籍或电影的评分进行排序，并用 NDCG 对自己的预测结果进行评分和进一步分析。**书籍和电影选一个完成就可以了**。

根据徐老师的最高指示，为了将反卷贯彻到底，这次实验我们不会像前几年一样给出评测平台，需要由同学们自行选择方法来完成任务并分析结果。

数据划分

你们可以按一定比例划分某些（或全部）用户的评分，例如以 5:5 划分，50% 用于分析，50% 拿来预测（或其他比例）。用于预测的数据为抹去了打分分值的数据，即：用户与这些电影/书籍交互过，但（假装）不知道得分。

有一些用户的评分数据过少，你们可以自行决定是否使用这些数据进行分析或预测。

评分排序

你们需要对上面抹去分值的对象进行顺序位置预测，即：若以升/降序排序用户的所有评价，那这些数据应该放在第几位。将你们预测出的对象顺序与实际的顺序进行比较，并用 NDCG 评估你们的预测效果。

同学们可能注意到了，在这里我们的用词是“顺序”，即不一定要预测用户的实际评分，给出合理的顺序即可（当然也可以先预测评分再排序）。如果同学需要预测评分，可以参考课件使用 kNN 或 SVD 等方法。*不需要预测评分直接给出顺序预测的方式助教暂时也不知道，但同学如果能实现也可以。*

我们给出的数据除了评分本身，还有社交关系/tag/时间戳，若有需要同学可以自行取用。

结果分析

你们需要根据上面的得分对自己的方法和结果进行一定分析，若采用了不同的方法，也可以比较不同方法的结果。同时你们需要保留预测结果和过程以备助教查验。

在实验报告中你们需要对以上几步里你们的分析、采用的方法、取得的效果进行举例和阐释。**同时你们需要保留本次实验的预测结果和数据集划分供助教查验，这些数据不用提交。**

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，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。