实验 1 信息获取与检索分析

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面向豆瓣平台，爬取指定的电影、书籍的主页，并解析其相关信息（Stage 1）；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（Stage 2）；在此基础上，结合用户的评价信息及其他边信息，进行个性化电影、书籍推荐（Stage 3）。

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）。
实验将分为爬虫、检索、个性化检索（推荐）三个阶段，持续时间约为 3-10 教学周，实验报告的具体提交时间将于第二阶段公布。

第一阶段任务如下：

1. **爬虫：**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；
b) 对于书籍数据，至少爬取其基本信息、内容简介、作者简介；
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。

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，助教组会每周更新下一阶段的实验数据。

1. **爬虫：**给定了需要爬取电影、书籍 ID 数据各 1000 条。基本信息如下：
   Movie_id.txt & Book_id.txt
   以电影数据为例，如第 0 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 0 行 ID 1046265 对应书籍《挪威的森林》
   https://book.douban.com/subject/1046265/

你可以在这里下载到本次实验的数据集：
链接：https://rec.ustc.edu.cn/share/821dd0b0-3762-11ed-a74b-47619140a7ac 密码：b2ok

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"学号 1-姓名 1-学号 2-姓名 2-学号 3-姓名 3-实验 1"格式。
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和截止日期时间。