HW5 Web搜索引擎（期末大作业）

作业描述

本次作业我们要求实现一个系统的Web搜索引擎（主题不限），为用户提供查询服务和个性化推荐。

PS：本次作业可以借助各种工具和包。

作业要求

实现本次作业主要有网页抓取、文本索引、链接分析、查询服务、个性化查询几个步骤，个性化推荐为扩展内容。

1.网页抓取

根据搜索引擎主题选择网站进行爬取，爬取网页数目不限制，**注意不要违法**，主题不会作为评分标准，不需要太过纠结爬什么。

参考资料：参考教材第20章-Web采集及索引

2.文本索引

对网页及其锚文本构建索引，可以按锚文本、网页标题、URL 等域构建索引。

PS：构建索引方法不做限制，可以使用Elasticsearch（调包）

3.链接分析

使用PageRank进行链接分析，评估网页权重。

参考资料：参考教材第21章-链接分析

4.查询服务

完成上述模块后，为用户提供**站内查询、短语查询、通配查询、查询日志、网页快照**等高级搜索功能。可以参考百度或者谷歌的高级搜索功能。

图 1: 百度的高级搜索功能

图 2: 谷歌的高级搜索功能

5.个性化查询

个性化查询为不同的用户提供不同的内容排序。

可以实现一个账号登录系统，通过用户完善的年龄性别等个人信息为其呈现不同的查询结果；或者是记录用户的查询历史，通过历史查询来提供个性化的查询结果。在 google 的查询中就会通过这些手段来优化用户的查

询体验。

6.Web页面，图形化界面

大家可能在“互联网数据库“课程中学习过如何使用yii框架搭建web页面，本次实验你也可以借用框架实现Web页面，但这有可能会让你本次实验重心偏移，因为实验重点应放在查询服务的具体原理上。

你在本次实验中不必详细区分前后端，但需要设计类似图形化界面的Web“前端”页面，并使用户与“前端”页面交互，能达到和你“后端”搜索引擎的核心逻辑进行交互的目标即可。

7.个性化推荐

本次作业的扩展内容为个性化推荐，个性化推荐系统通过用户的个人信息和查询历史获取用户可能的兴趣点，在用户查询时给用户推荐相关领域的其他内容。比如在百度上搜索 iphone，其会在查询结果的右侧为你推荐 ipad、iMac 等相关产品。

8.提交要求

这次作业的截止日期为2022.12.16日晚23:59，请同学们在截止日期前将代码、文档、演示视频（不超过15分钟）打包（命名“学号_姓名 _hw5”）发送到 nkulxb2022@163.com。

注意：不要发送爬取的网页

评分标准

本次作业截止日期之后，每迟交1天，扣除本次作业2%的起评分，扣到60%起评为止。抄袭现象，不再给补交机会，严肃处理。

- 代码内容
    - 资源抓取 10%
    - 索引构建 10%
    - 链接分析 10%
    - 提供查询服务 35% (第一项15%，每再做一项5%，上限35%)
    - 个性化查询服务 10%
    - Web页面 5%
    - 个性化推荐 10%
- 文档、演示视频
    - 文档 5%
    - 演示视频 5%