2021AIR-hw4: Web搜索引擎 - 南开资源站

本次作业的要求是针对**南开校内资源**构建一个Web搜索引擎，为用户提供南开信息的查询服务和个性化推荐。

本次作业是半开放性题目，你可以只针对某一方面资源构建搜索引擎作为本次实验的主题，如南开动漫资源站，南开新闻资源站；也可以制作综合性的资源搜索平台，如百度、谷歌，由你自己决定。**但至少要包括以下作业要求中的模块，具体的实现细节不做要求。**

本次作业可以借助各种工具和包，希望大家善于利用以减少重复工作量。在构建的时候，可以吸取上一次实验的经验，使用elastic search构建Web搜索引擎。

目录

具体实现

实现这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 代码模块要求

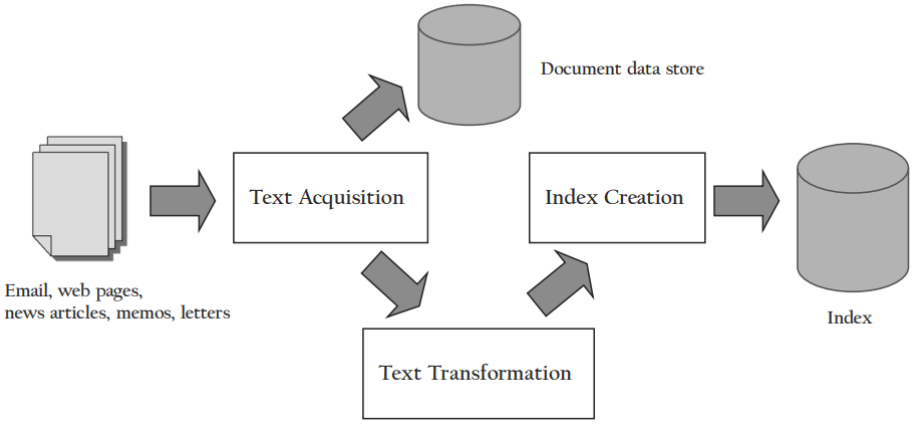
1.1 网页抓取

对南开大学校内资源内容进行抓取，可以包括新闻，文章，下载链接等等。

本部分内容原理你可以参考教材[1]第20章Web采集及索引，SEIRiP[2]第3章实现。

1.2 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。



> Tips: 可以复用第一次实验的索引构建部分代码，也可以使用elastic search构建索引，可以合理减少工作量。

1.3 链接分析

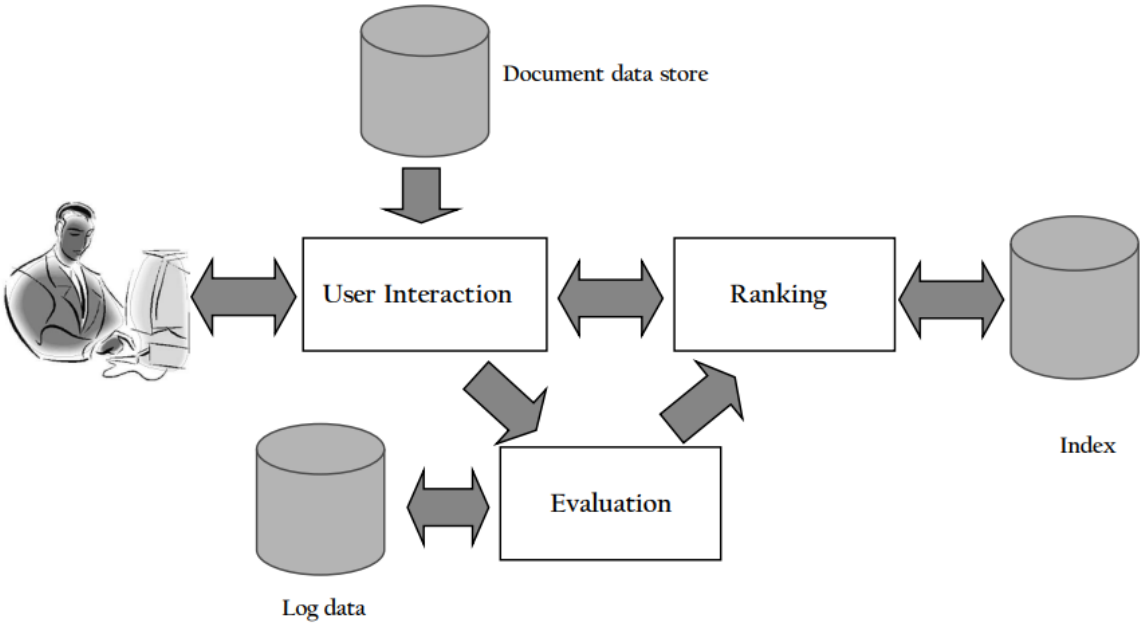
使用PageRank进行链接分析，评估网页权重。

本部分内容原理你可以参考教材[1]第21章链接分析，SEIRiP[2]第4.5节Link Analysis实现。

> Tips: Pagerank有对应的包实现，可以合理减少工作量。

1.4 查询服务

查询服务是本次实验重点，同样必然也是成绩占比最大的一部分。一个用户和搜索引擎交互过程如下图所示。



使用向量空间模型并结合链接分析对查询结果进行排序，为用户提供**站内查询、文档查询、短语查询、通配查询、查询日志、网页快照**等高级搜索功能。更多的内容可以参考百度或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

1.5 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，通过用户完善的学院专业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。

1.6 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。

2 作业提交

在这个学期剩下的时间里，大家还需完成包括这次作业在内的两次作业。这次作业的截止日期为12月12日 23:55，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_hw4"）发送到 nkuir2021fall@163.com。

3 评分标准

本次作业截止日期之后，期末考试之前，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 40% (前两项10%，每再做一项5%，上限40%)
  - 个性化查询 10%
  - Web页面 10%
- 文档、演示视频
  - 文档 5%
  - 演示视频 5%

4 Reference

所有有关架构的插图均引用自如下书籍，高级搜索功能的插图节选自百度和谷歌。

[1] 信息检索导论(课程参考教材)，人民邮电出版社，2010

[2] Search Engines Information Retrieval in Practice, W.B. Croft, D. Metzler, T. Strohman, 2015