

HW4 - NKU Web Search Engine

Xiaorui Qi 2025.05.07

1. 作业要求

本次作业针对**南开校内资源**构建 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。

作业为半开放性题目，你可以只针对某一方面的资源构建搜索主题，如南开动漫资源站，南开新闻资源站等。也可以制作综合性的资源搜索平台，类似 Baidu、Google 等。具体实现细节不做要求，自己制定主题，**但请至少包含本手册中涉及到的模块。**

作业**可以借助各种工具和包**，请大家善用并减少重复工作量。推荐采用 Elastic Search。

2. 具体实现

本次作业主要包含网页抓取、文本索引、查询服务、个性化查询四个步骤。

2.1 网页抓取

对南开校内资源进行抓取，依据自己主题而定。

抓取网页数量**至少为 10 万**，低于该数量将酌情扣分。

再次强调，请在校内抓取、礼貌抓取，遵循爬虫协议。

2.2 文本索引

对网页及其锚文本构建索引，可以设置多个索引域，比如：网页标题，URL，锚文本等。

2.3 查询服务

基于向量空间模型，结合链接分析结果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、网页快照等共计**六种**高级搜索功能。

2.3.1 站内查询

基本的查询操作，送分项。

2.3.2 文档查询

一些网页可能会携带或其本身就是附件下载链接，支持对文档（doc/docx，pdf，xls/xlsx 等）的查询操作。

2.3.3 短语查询

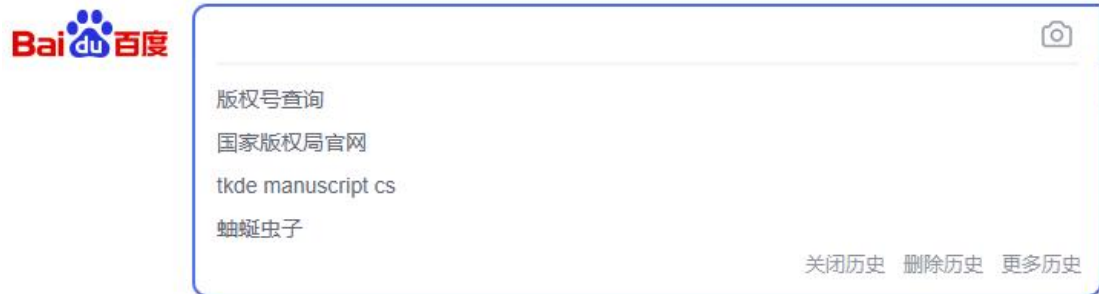
也是基本的查询操作，支持对多个 Term 的查询。注意，“南开¹大学²”和“南开¹是一所综合性大学²”的区别。

2.3.4 通配查询

用户可能并不知道自己要查什么，支持通配符（正则）查询操作。

例如：*代表多个字符，? 代表一个字符，“温*”用于查询所有姓温的老师，“计?”用于查询如“计网”、“计算”等词。

2.3.5 查询日志



参考百度的查询日志（历史），能够知道用户之前查了什么。

2.3.6 网页快照

网页快照是搜索引擎在收录网页时，对网页进行备份，存在自己的服务器缓存里，当用户在搜索引擎中点击链接时，搜索引擎将爬虫系统当时所抓取并保存的网页内容展现出来，称为网页快照。

例如：2008 年 7 月 20 日打开一个 Google 的网页快照，而这张快照上显示是 Google 在 7 月 10 日搜索并存档的。这表示：2008 年 7 月 20 日，这个网页或许已被删除或更新，但是，2008 年 7 月 10 日，当 Google 对该网页复制存档的时候，该网页是确实存在的。

2.4 个性化查询

针对不同登录用户提供不同的内容排序。

建议实现一个注册/登录系统，通过提供不同的用户信息实现内容排序算法的修正。

2.5 Web 界面

本次作业不强调界面美观度，以功能为主。

因此，可以 Terminal，可以 Web Page，二者不会存在得分区别。

2.6 个性化推荐

个性化推荐存在两类，一个是搜索上的联想关联，一个是内容分析后的推荐。任选其中一种实现即可。

×
📷

南开大学

南开大学在哪个城市

难开头的四字成语

南开中学

南开区

南开大学是211还是985

南开医院

南开大学滨海学院

南开大学医学院

南开大学大型仪器平台

反馈

相关院校

展开

 清华大学 红色工程师的摇篮	 复旦大学 自主创办的第一所高校	 厦门大学 综合类国家重点大学	 北京大学 中国第一所国立大学
 上海外国语大学 格高志远 学贯中外	 上海交通大学 享誉海内外的高等学府	 四川大学 综合性全国重点大学	 哈尔滨工业大学 世界一流大学建设高校

3. 评分标准

- 代码（90%）
 - 网页抓取（15%）
 - 文本索引（15%）
 - 查询服务（35%，完成一个获得 10%，后续每完成一个获得 5%，直到 35%）
 - 个性化查询（10%）
 - Web 界面（5%）
 - 个性化推荐（10%）

- 说明材料（10%）
 - 文档（5%）
 - 演示视频（5%）

4. 作业提交

本次作业持续时间为五周左右，截止时间为 6 月 11 日（周三）晚 23:59. 提交附件仅包含一个压缩包，优先推荐.zip/.rar 格式。邮件和文件命名格式均为学号_姓名_hw4，例如：1811412_戚晓睿_hw4.zip. 提交邮箱为：ir24fall@163.com

压缩包的目录格式如下：

- 1811412_戚晓睿_hw4.zip
 - 代码
 - 说明文档
 - 演示视频（不超过 15min，视频要求同前两次作业）