



哈爾濱工業大學
HARBIN INSTITUTE OF TECHNOLOGY

信息检索

实验三：企业搜索系统的设计与实现



School of Computer Science and Technology

Harbin Institute of Technology

1 实验目标

本次实验目的是对企业有所系统的设计与实现过程有一个全面的了解。本次实验设计的内容包括：对数据建立索引，实现文档的搜索，并对检索结果排序；实现企业搜索中的分权限访问。

2 实验环境

编程语言为：Python

其他无特殊要求

3 实验内容及要求

3.1 建立检索系统

任务描述：本次实验使用到的数据是实验 1 中爬取的网页数据。同学们首先要对 1000 个网页的网页内容建立索引，其次也要对爬取到的所有附件文档建立索引。然后实现一个简单的检索系统，实现数据和文档检索（文档检索要求同学们对文档内容建立索引，可以分开实现，也可以合二为一），并且能够精确的对检索结果进行排序。这一部分要求同学们做成简单的 UI 应用或者是 Web 应用，来保证易用性。

提交要求：无需提交文件

3.2 分权限访问

任务描述：同学们可以自己定义多种不同的“企业角色”（至少 4 种），这些角色对数据或文档的访问权限不同，然后为每条数据增加访问权限。然后在现有检索系统的基础上加入分权限访问功能，使得不同角色的用户在使用检索系统时，只能看到自己具有访问权限的那部分内容。同时在 3.1 的基础上对应用界面改进，便于切换企业角色。

分权限访问的实现一般有以下两种模式：

1. 检索条件控制：对检索条件加以限制，仅在用户访问权限内的文档中检索。
2. 检索结果过滤：检索结果的后处理，过滤掉用户不具备访问权限的结果。

提交要求：检索系统代码（retrival_system.py 或文件夹 retrival_system）

4. 实验提交

本次实验的实验报告请严格按照“信息检索实验报告模板.docx”的格式完成,并导出为 **PDF 格式**，按“学号+姓名+实验 3 企业检索系统设计与实现实验报告.pdf”命名。

请同学们按照各个实验模块的提交要求**正确命名提交文件**，然后将所有文件打包命名为“学号+姓名+实验 3 企业搜索系统的设计与实现.zip”，发送到邮箱：