哈尔滨工业大学

<<信息检索>>

实验报告

(2023 年度春季学期)

姓名：	李浩桢
学号：	7203610321
学院：	未来技术学院
教师：	张宇

实验三 企业搜索系统设计与实现

一、实验目的

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。

二、实验内容

1. 建立索引系统
2. 分权限访问

三、实验过程及结果

1. 建立索引系统

使用实验 1 爬取到的 craw.json 文件，并对其进行预处理，对 title 和 paragraphs 进行分词和停用词去除处理。同时处理其 file_name，利用模块 docx 读出每个文件内的内容，并对其进行分词。将页面检索和文件检索分开来实现，对于页面检索，对页面 title 和 paragraphs 建立索引，并使用 BM25 模型（bm25_page，语料为 title 和 paragraphs）检索；对于文件检索，我们对文件内容和文件所属页面的 title 建立索引，同样使用 BM25 模型（bm25_file，语料为分词后的文件内容和所属页面 title）检索。检索时先根据查询 Q，利用倒排索引找到包含查询 Q 中词的所有页面/文件，再计算每个页面/文件的 RSV 值，根据 RSV 值倒序排序得到查询结果。

2. 分权限访问

在预处理 craw.json 文件时，随机为每个页面和其页面下的附件设置一个访问权限值（1-4）。在检索时会检查当前用户的权限等级是否大于等于页面/文件的访问权限。若大于等于则将结果写入 table（允许展示给用户），若小于则不写入 table，用户检索不到这个结果。

3. UI 设计

使用 PyQt5 设计 UI，并设置了丰富的快捷键。

	页面标题	页面正文	权限要求
1	学期学历查询	2019年春季学期学历[201	权限等级1
2	通知公告	关于调停代课通知[2015-0	权限等级3
3	2015年春季学期"好大学...	2015年春季学期"	权限等级1
4	2021春季学期大学英语...	同学们：根据《大学英...	权限等级3
5	2021春季学期大学英语...	同学们：根据《大学英语...	权限等级2
6	关于C9合作交流项目201...	C9合作交流项目...	权限等级3
7	2015年春季学期港澳台...	有关院（系）及学生：	权限等级3
8	关于大学英语课程免修、...	同学们：根据《大学英语...	权限等级2
9	2015年春季学期学历	2015年春季学期学历.doc	权限等级2
10	2016年春季学期学历	2016年春季学期学历.doc	权限等级1
11	关于C9高校交流项目202...	各学院及学生：C9高校...	权限等级1
12	赴香港、澳门、台湾地区...	根据我校与香港、澳	权限等级2

四、实验心得

实验总体不难，将实验一和实验二的部分内容整合即可完成，细心处理好 craw.json 文件并规范编写 UI 页面即可。熟悉了 PyQt5 库的使用，docx 库的使用，win32com 库的使用。