信息检索实验报告——本地文件检索系统

王本亮

January 10, 2018

Contents

1 作业要求

无，与信息检索相关即可

2 选题构思

初拿到这个题目时，我犹豫了好几天究竟要选择什么题目。在我的心目中，一个完美的选题应该是既能与信息检索相关，又能在实际的生活中使用的项目。对于前面几次的作业，无论是莎士比亚全集的检索，还是南开内网的检索，都是虽然有用，但是日常生活中很少用到的作品。直到上周做操作系统作业时，为了查询某个特定的函数定义是什么，我只能打开 notepad++ 进行多文档搜索。然而这样不仅麻烦，而且需要在 notepad++ 中打开大量文件，污染编辑环境。因此，实现一个本地的文件夹内所有文件全文索引的想法油然而生。

3 实现流程

3.1 实现平台

在以往的三次作业中，均使用了 tomcat+ 阿里云的方式来构建服务器，考虑到该软件主要作用为本地操作，同时如果需要用户将文件上传到网上去，不仅效率上会因为网络状态不同而造成大幅下降，很多用户也将因为安全性和隐私性的考虑而不使用该软件。因此，本次实验我们使用了可在本地运行的控制台程序。考虑到平台的可移植性问题，我们采用了 Java 作为主要编程语言。

3.2 建立索引

考虑到用户的需求可能为同时在好几个目录下进行检索，因此只建立一个索引是远远不够的。因此本实验准备了十个文件夹存放索引，用户可自由选择存放索引的位置，也可随时切换查询的文件夹。同时，文件夹采用的是相对路径的方式来存储，便于在不同的电脑上运行

3.3 解析文档

在以往的几次实验中，我们都是采用的根据后缀名判断不同的文档格式，再引入不同的处理类库对文档进行处理。然而考虑到文件的后缀名可能被用户手动修改或不准确，同时手动对不同的文档格式进行处理过于麻烦，再考虑到用户本机的文档形式可能十分复杂，如编码方式等。因此本次实验采用了 apache 的 tika 类库，可以自动判断不同文件的类型及编码方式，并进行解析。我们采用 tika 类库来存储文档的类型及内容。

3.4 构建索引

对于索引的构建，本次实验仍然采用的 lucene 来进行索引的构建。在索引中将存放文件名、文件路径、文件内容及文档类型。

3.5 文档查询

用户搜索时，可自由选择希望查询的文件夹，输入关键字，将返回查询结果

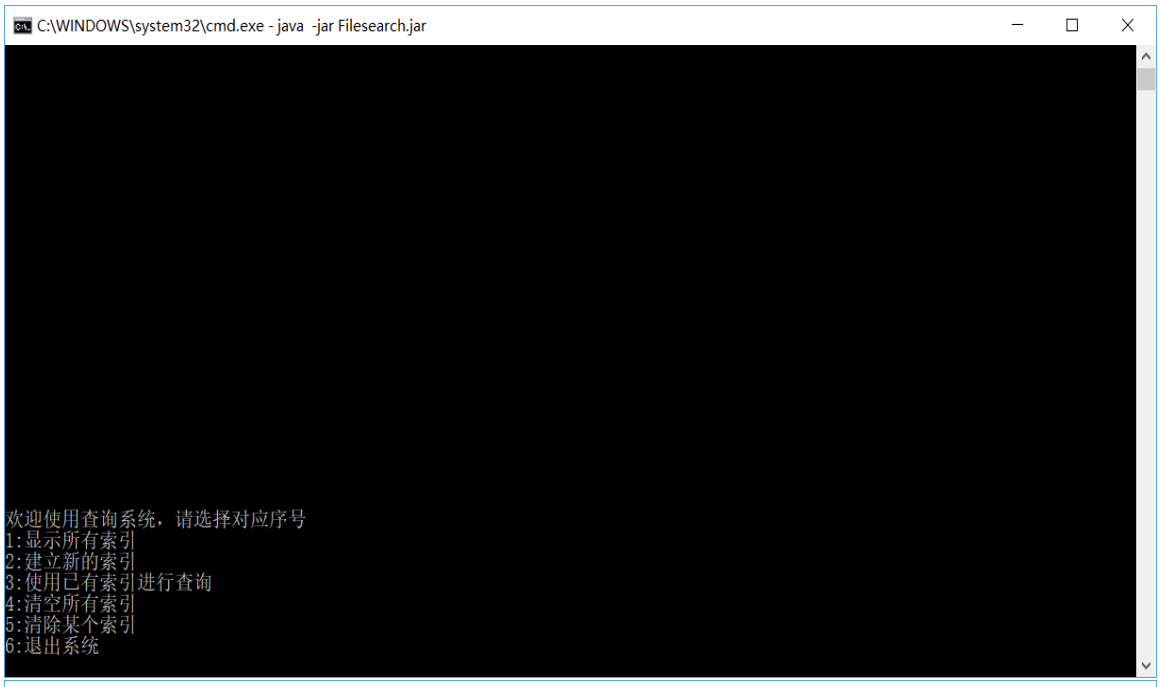
3.6 结果展示

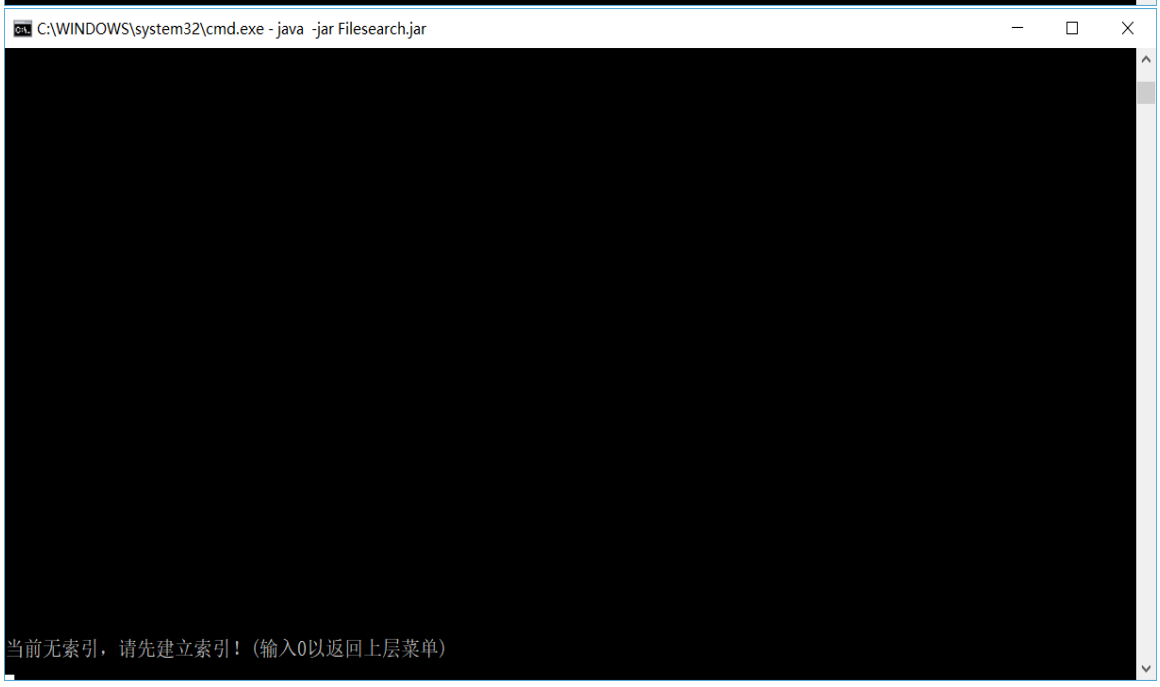
对于所有查询得出的结果，将展现在屏幕上，告知用户文件名及文件路径。本想加入清除控制台显示的功能，结果经过谷歌和百度以后，Java 并没有比较完美的清除控制台显示的功能，因此本实验中粗暴的采取了输入 20 个空行的方式。但这样造成的后果就是显示都在屏幕下方。

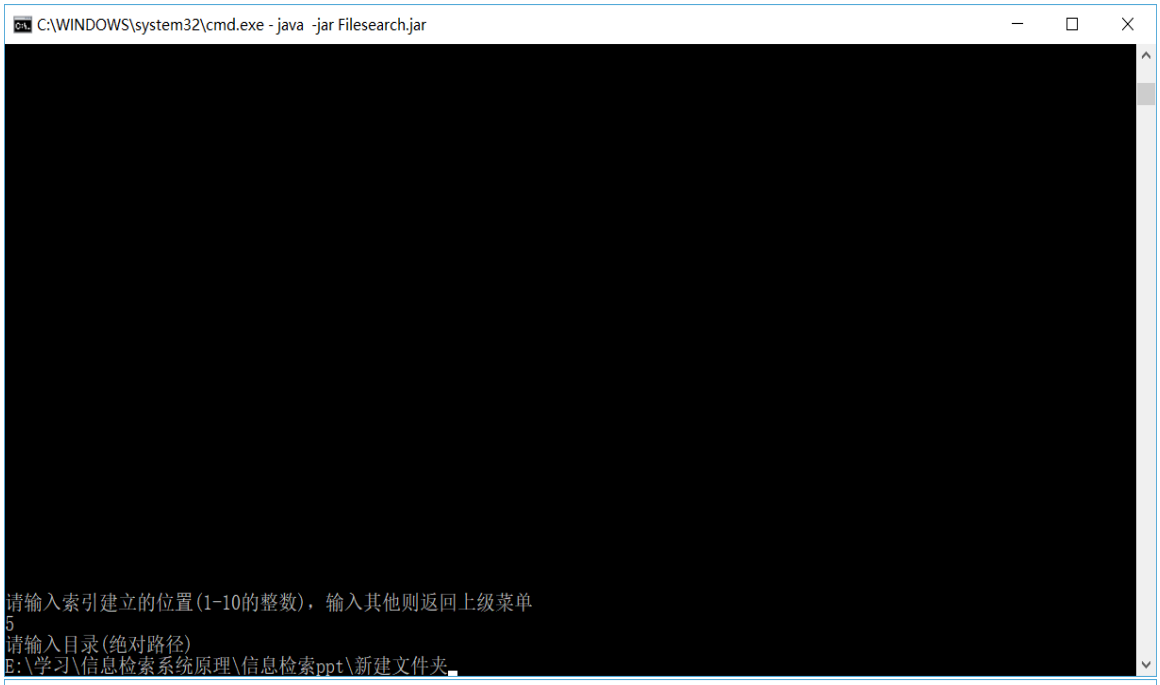
3.7 运行

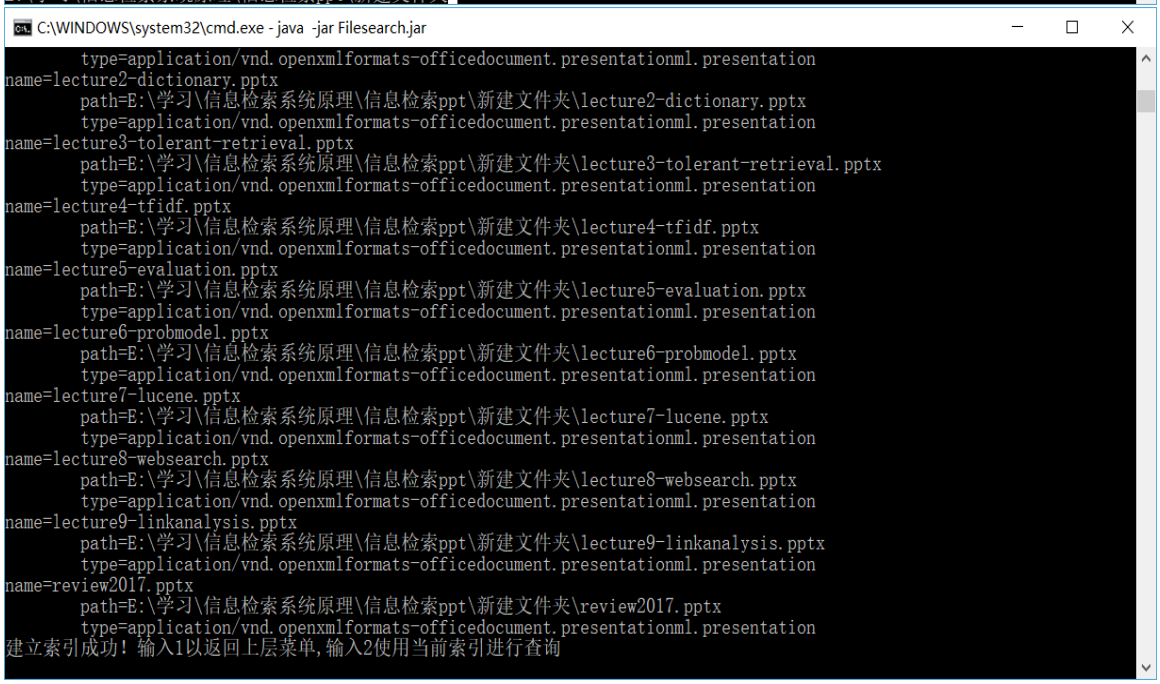
用户可直接输入 java -jar Filesearch.jar 即可运行
经过测试，在 windows 和 linux 下均可以运行

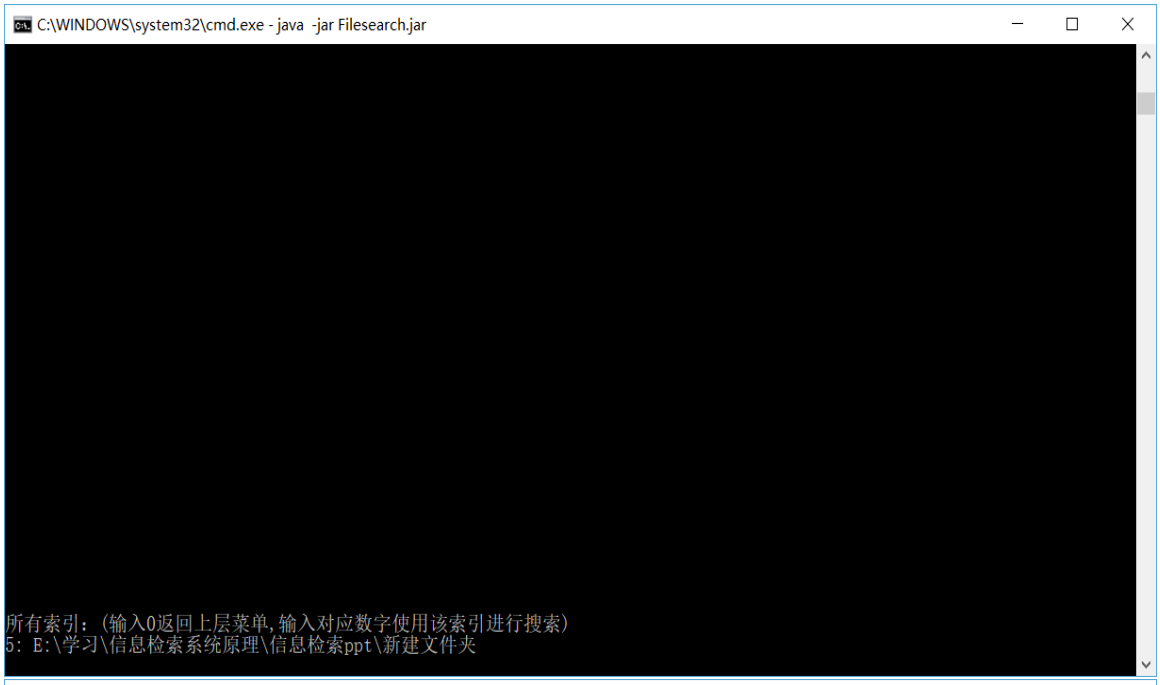
4 效果截图

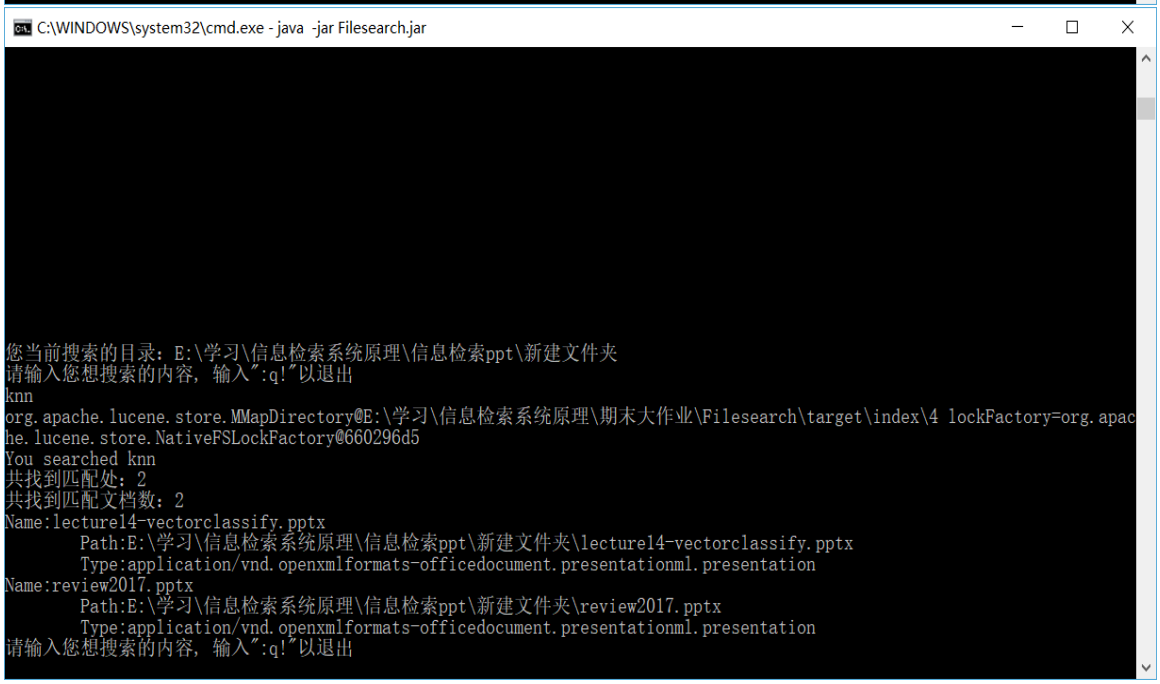


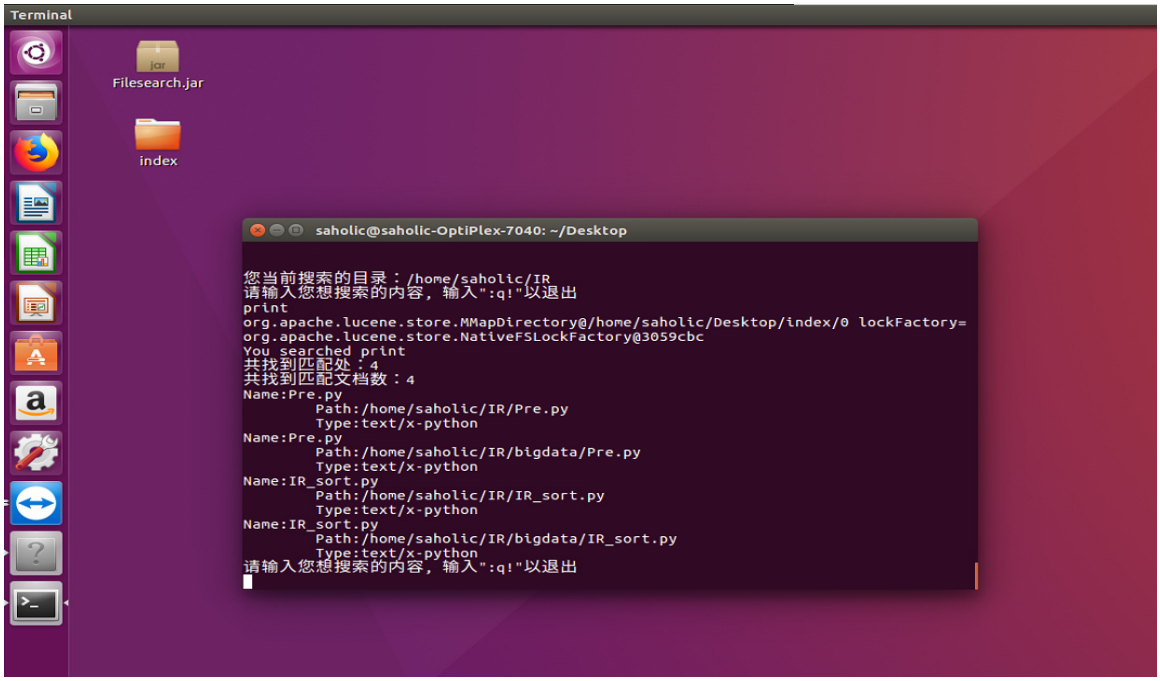



C:\WINDOWS\system32\cmd.exe - java -jar Filesearch.jar
请输入索引建立的位置(1-10的整数)，输入其他则返回上级菜单
5
请输入目录(绝对路径)
E:\学习\信息检索系统原理\信息检索ppt\新建文件夹


C:\WINDOWS\system32\cmd.exe - java -jar Filesearch.jar
type=application/vnd.openxmlformats-officedocument.presentationml.presentation
name=lecture2-dictionary.pptx
path=E:\学习\信息检索系统原理\信息检索ppt\新建文件夹\lecture2-dictionary.pptx
type=application/vnd.openxmlformats-officedocument.presentationml.presentation
name=lecture3-tolerant-retrieval.pptx
path=E:\学习\信息检索系统原理\信息检索ppt\新建文件夹\lecture3-tolerant-retrieval.pptx
type=application/vnd.openxmlformats-officedocument.presentationml.presentation
name=lecture4-tfidf.pptx
path=E:\学习\信息检索系统原理\信息检索ppt\新建文件夹\lecture4-tfidf.pptx
type=application/vnd.openxmlformats-officedocument.presentationml.presentation
name=lecture5-evaluation.pptx
path=E:\学习\信息检索系统原理\信息检索ppt\新建文件夹\lecture5-evaluation.pptx
type=application/vnd.openxmlformats-officedocument.presentationml.presentation
name=lecture6-probmodel.pptx
path=E:\学习\信息检索系统原理\信息检索ppt\新建文件夹\lecture6-probmodel.pptx
type=application/vnd.openxmlformats-officedocument.presentationml.presentation
name=lecture7-lucene.pptx
path=E:\学习\信息检索系统原理\信息检索ppt\新建文件夹\lecture7-lucene.pptx
type=application/vnd.openxmlformats-officedocument.presentationml.presentation
name=lecture8-websearch.pptx
path=E:\学习\信息检索系统原理\信息检索ppt\新建文件夹\lecture8-websearch.pptx
type=application/vnd.openxmlformats-officedocument.presentationml.presentation
name=lecture9-linkanalysis.pptx
path=E:\学习\信息检索系统原理\信息检索ppt\新建文件夹\lecture9-linkanalysis.pptx
type=application/vnd.openxmlformats-officedocument.presentationml.presentation
name=review2017.pptx
path=E:\学习\信息检索系统原理\信息检索ppt\新建文件夹\review2017.pptx
type=application/vnd.openxmlformats-officedocument.presentationml.presentation
建立索引成功！输入1以返回上层菜单,输入2使用当前索引进行查询


C:\WINDOWS\system32\cmd.exe - java -jar Filesearch.jar
所有索引：(输入0返回上层菜单,输入对应数字使用该索引进行搜索)
5: E:\学习\信息检索系统原理\信息检索ppt\新建文件夹


C:\WINDOWS\system32\cmd.exe - java -jar Filesearch.jar
您当前搜索的目录：E:\学习\信息检索系统原理\信息检索ppt\新建文件夹
请输入您想搜索的内容，输入":q!"以退出
knn
org.apache.lucene.store.MMapDirectory@E:\学习\信息检索系统原理\期末大作业\Filesearch\target\index\4 lockFactory=org.apache.lucene.store.NativeFSLockFactory@660296d5
You searched knn
共找到匹配处：2
共找到匹配文档数：2
Name:lecture14-vectorclassify.pptx
Path:E:\学习\信息检索系统原理\信息检索ppt\新建文件夹\lecture14-vectorclassify.pptx
Type:application/vnd.openxmlformats-officedocument.presentationml.presentation
Name:review2017.pptx
Path:E:\学习\信息检索系统原理\信息检索ppt\新建文件夹\review2017.pptx
Type:application/vnd.openxmlformats-officedocument.presentationml.presentation
请输入您想搜索的内容，输入":q!"以退出



更多截图，欢迎运行程序进行体验。