lucene只是全文检索的一个基础知识，现在已经用的很少了

一、什么是全文检索

1、数据的分类

1)结构化数据

格式固定、长度固定、数据类型固定。

例如数据库中的数据

2)非结构化数据

word文档、pdf文档、邮件、html、txt

格式不固定、长度不固定、数据类型不固定。

2、数据的查询

结构化查询语言

1）结构化数据的查询

SQL语句，查询结构化数据的方法。简单、速度快。

2）非结构化数据的查询

从文本文件中找出包含spring单词的文件。

1、目测

2、使用程序吧文档读取到内存中，然后匹配字符串。顺序扫描。

3、把非结构化数据变成结构化数据

先跟根据空格进行字符串拆分，得到一个单词列表，基于单词列表创建一个索引。

然后查询索引，根据单词和文档的对应关系找到文档列表。这个过程叫做全文检索。

索引：一个为了提高查询速度，创建某种数据结构的集合。

3、全文检索

先创建索引然后查询索引的过程叫做全文检索。

索引一次创建可以多次使用。表现为每次查询速度很快。

二、全文检索的应用场景

1、搜索引擎

百度、360搜索、谷歌、搜狗

2、站内搜索

论坛搜索、微博、文章搜索

3、电商搜索

淘宝搜索、京东搜索

4、只要是有搜索的地方就可以使用全文检索技术。

三、什么是Lucene

Lucene是一个基于Java开发全文检索工具包。

四、Lucene实现全文检索的流程

1、创建索引

1）获得文档

原始文档：要基于那些数据来进行搜索，那么这些数据就是原始文档。

搜索引擎：使用爬虫获得原始文档

站内搜索：数据库中的数据。

案例：直接使用io流读取磁盘上的文件。

2）构建文档对象

对应每个原始文档创建一个Document对象

每个document对象中包含多个域（field）

域中保存就是原始文档数据。

域的名称

域的值

每个文档都有一个唯一的编号，就是文档id

3）分析文档

就是分词的过程

1、根据空格进行字符串拆分，得到一个单词列表

2、把单词统一转换成小写。

3、去除标点符号

4、去除停用词

停用词：无意义的词 没有意义的词，像 and the is之类的

每个关键词都封装成一个Term对象中。

Term中包含两部分内容：

关键词所在的域

关键词本身

不同的域中拆分出来的相同的关键词是不同的Term。

4）创建索引

基于关键词列表创建一个索引。保存到索引库中。

索引库中：

索引

document对象

关键词和文档的对应关系

通过词语找文档，这种索引的结构叫倒排索引结构。

2、查询索引

1）用户查询接口

用户输入查询条件的地方

例如：百度的搜索框

2）把关键词封装成一个查询对象

要查询的域 这两个必不可少

要搜索的关键词

3）执行查询

根据要查询的关键词到对应的域上进行搜索。

找到关键词，根据关键词找到 对应的文档

4）渲染结果

根据文档的id找到文档对象

对关键词进行高亮显示

分页处理

最终展示给用户看。

五、入门程序

1、创建索引

环境：

需要下载Lucene
http://lucene.apache.org/
最低要求jdk1.8

工程搭建：

创建一个java工程
添加jar：

lucene-analyzers-common-7.4.0.jar
lucene-core-7.4.0.jar
commons-io.jar

步骤：

1、创建一个Director对象，指定索引库保存的位置。
2、基于Directory对象创建一个IndexWriter对象
3、读取磁盘上的文件，对应每个文件创建一个文档对象。
4、向文档对象中添加域
5、把文档对象写入索引库
6、关闭indexwriter对象

2、使用luke查看索引库中的内容

3、查询索引库

步骤：

1、创建一个Director对象，指定索引库的位置
2、创建一个IndexReader对象
3、创建一个IndexSearcher对象，构造方法中的参数indexReader对象。
4、创建一个Query对象，TermQuery
5、执行查询，得到一个TopDocs对象
6、取查询结果的总记录数
7、取文档列表
8、打印文档中的内容
9、关闭IndexReader对象

六、分析器

所有分析器的最终父类都是Analyzer

默认使用的数标准分析器StandardAnalyzer

1、查看分析器的分析效果

使用Analyzer对象的tokenStream方法返回一个TokenStream对象。词对象中包含了最终分词结果。

实现步骤：

1）创建一个Analyzer对象，StandardAnalyzer对象
2）使用分析器对象的tokenStream方法获得一个TokenStream对象
3）向TokenStream对象中设置一个引用，相当于数一个指针
4）调用TokenStream对象的rest方法。如果不调用抛异常
5）使用while循环遍历TokenStream对象
6）关闭TokenStream对象

2、IKAnalyze的使用方法

1）把IKAnalyzer的jar包添加到工程中

2）把配置文件和扩展词典添加到工程的classpath下

注意：扩展词典严禁使用windows记事本编辑保证扩展词典的编码格式是utf-8

扩展词典：添加一些新词

停用词词典：无意义的词或者是敏感词汇

七、索引库维护

1、添加文档

2、删除文档

1）删除全部

2）根据查询、关键词删除文档

indexWriter：操作索引库
indexReader：读取索引库
indexSearcher：查询索引库

3、修改文档

修改的原理是先删除后添加

八、索引库查询

1、使用Query的子类

1）TermQuery 直接new一个TermQuery对象，里面穿第一个Term对象即可

根据关键词进行查询。

需要指定要查询的域及要查询的关键词

2）RangeQuery 需要使用LongPoint类的静态方法newRangeQuery获取

范围查询 长整型使用LongPoint创建，短整型使用intPoint创建

2、使用QueryPaser进行查询

可以对要查询的内容先分词，然后基于分词的结果进行查询。

添加一个jar包

lucene-queryparser-7.4.0.jar