Web 信息处理与应用实验一程序说明——邢宇（PB15000195）

简单的搜索引擎实现

【实验内容】

用 java 实现简单的搜素引擎

【实验环境】

编程语言：java
编程环境：Eclipse j2EE
运行环境：windows10 tomcat7.0
使用工具：tomcat、lucene 、jsoup

【实验步骤及方法】

写出实验的主要步骤及实现方法，给出关键部分的代码实现，参考实验要求中需要实现的几点内容。

主要步骤及实现方法

1、用 jsoup 实现文本预处理

使用 getElementByXXX 等接口函数提取相应标签

2、用 github 上 NLPIR 自带的 lucene 工具包导入，其中 NLPIR 有派生类代码，可以不用自己手动实现。

3、用 indexWriter 将每一篇 doc 的关键字条目组织成字典写入索引文件

4、实现获取索引字符串即可返回结果的接口函数

5、配置 tomcat 运行环境

6、建立 dynamic web project，导入之前的 java 工程，编写 jsp 代码，调用上述 package，前端输入索引字符串

7、获取返回值并在页面 out.Println

8、添加其他一些功能

关键接口函数：

```java
public Info Parse(Element doc) throws Exception{
        Info docInfo = new Info();
        String url = doc.getElementsByTag("url").get(0).ownText();
        String title = doc.getElementsByTag("title").get(0).ownText();
        String content = doc.ownText();
        //注意publishid有可能为空
        String publishid = "";
        try{
        publishid = doc.getElementsByAttributeValue("name","publishid").get(0).attr("content"); //publishid
        }catch(Exception e){
            ;
        }
        String subjectid = "";
        try{
            subjectid = doc.getElementsByAttributeValue("name", "subjectid").get(0).attr("content");
        }catch(Exception e){
            ;
        }

        //注意keywords有可能为空
        String keywords = "";
        try{
            keywords = doc.getElementsByAttributeValue("name","keywords").get(0).attr("content");
        }catch(Exception e){
```

<页面提取索引>

```java
public buildIndex() {

    analyzer = new NLPIRTokenizerAnalyzer("", 1, "", "", false);
    IndexWriterConfig indexWriterConfig = new IndexWriterConfig(analyzer);
    indexWriterConfig.setOpenMode(OpenMode.CREATE_OR_APPEND);
    try{
        File file = new File(PATH_INDEX);
        directory = FSDirectory.open(FileSystems.getDefault().getPath(".",PATH_INDEX))
        System.out.println(FileSystems.getDefault().getPath(PATH_INDEX));
        indexWriter = new IndexWriter(directory, indexWriterConfig);
    }
    catch(IOException e){
        ;
    }
}

public Boolean IndexWriterClose(){
    try{
        indexWriter.close();
        return true;
    }
    catch(IOException e){
        return false;
    }
}

public boolean Build(Info info) throws Exception{
    if(indexWriter == null) return false;
    Document document = new Document();
    document.add(new StringField("url", info.url, Store.YES));
    document.add(new StringField("publishid", info.publishid, Store.YES));
    document.add(new StringField("subjectid", info.subjectid, Store.YES));
    document.add(new StringField("description", info.description, Store.YES));
    document.add(new StringField("keywords", info.keywords, Store.YES));
    document.add(new StringField("title", info.title, Store.YES));
    document.add(new TextField("content", info.content,Store.YES));
    System.out.println(document);
```

<建立索引>

```java
public static void IndexBuilding() throws Exception{
    HtmlParser htmlParser = new HtmlParser();
    buildIndex BuildIndex = new buildIndex();
    String[] materials = {"material/2012.q1.txt","mat/2012.q2.txt","mat/2012.q3.txt","mat/2012.q4.txt","mat/2013.q1.txt"};
    System.out.println("IndexBuilding...\n");
    for(int i = 0; i < 5; ++i){
        System.out.println("Processing  "+materials[i]);
        File newsfile = new File(materials[i]);
        Document document = Jsoup.parse(newsfile,"utf-8");
        Elements docs = document.getElementsByTag("doc");
        int counter = 0;
        float whole_num = htmlParser.GetDocNum(docs);
        for(Element doc:docs){
            counter+=1;
            Info docinfo = htmlParser.Parse(doc);
            BuildIndex.Build(docinfo);
            if(counter % 100 == 0)
                System.out.println("[%" + counter / whole_num + "]");
        }
    }
}
```

```
Console  Tasks  Display  Debug Output  Browser Output
<terminated> main_prog (2) [Java Application] F:\jdk\jre1.8.0_121\bin\javaw.exe (2017年12月31日 下午9:00:31)
index
IndexBuilding...

Processing  material/2012.q1.txt
[%0.00231954]
[%0.00463908]
[%0.006958619]
[%0.00927816]
[%0.011597699]
[%0.013917238]
[%0.016236778]
[%0.01855632]
[%0.020875858]
[%0.023195397]
```

<主函数生成索引文件>

```java
package main_prog;

import org.apache.lucene.document.Document;


public class TestSearch {
    public static void main(String[] args)throws Exception
    {
        IndexSearch indexSearch = new IndexSearch();

        indexSearch.Search("朝鲜");
        int num = indexSearch.GetResultNum();
        System.out.println(num);
        for(int i = 0; i < (num > 20 ? 20 : num); ++i){
            Document document = indexSearch.GetResults(i);
            System.out.println(document.get("title"));
            System.out.println(document.get("url"));
        }
    }
}
```

```
Console  Tasks  Display  Debug Output  Browser Output
<terminated> TestSearch (1) [Java Application] F:\jdk\jre1.8.0_121\bin\javaw.exe (2017年12月31日 下午9:06:06)
5506
朝鲜人民军称誓死保卫金正恩|朝鲜_新浪新闻
http://go.rss.sina.com.cn/redirect.php?url=http://news.sina.com.cn/w/2012-06-15/164124599046.shtml
金正恩电贺朝总联支部干部大会 向爱国同胞致敬|朝鲜_新浪新闻
http://go.rss.sina.com.cn/redirect.php?url=http://news.sina.com.cn/w/2012-07-21/080624817171.shtml
朝鲜宣布全面废除朝韩间所有互不侵犯协议|朝鲜_新浪新闻
http://go.rss.sina.com.cn/redirect.php?url=http://news.sina.com.cn/w/2013-03-08/101326469006.shtml
朝鲜官方向民众推荐发型：女性18款男性10款|朝鲜_新浪新闻
http://go.rss.sina.com.cn/redirect.php?url=http://news.sina.com.cn/w/2013-02-22/185326332665.shtml
朝鲜军方宣布应对美韩敌朝政策三大举措|朝鲜|三大举措_新浪新闻
http://go.rss.sina.com.cn/redirect.php?url=http://news.sina.com.cn/w/2013-03-05/212526437262.shtml
分析称朝卫星发射成功与否不会影响金正恩地位_新闻中心_新浪网
http://go.rss.sina.com.cn/redirect.php?url=http://news.sina.com.cn/w/2012-04-14/060124270951.shtml
朝鲜宣布将不承认朝鲜停战协定|朝鲜|高强度对应措施_新浪新闻
http://go.rss.sina.com.cn/redirect.php?url=http://news.sina.com.cn/w/2013-03-05/191626436641.shtml
美国5名参议员批评政府屡次接受朝鲜空洞承诺_新闻中心_新浪网
http://go.rss.sina.com.cn/redirect.php?url=http://news.sina.com.cn/w/2012-03-18/061224132515.shtml
```

<测试输出结果>

<搜索框>

```
int pageNum;
try{
    int return_num = indexSearch.GetResultNum() > 65 ? 65 :indexSearch.GetResultNum();
    pageNum = return_num/13;
}
catch(Exception e)
{
    pageNum = 0;
}
int currentPage = (request.getParameter("p") != null)?Integer.parseInt(request.getParameter("p")):0;
if(request.getParameter("p") != null)
{
    String requestStr = request.getRequestURL() + "?" +request.getQueryString();
    String[] requestStrArr = requestStr.split("&");
    String newRequestStr = "";
    for(int i = 0;i<requestStrArr.length - 2;i++)
    {
        newRequestStr += requestStrArr[i];
        newRequestStr += "&";
    }
    for(int i = currentPage -5;i<currentPage+5 && i<pageNum;i++)
    {
        if(i >= 0&&i!=currentPage)
        {
            out.println("<div style=\"border-style:dotted;display:inline\" >" + "<a  href=\""+ newRequestStr + "p=" + Integer.toString(i)
                + "&c=1\">" + "    " +  Integer.toString(i + 1) + "</a>" + "</div>");
        }
        if(i==currentPage)
        {
```

【实验结果说明及演示】

搜索框

网页正文

【实验总结】

不足：页面布局写的太差，没能完全实现下拉框和最近历史。据说需要使用 cookie。