北在北方
太白枝头看，花开不计年，杯中浮日月，楼外是青天。

posts - 160, comments - 209, trackbacks - 0, articles - 0

导航
博客园
首页
新随笔
联系
订阅
管理

公告

昵
称：CN.programmer.Luxh
园龄：1年8个月
粉丝：106
关注：8
+加关注

< 2014年1月 >

日	一	二	三	四	五	六
29	30	31	1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	**24**	25
26	27	28	29	30	31	1
2	3	4	5	6	7	8

搜索
找找看
谷歌搜索

常用链接
我的随笔
我的评论
我的参与
最新评论
我的标签

随笔分类
Android(6)
Axure RP(7)
FTP(1)
Hadoop(19)
JavaEE(9)
JavaSE(11)
JBPM(6)
JDBC Pool(1)
JPA(18)
jQuery(6)
Linux(20)
Lucene(7)
Maven(7)
MongoDB(5)
MySQL(5)
Oracle(8)
Permissions(1)
Spring(5)
Spring Data JPA(1)
Spring Security(2)
SpringMVC(4)
Struts2(7)
Terracotta(2)
Web Services(3)
负载均衡/集群(3)
框架整合(5)
图表绘制(2)
一些问题记录(7)
应用服务器(5)

随笔档案
2014年1月 (1)
2013年12月 (6)
2013年11月 (7)
2013年10月 (1)
2013年9月 (5)
2013年7月 (4)
2013年6月 (3)

Lucene的中文分词器IKAnalyzer

Posted on 2012-06-23 13:55 CN.programmer.Luxh 阅读(6689) 评论(12) 编辑 收藏

分词器对英文的支持是非常好的。

一般分词经过的流程：

1）切分关键词

2）去除停用词

3）把英文单词转为小写

但是老外写的分词器对中文分词一般都是单字分词，分词的效果不好。

国人林良益写的IK Analyzer应该是最好的Lucene中文分词器之一，而且随着Lucene的版本更新而不断更新，目前已更新到IK Analyzer 2012版本。

IK Analyzer是一个开源的，基于java语言开发的轻量级的中文分词工具包。到现在，IK发展为面向Java的公用分词组件，独立于Lucene项目，同时提供了对Lucene的默认优化实现。在2012版本中，IK实现了简单的分词歧义排除算法，标志着IK分词器从单纯的词典分词向模拟语义分词衍化。

在系统环境：Core2 i7 3.4G双核，4G内存，window 7 64位， Sun JDK 1.6_29 64位 普通pc环境测试，IK2012具有160万字/秒（3000KB/S）的高速处理能力。

特别的，在2012版本，词典支持中文，英文，数字混合词语。

IK Analyzer 2012版本的分词效果示例：

IK Analyzer2012版本支持 细粒度切分 和 智能切分。

我们看两个演示样例：

1）文本原文1：

IKAnalyzer是一个开源的，基于java语言开发的轻量级的中文分词工具包。从2006年12月推出1.0版本开始，IKAnalyzer已经推出了3个大版本。

智能分词结果：

ikanalyzer | 是 | 一个 | 开源 | 的 | 基于 | java | 语言 | 开发 | 的 | 轻量级 | 的 | 中文 | 分词 | 工具包 | 从 | 2006年 | 12月 | 推出 | 1.0版 | 开始 | ikanalyzer | 已经 | 推 | 出了 | 3个 | 大 | 版本

最细粒度分词结果：

ikanalyzer | 是 | 一个 | 一 | 个 | 开源 | 的 | 基于 | java | 语言 | 开发 | 的 | 轻量级| 量级 | 的 | 中文 | 分词 | 工具包 | 工具 | 包 | 从 | 2006 | 年 | 12 | 月 | 推出 | 1.0 | 版 | 开始 | ikanalyzer | 已经 | 推出 | 出了 | 3 | 个 | 大 | 版本

2）文本原文2：

张三说的确实在理。

智能分词结果：

张三 | 说的 | 确实 | 在理

最细粒度分词结果：

张三 | 三 | 说的 | 的确 | 的 | 确实 | 实在 | 在理

IKAnalyzer的使用

1）下载地址：

GoogleCode开源项目：http://code.google.com/p/ik-analyzer/

GoogleCode下载地址：http://code.google.com/p/ik-analyzer/downloads/list

2）兼容性：

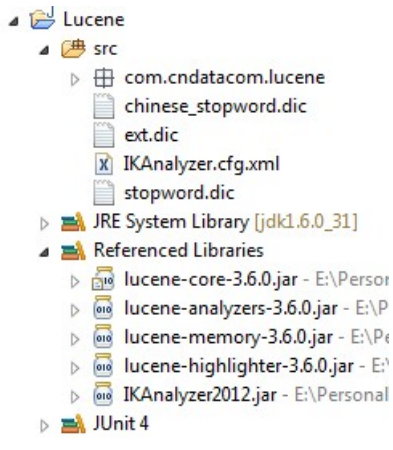
IKAnalyzer 2012版本兼容Lucene3.3以上版本。

3）安装部署：

十分简单，只需要将IKAnalyzer2012.jar引入项目中就可以了。对于"的"、"了"、"着"之类的停用词，它有一个词典stopword.dic。把stopword.dic和IKAnalyzer.cfg.xml复制到class根目录就可以启用停用词功能和扩展自己的词典。

4）测试例子：

新建一个Java Project，引入Lucene所需的jar文件和IKAnalyzer2012.jar文件，把stopword.dic和IKAnalyzer.cfg.xml复制到class根目录，建立一个扩展词典ext.dic和中文停用词词典chinese_stopword.dic。



IKAnalyzer2012发布包自带的stopword.dic里面存的是英文的停用词。所以我们新建一个chinese_stopword.dic，用来存放中文停用词。chinese_stopword.dic需要使用UTF-8编码。词典中，每个中文词汇独占一行。

chinese_stopword.dic内容格式：

2013年5月 (3)
2013年4月 (10)
2013年3月 (5)
2013年1月 (8)
2012年12月 (4)
2012年11月 (26)
2012年10月 (3)
2012年9月 (6)
2012年8月 (7)
2012年7月 (16)
2012年6月 (25)
2012年5月 (20)

最新评论

1. Re:Servlet3.0-文件上传

//把文件写到指定路径

part.write(storePath+File.separator+fileName);
难道就不能写到一个指定路径吗？为什么要弄这么复杂。。比如

part.write("D:/kkk/kk.txt");

--沙漠之狐ph

2. Re:Spring Data JPA初使用

谢谢你这么详细的分享，对我帮助很大

--delphi日记

3. Re:Lucene索引库的简单优化

@教父右手
正常的，这些都是索引文件的格式。

--CN.programmer.Luxh

4. Re:Struts2多文件下载

@-六月飞雪-
这种一般适合文档资料下载。不适合很多的大文件。

--CN.programmer.Luxh

5. Re:Lucene索引库的简单优化

楼主好，我在搞lucene时候发现创建的索引文件 很多。每次存一个对象时候就对他的部分字段创建索引IndexWriterConfig.OpenMode.APPEND 增量模式。然后索引文件夹下就会创建一组文件都是1KB，这样正常吗？500)this.width=500;"/>

--教父右手

阅读排行榜

1. Spring Data JPA初使用(15867)
2. jQuery操作<input type="radio">(8669)
3. Lucene的中文分词器IKAnalyzer(6689)
4. JPA的persistence.xml文件(6000)
5. Servlet3.0-使用注解定义Servlet(5969)

评论排行榜

1. Struts2多文件下载(20)
2. JPA的多对多映射(15)
3. Lucene的中文分词器IKAnalyzer(12)
4. Spring3+Struts2+JPA2.0(10)
5. Web项目中使用Spring整合CXF发布Web Services(10)

推荐排行榜

1. Lucene的中文分词器IKAnalyzer(4)
2. 配置Tomcat数据源(4)
3. Struts2多文件下载(4)
4. 使用spring的邮件发送功能(3)
5. jQuery操作<input

chinese_stopword.dic

```
是
啊
的
了
年

```

IKAnalyzer.cfg.xml：

```xml
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE properties SYSTEM "http://java.sun.com/dtd/properties.dtd">
<properties>
    <comment>IK Analyzer 扩展配置</comment>
    <!--用户可以在这里配置自己的扩展字典 -->
    <entry key="ext_dict">ext.dic;</entry>

    <!--用户可以在这里配置自己的扩展停止词字典-->
    <entry key="ext_stopwords">stopword.dic;chinese_stopword.dic</entry>

</properties>
```

可以配置多个词典文件，文件使用";"号分隔。文件路径为相对java包的起始根路径。

扩展词典ext.dic需要为UTF-8编码。

ext.dic内容：

ext.dic

```
2012
欧洲杯四强赛

```

我把"2012"作为一个词，"欧洲杯四强赛"作为一个词。

测试分词代码：

```java
package com.cndatacom.lucene.test;

import java.io.StringReader;

import org.apache.lucene.analysis.Analyzer;
import org.apache.lucene.analysis.TokenStream;
import org.apache.lucene.analysis.tokenattributes.CharTermAttribute;
import org.junit.Test;
import org.wltea.analyzer.lucene.IKAnalyzer;


/**
 * IKAnalyzer 分词器测试
 * @author Luxh
 */
public class IKAnalyzerTest {

    @Test
    public void testIKAnalyzer() throws Exception {

        String keyWord = "2012年欧洲杯四强赛";

        IKAnalyzer analyzer = new IKAnalyzer();

        //使用智能分词
        analyzer.setUseSmart(true);

        //打印分词结果
        printAnalysisResult(analyzer,keyWord);

    }

    /**
     * 打印出给定分词器的分词结果
     * @param analyzer 分词器
     * @param keyWord 关键词
     * @throws Exception
     */
    private void printAnalysisResult(Analyzer analyzer, String keyWord) throws Exception {
        System.out.println("当前使用的分词器：" + analyzer.getClass().getSimpleName());
        TokenStream tokenStream = analyzer.tokenStream("content", new StringReader(keyWord));
        tokenStream.addAttribute(CharTermAttribute.class);
        while (tokenStream.incrementToken()) {
            CharTermAttribute charTermAttribute = tokenStream.getAttribute(CharTermAttribute.class);
            System.out.println(new String(charTermAttribute.buffer()));
        }
    }
}
```

打印出来的分词结果：

type="checkbox">(3)

```
当前使用的分词器：IKAnalyzer
加载扩展词典：ext.dic
加载扩展停止词典：stopword.dic
加载扩展停止词典：chinese_stopword.dic
2012□□□□□□□□□□
欧洲杯四强赛□□□□□□□□
```

可以看到”2012“作为一个词，“欧洲杯四强赛”也是作为一个词，停用词”年“已被过滤掉。

分类：Lucene

绿色通道： 好文要顶 关注我 收藏该文 与我联系

CN.programmer.Luxh
关注 - 8
粉丝 - 106
+加关注

4 0

(请您对文章做出评价)

« 上一篇：Lucene索引库的简单优化
» 下一篇：Lucene的高亮器Highlighter

Feedback

#1楼 回复 引用
2012-07-13 15:13 by Judas.n

这里很多跟我一样的新手没办法成功的原因就是被无bom的UTF-8格式给折磨的...这个一定要做好...
IK作者自己也这样说了:3.如果你不知道啥叫无BOM，也不确定自己的文件是不是UTF-8无bom，那么请在第一行使用回车换行，从第二行开始添加停止词

支持(0) 反对(0)

#2楼 回复 引用
2012-07-13 15:47 by Judas.n

作者我能跟你交流一下lucene...本来才刚刚起步还有很多东西要学习呢...jn3.141592654@163.com

支持(0) 反对(0)

#3楼[楼主] 回复 引用
2012-07-13 22:06 by programmer_luxh

@Judas.n
我也是刚学，一起学习交流。

支持(0) 反对(0)

#4楼 回复 引用
2013-03-15 13:27 by 易志娃

博主你好，我测试的不稳定，能否给解释一下，非常感谢
需要拆分的字符串：生地黄水蜜丸
默认分词显示为：生地黄|水|蜜|丸
在ext.dic文件配置如下：
地黄
水蜜丸
配置后分词显示为：生地黄|水蜜丸

问题：“水蜜丸”被成功拆分，“生地黄”为什么没有被拆分成“生”和“地黄”？

支持(0) 反对(0)

#5楼[楼主] 回复 引用
2013-03-15 17:58 by CN.programmer.Luxh

@易志娃
我猜想应该是正向最大匹配的分词方法，生地黄 这个词在主词典已经存在。

支持(0) 反对(0)

#6楼 回复 引用
2013-03-15 18:06 by 易志娃

@CN.programmer.Luxh
即使主字典中已经包含，但应该是以我们配置的覆盖原有的，不然配置就没有意义了

支持(0) 反对(0)

#7楼[楼主] 回复 引用
2013-03-15 19:09 by CN.programmer.Luxh

@易志娃
配置的词典是扩展，不是覆盖。

支持(0) 反对(0)

#8楼 回复 引用
2013-07-16 20:19 by comeonniu

@CN.programmer.Luxh
那为什么不干脆直接把自己配置的字典写进main2012自带字典中呢？

支持(0) 反对(0)

#9楼 回复 引用
2013-07-16 20:30 by comeonniu

main2012中有公安、分局、公安分局的字典项，如果查询的是公安分局，结果是公安|分局|，可我希望是公安分局，该怎么做？

支持(0) 反对(0)

#10楼[楼主] 回复 引用
2013-07-16 22:17 by CN.programmer.Luxh

@comeonniu
main2012中有公安、分局、公安分局的字典项,如果查询的是公安分局，结果就是公安分局，不会是公安|分局|

支持(0) 反对(0)

#11楼[楼主] 回复 引用
2013-07-16 22:24 by CN.programmer.Luxh

@comeonniu
引用
@CN.programmer.Luxh
那为什么不干脆直接把自己配置的字典写进main2012自带字典中呢？

你可以这样做，但是修改了分词器自带的词典，你要重新打包，每添加一次，都要这样做，你觉得方便吗。

支持(0) 反对(0)

#12楼 回复 引用
2013-07-20 02:15 by JavaSmart

好强大

支持(0) 反对(0)

刷新评论 刷新页面 返回顶部

发表评论

昵称: thinkbase

评论内容:

提交评论 注销 订阅评论

[使用Ctrl+Enter键快速提交]

程序员问答平台，解决您的技术难题

博客园首页 博问 新闻 闪存 程序员招聘 知识库

最新**IT**新闻:
· 阿里巴巴投资中信21世纪的逻辑
· Tizen终于来了：三星远离Android的希望
· 谷歌高层在邮件中向网友确认：支付服务中支持比特币
· 亚马逊Web服务发布2013年推荐技术内容列表
· Mac 版微信客户端开始内测，只保留 IM 功能
» 更多新闻...

最新知识库文章:
· 深入剖析阿里巴巴云梯YARN集群
· 关于技术团队管理的胡言乱语
· 浅谈TCP优化
· SSL与TLS的区别以及介绍
· DDD & DDDLib在恒拓开源的发展历程与推广经验
» 更多知识库文章...

Powered by:
博客园
Copyright © CN.programmer.Luxh