



NLPIR 大数据搜索与挖掘共享开发平台

NLPIR Big Data Search and Mining Development Platform

用户手册

Manual



April 16, 2013

For the latest information about NLPIR, please visit [Http://ICTCLAS.nlpir.org/](http://ICTCLAS.nlpir.org/)

Document Information

Document ID	NLPIR-ICTCLAS-2013-WHITEPAPE R	Version	V3.0
Security level	Public 公开	Status	Creation and first draft for comment
Author	张华平	Date	Jan 5, 2012
Publisher	/	Approved by	

Version History

Note: The first version is "v0.1". Each subsequent version will add 0.1 to the exiting version. The version number should be updated only when there are significant changes, for example, changes made to reflect reviews. The first figure in the version 1.x denotes current review status by. 1. x denotes review process has passed round 1 etc .Anyone who create, review or modify the document should describe his action.

Version	Author/Reviewer	Date	Description
V1.0	Kevin Zhang	2011-8-21	first complete draft for comment. ICTCLAS2010
V2.0	Kevin Zhang	2012-8-21	complete draft for comment.ICTCLAS2012
V3.0	Kevin Zhang	2012-12-19	complete draft for comment.ICTCLAS2013

目 录

一、NLPIR 大数据搜索与挖掘共享开发平台简介.....	4
二、NLPIR 开发平台可视化软件操作指南.....	5
2.1: 全文精准检索.....	5
2.2: 新词发现.....	6
2.3: 导入用户词典, 对语料进行分词及词性标注.....	6
2.4: 词频统计及翻译.....	8
2.5: 文本聚类及热点内容分析.....	11
2.6: 分类过滤.....	11
2.7: 文本摘要与关键词提取.....	13
2.8: 文档去重.....	14
2.9: HTML 正文解析.....	15
2.10: 正负面分析.....	15
2.11: 编码转换.....	18
三、作者简介.....	18

一、NLPIR 大数据搜索与挖掘共享开发平台简介

NLPIR 内容搜索与挖掘开发平台针对互联网内容处理的需要，融合了自然语言理解、网络搜索和文本挖掘的技术，提供了用于技术二次开发的基础工具集。开发平台由多个中间件组成，各个中间件 API 可以无缝地融合到客户的各类复杂应用系统之中，可兼容 Windows, Linux, Android, Maemo5, FreeBSD 等不同操作系统平台，可以供 Java, C, C# 等各类开发语言使用。

NLPIR 是一套专门针对原始文本集进行处理和加工的软件，提供了中间件处理效果的可视化展示，也可以作为小规模数据的处理加工工具。用户可以使用该软件对自己的数据进行处理。

NLPIR 内容搜索与挖掘开发平台的十大功能：

■ 1. 全文精准检索

支持文本、数字、日期、字符串等各种数据类型，多字段的高效搜索，支持 AND/OR/NOT 以及 NEAR 邻近等查询语法，支持维语、藏语、蒙语、阿拉伯、韩语等多种少数民族语言的检索。可以无缝地与现有文本处理系统与数据库系统融合。

■ 2. 新词发现：

从文件集合中挖掘出内涵的新词语列表，可以用于用户专业词典的编撰；还可以进一步编辑标注，导入分词词典中，从而提高分词系统的准确度，并适应新的语言变化。

■ 3. 分词标注：

对原始语料进行分词、自动识别人名地名机构名等未登录词、新词标注以及词性标注。并可在分析过程中，导入用户定义的词典。

■ 4. 统计分析与术语翻译

针对切分标注结果，系统可以自动地进行一元词频统计、二元词语转移概率统计（统计两个词左右连接的频次即概率）。针对常用的术语，会自动给出相应的英文解释。

■ 5. 文本聚类及热点分析

能够从大规模数据中自动分析出热点事件，并提供事件话题的关键特征描述。同时适用于长文本和短信、微博等短文本的热点分析。

■ 6. 分类过滤

针对事先指定的规则和示例样本，系统自动从海量文档中筛选出符合需求的样本。

■ 7. 自动摘要

能够对单篇或多篇文章，自动提炼出内容的精华，方便用户快速浏览文本内容。

■ 8. 关键词提取

能够对单篇文章或文章集合，提取出若干个代表文章中心思想的词汇或短语，可用于精化阅读、语义查询和快速匹配等。

■ 9. 文档去重

能够快速准确地判断文件集合或数据库中是否存在相同或相似内容的记录，同时找出所有的重复记录。

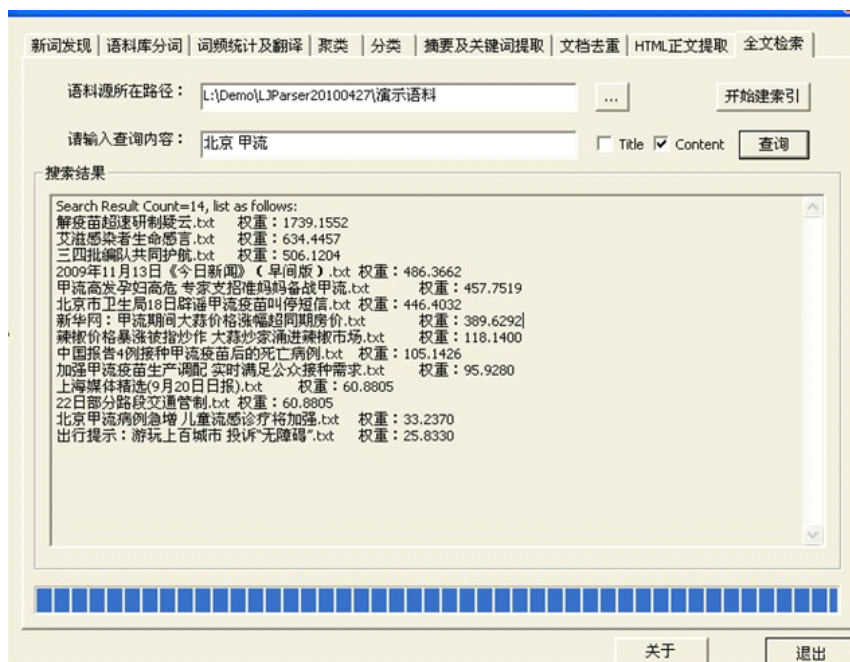
■ 10. HTML 正文提取

自动剔除导航性质的网页，剔除网页中的 HTML 标签和导航、广告等干扰性文字，返回有价值的正文内容。适用于大规模互联网信息的预处理和分析。

二、NLPIR 开发平台可视化软件操作指南

按照功能依次介绍如下：

2.1：全文精准检索



选择语料文件夹，点击“开始建索引”按钮，系统对语料快速建立压缩索引；输入查询关键词，点击查询按钮，系统返回查询结果，并配以权重。

全文精准搜索的特色在于：

1、支持无词典索引，支持搜索维语、藏语、蒙语、阿拉伯、韩语等多种少数民族语言；

当前的搜索大部分都需要内置一部核心词库，而维语、藏语、蒙语、阿拉伯、韩语等多种少数民族语言往往缺乏相关的电子资源，整理一部词典往往费时费力。JZSearch 全文精准搜索引擎支持词典与无词典两种模式，无词典时，采用 N-Gram 模型，同样可以构建高速的索引与搜索。

2、支持文本、数字、日期、字符串等各种数据类型，多字段的高效搜索；

3、内置多种检索模型，支持多种排序策略，包括相关度、时序等；

4、全文索引压缩比约为 1/4，大大减少了索引的开销，提高了所有效率；

5、支持丰富的查询语法，支持与、或、非以及邻近运算；

支持的典型查询语法包括：

Sample1: [FIELD] title [AND] 解放军

Sample2: [FIELD] title [AND] 解放军某部发生数百人感染甲流疫情

Sample3: [FIELD] content [AND] 甲型 H1N1 流感

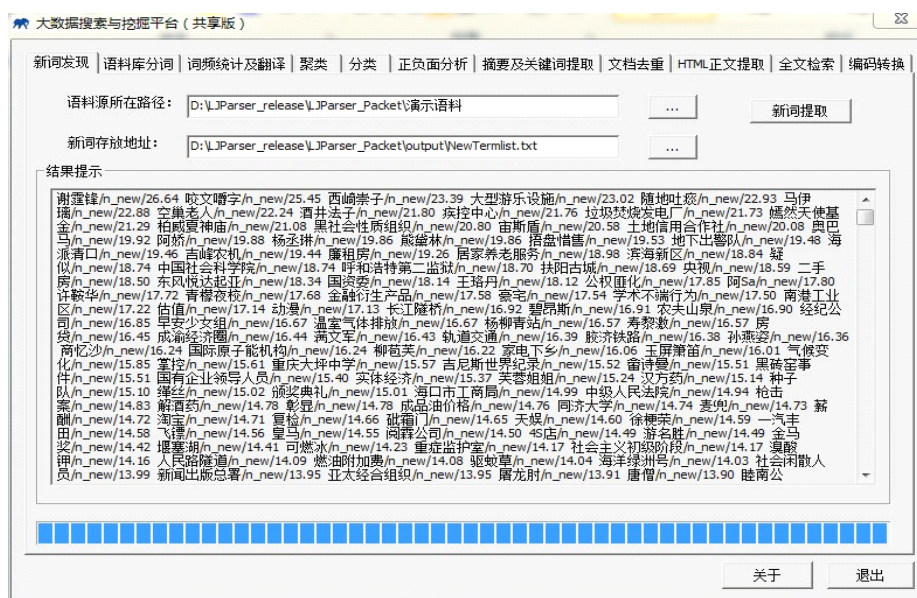
Sample4: [FIELD] content [NEAR] 张雁灵 解放军

Sample5: [FIELD] content [OR] 解放军 甲流

Sample6: [FIELD] title [AND] 解放军 [FIELD] content [NOT] 甲流

6、可扩展性强：支持数据库的全文搜索，以及 word, ppt, pdf, email 等各种文档格式的搜索；可以便利地构建各类网络搜索引擎服务。

2.2: 新词发现



1) 在“语料源所在路径”输入框中输入需要提取新词的语料所在路径，语料须以 txt 文件的方式存储在输入的语料源目录下。

2) 如果“语料源所在路径”是通过选择文件夹方式确定，则系统会缺省指定“新词存放地址”为当前工作目录\output\NewTermlist.txt；如果“语料源所在路径”是由手动输入，则需要指定输出的“新词存放地址”。

3) 点击“新词提取”按钮，系统开始进行发现新词的过程。结果输出到“新词存放地址”所指定的文件，另外也会输出到结果提示框中。

本步骤所得到的新词，可以作为分词标注器的用户词典导入，从而使分词结果更加准确。对于不需要导入新词的用户，本步骤可以跳过。

2.3: 导入用户词典，对语料进行分词及词性标注

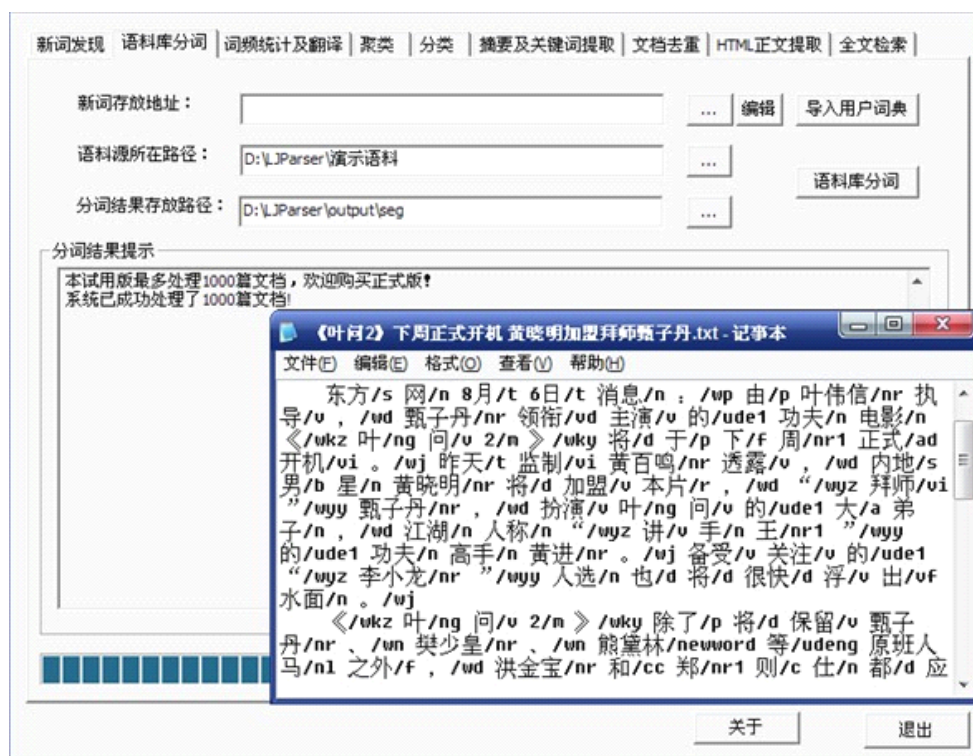
1) 导入用户词典



指定新词文件，用户可以对新词列表进行编辑（编辑见小图，注：系统给出的标注默认为newword，用户可以根据实际情况进行校对，词性可以标注为任意字符串，系统不做限制）后，再点击“导入用户词典”，在结果提示框中会显示是否导入成功。

对于不需要导入新词的用户，本步骤可以跳过。

2) 语料标注分词

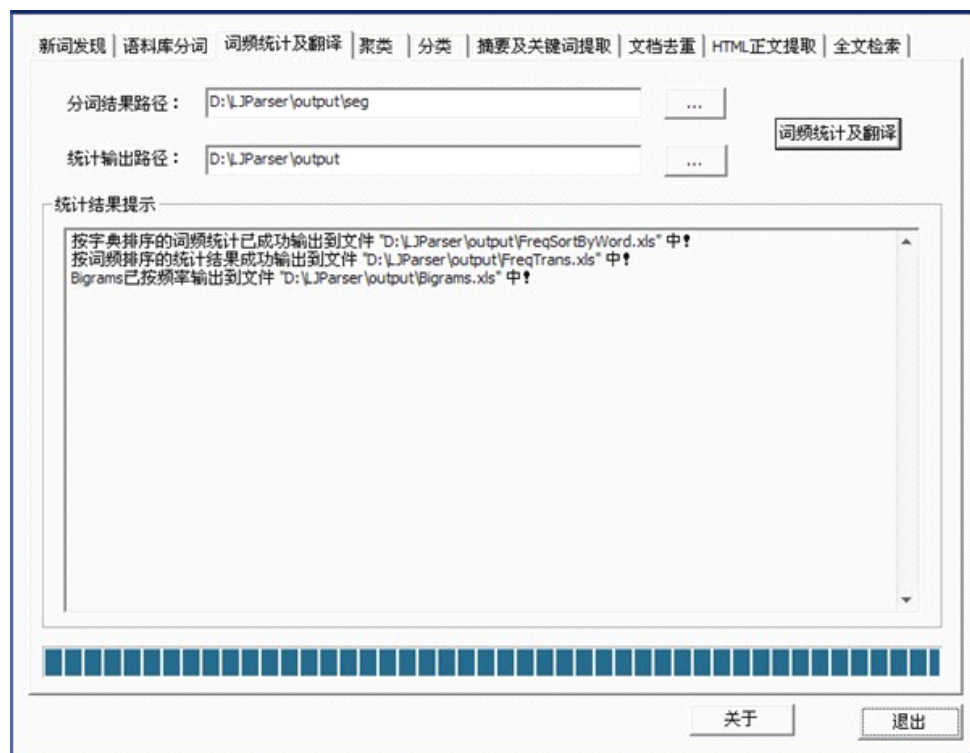


首先指定语料源所在路径，该目录下的语料可以与新词发现中所使用的语料相同，也可以不同，根据用户需求确定。

同第一步一样，选择语料源所在路径后，系统会指定默认的“分词结果存放路径”为：当前工作目录\output\seg。用户也可以指定其它输出路径。分词及词性标注结果以 txt 格式文件存放，文件名与源语料中的文件名一致。

点击“语料库”分词，系统开始分词与词性标注。处理完成后，结果输出到“分词结果存放路径”目录下，系统会在完成时自动为用户打开该目录。

2.4: 词频统计及翻译



1) 输入“分词结果路径”，该目录下的文件为第二步分词标注的结果。

2) 同样的，指定“分词结果路径”之后，系统会指定一个默认的“统计输出路径：当前工作目录\output”。用户也可以指定其它输出路径。

3) 点击“词频统计及翻译”按钮，系统开始统计词频、共现词对频率等信息。输出结果分别为：按照词典序排列的词频统计；按照词频大小排序的词频统计，该输出文件包含了词的英文翻译（如下图所示）；按照共现词对频率排列的共现词对统计文件（如下图所示）。

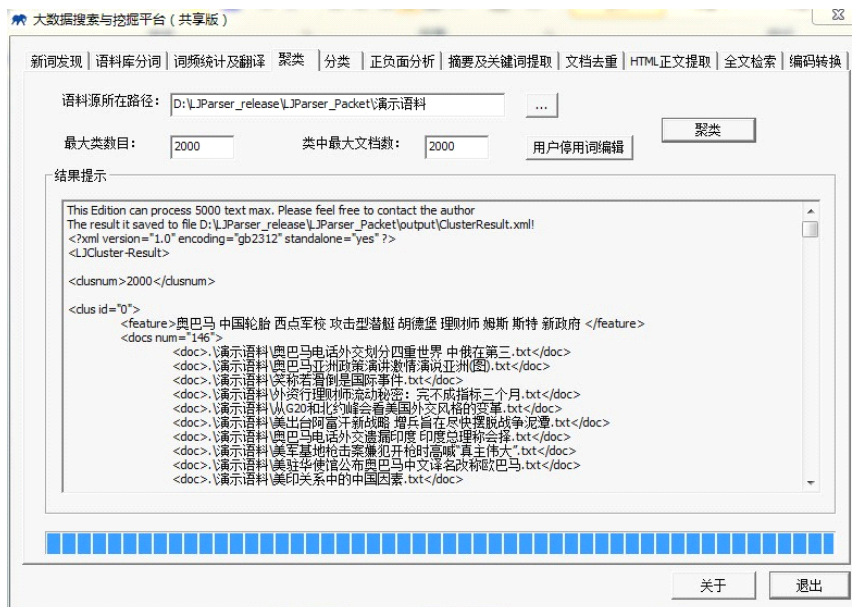
Microsoft Excel - FreqTrans								
文件(F) 编辑(E) 视图(V) 插入(I) 格式(O) 工具(T) 数据(D) 窗口(W) 帮助(H)								
Σ 宋体 12 B I								
E2	A	B	C	D	E	F	G	H
1	总词数为: 65259, 所有词的平均频率为: 23.121746							
2	词语	词频	一元概率	译文				
3	,	65410	0.043349					
4	的	48977	0.032459	target; bull's-eye	有~放矢	shoot the arrow		
5	.	32846	0.021768					
6	。	29837	0.019774					
7	、	12285	0.008142					
8	在	11929	0.007906	① (存在; 生存)	exist; be living	② (表示位置)		
9	是	11259	0.007462	① (对; 正确)	correct; right	② (表示答应)		
10	”	10531	0.006979					
11	“	10481	0.006946					
12	了	8479	0.005619	① (明白)	know clearly; understand	② (完了)		
13	中国	8090	0.005362	China; Chinese	(adj.)			
14	新加坡	7463	0.004946					
15	不	7239	0.004798	<副> ① (表示否定)	not; no	~完全	not comp	
16	&	7176	0.004756					
17	(7145	0.004735					
18	:	7059	0.004678					
19	和	6897	0.004571	mix; blend				
20	一	6638	0.004399	① (数目)	one; first	② (同一)	same; one	咱
21	,	6305	0.004179					
22	;	6256	0.004146					
23		6212	0.004117					
24	有	5356	0.00355	① (拥有)	have; possess	~空	be free	你~笔
25	中	5071	0.003361	① (正对上)	fit exactly; hit	~靶	hit the te	
26		4863	0.003223					
27	[4638	0.003074					
28]	4632	0.00307					
29	#	4512	0.00299					

图：词频统计分析及翻译结果

Microsoft Excel - Bigrams				
文件(F) 编辑(E) 视图(V) 插入(I) 格式(O)				
D2 转移概率				
	A	B	C	D
1	二元词对总数为: 497909			
2	前一个词	后一个词	频次	转移概率
3	.	.	31848	0.969616
4	&	#	4245	0.591555
5	;	&	2204	0.352302
6	*	*	1741	0.765275
7	,	但	1433	0.021908
8	。	”	1403	0.047022
9	,	在	1250	0.01911
10	'	s	1137	0.538097
11	的	,	1123	0.022929
12	>	>	1106	0.372391
13	说	,	1083	0.412571
14	”	,	1062	0.100845
15	”	的	1033	0.098091
16	,	这	1007	0.015395
17	的	“	1000	0.020418
18	更	多	879	0.398278
19	”	。	879	0.083468
20	,	并	854	0.013056
21	,	也	848	0.012964
22	首	页	832	0.547368
23	。	在	818	0.027416
24	后	,	812	0.354585
25	。	“	800	0.026812
26	#	58	776	0.171986
27	58	;	776	0.979798
28	,	“	744	0.011374
29	:	“	739	0.104689

图：二元词对的统计结果

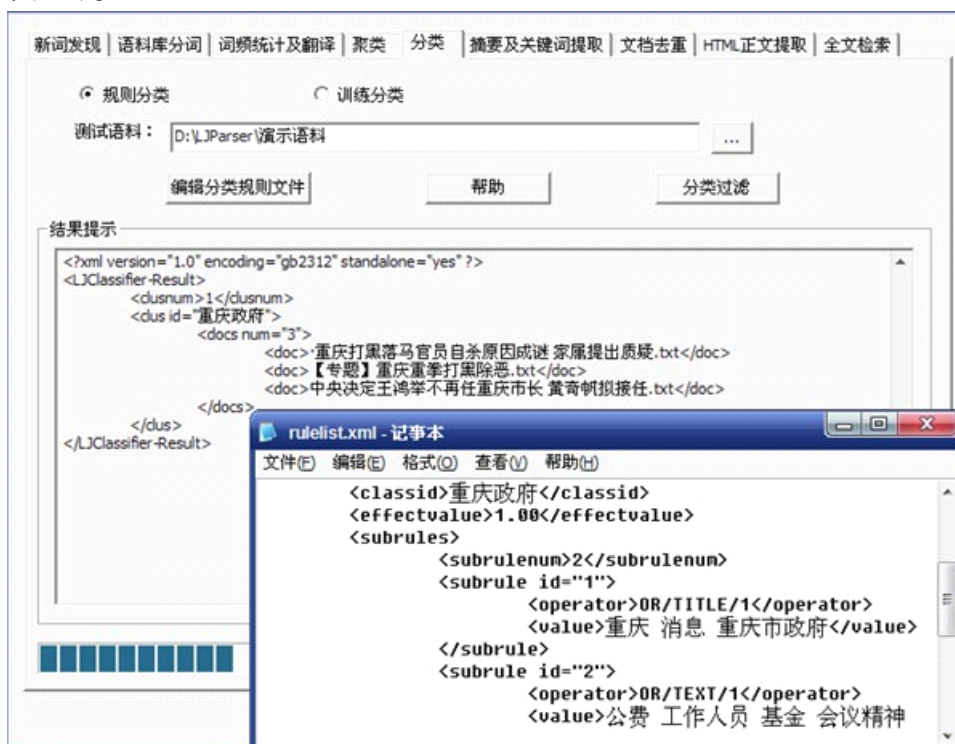
2.5: 文本聚类及热点内容分析



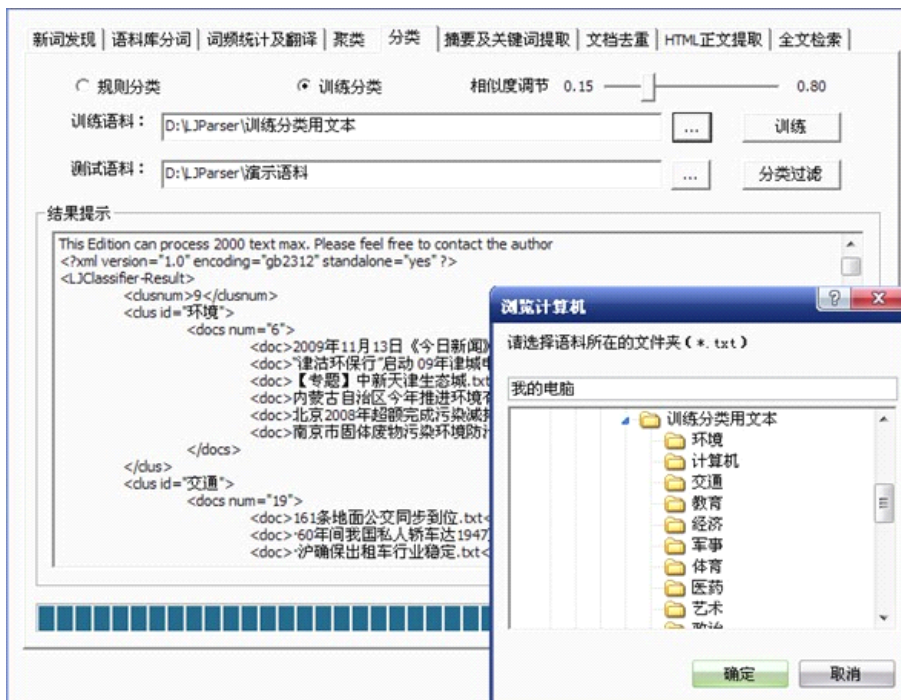
选择语料文件夹，设置参数和频繁出现的领域干扰词，点击聚类，系统返回语料所描述的热点事件话题。

2.6: 分类过滤

1) 规则分类过滤



- 2) 选择语料文件夹，编辑分类规则文件（如图所示），点击“分类过滤”按钮，系统返回规则过滤的结果。
- 3) 训练分类过滤



选择训练语料（各个类别需要按子文件夹排放，如图），点击“训练”按钮，系统进行类别特征的自学习；选择测试语料文件夹，点击“分类过滤”按钮，系统返回分类过滤的结果。可以通过调节相似度，来控制分类过滤的内容模糊匹配程度。

2.7: 文本摘要与关键词提取

大数据搜索与挖掘平台 (共享版)

新词发现 | 语料库分词 | 词频统计及翻译 | 聚类 | 分类 | 正负面分析 | 摘要及关键词提取 | 文档去重 | HTML正文提取 | 全文检索 | 编码转换

语料源目录: D:\JParser_release\JParser_Packet\演示语料 ... 获取摘要及关键词 上一篇 下一篇

最大摘要长度: 250 摘要最大压缩率(0.0~1.0): 0

源文档内容

北杜市 (日本), 2009年10月16日 探访日本大型太阳能电池试验场 10月15日, 在日本山梨县北杜市, 一名日本技术人员正在介绍北杜试验场。北杜市的大规模太阳光发电研究所北杜试验场, 是日本最主要的太阳能电池测试场所之一, 这里对来自中国在内的十几个国家的太阳能电池生产厂商的产品进行对比试验, 也为日本研究新型太阳能电池提供参考依据。新华社记者刘华摄

摘要

北杜市 (日本), 2009年10月16日 探访日本大型太阳能电池试验场 10月15日, 在日本山梨县北杜市, 一名日本技术人员正在介绍北杜试验场。北杜市的大规模太阳光发电研究所北杜试验场, 是日本最主要的太阳能电池测试场所之一, 这里对来自中国在内的十几个国家的太阳能电池生产厂商的产品进行对比试验, 也为日本研究新型太阳能电池提供参考依据。

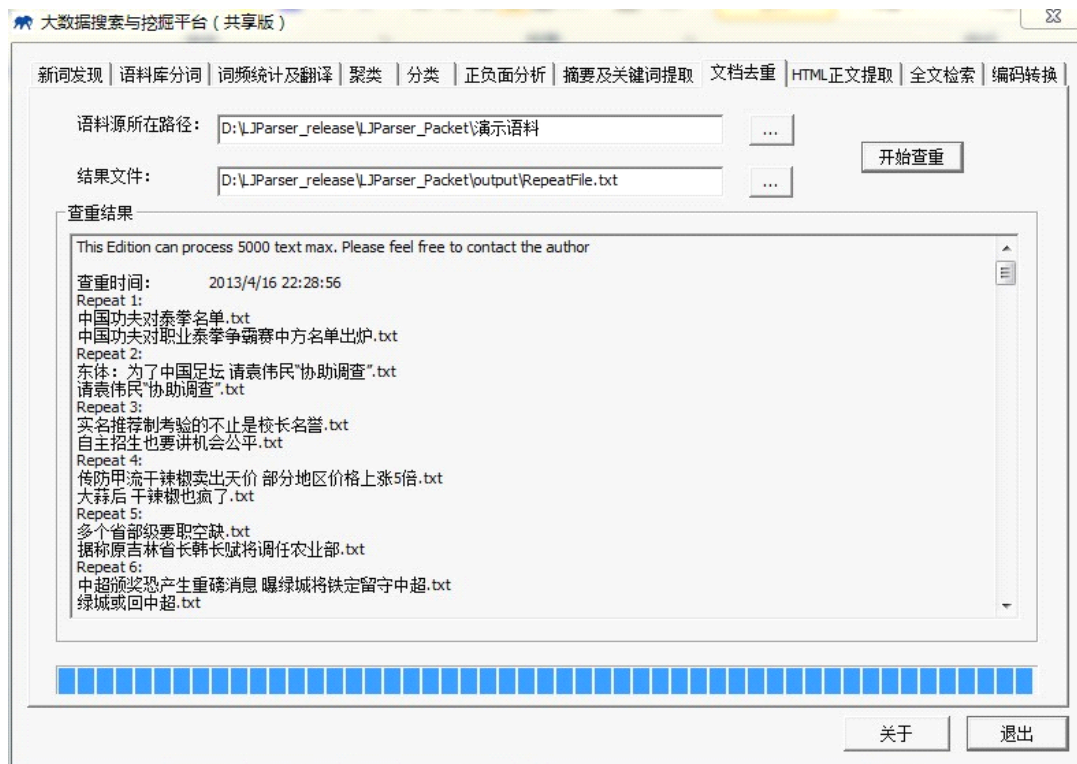
关键词

太阳能电池 北杜试验场 日本 试验场

关于 退出

选择语料文件夹, 设置参数, 点击获取按钮, 系统自动显示摘要和关键词的结果。通过点击“上一篇”、“下一篇”按钮, 可实现结果的快速浏览。

2.8: 文档去重



选择语料文件夹，选择结果文件存放路径，点击“开始查重”按钮，系统返回查重的结果。

2.9: HTML 正文解析

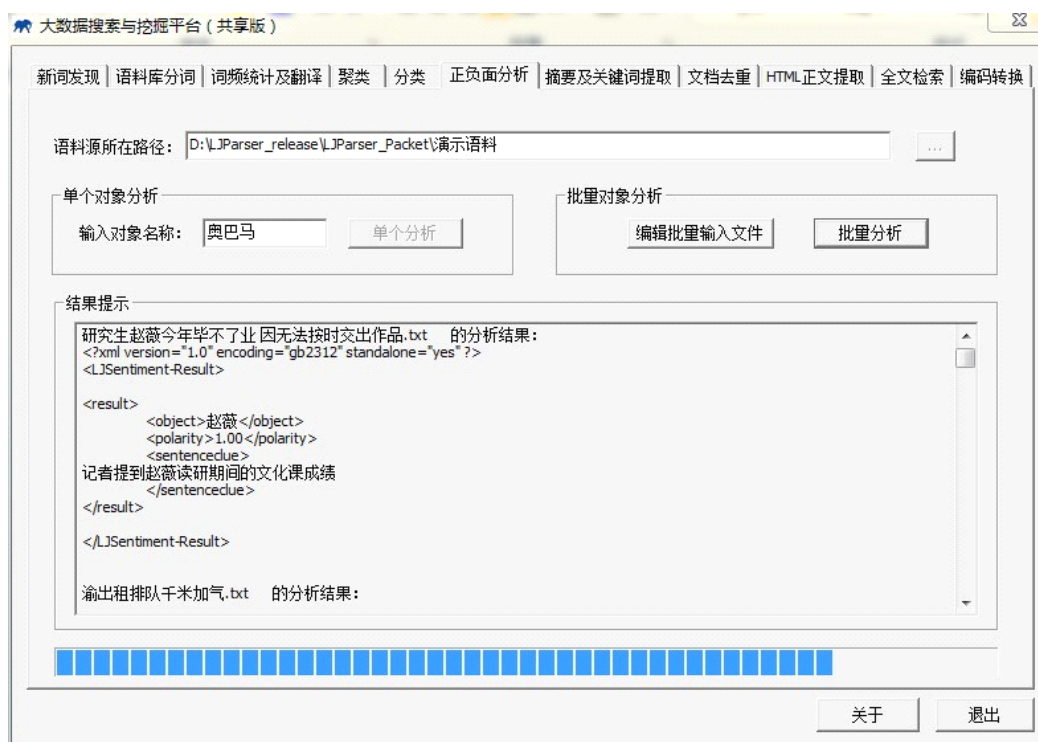


输入 URL，点击抓取按钮，下载网页源文件；然后点击提取正文按钮，系统显示正文结果，去除了大量的垃圾干扰信息。

2.10: 正负面分析

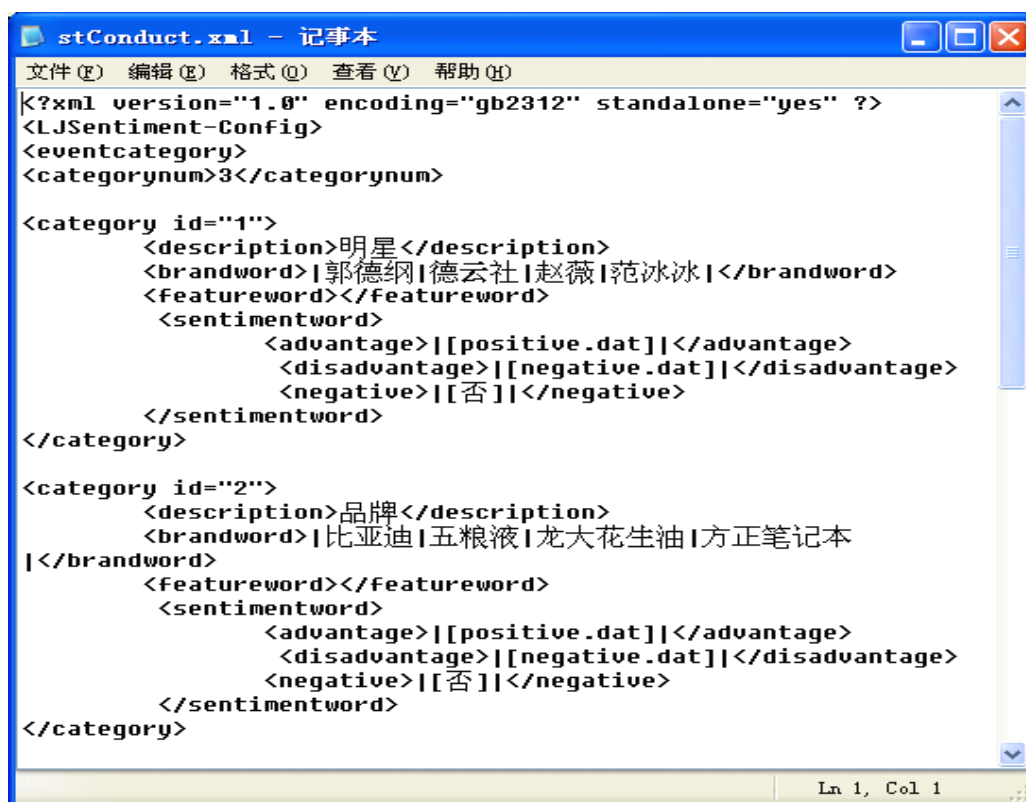
正负面分析可以单个分析，也可以批量分析。

1、单个分析，对某人或某事的正负面判断，只需要在“单个对象分析”中输入要分析的人、事或机构等，点击“单个分析”即可，如图所示：

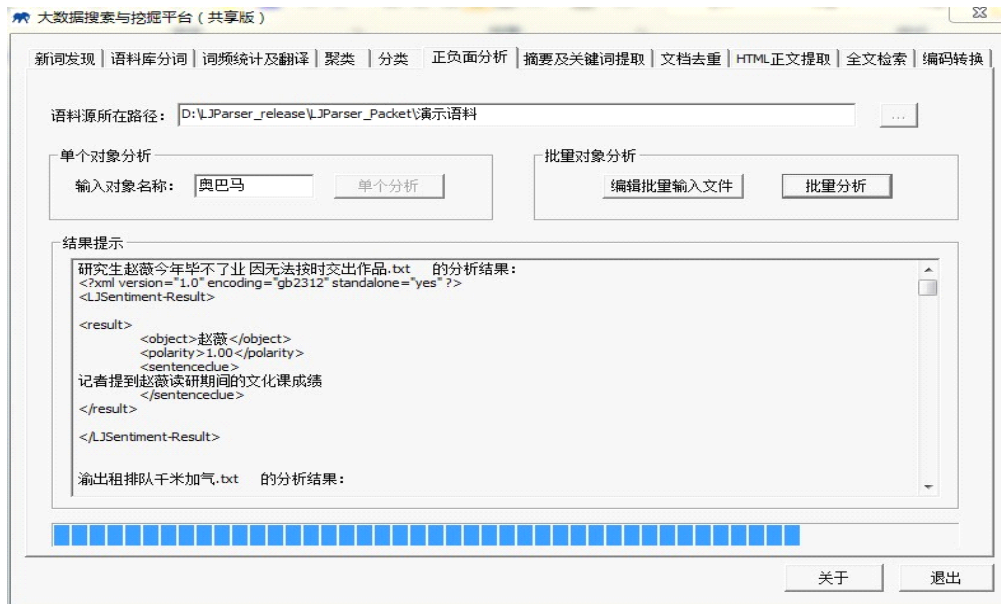


2、批量分析，对多人或几件事进行分析。具体操作为：

1) 点击编辑批量输入文件，填写要分析的所有人、事或物，如图所示：



2) 点击批量分析，如图所示：

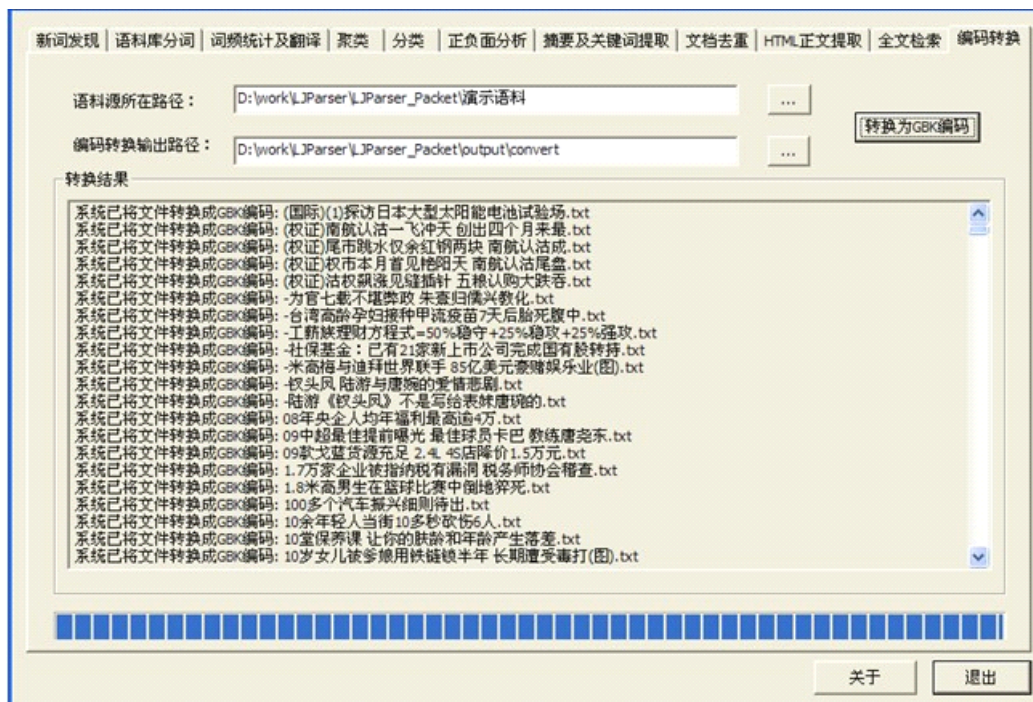


2.11: 编码转换

编码转换基础件可以实现批量将 utf8、gbk、big5、unicode 等编码的文件自动识别并转换为指定编码(相互转换)的文件，试用版本中只集成了其他编码转换为 GBK 的功能，具体操作如下：

- 1) 选择语料文件夹
- 2) 选择输出文件夹
- 3) 点击“转换为 GBK 编码”

如下图所示：



三、作者简介



张华平 博士 副教授 硕导
北京理工大学计算机学院 院长助理
北京理工大学网络搜索挖掘与安全实验室 主任
地址: 北京海淀区中关村南大街 5 号 100081
电话: +86-10-68918642
Email: kevinzhang@bit.edu.cn
MSN: pipy_zhang@msn.com;
网站: <http://ictclas.nlpir.org> (自然语言处理与信息检索共享平台)
博客: <http://hi.baidu.com/drkevinzhang/>
微博: @ICTCLAS 张华平博士

Dr. Kevin Zhang (张华平, Zhang Hua-Ping)
Associate Professor, Graduate Supervisor
Dean Assistant, School of Computer
Director, Web Search, Mining and Security Lab.
Beijing Institute of Technology
Add: No.5, South St., Zhongguancun, Haidian District, Beijing, P.R.C PC: 100081
Tel: +86-10-68918642
Email: kevinzhang@bit.edu.cn
MSN: pipy_zhang@msn.com;
Website: <http://ictclas.nlpir.org> (Natural Language Processing and Information Retrieval Sharing Platform)
Blog: <http://hi.baidu.com/drkevinzhang/>
Twitter: @ICTCLAS 张华平博士