

NLPIR 大数据搜索与挖掘使用手册



自然语言处理与信息检索共享平台
Natural Language Processing & Information Retrieval Sharing Platform

<http://www.nlpir.org/>

@ICTCLAS 张华平博士

2017-3

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

访问 <http://ictclas.nlpir.org/>(自然语言处理与信息检索共享平台), 您可以获取 NLPIR 系统的最新版本, 并欢迎您关注张华平博士的新浪微博 @ICTCLAS 张华平博士 交流。

目前提交的 NLPIR-Parser 共享版可满足一般科研与非商业用途, 能处理 200 篇文章, 总量 500KB。正式版本不受任何限制, 需要联系实验室助理洽谈正式版授权。

Document Information

Document ID	NLPIR- 2014-WHITEPAPER	Version	V4.0
Security level	Public 公开	Status	Creation and first draft for comment
Author	张华平	Date	Dec 19, 2013
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. NLPIR 2011
V2.0	Kevin Zhang	2012-8-21	complete draft for comment. NLPIR 2012
V3.0	Kevin Zhang	2012-12-19	complete draft for comment. NLPIR 2013
V4.0	Kevin Zhang	2013-12-19	complete draft for comment. NLPIR 2014
V5.0	Kevin Zhang	2017-3-22	complete draft for comment. NLPIR 2017

目录

目录

NLPIR 大数据搜索与挖掘使用手册	1
目录	4
NLPIR 大数据搜索与挖掘共享开发平台	5
1、NLPIR 大数据搜索与挖掘共享开发平台简介	5
2、开发平台工具软件操作指南	6
2.1: JZSearch 全文精准检索	6
2.2: 新词发现	7
2.3: 分词及词性标注	8
2.4: 词频统计及翻译	11
2.5: 大数据聚类及热点内容分析	14
2.6: 大数据过滤与分类	15
2.7: 文本摘要与关键词提取	17
2.8: 文档去重	18
2.9: HTML 正文解析	19
2.10: 编码识别与转换	19
2.11: 敏感关键词实时智能扫描	20
3 作者简介	23

NLPIR 大数据搜索与挖掘共享开发平台

1、NLPIR 大数据搜索与挖掘共享开发平台简介

NLPIR 大数据搜索与挖掘共享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 大数据搜索与挖掘共享开发平台的 11 种功能：

■ 1. JZSearch 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. 大数据聚类及热点分析

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. 大数据分类过滤

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. 自动摘要与关键词提取

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。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。

■ 8. 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。

■ 9. 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。

■ 10. 编码自动识别与转换

自动识别文档内容的编码，并进行自动转换，目前支持 Unicode/BIG5/UTF-8 等编码自动转换为简体的 GBK 或者 UTF8，同时将繁体 BIG5 和繁体 GBK 进行繁简转化。

■ 11. 敏感关键词实时智能扫描

根据配置的关键词，实时智能扫描关键词及各种变种（编码变种、音变、形变、噪音干扰）等。

2、开发平台工具软件操作指南

大家首先根据自己操作系统的情况选择 bin-win32 或者 bin-win64 目录，再点击运行 NLPIR-Parser.exe

按照功能依次介绍如下：

2.1: JZSearch 全文精准检索

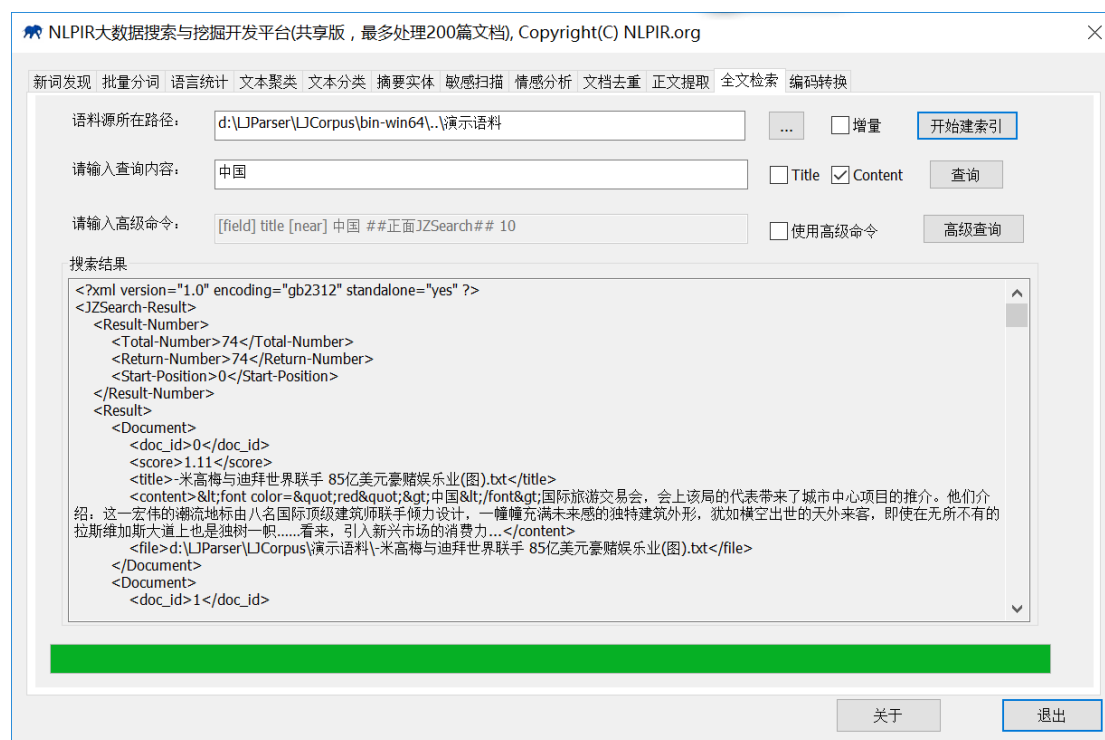


图 1 NLPIR 平台之 JZSearch 大数据全文搜索功能

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

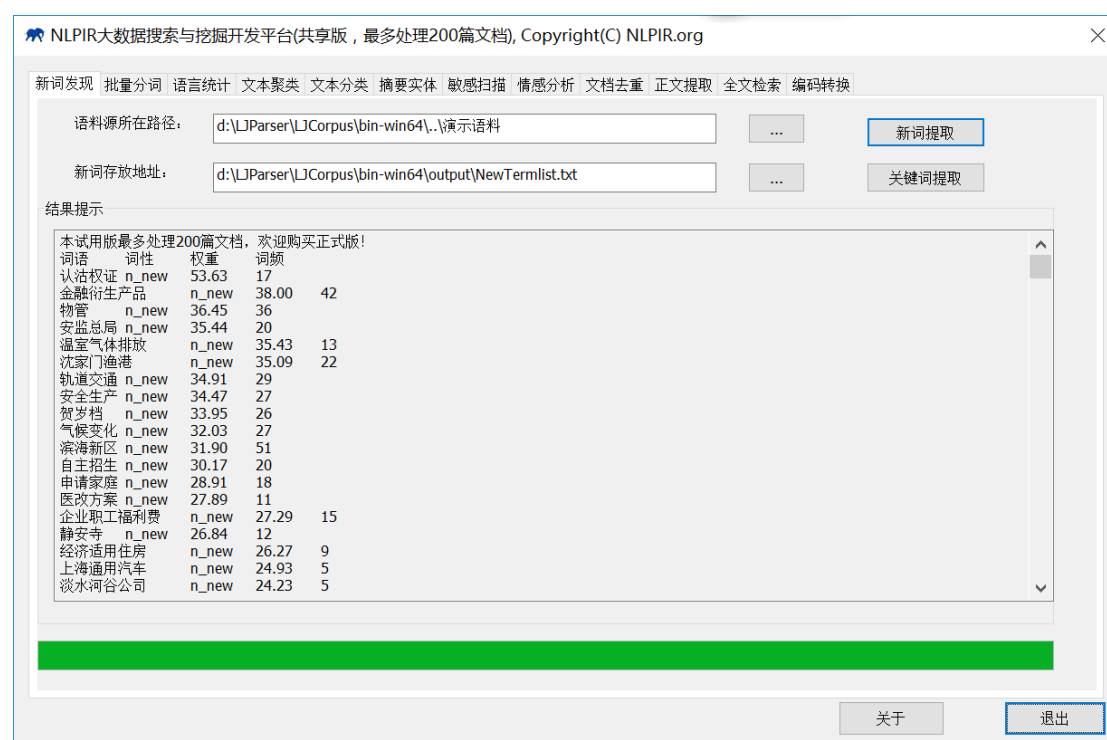


图 2 NLPIR 平台之批量新词发现功能

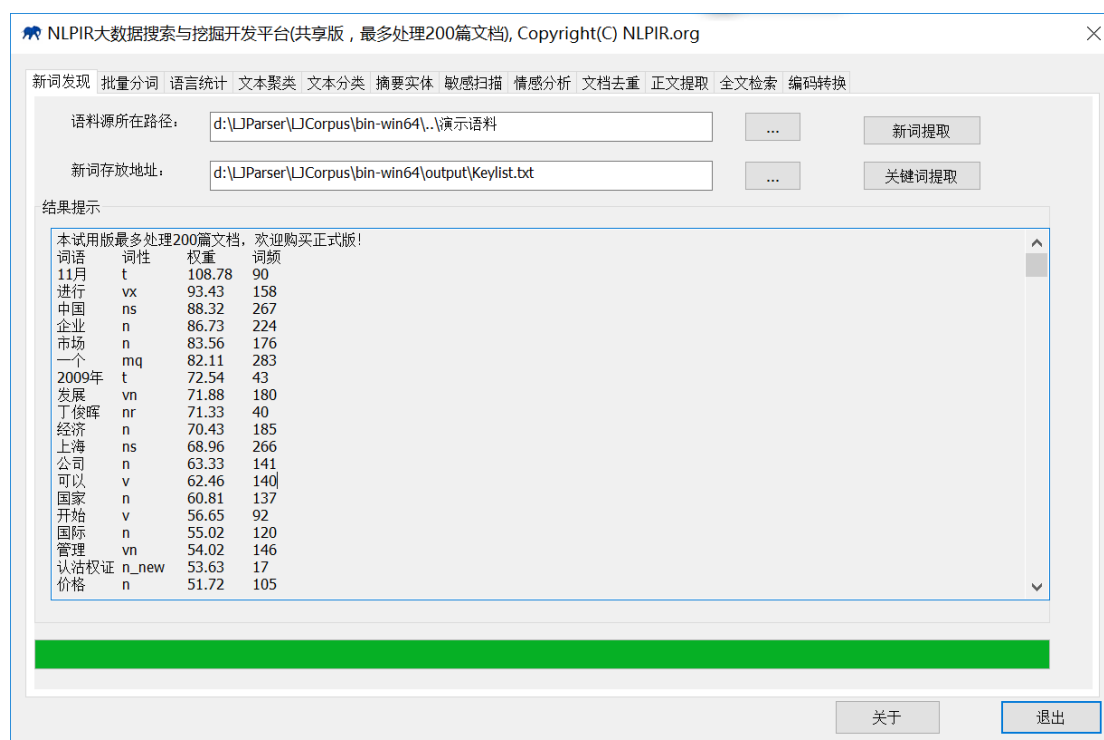


图 3 NLPIR 平台之批量关键词提取功能

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（如果是关键词提取则为 KeyList.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: 分词及词性标注

1) 导入用户词典

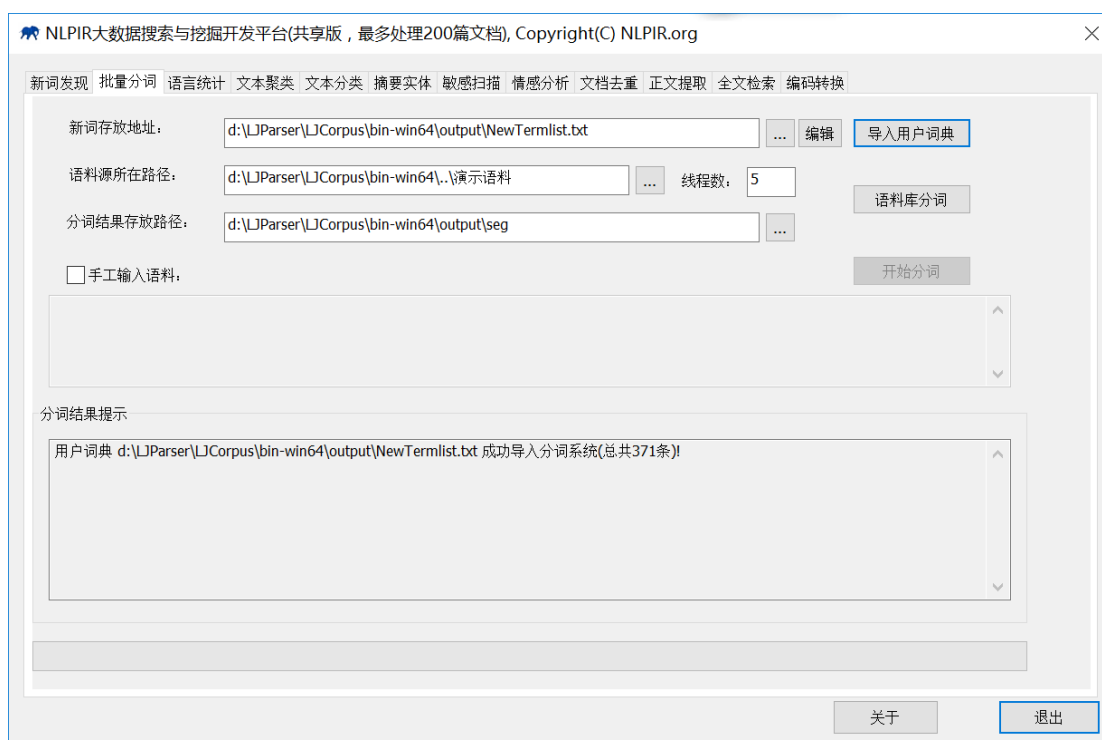


图 4 NLPIR 平台之用户词典导入功能

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：每行一个用户词与词性，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) 语料库批量分词与词性标注

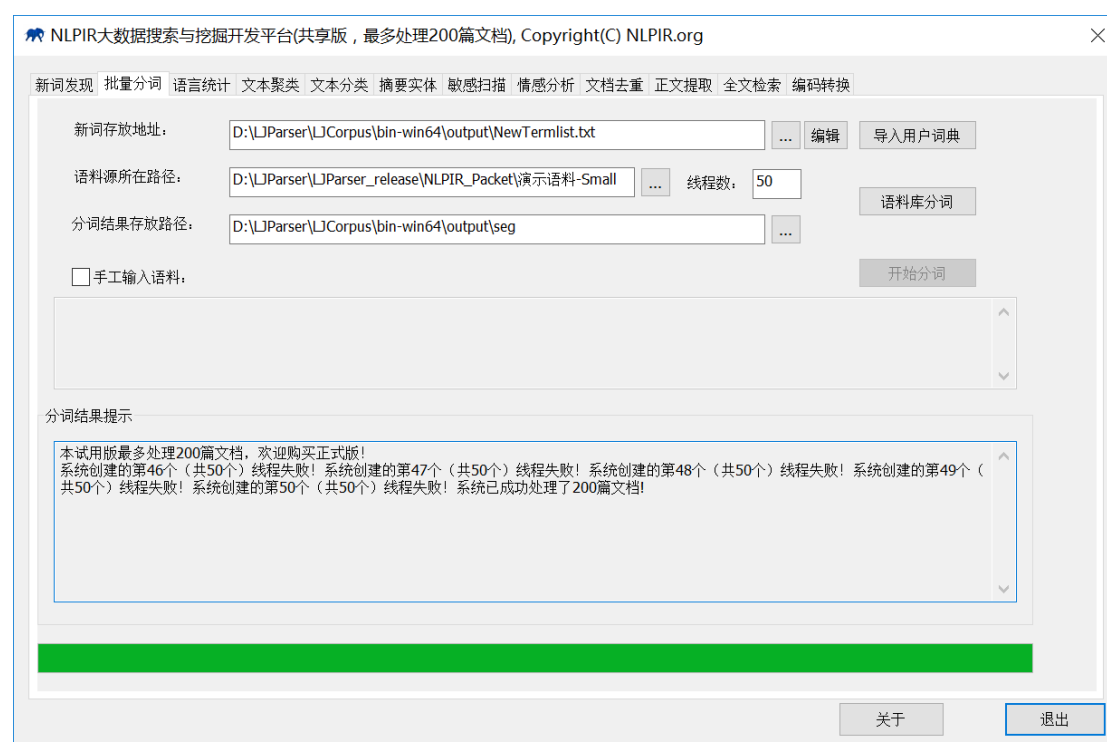


图 5 NLPIR 平台之语料库批量分词

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。

可选择多线程数目，用于多线程并行分词处理。

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。

3) 手工分词

支持用户手动输入分词，如下图所示：

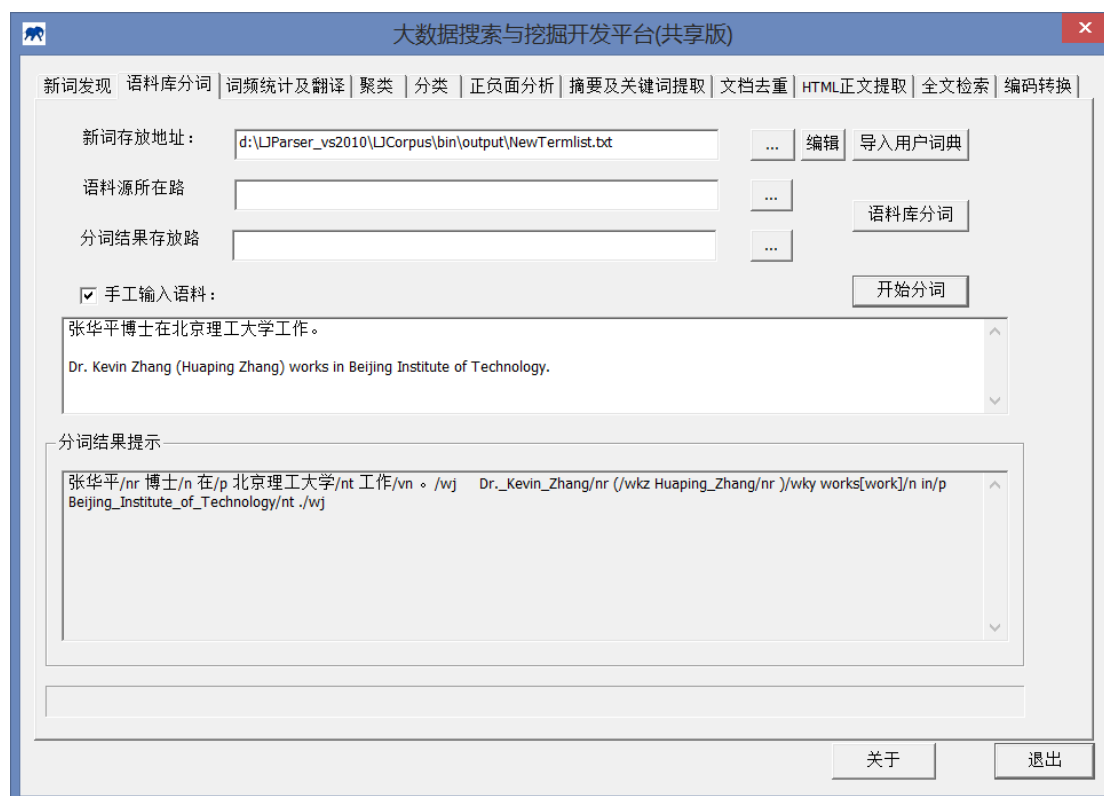


图 6 NLPIR 平台之手工输入分词

2.4: 词频统计及翻译

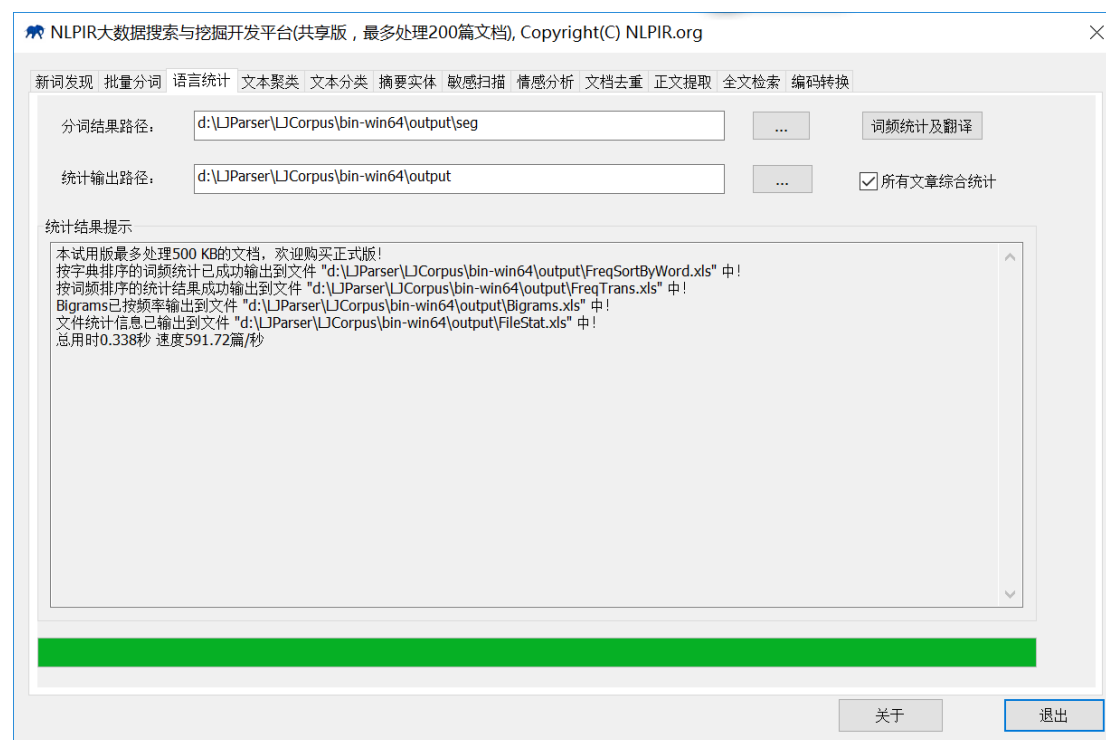
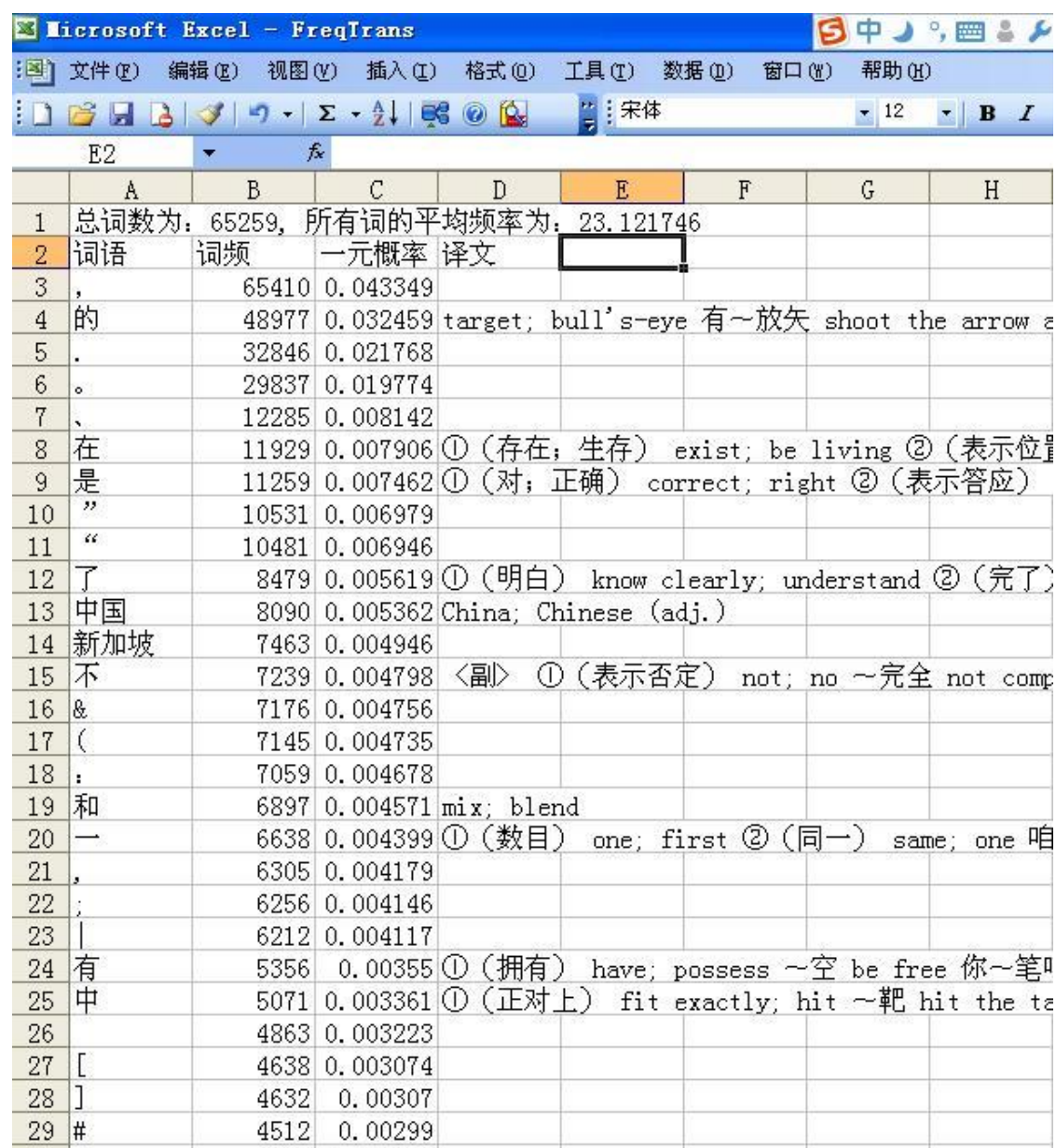


图 7 NLPIR 平台之语言模型统计

- 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）。



	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 a				
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~笔叫				
25	中	5071	0.003361	① (正对上) fit exactly; hit ~靶 hit the ta				
26		4863	0.003223					
27	[4638	0.003074					
28]	4632	0.00307					
29	#	4512	0.00299					

图 8: 词频统计分析及翻译结果

	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 9：二元词对的统计结果

文档名	总词频	总词数	用户词典总词频	用户词典总词数
-米高梅与迪拜世界联手 85亿美	606	376	7	6
08年央企人均年福利最高逾4万	457	238	35	13
12教师举报校长造假 称不处理恐	793	402	11	7
12月4日天津北风逼走大雾 最低	171	125	3	1
15%居民有机会选择就近安置	267	151	16	6
15岁少女惨遭7名90后三次轮奸	418	229	3	3
161条地面公交同步到位	248	146	23	5
19岁男孩的生命绝问	1832	775	16	6
4名女生殴打扒光女同学	363	203	4	3
50寸平板电视惨遭甩卖	204	144	2	2
52届格莱美奖入围名单揭晓 碧昂	437	214	18	6
5人罪行极其严重被判死刑 2人获	875	356	33	13
68家央企涉足金融衍生产品业务	3038	1052	117	36
76人重签艾弗森	349	222	7	4
7批次人用狂犬病疫苗存质量问题	144	93	8	3
80后女孩的求助	783	434	1	1
90人花3400小时装饰白宫 焕然一	132	91	4	2
90年后德国仍在为一战“埋单”	339	206	2	2
94岁老人步行10里取劳保费 - 昆	1417	549	12	4
94岁老人步行10里取劳保费	1417	550	12	4
980元起 12月必“火”的10款手	432	259	1	1

图 10: FileStat.xls 文件的统计结果（包括词频、用词数、用户词典词频、用户词典词数）

2.5: 大数据聚类及热点内容分析

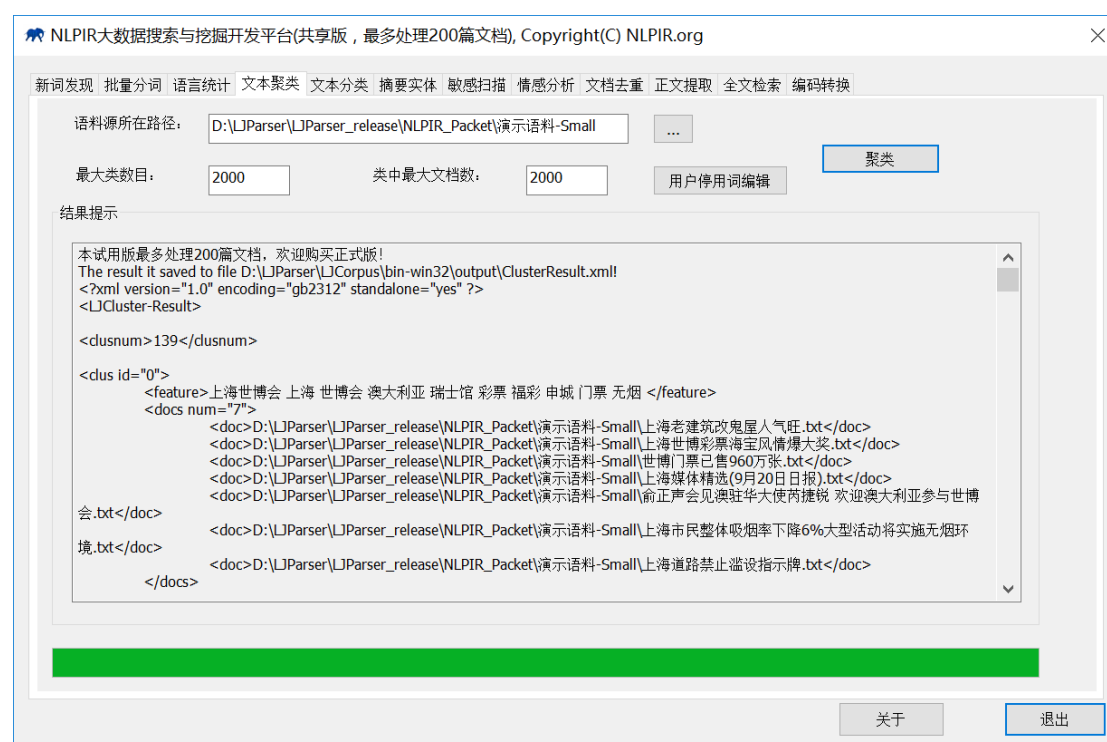


图 11: NLPIR 平台之大数据聚类功能

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: 大数据过滤与分类

大数据规则过滤

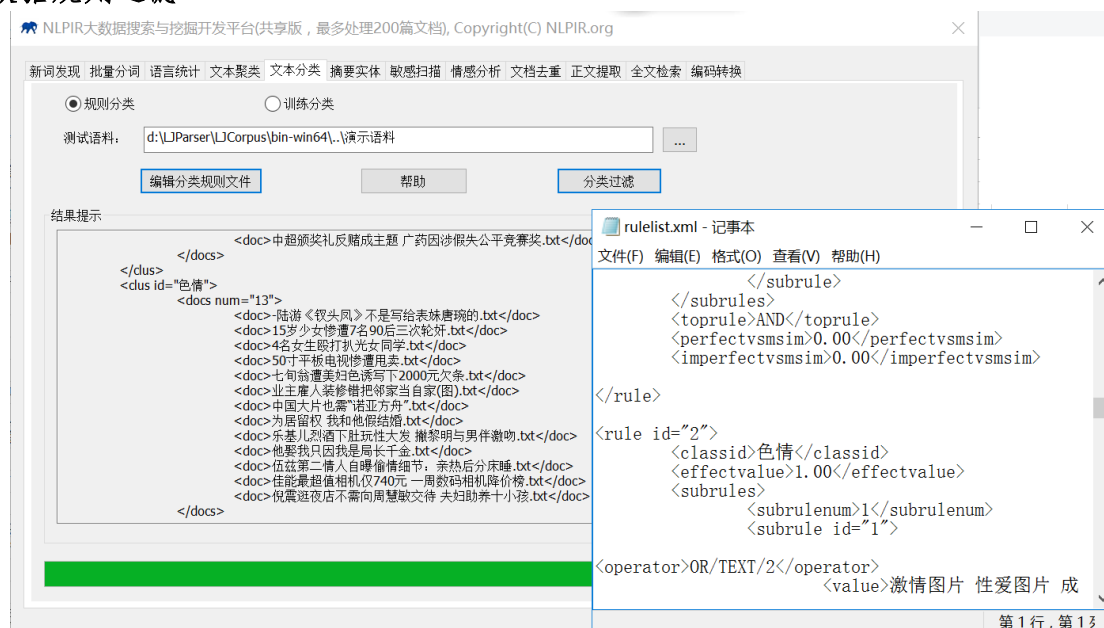


图 9: NLPIR 平台之大数据过滤功能

- 1) 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。

大数据深度机器学习自动分类

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；输出的是系统自动选择的分类特征极其权重信息。

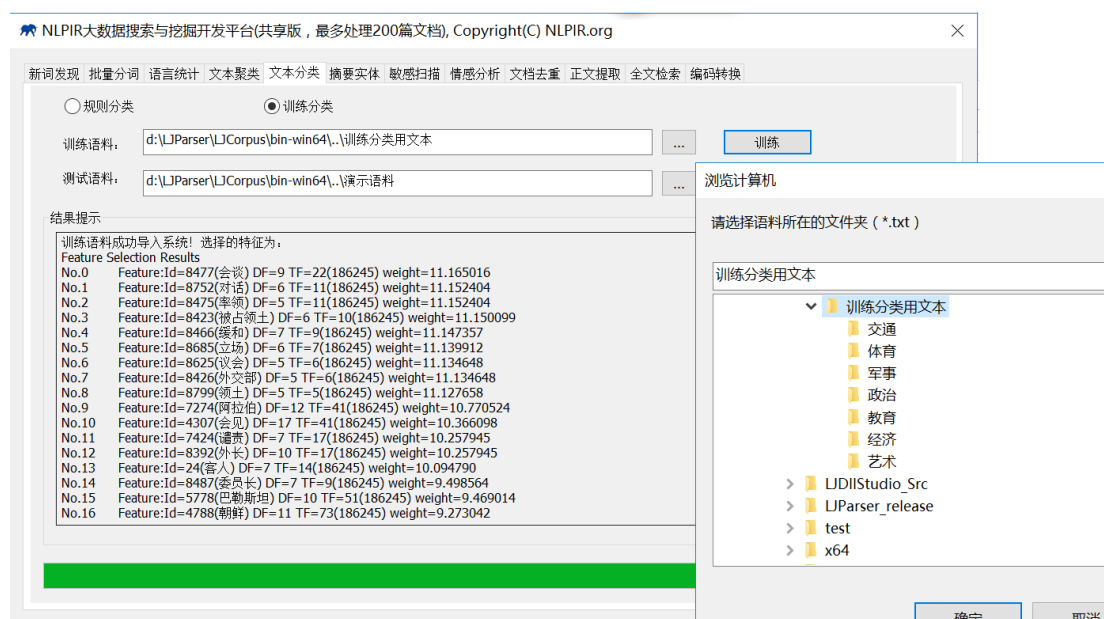


图 10: NLPIR 平台之大数据自动分类的训练过程

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。

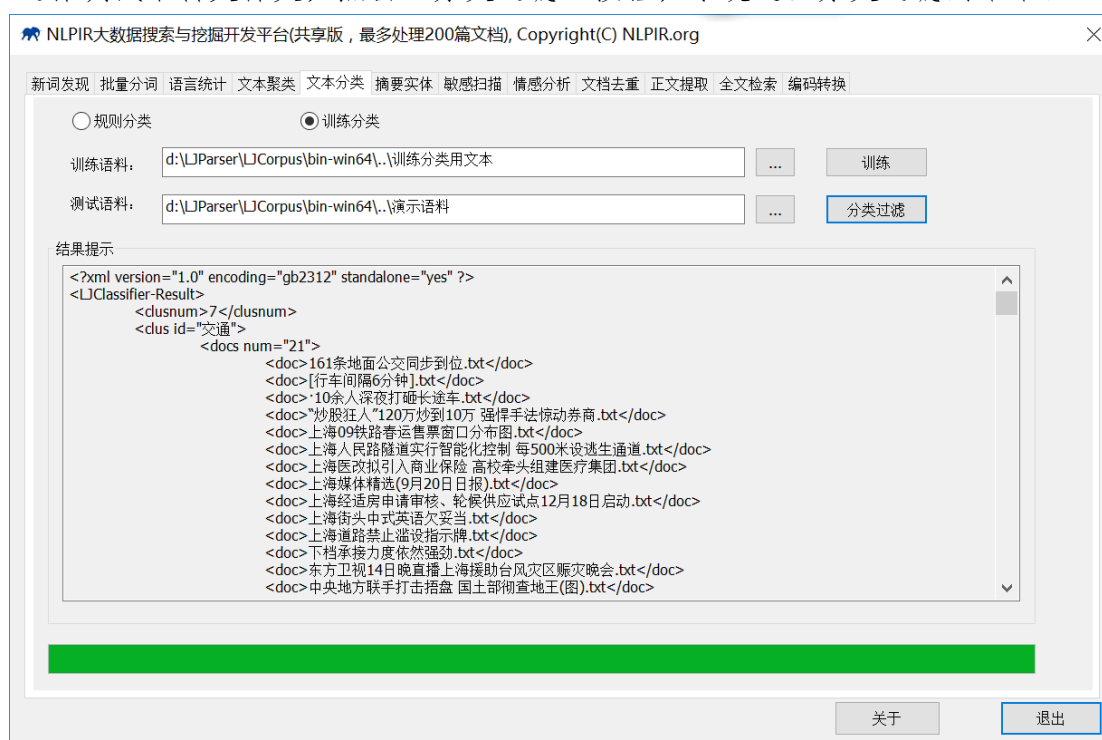


图 11: NLPIR 平台之大数据自动分类的分类过滤过程

2.7: 文本摘要与关键词提取



图 12: NLPIR 平台之大数据自动摘要与关键词提取功能

选择语料文件夹，设置参数，点击获取按钮，系统自动显示摘要关键词和命名实体抽取的结果。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

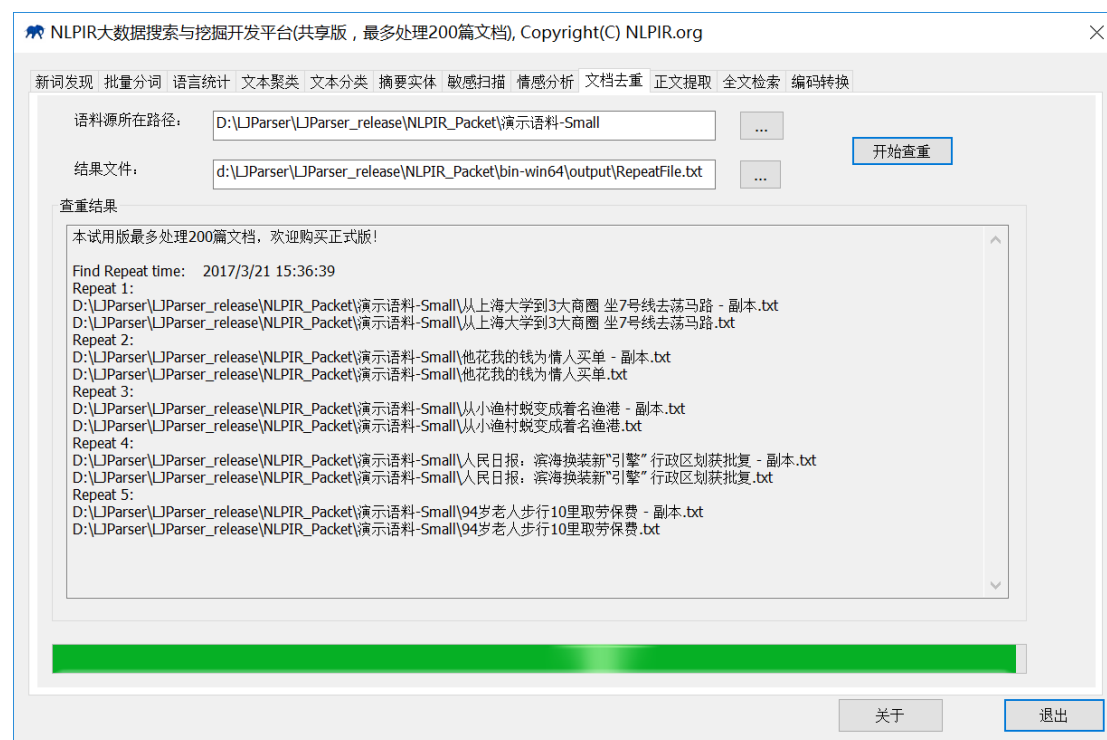


图 13: NLPIR 平台之大数据相似去重

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

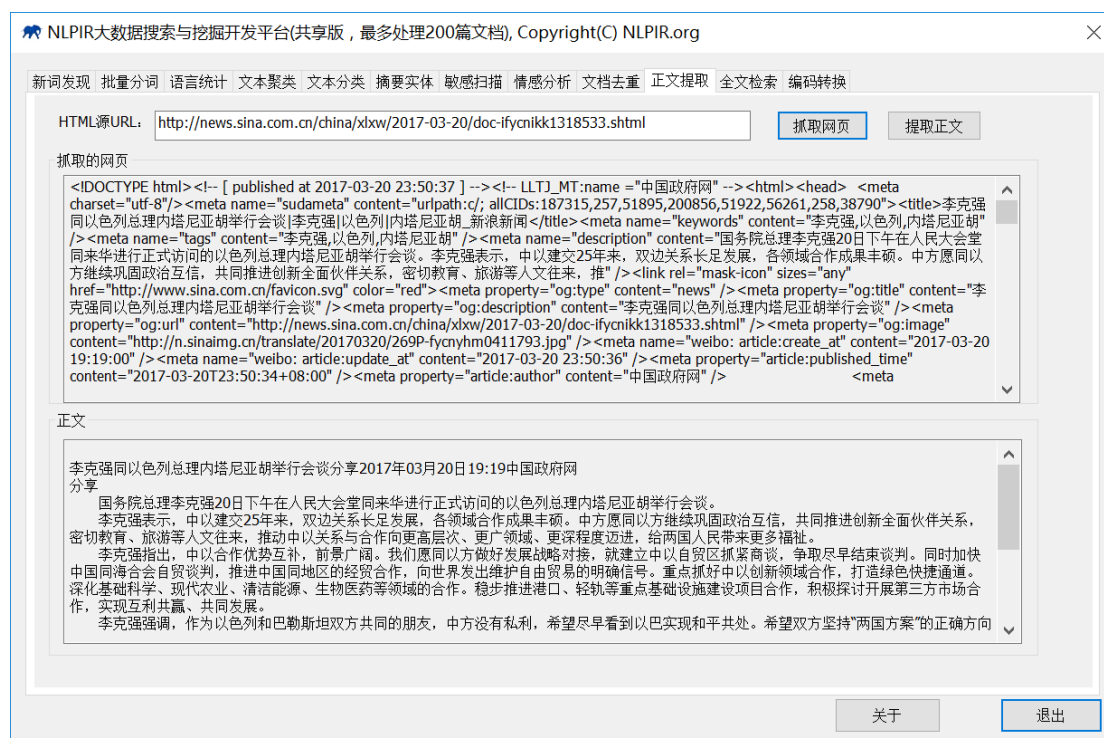


图 14: NLPIR 平台之 HTML 正文提取

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: 编码识别与转换

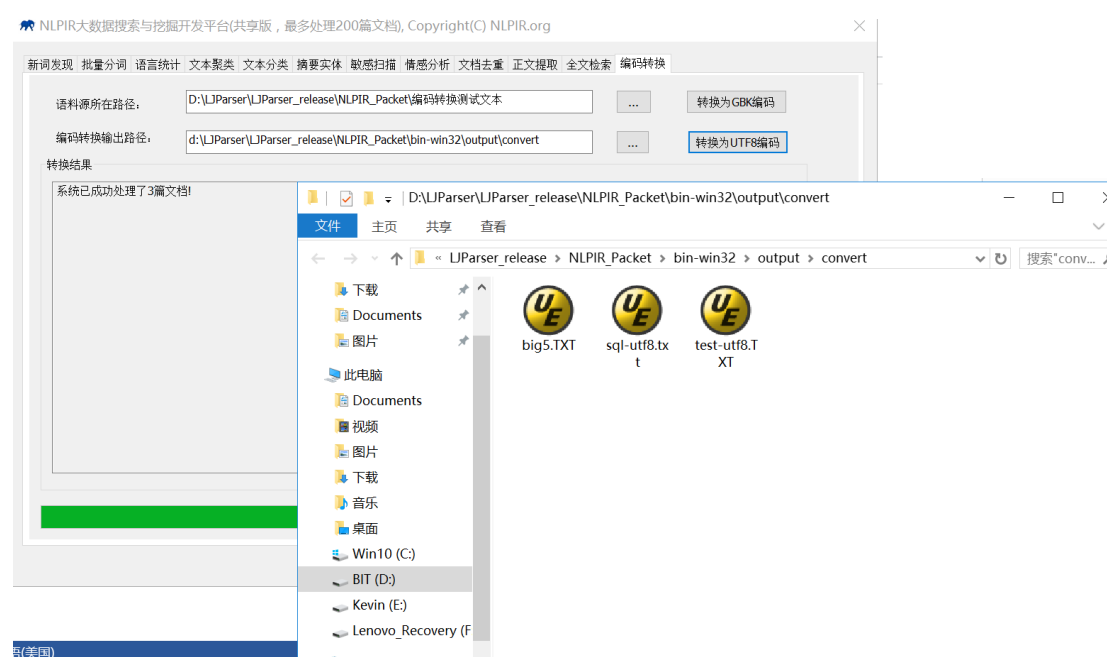


图 15: NLPIR 平台之编码自动识别与转换

选择待处理的文件夹，点击转换按钮，系统自动识别给定的 BIG5 文件，GBK 以及 UTF-8,Unicode 文件，最终转化为简体 GBK。同样可以转换为 UTF8

2.11: 敏感关键词实时智能扫描

关键词配置与导入：总共四列，第一列是关键词，第二列是类别，第三列是权重，第四列表示是否匹配同音词（1 为是，不写默认为不匹配）

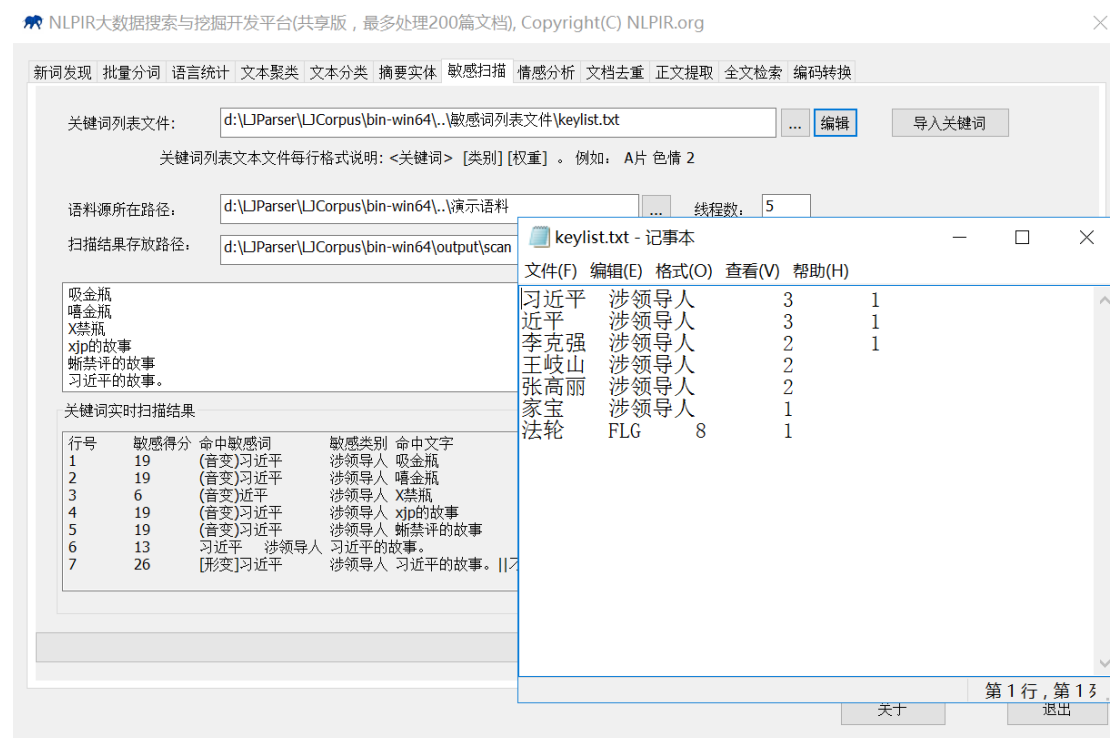


图 16: NLPIR 平台之敏感关键词实时智能扫描

实时智能扫描结果:

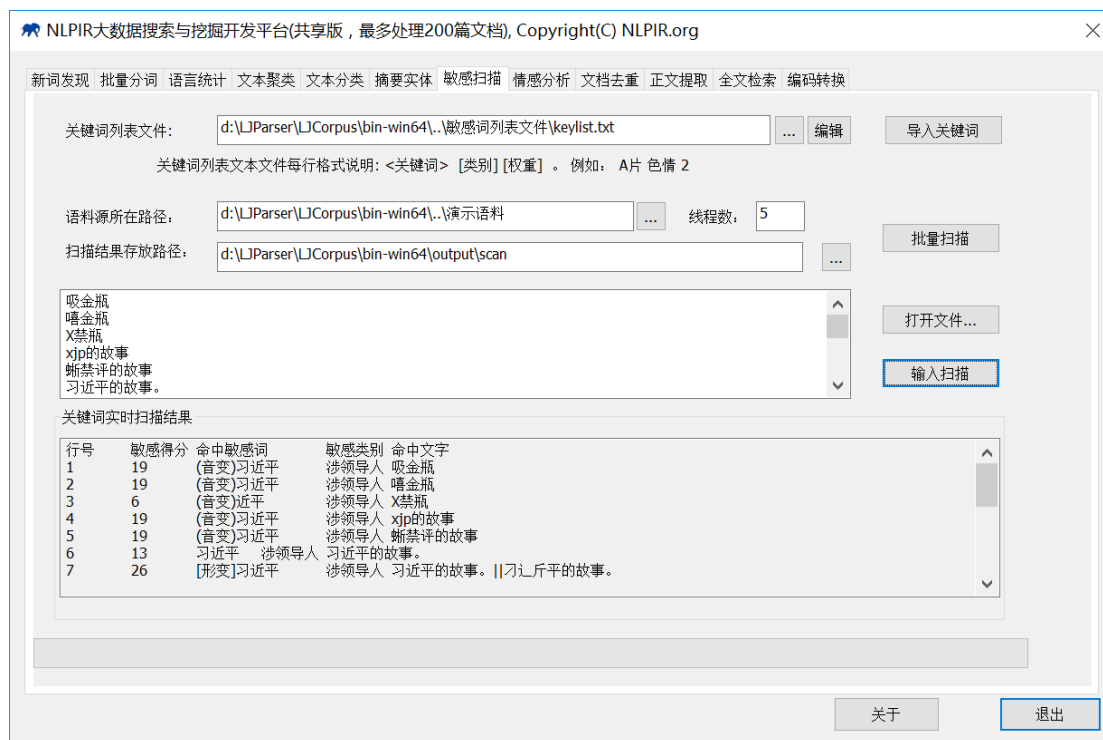


图 17: NLPIR 平台之敏感关键词针对输入的实时智能扫描

批量扫描结果如下:

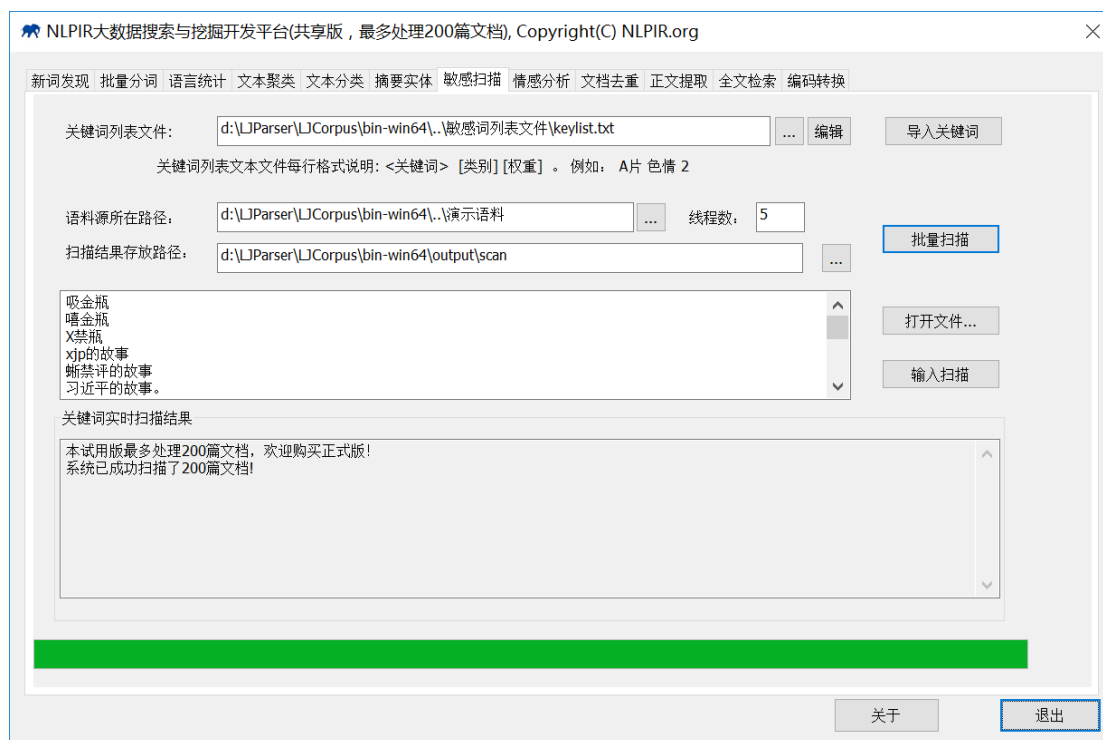


图 17: NLPIR 平台之敏感关键词实时智能扫描结果

具体文件扫描结果:

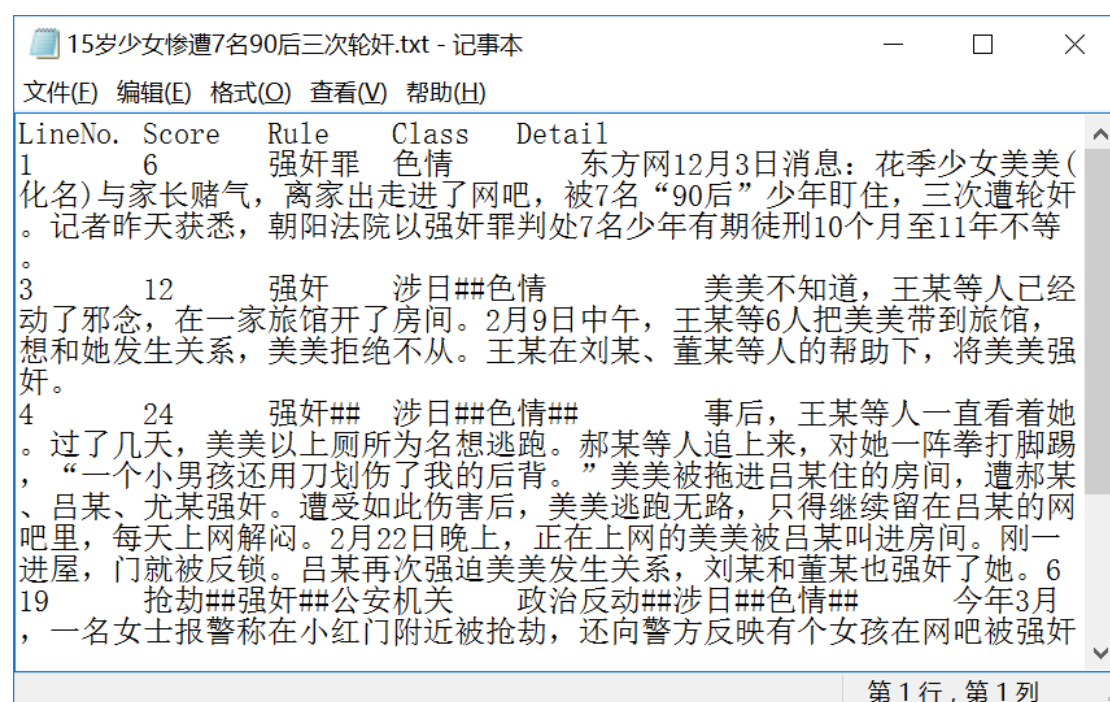


图 18: NLPIR 平台之敏感关键词实时智能扫描结果（单个具体文件的分析）

说明：第一列是扫描到的文件的行数，第二列是敏感得分（超过 5 分就是敏感，背后有敏感计算模型），第三列是命中的关键词，第四列是敏感类别；第五列是命中的原文信息（如果文字超过 1KB，仅给出摘要，并显示出命中部分，如果比较短，则只给原文）

所有命中结果的统计报告如下：

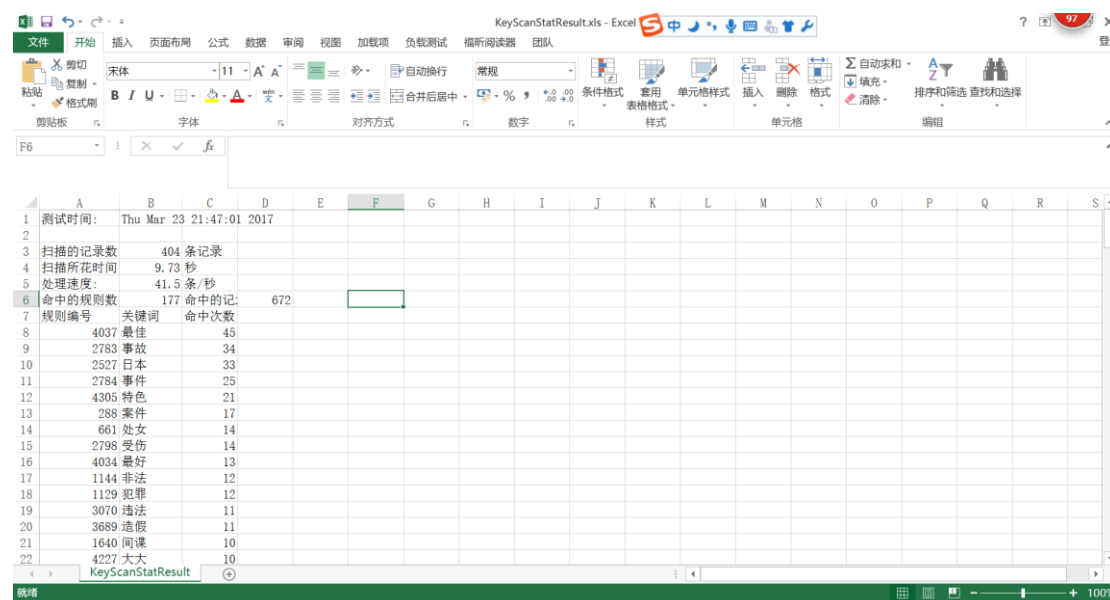


图 19: NLPIR 平台之敏感关键词实时智能扫描结果（宏观统计分析报告）

说明：这里给出的是命中的关键词及命中次数，通过这个统计结果，可以排查高频的关键词设置是否合理，还有那些关键词实际上从来就没有被命中过。

3 作者简介



张华平 博士 副教授 研究生导师

大数据搜索挖掘实验室 主任

地址：北京海淀区中关村南大街 5 号 100081

电话：+86-10-68918642 13681251543(助理)

Email: kevinzhang@bit.edu.cn

MSN: pipy_zhang@msn.com;

网站: <http://www.nlpir.org> (自然语言处理与信息检索共享平台)

<http://www.bigdataBBS.com> (大数据论坛)

微信公众号：大数据千人会

微博: <http://www.weibo.com/drkevinzhang/>

Dr. Kevin Zhang (张华平, Zhang Hua-Ping)

Associate Professor, Graduate Supervisor

Director, Big Data Search and Mining Lab.

Beijing Institute of Technology

Add: No.5, South St., Zhongguancun, Haidian District, Beijing, P.R.C PC: 100081

Tel: +86-10-68918642 13681251543 (Assistant)

Email: kevinzhang@bit.edu.cn

MSN: pipy_zhang@msn.com;

Website: <http://www.nlpir.org> (Natural Language Processing and Information Retrieval Sharing Platform)

<http://www.bigdataBBS.com> (Big Data Forum)

Subscriptions: Thousands of Big Data Experts Twitter:

<http://www.weibo.com/drkevinzhang/>



自然语言处理与信息检索共享平台
Natural Language Processing & Information Retrieval Sharing Platform