NLPIR 大数据语义智能分析平台

用户手册

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/

目 录

一、NLPIR 平台简介

NLPIR 大数据语义智能分析平台，针对大数据内容处理的需要，融合了网络精准采集、自然语言理解、文本挖掘和网络搜索的技术，提供客户端工具、云服务、二次开发接口。平台先后历时十八年，服务了全球四十万家机构用户，是大数据时代语义智能分析的一大利器。

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。

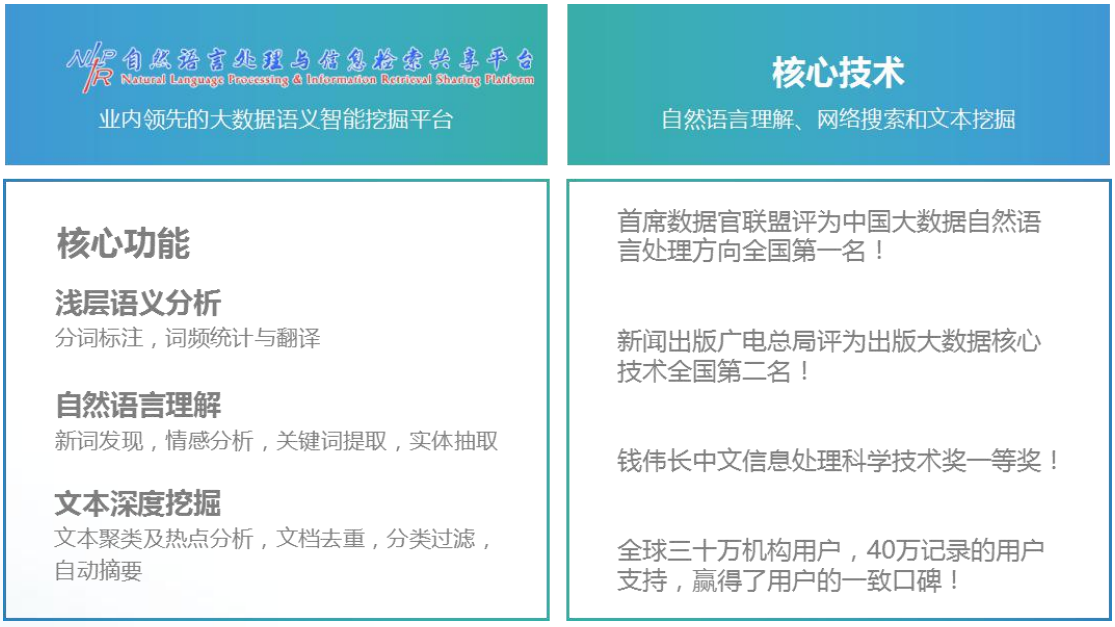


图 1.1　NLPIR 大数据语义智能分析平台简介

NLPIR 大数据语义智能分析平台的十三大功能：

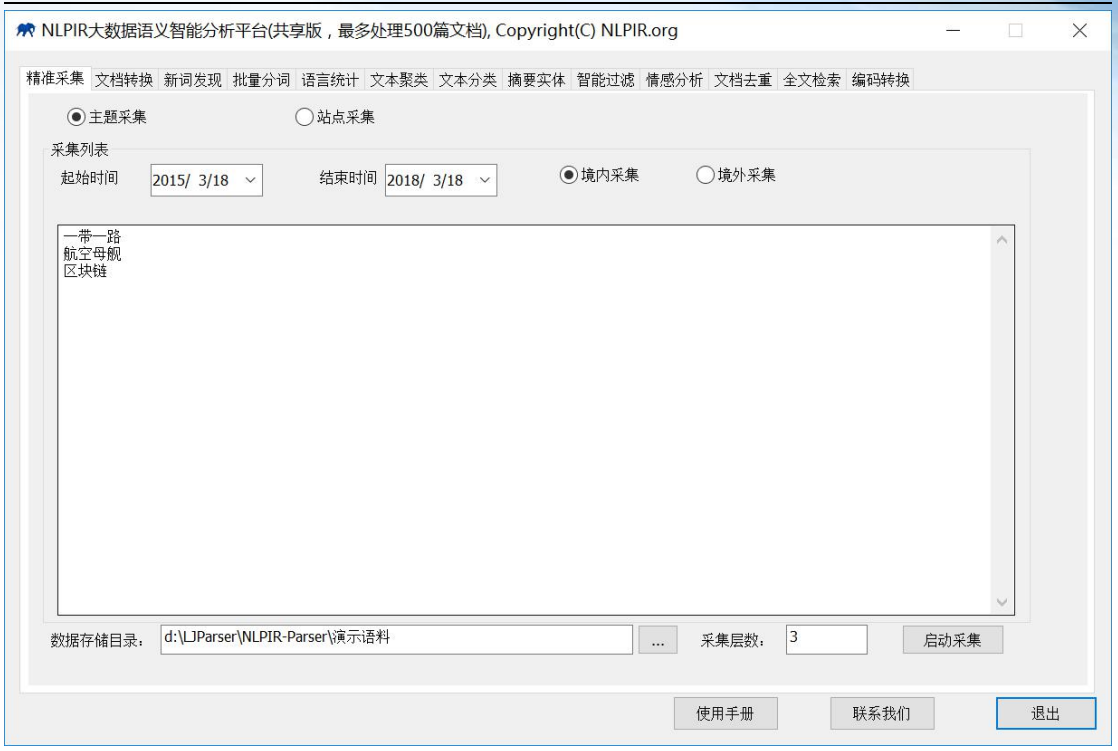


图 1.2 NLPIR 大数据语义智能分析平台客户端

1. 精准采集

对境内外互联网海量信息实时精准采集，有主题采集（按照信息需求的主题采集）与站点采集两种模式（给定网址列表的站内定点采集功能）。可帮助用户快速获取海量信息。

2. 文档转换

对 doc、excel、pdf 与 ppt 等多种主流文档格式，进行文本信息格式转换，信息抽取准确率极高，效率达到大数据处理的要求。

3. 新词发现

新词发现能从文本中挖掘出具有内涵新词、新概念，用户可以用于专业词典的编撰，还可以进一步编辑标注，导入分词词典中，提高分词系统的准确度，并适应新的语言变化。

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。

3. 批量分词

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。可在分析过程中，导入用户定义的词典。

5. 语言统计

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。

6. 文本聚类

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。

7. 文本分类

针对事先指定的规则和示例样本，系统自动从海量文档中识别并训练分类。NLPIR 深度文本分类，可以用于新闻分类、简历分类、邮件分类、办公文档分类、区域分类等诸多方面。

8. 摘要实体

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。实体提取能够对单篇或多篇文章，自动提炼出内容摘要，抽取人名、地名、机构名、时间及主题关键词；方便用户快速浏览文本内容。

9. 智能过滤

对文本内容的语义智能过滤审查，内置国内最全词库，智能识别多种变种：形变、音变、繁简等多种变形，语义精准排歧。

10. 情感分析

情感分析，针对事先指定的分析对象，系统自动分析海量文档的情感倾向：情感极性及情感值测量，并在原文中给出正负面的得分和句子样例。

11. 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。

12. 全文检索

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。

13. 编码转换

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，同时将繁体 BIG5 和繁体 GBK 进行繁简转化。

二、文件下载与说明

2.1 文件下载

GitHub 下载地址:

https://github.com/NLPIR-team/NLPIR/tree/master/NLPIR-Parser

下载教程参见附录。【有可能国内访问国外网址受限】

注：用户在 github 上下载 NLPIR-Parser 文件时需要专门的下载工具，建议使用 svn 工具下载文件。

2.2 文件说明

NLPIR-Paser 文件目录如下：

图 2.1 文件目录

文件说明：

├─bin-win32　　Windows 32bit 环境下的可执行程序和库文件，也可运行于 Win64；点击 NLPIR-Parser.exe 即可运行。
│　└─output　　运行结果存放路径
├─bin-win64　　Windows 64bit 环境下的可执行程序和库文件；点击 NLPIR-Parser.exe 即可运行。
│　└─output　　运行结果存放路径
├─Data　　整个系统运行需要的数据文件
│　├─Cluster　　聚类系统运行需要的数据文件
│　│　└─Data
│　├─DeepClassifier　　机器学习分类运行需要的数据文件
│　├─English　　英语处理需要的数据文件
│　├─JZSearch　　JZSearch 精准语义搜索引擎处理需要的数据文件
│　├─KeyScanner　　JZSearch 精准语义搜索引擎处理需要的数据文件
│　├─RedupRemover　　去重需要的数据文件

```
│   └─SentimentNew                情感分析需要的数据文件
│       └─Data
│           └─English
├─doc                             NLPIRParser 使用手册与各模块接口文档文件
│   └─大数据组件接口文档
│       ├─Classifier
│       ├─Cluster
│       ├─DocExtractor
│       ├─DupRemove
│       ├─JZSearch
│       ├─KeyExtract
│       ├─LJSentimentAnalysis
│       ├─Summary
│       └─WordFreq
├─演示语料                                    NLPIR-Parser 提供
的测试语料，可以自行替换
├─编码转换测试文本                            NLPIR-Parser 提供的编码
转换测试语料，可以自行替换
└─训练分类用文本                              NLPIRParser 提供
的分类训练语料，可以自行替换
        ├─交通
        ├─体育
        ├─军事
        ├─政治
        ├─教育
        ├─经济
        └─艺术
```

1. NLPIR-Parser.exe 可执行文件，本版本为共享版本（只能处理 200 个文件，总量不超过 500KB 纯文本），大规模语料处理需要购买正式版

2. 演示语料，用户可替换，必须为文本文件，如果为 GBK 以外的编码，必须先进行编码识别与转换后方可进行其他操作。

3. 各种 dll 为各组件的调用接口，本演示程序全部基于已有的调用接口实现；

三、各个功能操作指南

首先，启动程序。

用户需要点击 C:\Users\Administrator\Desktop\NLPIR-paser/bin-win64/ 路径下的 NLPIR-Parser.exe 程序，即可打开软件，平台界面如下：

图 3.1 NLPIR 大数据语义智能分析平台界面

然后，平台界面介绍。

平台包括三大模块：“功能导航”（点击功能名称即可完成功能切换）、“功能操作”和“基础功能”（使用手册、联系我们与退出）。

平台的十三大功能（由左至右）：精准采集，文档转换、新词发现、批量分词、语言统计、文本聚类、文本分类、摘要实体、智能过滤、情感分析、文档去重、全文检索和编码转换。用户可根据需要选

择使用。

点击“使用手册”，即可打开平台使用手册文档，帮助用户了解平台，指导用户进行各项功能操作。

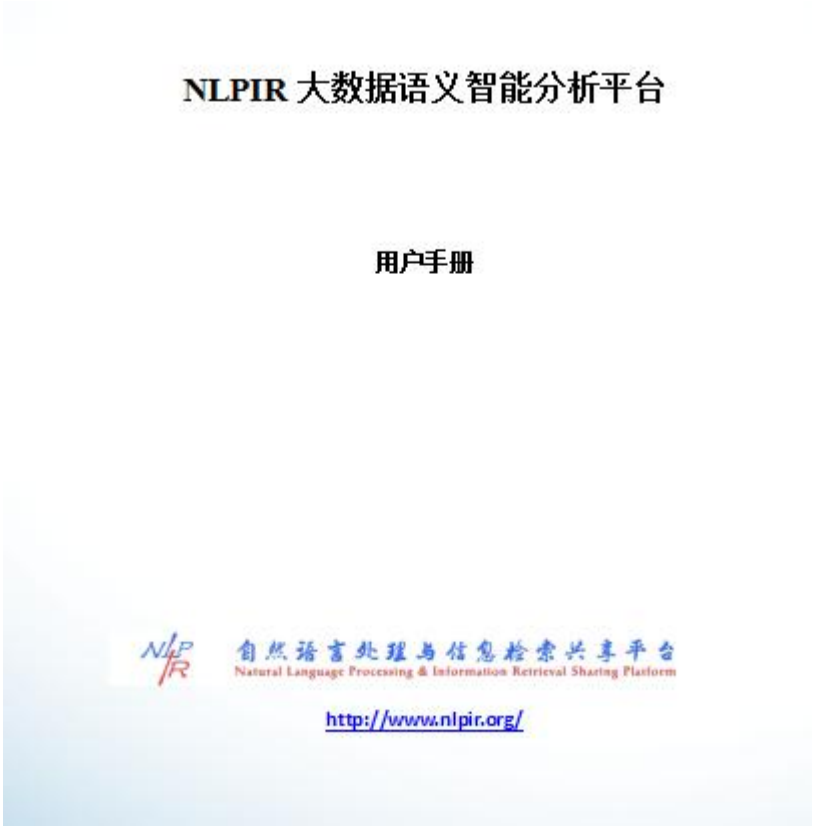


图 3.2 使用手册

点击“联系我们”，联系信息框弹出，用户可查看咨询。

图 3.3 联系我们

注：平台内置测试语料，但用户仍可定义自己的语料（新建文件夹放入自己的语料）。

3.1 精准采集

用户点击“精准采集”（第一个功能模块），进入精准采集模块。

精准采集功能可实现对境内外互联网海量信息的实时精准采集。精准采集包括主题采集（按照信息需求的主题采集）与站点采集两种模式（给定网址列表的站内定点采集功能）。可帮助用户快速获取海量信息。用户可自定义采集模式、采集时间区域、采集主题站点与采集存储。

- 主题采集

按照给定的关键词或主题词进行信息采集。

Step1：定义主题词。选择“主题采集”，在采集模块输入关键词，例如“一带一路”、“航空母舰”与“区块链”等三个主题，系统将按此关键词进行主题采集，获取主题相关的主流新闻报道、BBS与博客等内容。

Step2：采集设置。用户可自定义采集时间（系统默认采集时段为近 3 年，用户可在此时间段内自定义自己的采集时间）。选择采集区域“境内采集”（或境外采集，需要启动翻墙措施方可使用）。

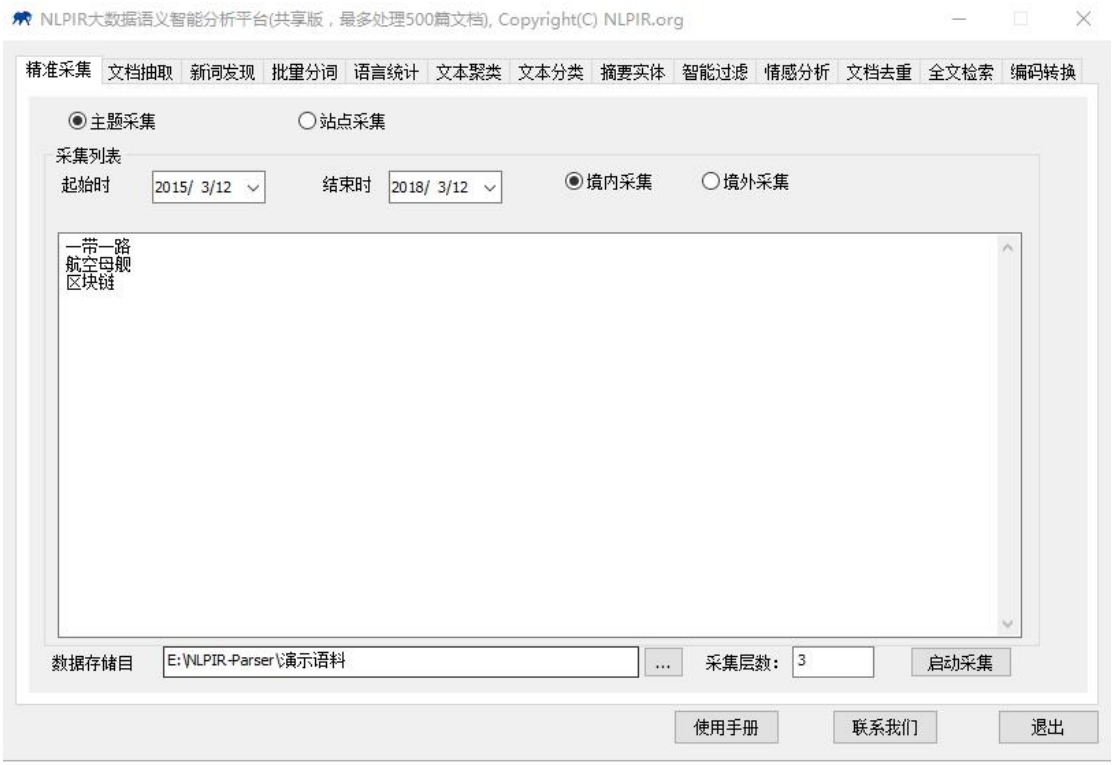


图 3.4 主题采集

Step3：定义采集存储。选择语料存放路径（默认路径：NLPIR-Parser\演示语料）。点击“启动采集”，系统弹出信息采集窗口，开始采集信息，用户了解信息采集过程与详情。

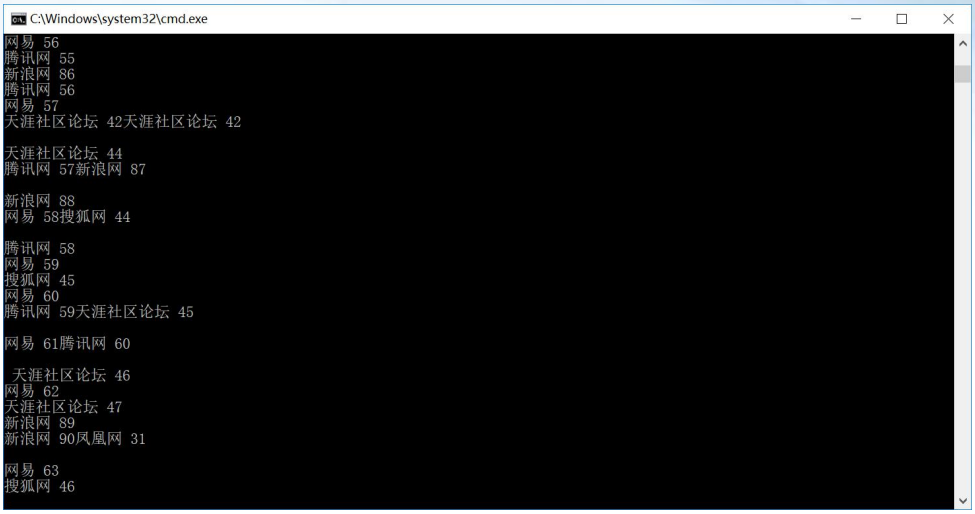


图 3.5 信息采集过程

系统提示采集全部结束，用户可关闭此窗口。

图 3.6 采集结束

采集完成以后，用户可查看采集结果（默认：\NLPIR-Parser\演示语料），采集结果文件夹包括：境内新闻、境外新闻与 bbs 以及通用采集。其中的子目录中的数字指的是文章发布的日期，如 境内新闻 20180301：指的是 2018 年 3 月 1 日的境内新闻。

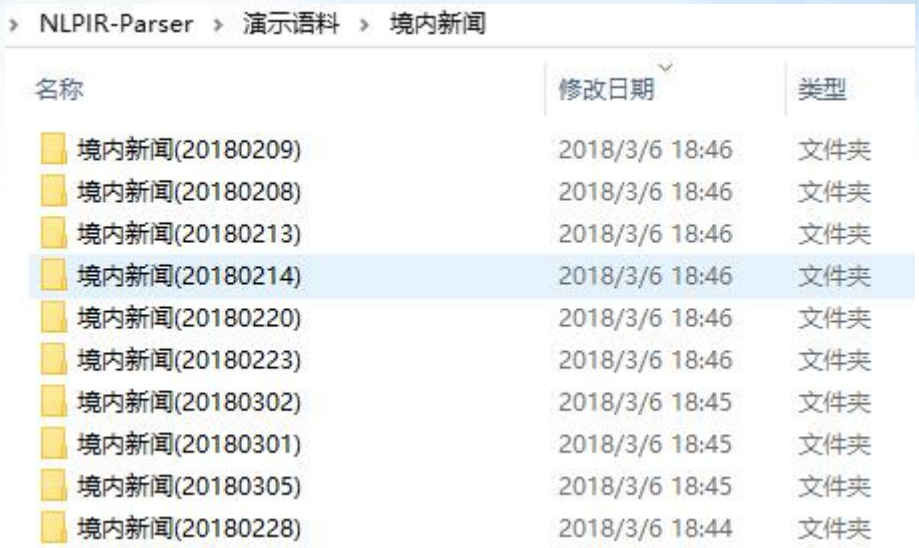


图 3.7 采集结果文件

- 站点采集

站点采集指的是按照给定的网址，在该网址内部垂直采集。

Step 1：选择“站点采集”，输入站点地址，例如：http://news.sina.com.cn/。

Step 2：定义采集时间、区域与采集结果存放路径，点击“启动采集”，系统开始采集任务。

系统将站点采集的结果保存在“通用采集”文件夹中，文件目录：\NLPIR-Parser\演示语料。

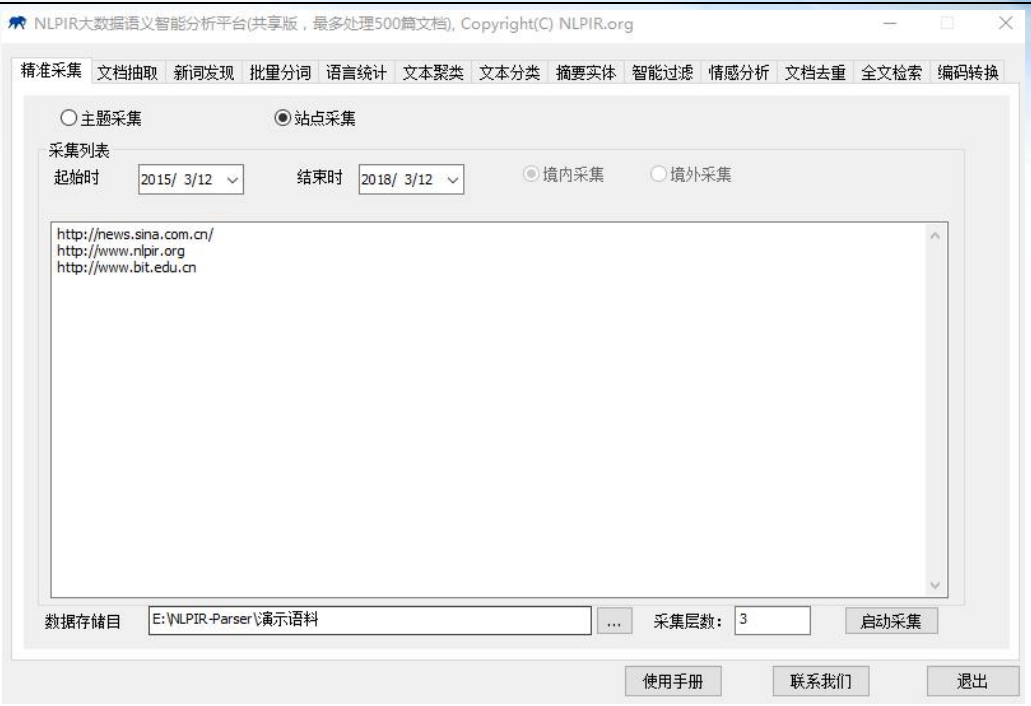


图 3.8 站点采集

站点采集程序启动后，显示采集过程与详情，采集完毕后窗口自动关闭。

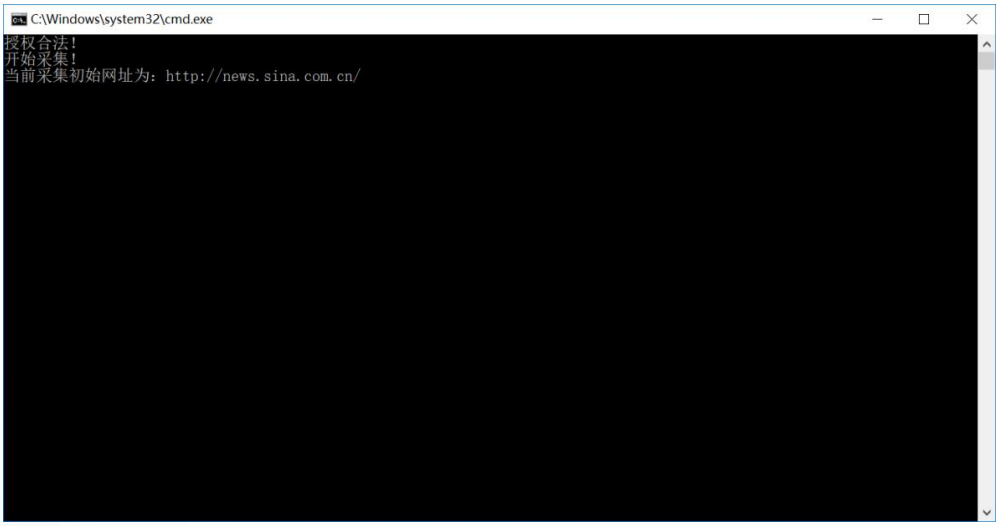


图 3.9 站点采集

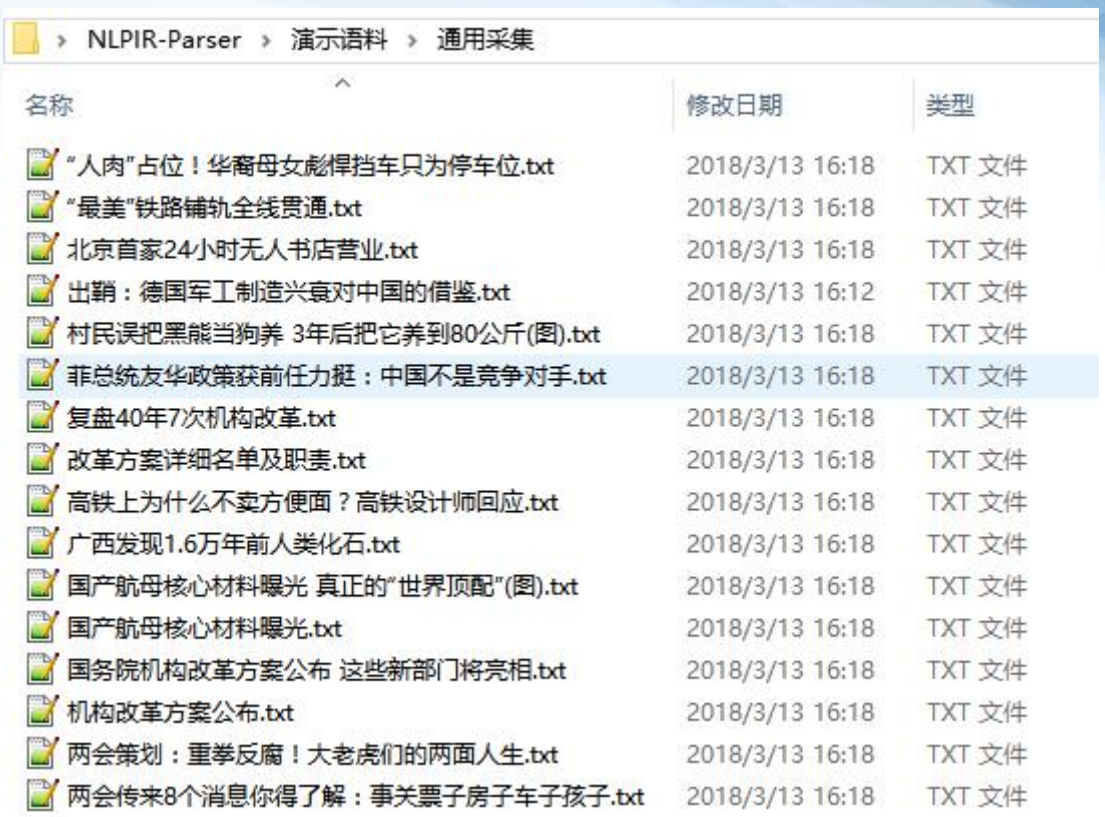


图 3.10 站点采集结果文件

3.2 文档转换

用户点击功能导航栏“文档转换”，系统进入“文档转换”模块。

文档转换功能对 doc、excel、pdf 与 ppt 等多种主流文档格式，进行文本信息抽取，信息抽取准确率极高，达到大数据处理的要求。

Step1：选择待处理文件。在“文档所在路径”输入框中输入或选择需要抽取的文档文件（用户可选择电脑中的任何文档），例如:\NLPIR-Parser\文档转换。

Step2：定义结果存储。在“结果存放路径”选择文档转换完成文件存放的地址路径，例如:\NLPIR-Parser\文档转换。

Step3：点击“文档解析抽取”，系统弹出文档转换处理窗口，

开始文档转换。

图 3.11 文档转换

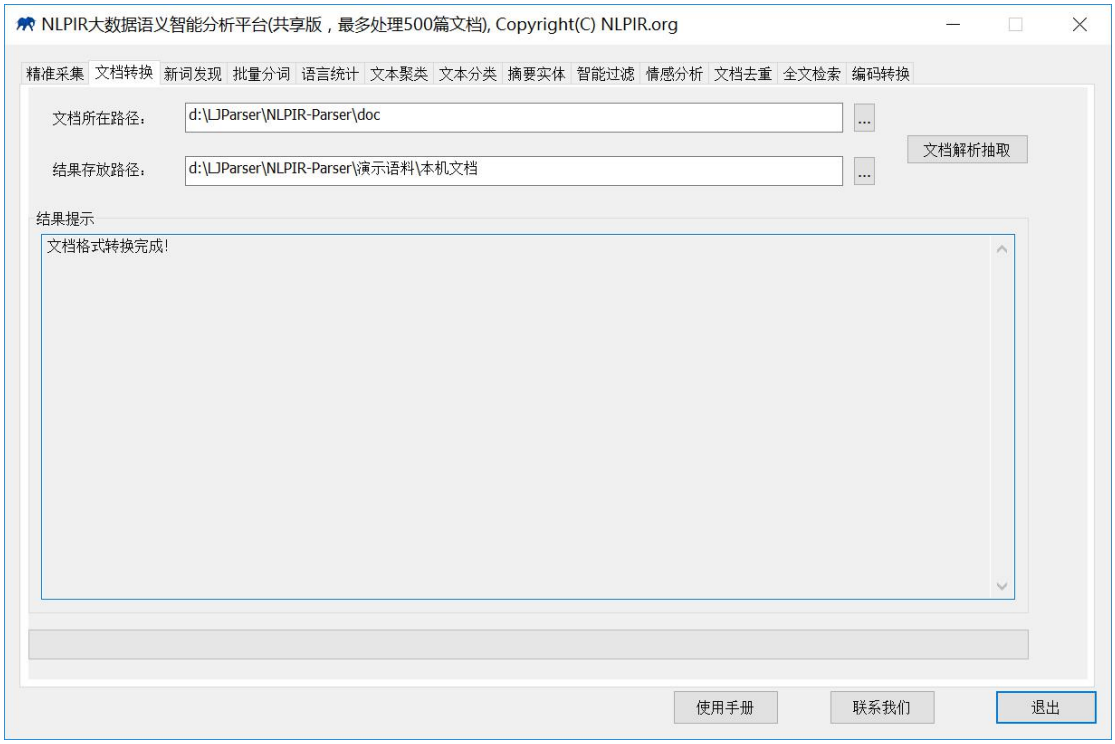


图 3.12 文档转换完成

文档转换结果文件会自动打开（用户也可打开文档转换结果存储目录查看结果文件），抽取完成的文档以文本文件的格式保存。通过结果文件与文件原文的对比，可发现文件抽取具有非常高的准确率。

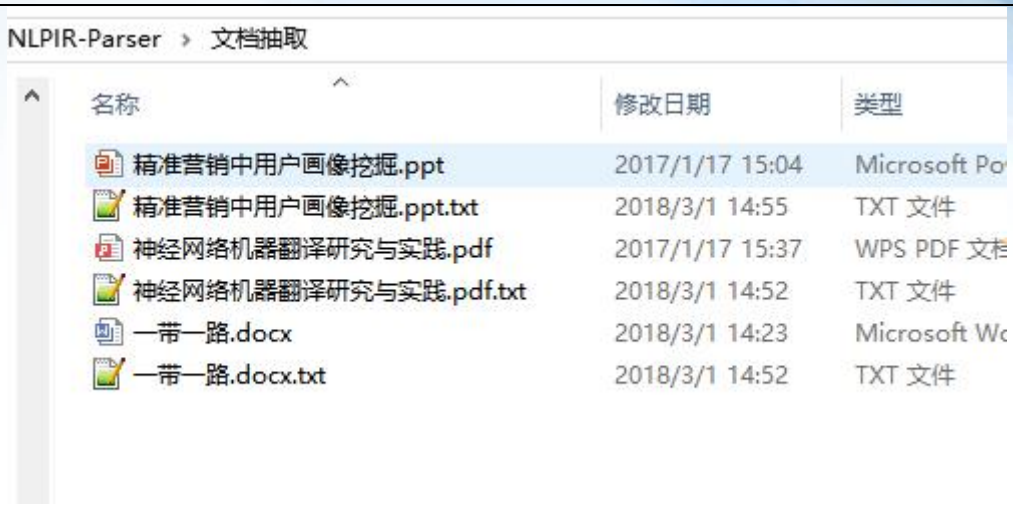


图 3.13 文档转换结果文件

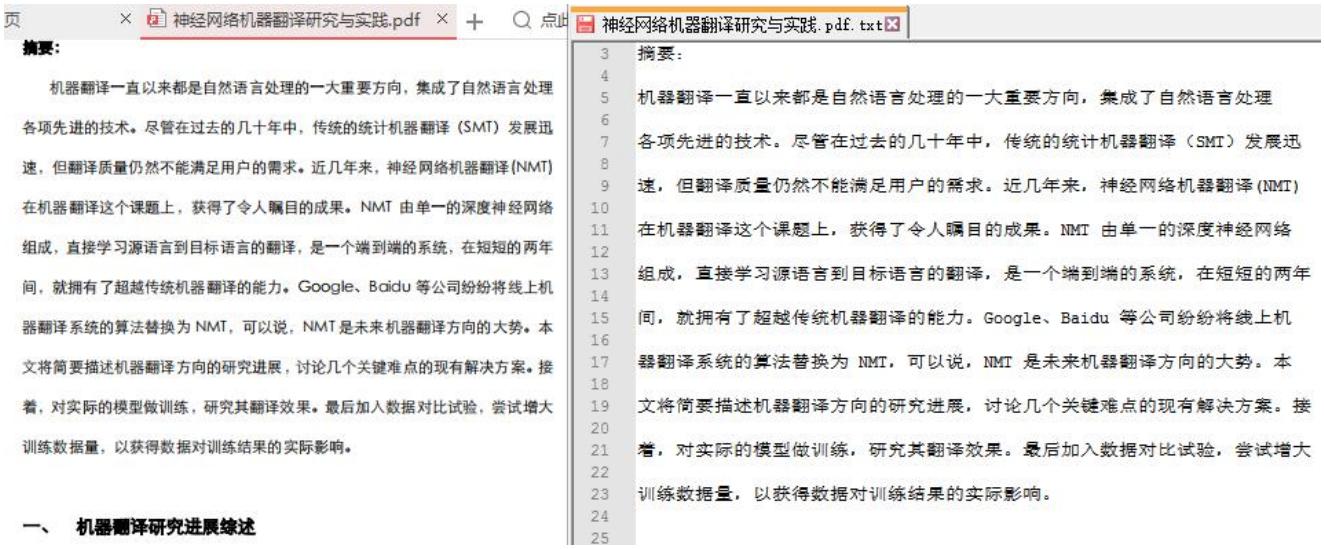


图 3.14 文档转换效果对比

3.3 新词、关键词提取

用户点击“新词发现”，系统切换进入“新词发现”功能模块。新词发现模块包括新词发现与关键词抽取两个功能。

3.3.1 新词发现

新词发现能从文本中挖掘出具有内涵新词、新概念，用户可以用

于专业词典的编撰，还可以进一步编辑标注，导入分词词典中，提高分词系统的准确度，并适应新的语言变化。

Step1：选择语料源。在“语料源所在路径”输入框中输入或选择需要提取新词的语料所在路径，用户需要事先定义并存放需要处理的语料源，比如：\NLPIR-Parser\十九大报告。在“新词存放地址”选择结果文件的存储路径。

如果“语料源所在路径”是通过选择文件夹方式确定，则系统会自动指定“新词存放地址”为\NLPIR-Parser\output\关键词分析\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”。

Step2：点击“新词提取”，系统开始进行发现新词任务。

新词提取结果输出到“新词存放地址”所指定的文件路径，也会输出到结果提示框中。

图 3.15 新词提取

新词提取完成后，系统会自动打开结果文件。

NewTermlist.txt

	Word	Part-Of-Speech	Weight	Frequency
2	中国特色社会主义	n_new	68.48	70
3	中华民族伟大复兴	n_new	24.76	27
4	社会主义核心价值观	n_new	24.52	6
5	人民军队	n_new	21.94	12
6	团结带领人民	n_new	18.48	7
7	坚定理想信念	n_new	17.71	3
8	中国梦	n_new	17.01	13
9	人类命运共同体	n_new	15.15	6
10	实体经济	n_new	14.72	5
11	美好生活需要	n_new	14.65	6
12	全体人民共同	n_new	14.61	7
13	发展道路	n_new	12.68	8
14	文化自信	n_new	12.68	6
15	一百年	n_new	12.39	6
16	伟大工程	n_new	10.91	6
17	医疗卫生	n_new	10.70	4
18	保险制度	n_new	10.69	4
19	不懈奋斗	n_new	10.69	4

图 3.16 新词结果文件

NewTermlist 是新词提取结果文件。新词提取内容包括：词语、词性、权重和词频统计。

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。

3.3.2 关键词提取

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。

Step1：选择语料源文件夹，以十九大报告为例：\NLPIR-Parser\十九大报告。

Step2：点击“关键词提取”，系统即可开始进行关键词提取。

关键词存放路径默认为：\NLPIR-Parser\output\关键词分析\Keylist.txt。

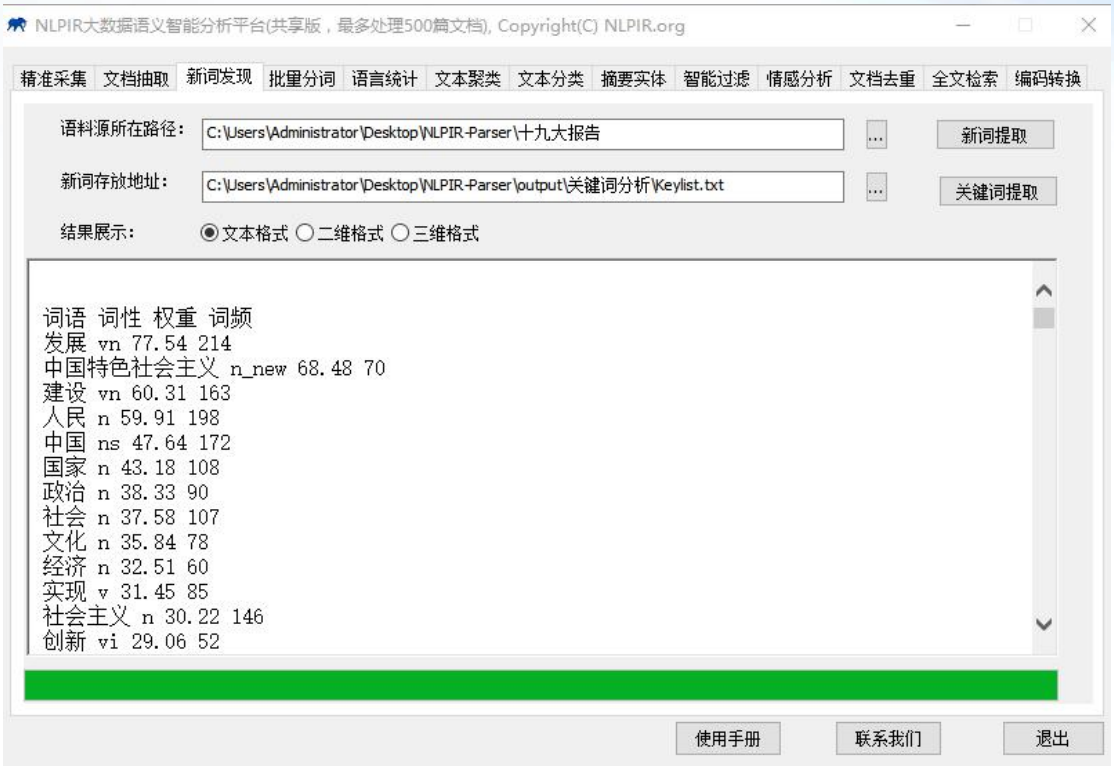


图 3.17 关键词提取

关键词提取完成以后，系统自动打开结果文件 keylist。

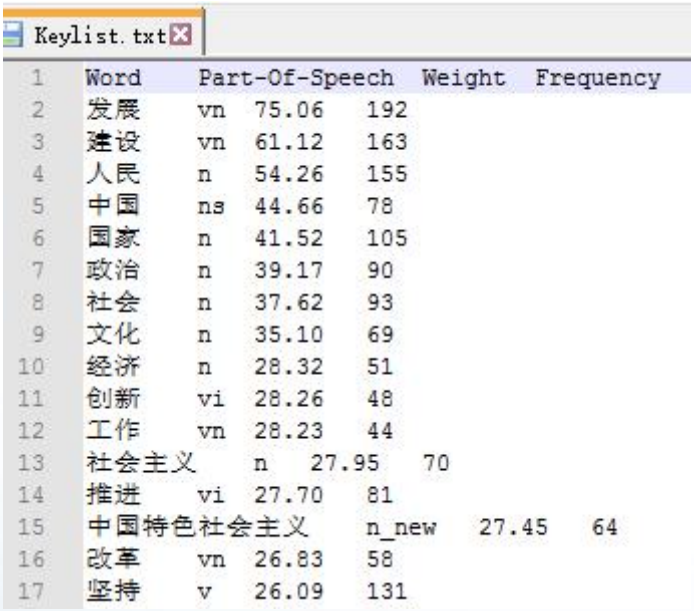


图 3.18 keylist

关键词分析内容包括：词语、词性、权重和词频统计。系统默认

词汇以权重值高低排序。

3.3.3 可视化展示

系统可实现对于新词、关键词提取结果的高维可视化展示，可视化形式有三种：文本格式、二维格式与三维格式。用户可根据需要直接使用，无需再次设计美化。

- 文本格式：以文本的形式展示提取结果

图 3.19 文本格式

- 二维格式：top42 词汇的词云形式展示效果，非常直观。

图 3. 20 二维格式

- 三维格式：top20 词汇的三维动态展示，简洁美观。

图 3. 21 三维格式

3.4 批量分词

用户点击“批量分词”，系统切换进入“批量分词”功能模块。

批量分词能够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。

1）导入用户词典

用户可自定义自己的词典，并将词典导入，分词过程将会融合用户的自定义词典。例如，将十九大报告提取新词作为用户新词导入。

Step1：在“新词存放地点”指定新词文件，选择新词提取 new termlist 文件（默认）。点击“编辑”，系统弹出词典文件，用户可对新词文件进行编辑（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）。编辑完成后保存新词文件并关闭。

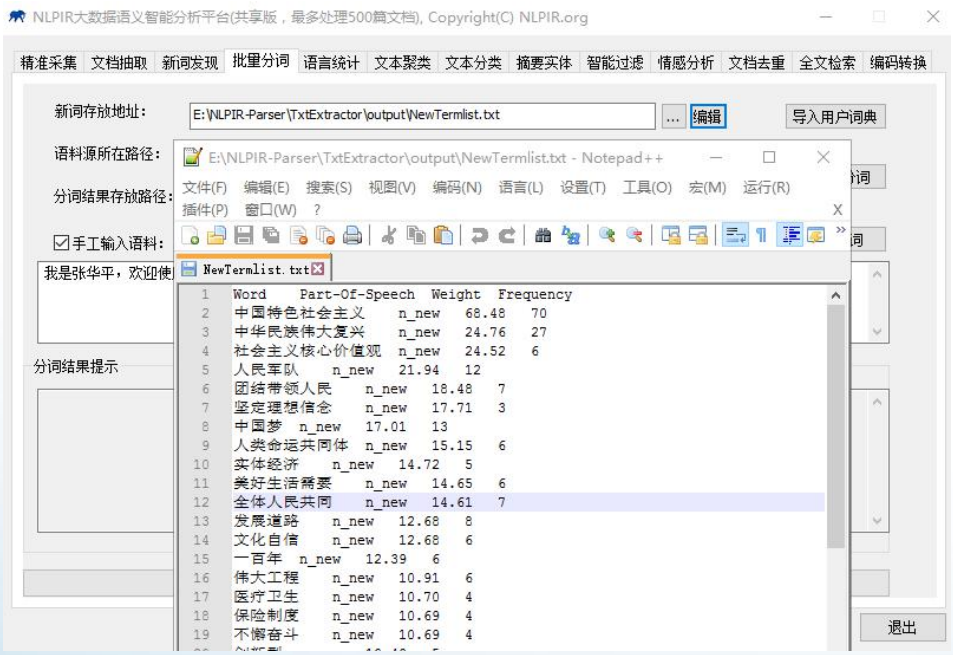


图 3. 22 编辑用户词典

Step2：点击“导入用户词典”，系统开始导入用户词典，并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。

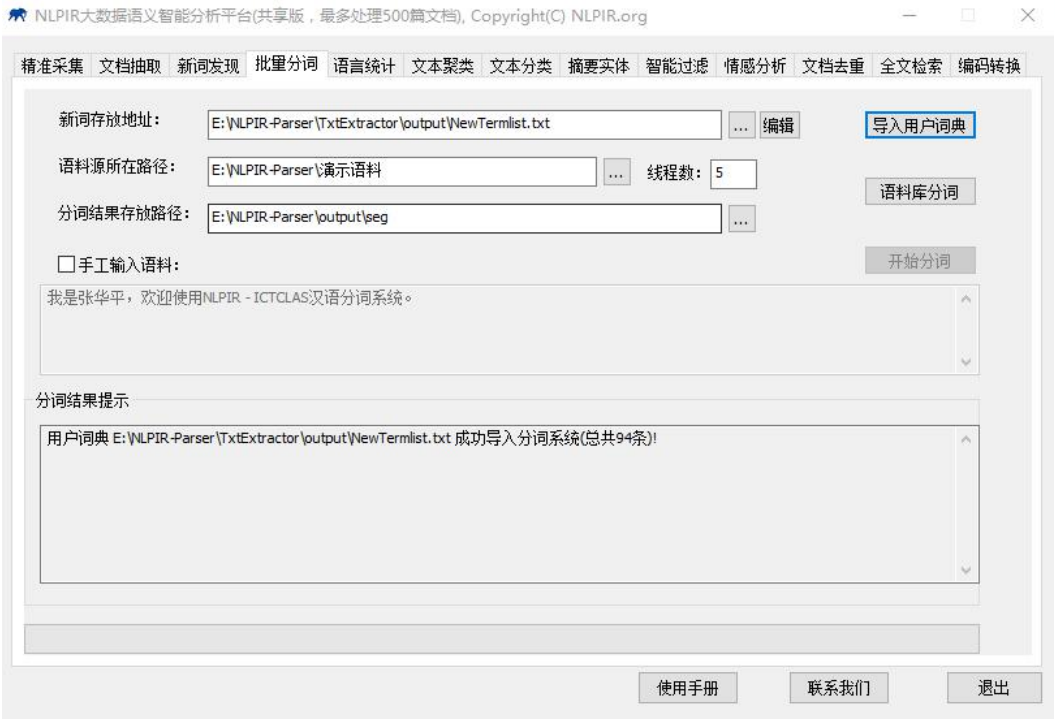


图 3. 23 导入用户词典

2）批量分词

Step1：选择待分词文件，定义“语料源所在路径”（以十九大报告为例），文件路径：NLPIR-Parser\十九大报告。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。

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：NLPIR-Parser\bin-win64\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。

Step2：点击“语料库分词”，系统开始分词与词性标注。系统会在完成时自动为用户打开分词结果目录。

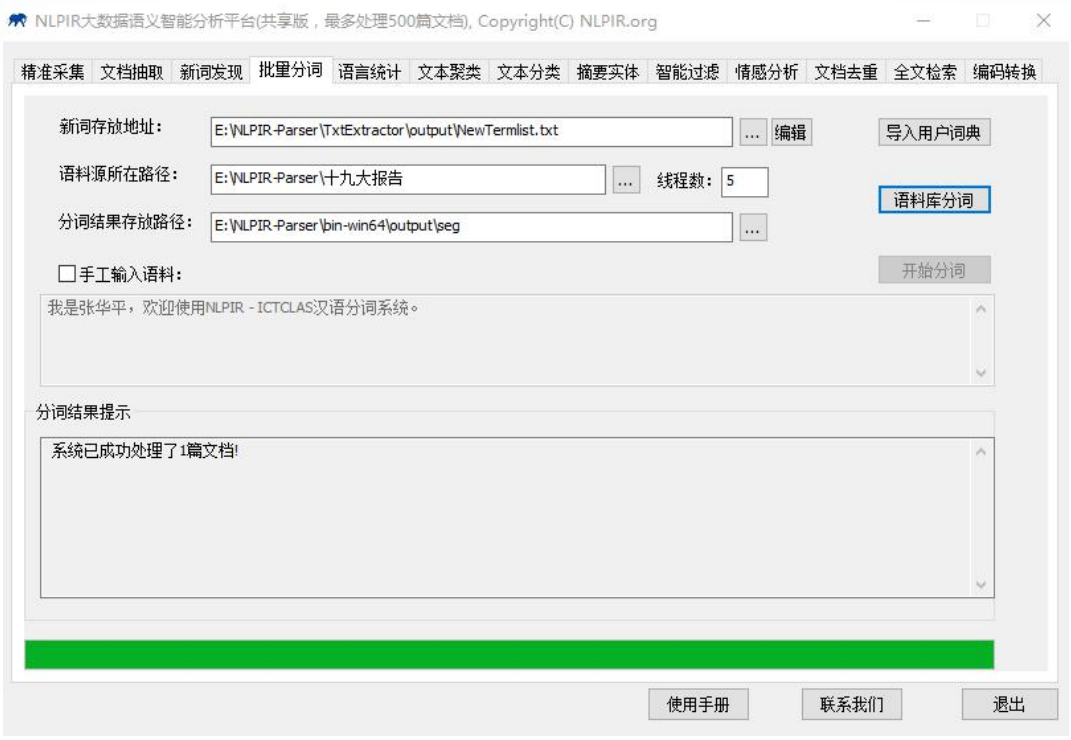


图 3. 24 分词成功

分词结果文件地址：NLPIR-Parser\bin-win64\output\seg。

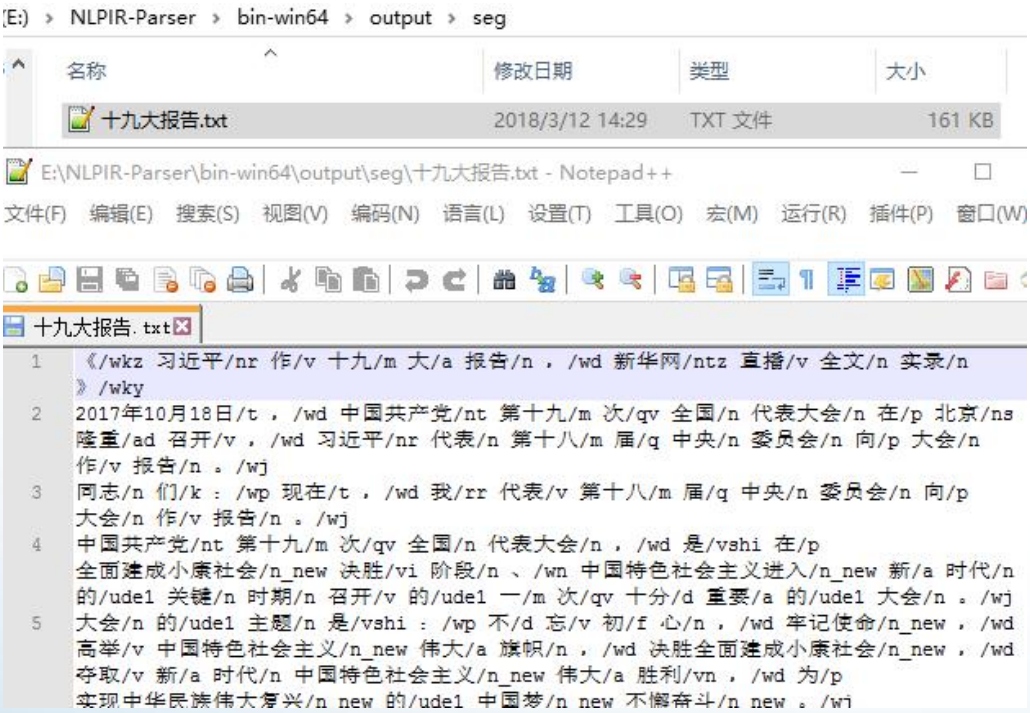


图 3. 25 分词结果文件

3）手动输入语料分词

系统支持用户手动输入语料进行分词。

选择“手动输入语料”，输入语料，点击“开始分词”，系统进行分词。分词结果会呈现在分词结果提示框中。

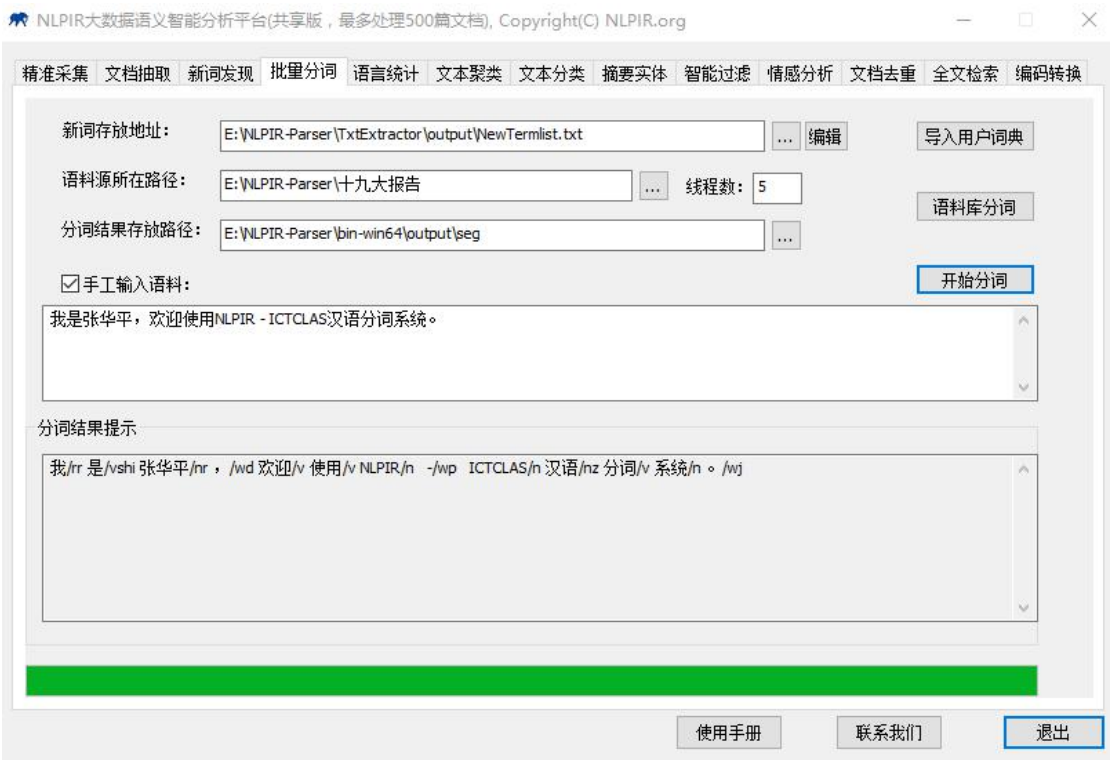


图 3. 26 手动输入分词

“我是张华平，欢迎使用 NLPIR - ICTCLAS 汉语分词系统”的分词结果为：我/rr 是/vshi 张华平/nr ，/wd 欢迎/v 使用/v NLPIR/n -/wp ICTCLAS/n 汉语/nz 分词/v 系统/n 。/wj

3.5 语言统计

语言统计功能针对切分标注结果，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。

用户点击“语言统计”，进入系统语言统计功能模块。

Step1：选择分词结果文件作为语言统计的输入文件 NLPIR-Parser\bin-win64\output\seg。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。

Step2：点击“词频统计与翻译”，系统开始统计词频、共现词对频率等信息。

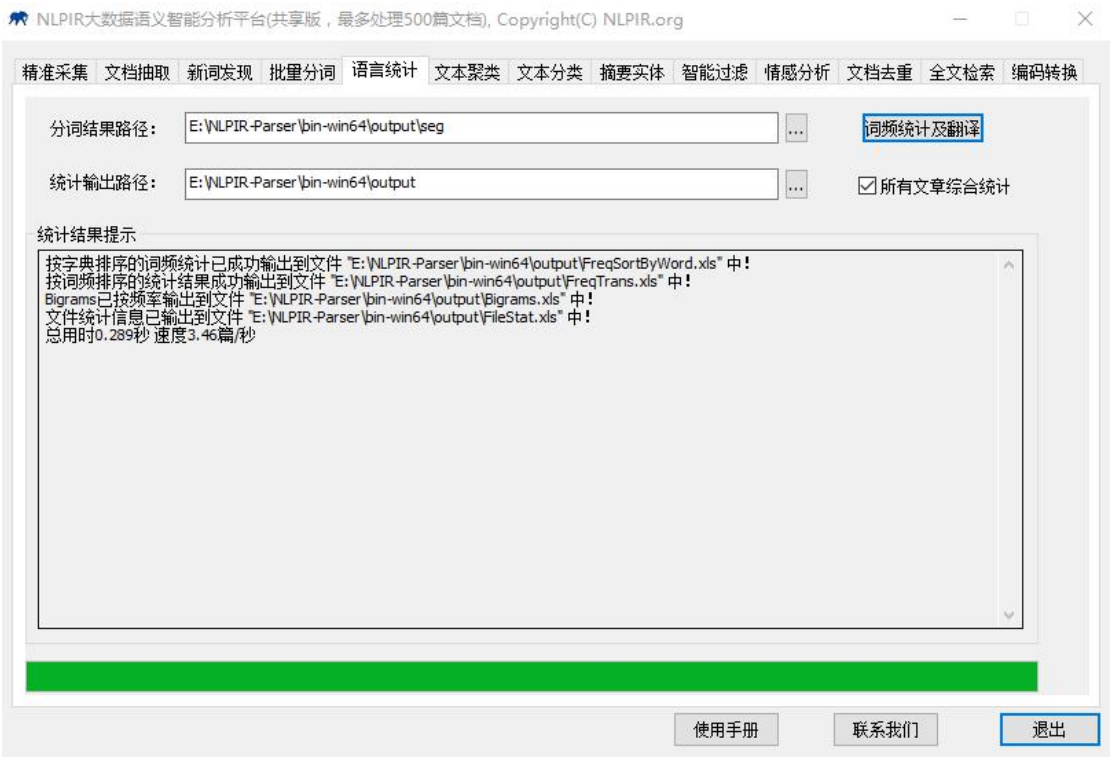
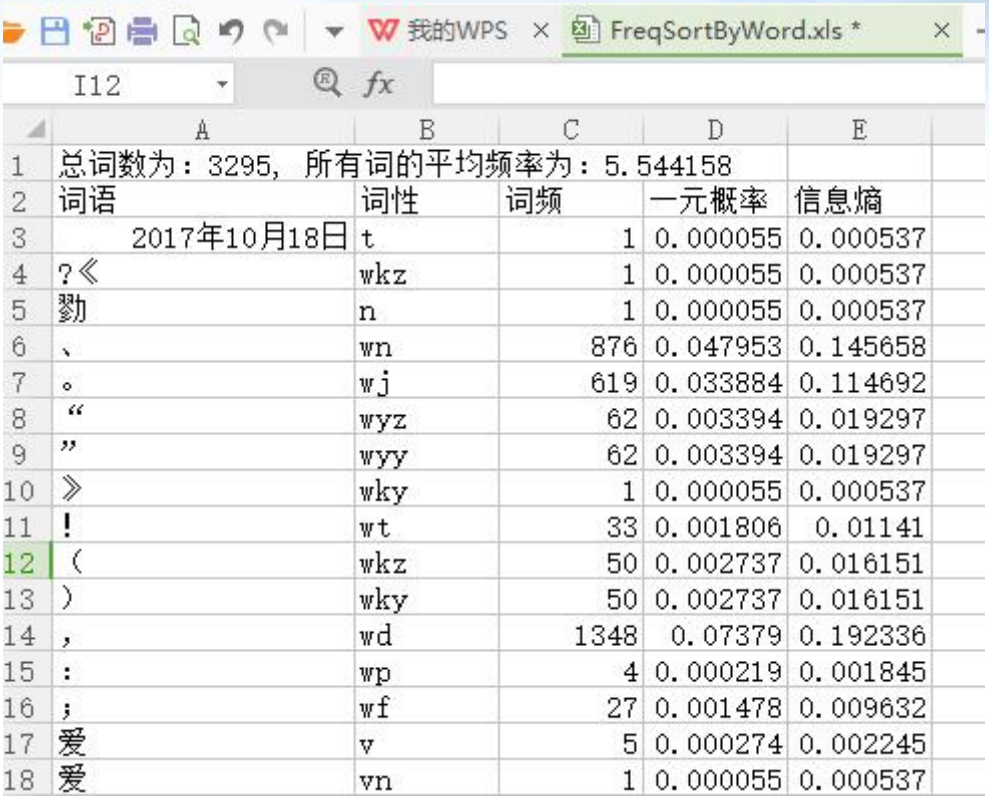


图 3. 27 词频统计

词频统计及翻译分析结果有四个输出文件，分别为：

✧ 按字典排序的词频统计文件

"E:\NLPIR-Parser\bin-win64\output\FreqSortByWord.xls"



我的WPS | FreqSortByWord.xls *

I12

	A	B	C	D	E
1	总词数为：3295， 所有词的平均频率为：5.544158				
2	词语	词性	词频	一元概率	信息熵
3	2017年10月18日	t	1	0.000055	0.000537
4	?《	wkz	1	0.000055	0.000537
5	勠	n	1	0.000055	0.000537
6	、	wn	876	0.047953	0.145658
7	。	wj	619	0.033884	0.114692
8	“	wyz	62	0.003394	0.019297
9	”	wyy	62	0.003394	0.019297
10	》	wky	1	0.000055	0.000537
11	!	wt	33	0.001806	0.01141
12	（	wkz	50	0.002737	0.016151
13	）	wky	50	0.002737	0.016151
14	,	wd	1348	0.07379	0.192336
15	:	wp	4	0.000219	0.001845
16	;	wf	27	0.001478	0.009632
17	爱	v	5	0.000274	0.002245
18	爱	vn	1	0.000055	0.000537

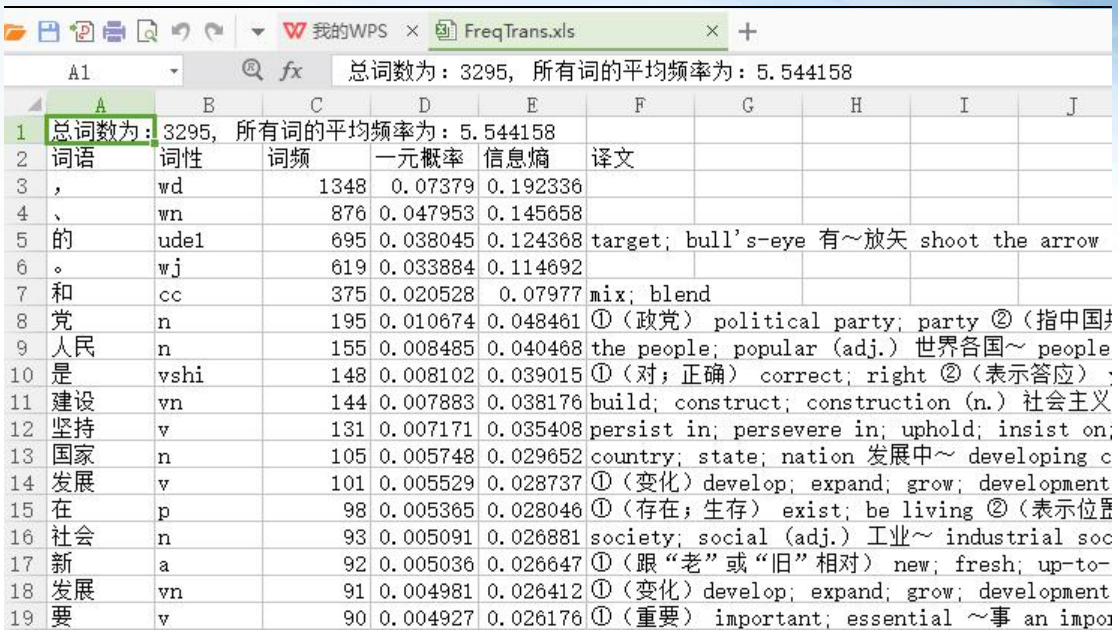
图 3.28 FreqSortByWord

按字典排序词频统计结果包括：词频统计结果（总词数与平均频率）、词语、词性、词频、一元概率与信心熵。其中，一元概率指的是单个词独立出现的概率，信息熵指的是该词包含的信息广度，其公式为： $H(X)=-\sum_{1}^{n}P(X)\log P(X)$ 。

✧ 按词频排序的统计结果文件

"E:\NLPIR-Parser\bin-win64\output\FreqTrans.xls"

按词频排序的统计内容如下，包括：词语、词性、词频、一元概率、信息熵与译文。



我的WPS | FreqTrans.xls

A1 | 总词数为：3295， 所有词的平均频率为：5.544158

	A	B	C	D	E	F
1	总词数为：	3295， 所有词的平均频率为： 5.544158				
2	词语	词性	词频	一元概率	信息熵	译文
3	，	wd	1348	0.07379	0.192336	
4	、	wn	876	0.047953	0.145658	
5	的	ude1	695	0.038045	0.124368	target; bull's-eye 有～放矢 shoot the arrow
6	。	wj	619	0.033884	0.114692	
7	和	cc	375	0.020528	0.07977	mix; blend
8	党	n	195	0.010674	0.048461	①（政党） political party; party ②（指中国共
9	人民	n	155	0.008485	0.040468	the people; popular (adj.) 世界各国～ people
10	是	vshi	148	0.008102	0.039015	①（对；正确） correct; right ②（表示答应）
11	建设	vn	144	0.007883	0.038176	build; construct; construction (n.) 社会主义
12	坚持	v	131	0.007171	0.035408	persist in; persevere in; uphold; insist on;
13	国家	n	105	0.005748	0.029652	country; state; nation 发展中～ developing c
14	发展	v	101	0.005529	0.028737	①（变化）develop; expand; grow; development
15	在	p	98	0.005365	0.028046	①（存在；生存） exist; be living ②（表示位置
16	社会	n	93	0.005091	0.026881	society; social (adj.) 工业～ industrial soc
17	新	a	92	0.005036	0.026647	①（跟“老”或“旧”相对） new; fresh; up-to-
18	发展	vn	91	0.004981	0.026412	①（变化）develop; expand; grow; development
19	要	v	90	0.004927	0.026176	①（重要） important; essential ～事 an impor

图 3.29 FreqTrans.xls

“党”的译文：①（政党） political party; party ②（指中国共产党）the Party (the Communist Party of China) 入～ join the Party 整～ Party consolidation ③（集团） clique; faction; gang 死～ sworn follower ④（偏袒） be partial to; take sides with ⑤（亲族） kinsfolk; relatives 父～ father's kinsfolk。

✧ Bigrams 输出文件

"E:\NLPIR-Parser\bin-win64\output\Bigrams.xls"

Bigrams 结果包括：二元词对总数、前一个词、后一个词、共现频次与二元词对信息熵。共现频次指的是两个词以前后顺序同时出现的频率，二元词对信息熵指的是这两个词包含的信息广度。如下：“党”和“的”以“党的”共现形式出现了 87 词，频率为 0.446154，其信息熵值为 0.025465。

我的WPS | Bigrams.xls

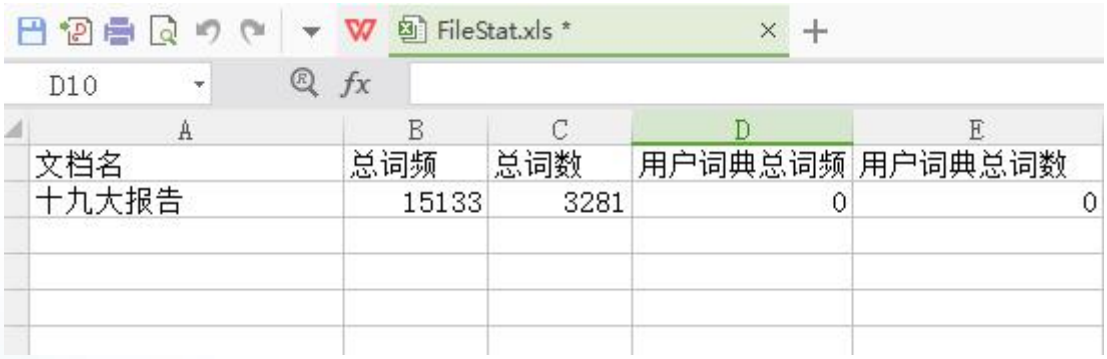
G19

	A	B	C	D	E	F
1	二元词对总数为：11926					
2	前一个词	后一个词	共现频次	二元概率	二元词对信息熵	
3	党	的	87	0.446154	0.025465	
4	。	（	50	0.080775	0.016151	
5	，	坚持	43	0.031899	0.014245	
6	，	是	38	0.02819	0.012846	
7	。	要	38	0.061389	0.012846	
8	新	时代	35	0.380435	0.011989	
9	建设	，	34	0.236111	0.0117	
10	，	推动	31	0.022997	0.010825	
11	。	我们	30	0.048465	0.010529	
12	。	加强	29	0.04685	0.010232	
13	体系	，	28	0.363636	0.009933	
14	我们	党	28	0.4375	0.009933	
15	，	加强	26	0.019288	0.009329	
16	制度	，	26	0.292135	0.009329	
17	，	必须	25	0.018546	0.009024	
18	，	在	24	0.017804	0.008717	

图 3.30 Bigrams.xls

✧ 文件统计信息输出文件

"E:\NLPIR-Parser\bin-win64\output\FileStat.xls"



FileStat.xls *

D10

	A	B	C	D	E
1	文档名	总词频	总词数	用户词典总词频	用户词典总词数
2	十九大报告	15133	3281	0	0

图 3.31 FileStat.xls

文件统计结果包括：文档名、总词频、总词数、用户词典总词频与用户词典总词数。

3.6 文本聚类

文本聚类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。文本聚类适用于长文本和短信、微博等短文本的热点分析。

用户点击“文本聚类”，进入系统文本聚类功能模块。

Step1：“在语料源所在路径”选择语料源文件夹（乐视相关新闻）。

Step2：设置聚类参数（最大类数目与类中最大文档数），点击“聚类”，系统进行聚类。聚类结果自动保存在 output 目录下：\NLPIR-Parser\output\聚类结果，并于结果提示框呈现聚类结果。如下所示：

共有 24 个聚类类别，第一类别主题词汇：贾跃亭、甘薇、孙宏斌、乐视影业、刘江峰、乐漾影业、乐视手机、乐视商城、估值，第一类别文档总数 116 篇。

图 3. 32 聚类

聚类结果文件有两种形式：网页和文件，都保存到本地文件夹：\NLPIR-Parser\output\聚类结果，文件夹按照文件数量排序，名称包含信息：文件数量与聚类特征词。

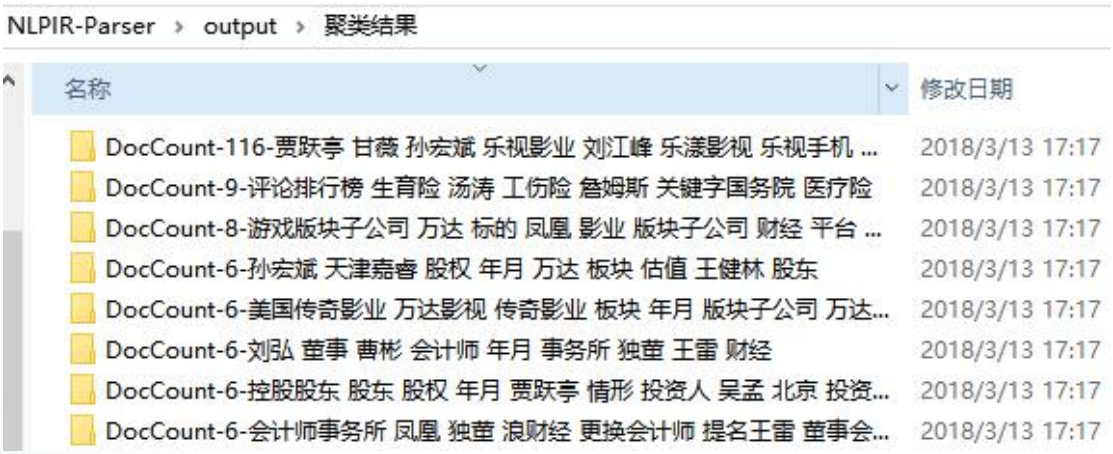


图 3. 33 聚类结果文件

用户可查看同属一个类别的多个文件。聚类详情文件名称包含：聚类特征词、媒体来源与新闻标题。

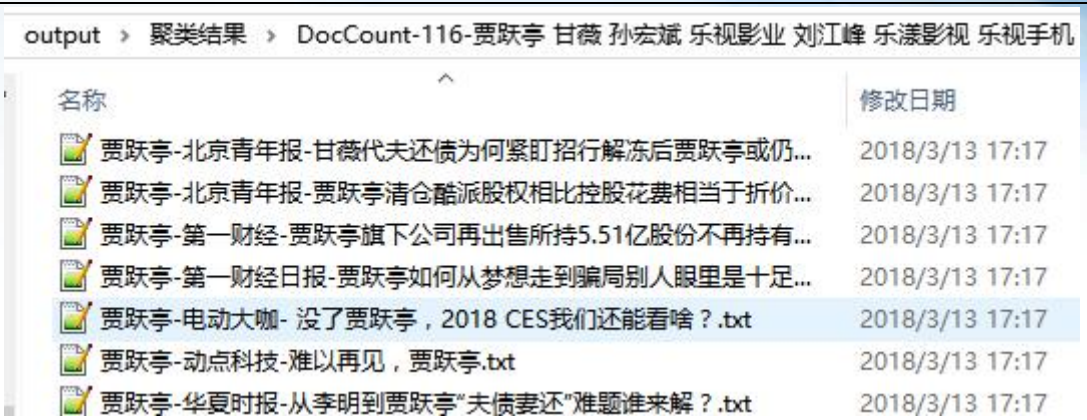


图 3. 34 聚类详情文件

3.7 文本分类

文本分类能够根据事先指定的规则和示例样本，自动从海量文档中识别并训练分类。NLPIR 深度文本分类，可以用于新闻分类、简历分类、邮件分类、办公文档分类、区域分类等诸多方面。此外还可以实现文本过滤，能够从大量文本中快速识别和过滤出符合特殊要求的信息，可应用于品牌报道监测、垃圾信息屏蔽、敏感信息审查等领域。

NLPIR 采用深度神经网络对分类体系进行了综合训练。演示平台目前训练的类别只是新闻的政治、经济、军事等。我们内置的算法支持类别自定义训练，该算法对常规文本的分类准确率较高，综合开放测试的 F 值接近 86%。

用户点击“文本分类”，进入系统文本分类功能模块。

文本分类有两种模式：专家规则分类与机器学习分类。

专家规则分类指的是根据事先人为制定的分类规则进行分类，比如“毒品”类别，我们可定义该类别的规则：“海洛因 快乐丸 罂粟 可卡因 Pethidine 摇头丸 K 粉”，系统会根据文本中出现规（词）

判定文本类别为：毒品。

机器学习分类是利用机器自动学习的能力，通过大量文本的训练，是系统具有分类的能力。比如我们准备军事、政治类别的大量语料，通过训练，机器自动学习类别特征，经过不断的语料训练，分类效果越来越精准。

➢ 专家规则分类

Step1:选择测试语料(分类测试语料为例)，点击“帮助”。

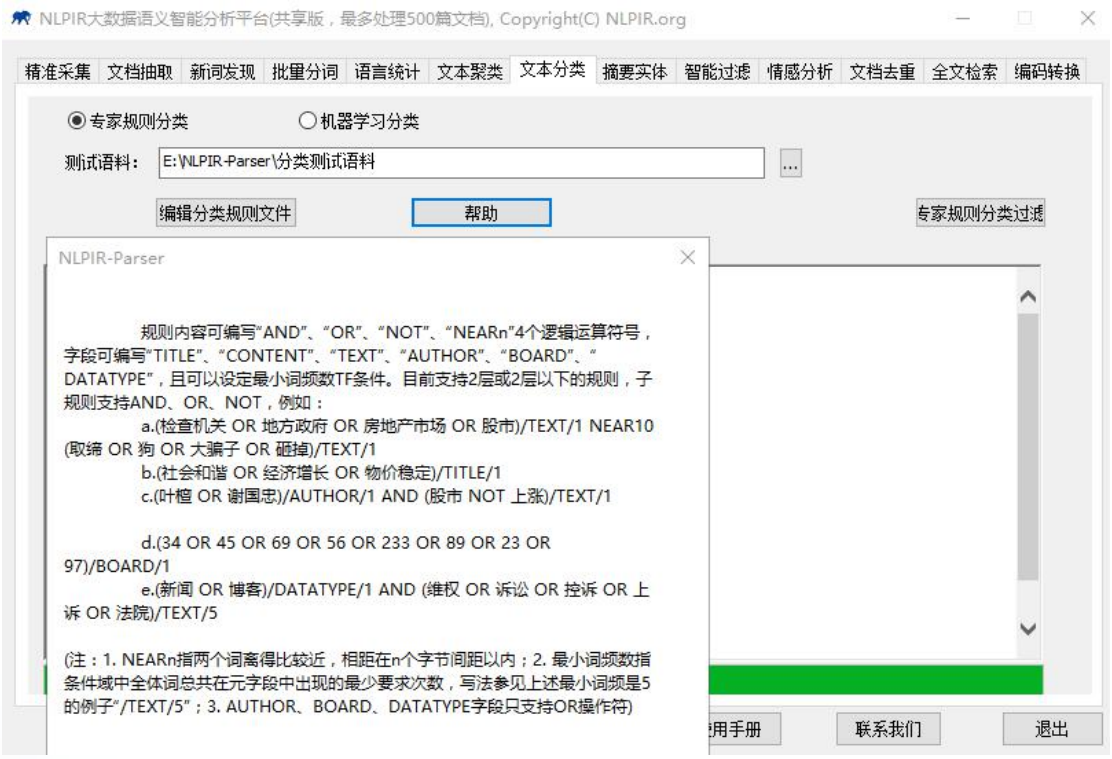


图 3. 35 分类规则“帮助”

“帮助”里面详细介绍了分类规则的书写格式。例如：

(检察机关 OR 地方政府 OR 房地产市场 OR 股市) /TEXT/1 NEAR10(取缔 OR 狗 OR 大骗子 OR 砸掉)/TEXT/1

表示：在原文中，这两组词（组内任意一词与另一组内任意一词）距离在 10 个字节（5 个中文字）以内，共现现频率至少为 1 次。

Step2:点击“编辑分类规则文件”，系统弹出规则分类文件。用户可自定义分类规则。系统有默认的 rulelist，用户可在此基础上添加、修改、删除分类的规则。用户编辑完分类规则文件需要保存。

图 3.36 编辑分类规则文件

Step3:点击“专家规则分类过滤”，系统进行分类分析。分类结果同样会呈现在结果提示框中，如下所示：两篇语料，一篇分类为政治类，一篇分类为经济类。

图 3.37 专家规则分类过滤

系统会将分类结果以网页和文件的同时自动保存至：\NLPIR-Parser\output\专家规则分类结果文件夹，并在分类结束后自动打开，用户可直接查看与利用分类结果。

图 3.38 分类过滤结果文件

➢ 机器学习分类

Step1:选择训练分类，点击“训练”按钮，系统进行类别特征的自学习；系统目前已有 7 大类分类训练语料，用户仍可自定义在在此基础上进行语料的更新或类别的定义。

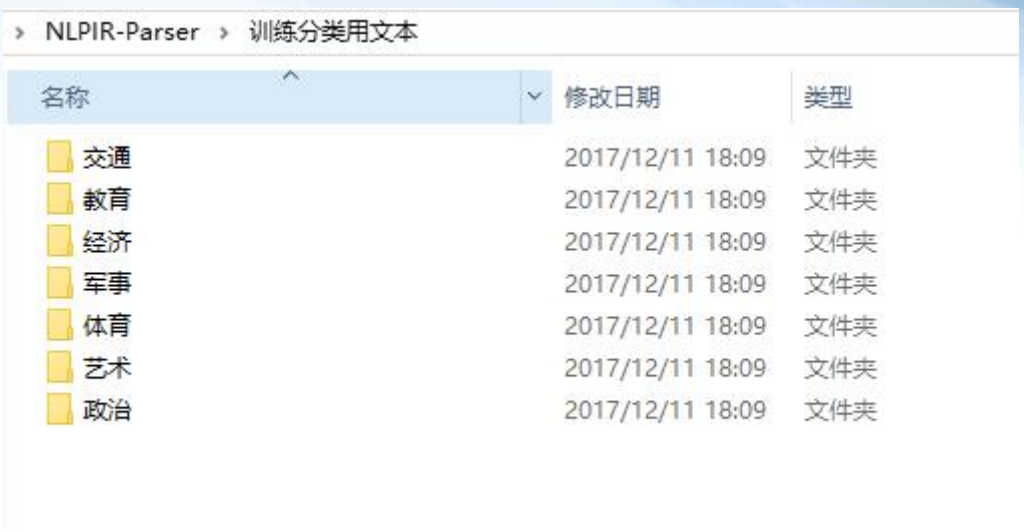


图 3. 39 训练分类用文本

图 3. 40 训练

如上所示，系统将训练结果以网页的形式呈现在提示框中，总计频率为 186964，共有 1000 个特征词，第一个特征词为“会谈”，在 9 篇文档中出现共 22 次，权重值为 11。

Step2：选择测试语料（以乐视新闻语料为例），点击“机器学习分类过滤”，系统进行分类分析。分类结果如下：共有 6 个类别，

交通类四篇文档，教育类 4 篇文档，...

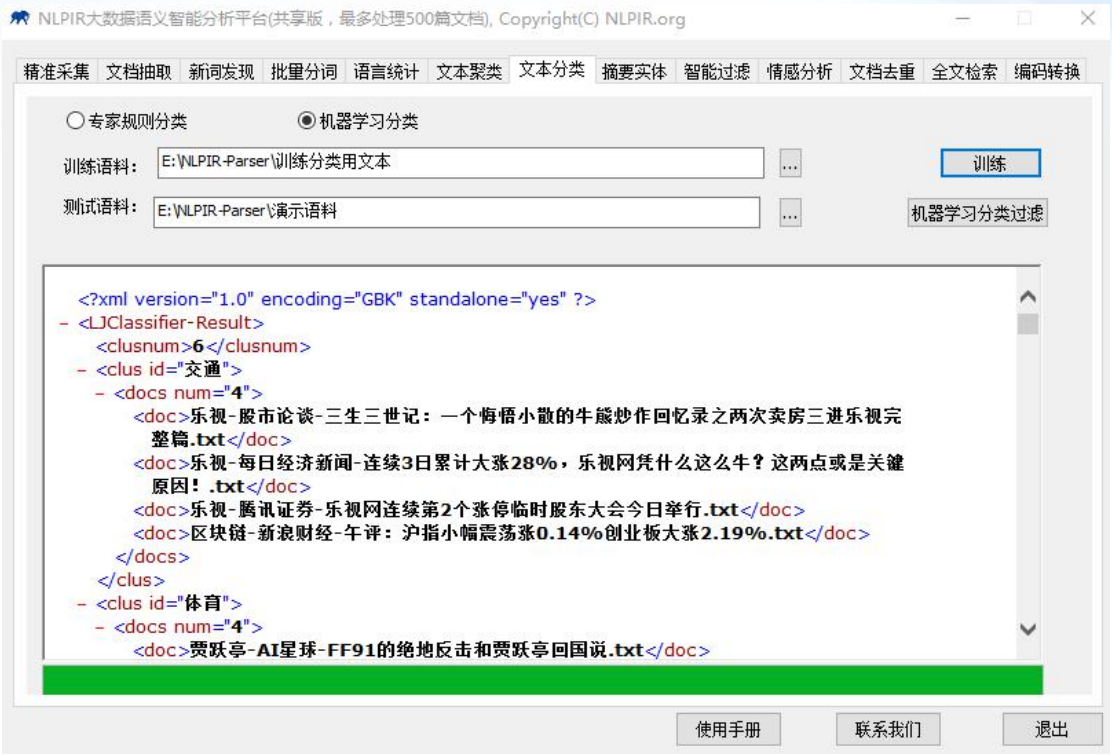


图 3.41 分类过滤

系统会将分类结果以网页和文件的同时自动保存至：\NLPIR-Parser\output\机器学习分类结果，并自动打开文件夹。

图 3.42 分类结果文件

3.8 摘要实体

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。实体提取能够对单篇或多篇文章，自动提炼出内容摘要，抽取人名、地名、机构名、时间及主题关键词；方便用户快速浏览文本内容。

用户首先点击“摘要实体”，进入系统摘要实体功能模块。

Step1：选择语料源目录（以十九大报告为例）；用户可自定义摘要长度（默认最大为 250），摘要最大压缩率和关键词数量；

Step2：点击“摘要与实体抽取”，系统进行提取与分析，并显示摘要和关键词的结果。点击“上一篇”、“下一篇”按钮，可实现结果的快速浏览。

图 3.43 摘要与实体抽取

摘要实体结果包括：自动摘要和实体抽取（关键词、人、时间、

地点、国家与机构）。

十九大报告分析结果：

摘要（摘要长度定义为 300 的结果）：要长期坚持、不断发展我国社会主义民主政治，积极稳妥推进政治体制改革，推进社会主义民主政治制度化、规范化、法治化、程序化，保证人民依法通过各种途径和形式管理国家事务，管理经济文化事业，管理社会事务，巩固和发展生动活泼、安定团结的政治局面。成立中央全面依法治国领导小组，加强对法治中国建设的统一领导。

实体抽取：

关键词（关键词数量定义为 10 的分析结果）：发展#建设#人民#中国#国家#政治#社会#文化#经济#创新#

时间：2017 年 10 月 18 日#现在#当前#近代#一九二一年#一九四九年#今天#未来#本世纪中叶#千年#二〇二〇年#二〇三五年#现代#当今#冬#当代#清明#

国：中国#

人物：习近平#金山银#向发力#德治相#言代法#安邦定#强国强#来海#晏河清#高强#

地点：中国#台湾#北京#京津冀#中华人民共和国#惠民#澳门#香港#长江#澳门特别行政区#亚洲#杭州#香港特别行政区#厦门#南海#古田#亚丁#安新#

机构：中国共产党#党中央#联合国#中共中央#

3.9 智能过滤

智能过滤能够对文本内容进行语义智能过滤审查，内置国内最全词库，智能识别多种变种：形变、音变、繁简等多种变形，且实现语义精准排歧。

用户首先点击“智能过滤”，进入系统智能过滤功能模块。

（1）导入关键词

系统已内置约 10 类近 4 万关键词，用户仍可根据需求添加自己的关键词。

Step1：选择关键词文件，在“关键词列表文件”中选择文件，点击“编辑”，系统弹出关键词文件，用户可编辑关键词列表，编辑完成后保存并关闭文件。

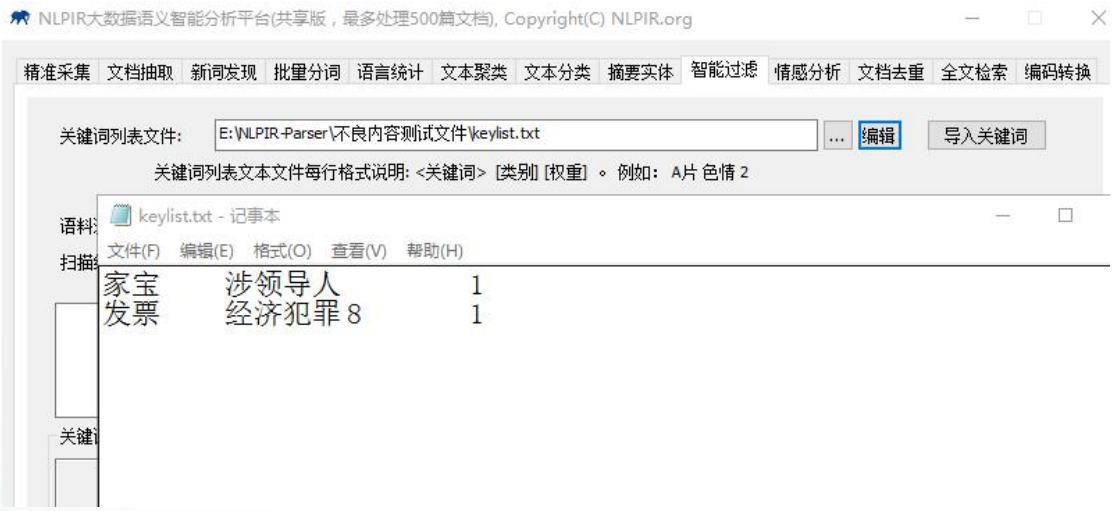


图 3. 44 编辑关键词

Step2：点击“导入关键词”，系统显示导入关键词成功。

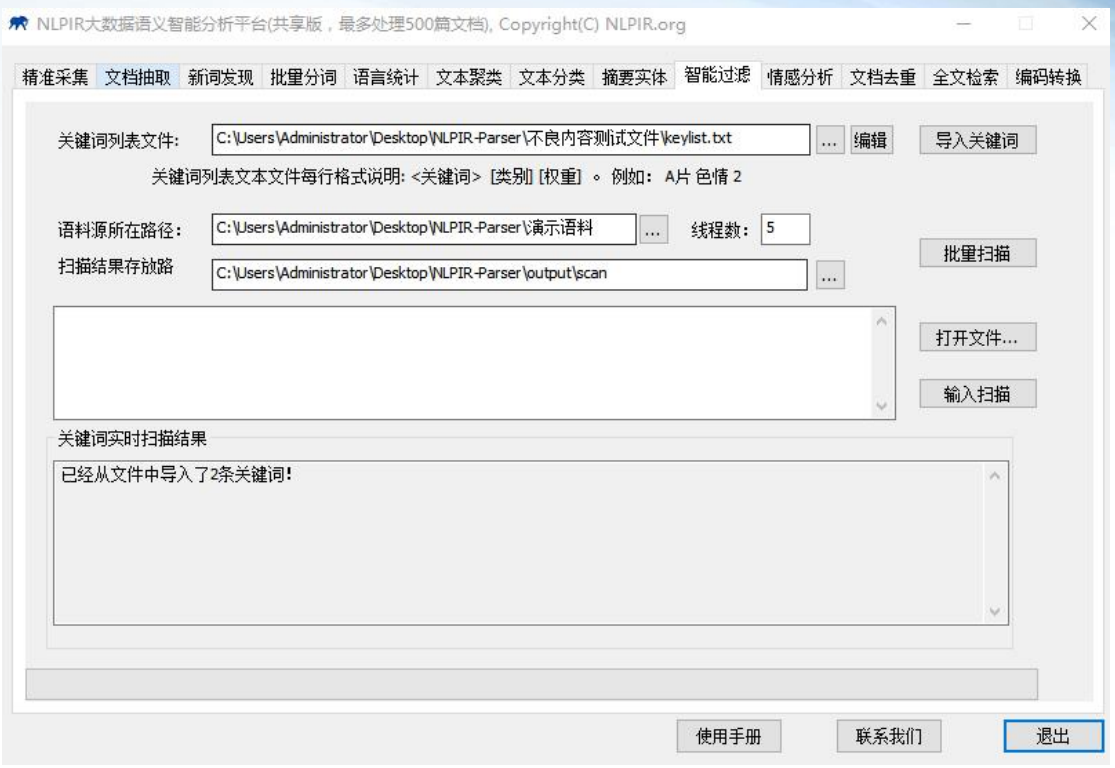


图 3. 45 导入关键词成功

（2）批量扫描

Step1：选择语料源：\NLPIR-Parser\不良内容测试文件（系统默认，用户可定义自己的过滤语料），系统会指定默认的“扫描结果存放路径”为：\NLPIR-Parser\output\scan。用户也可以指定其它输出路径。

Step2：点击“批量扫描”，系统开始进行不良信息过滤。

智能过滤扫描结果以 txt 格式文件存放，文件名与源语料中的文件名一致。扫描统计结果 KeyScanStatResult.xls 放入 NLPIR-Parser\output 目录下并自动打开。扫描详情结果存放路径：\NLPIR-Parser\output\scan，扫描完成时自动为用户打开该目录。

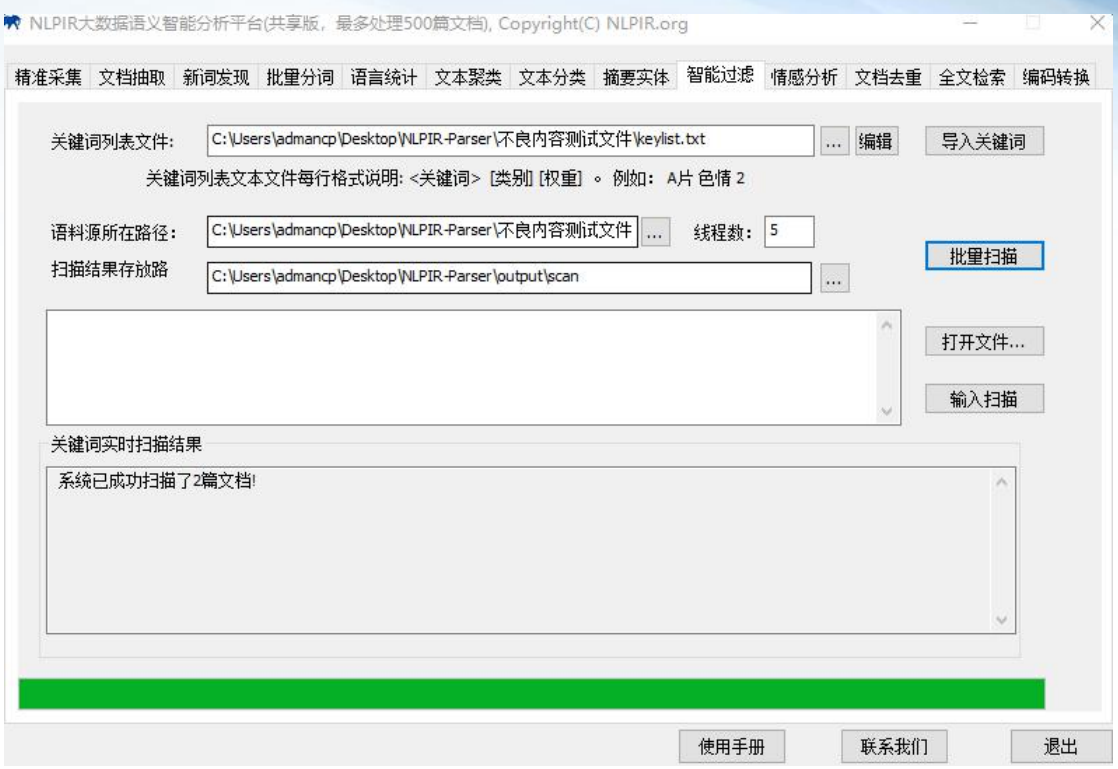


图 3.46 批量扫描

KeyScanStatResult.xls 包括：关键词、类别、权重与命中次数。扫描详情文件会在原文中标出扫描结果。

A	B	C	D	E
测试时间:	Mon Mar 12 22:16:38 2018			
扫描的记录	0	条记录		
扫描所花时	674.9	秒		
处理速度:	0	条/秒		
命中的规则	6	命中的记录	0	疑似敏感
规则编号	关键词	类别	权重	命中次数
36852	性交	色情	4	2
777	sm	色情	2	1
11764	家宝	涉领导人	3	1
14120	领导人	涉领导人	2	1
41348	发票	经济犯罪	4	1

图 3.47 扫描过滤结果统计

图 3.48 原文扫描结果

（3）输入扫描

Step1：点击“打开文件”或者直接将扫描文本粘贴至文本框中；

Step2：点击“输入扫描”，结果如下：

输入文本：六人一口彩（六合彩的形变）和法轮功

扫描结果：不良得分、命中不良内容、不良类别与命中文字

六人一口彩：

不良得分：16，命中不良内容：[形变]六合彩→六人一口彩，不良类别：涉赌，命中文字：六人一口彩。

图 3. 49 输入扫描

3.10 情感分析

情感分析，针对事先指定的分析对象，系统自动分析海量文档的情感倾向：情感极性及情感值测量，并在原文中给出正负面的得分和

句子样例。NLPIR 情感分析的情感分类丰富，不仅包括正、负两面，还包括好、乐、惊、怒、恶、哀和惧的具体情感属性。NLPIR 还提供关于特定人物的情感分析，并能计算正负面的具体得分。

用户首先点击“情感分析”，进入系统情感分析功能模块。

- 单个分析：对单个对象做情感分析

Step1：选择语料源（以乐视新闻报道为例）；输入分析对象：乐视；

Step2：点击“单个分析”，系统开始以“乐视”为分析对象进行情感分析。

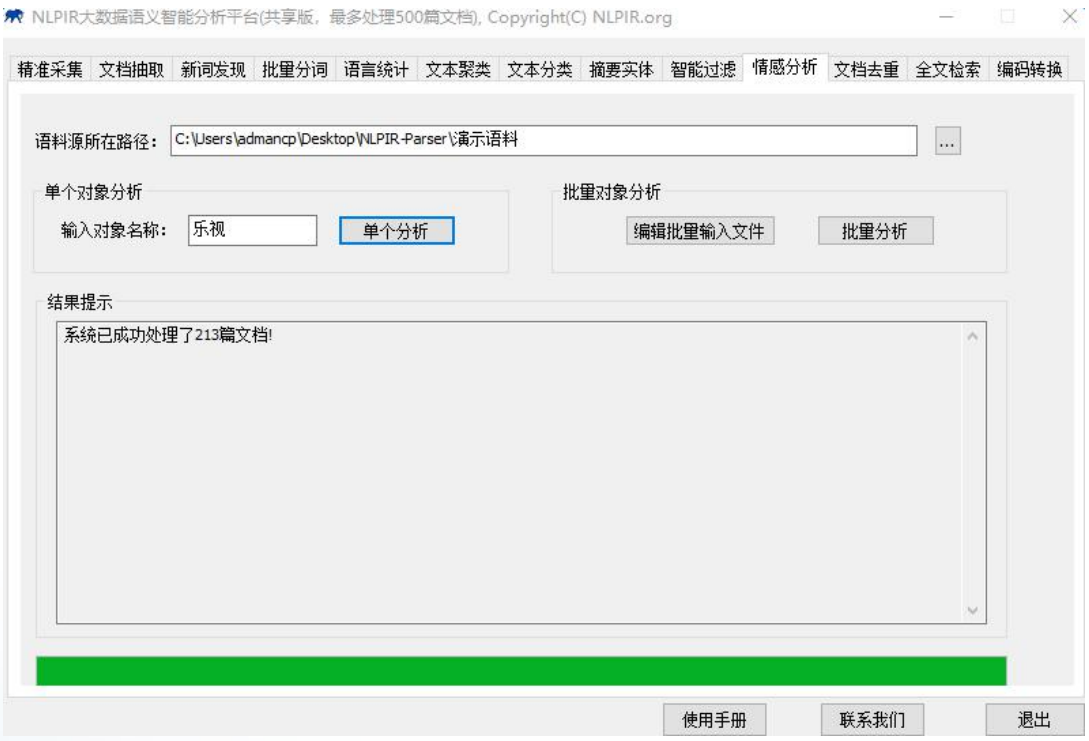
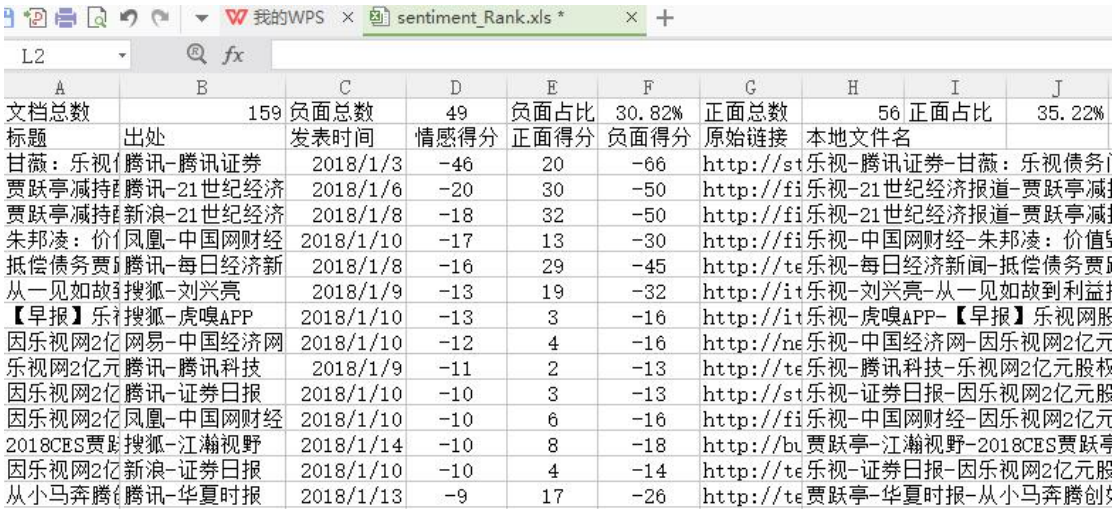


图 3.50 情感分析

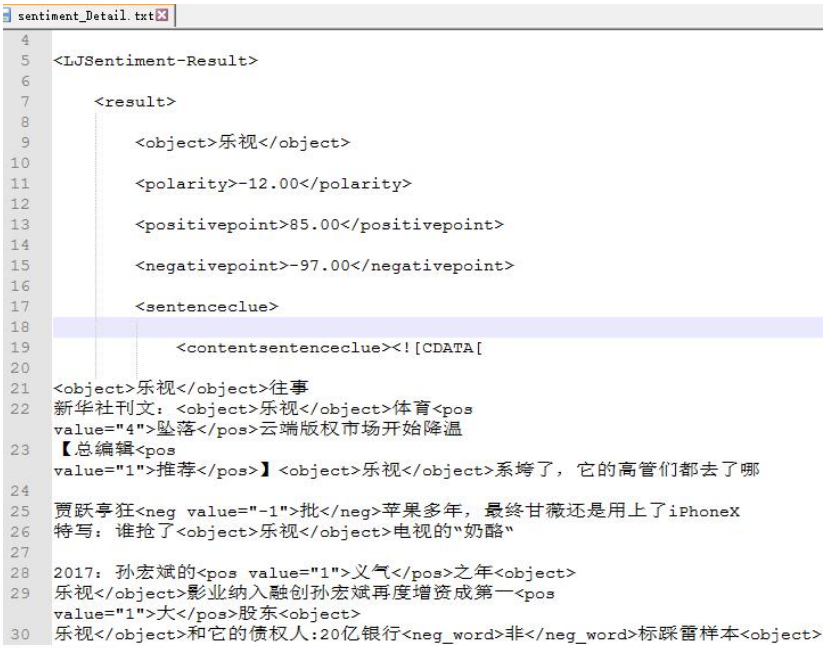
情感分析结果默认存放路径：NLPIR-Parser\output，情感分析有两个分析结果，sentiment-rank.xls(系统分析完毕后自动打开)和 sentiment-detail.txt，前者是统计结果，后者是分析详情结果。

情感分析统计结果包括：文档总数、正面数量及占比，每一篇文档的正负面得分与排序。情感分析详情结果会在原文本中显示情感分析的详情：对象、得分、原文等。



A	B	C	D	E	F	G	H	I	J
文档总数	159	负面总数	49	负面占比	30.82%	正面总数	56	正面占比	35.22%
标题	出处	发表时间	情感得分	正面得分	负面得分	原始链接	本地文件名		
甘薇：乐视(	腾讯-腾讯证券	2018/1/3	-46	20	-66	http://st	乐视-腾讯证券-甘薇：乐视债务(		
贾跃亭减持[	腾讯-21世纪经济	2018/1/6	-20	30	-50	http://fi	乐视-21世纪经济报道-贾跃亭减[		
贾跃亭减持[	新浪-21世纪经济	2018/1/8	-18	32	-50	http://fi	乐视-21世纪经济报道-贾跃亭减[		
朱邦凌：价(	凤凰-中国网财经	2018/1/10	-17	13	-30	http://fi	乐视-中国网财经-朱邦凌：价值[		
抵偿债务贾[	腾讯-每日经济新	2018/1/8	-16	29	-45	http://te	乐视-每日经济新闻-抵偿债务贾[		
从一见如故[	搜狐-刘兴亮	2018/1/9	-13	19	-32	http://it	乐视-刘兴亮-从一见如故到利益[		
【早报】乐[	搜狐-虎嗅APP	2018/1/10	-13	3	-16	http://it	乐视-虎嗅APP-【早报】乐视网股		
因乐视网2亿	网易-中国经济网	2018/1/10	-12	4	-16	http://ne	乐视-中国经济网-因乐视网2亿元		
乐视网2亿元	腾讯-腾讯科技	2018/1/9	-11	2	-13	http://te	乐视-腾讯科技-乐视网2亿元股权		
因乐视网2亿	腾讯-证券日报	2018/1/10	-10	3	-13	http://st	乐视-证券日报-因乐视网2亿元股		
因乐视网2亿	凤凰-中国网财经	2018/1/10	-10	6	-16	http://fi	乐视-中国网财经-因乐视网2亿元		
2018CES贾跃	搜狐-江瀚视野	2018/1/14	-10	8	-18	http://bu	贾跃亭-江瀚视野-2018CES贾跃亭		
因乐视网2亿	新浪-证券日报	2018/1/10	-10	4	-14	http://te	乐视-证券日报-因乐视网2亿元股		
从小马奔腾(	腾讯-华夏时报	2018/1/13	-9	17	-26	http://te	贾跃亭-华夏时报-从小马奔腾创始		

图 3.51 sentiment-rank



```
sentiment_Detail.txt
4
5   <LJSentiment-Result>
6
7       <result>
8
9           <object>乐视</object>
10
11          <polarity>-12.00</polarity>
12
13          <positivepoint>85.00</positivepoint>
14
15          <negativepoint>-97.00</negativepoint>
16
17          <sentenceclue>
18
19              <contentsentenceclue><![CDATA[
20
21  <object>乐视</object>往事
22  新华社刊文：<object>乐视</object>体育<pos
    value="4">坠落</pos>云端版权市场开始降温
23  【总编辑<pos
    value="1">推荐</pos>】<object>乐视</object>系垮了，它的高管们都去了哪
24
25  贾跃亭狂<neg value="-1">批</neg>苹果多年，最终甘薇还是用上了iPhoneX
26  特写：谁抢了<object>乐视</object>电视的"奶酪"
27
28  2017：孙宏斌的<pos value="1">义气</pos>之年<object>
29  乐视</object>影业纳入融创孙宏斌再度增资成第一<pos
    value="1">大</pos>股东<object>
30  乐视</object>和它的债权人:20亿银行<neg_word>非</neg_word>标踩雷样本<object>
```

图 3.52 sentiment-detail

对象：乐视，情感得分：-12，正面得分：85，负面得分：-97

- 批量对象分析

Step1：选择语料源（以乐视新闻报道为例）；点击“编辑批量

输入文件”，用户可自定义多个分析对象与分析条件。

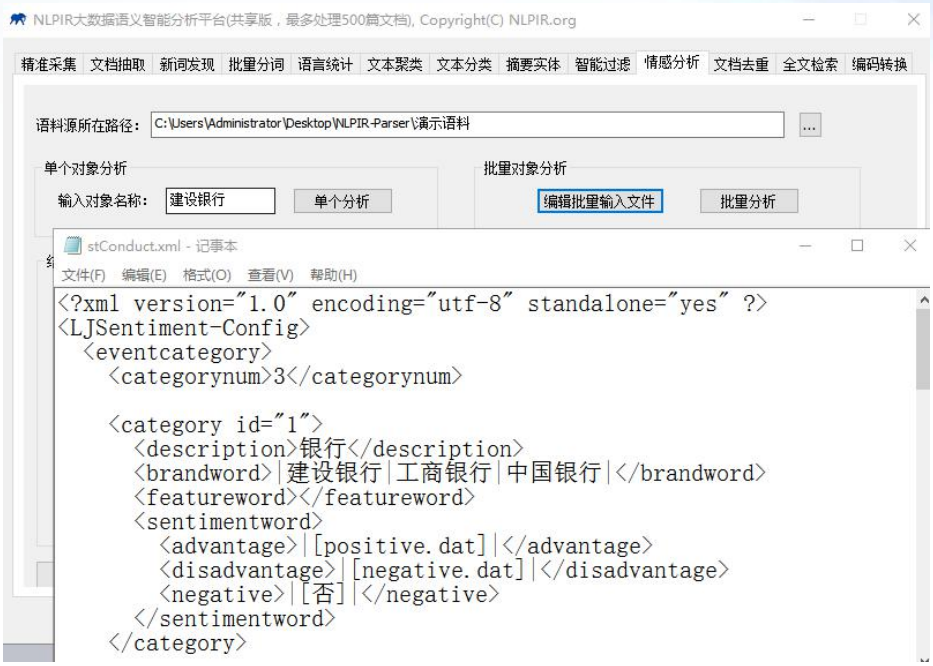


图 3.53 定义批量对象

Step2：点击“批量分析”，系统开始对多个对象进行情感分析。

批量分析同样有两个结果文件（\NLPIR-Parser\output），sentiment-rank.xls(系统分析完毕后自动打开)和 sentiment-detail.txt，前者是统计结果，后者是分析详情结果。

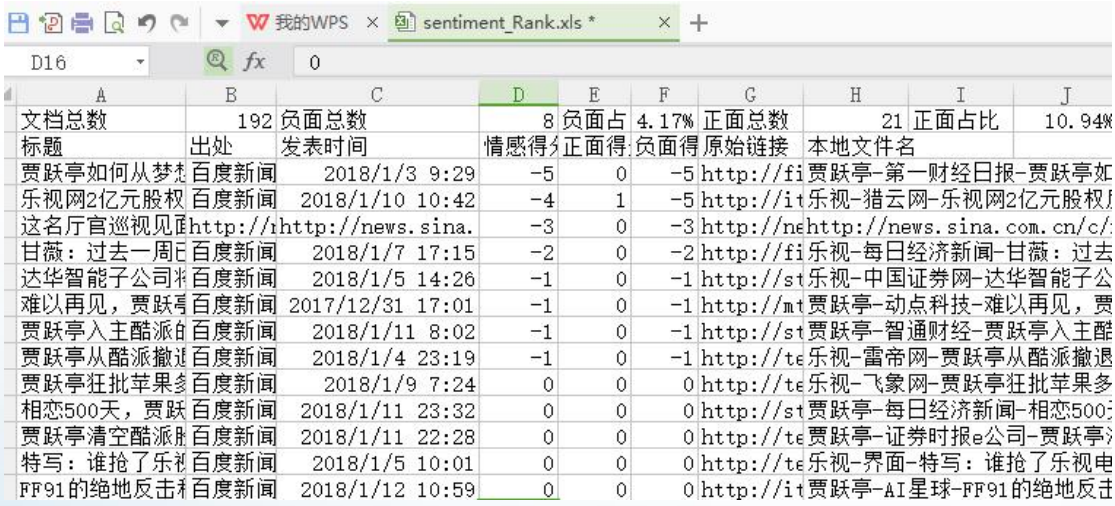


图 3.54 情感批量分析结果

3.11 文档去重

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。

用户首先点击“文档去重”，进入系统文档去重功能模块。

Step1：选择语料源；选择结果文件存放路径。

Step2：点击“开始查重”，系统即刻开始查重处理，并输出查重结果文件 RepeatFile（NLPIR-Parser\bin-win64\output\RepeatFile.txt）查重结果会显示在结果提示框中，如下图所示：

图 3.55 文档去重

RepeatFile 文档去重分析结果包括：重复文档数量统计(共有 5 片文档重复)，重复文档标题与重复文档路径。

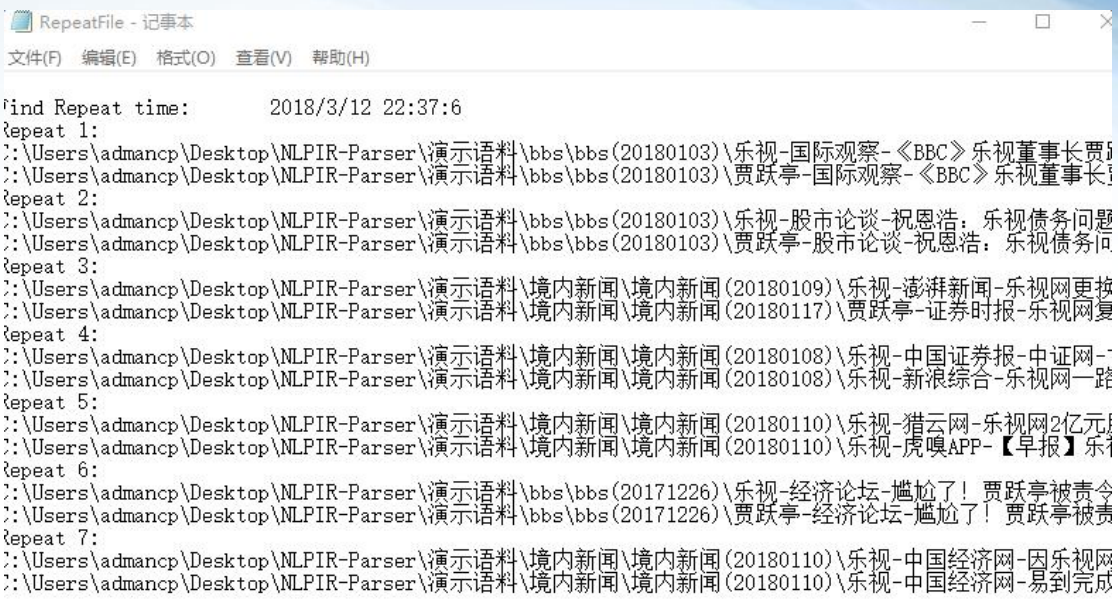


图 3.56 RepeatFile

3.12 全文检索

全文检索支持文本、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。

支持的典型查询语法包括：

Sample1: [FIELD] title [AND] 解放军

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

用户首先点击“全文检索”，进入系统全文检索功能模块。

➢ 建立索引

Step1：选择语料文件夹（以十九大报告为例）；

Step2：选择是否“增量”，增量是指在历史索引的基础上需要对新增部分文件的内容建立索引。系统在历史索引基础上新增索引，不选择增量，系统将以预料源为基础重新建立索引。点击”建立索引”，系统对语料快速建立压缩索引。

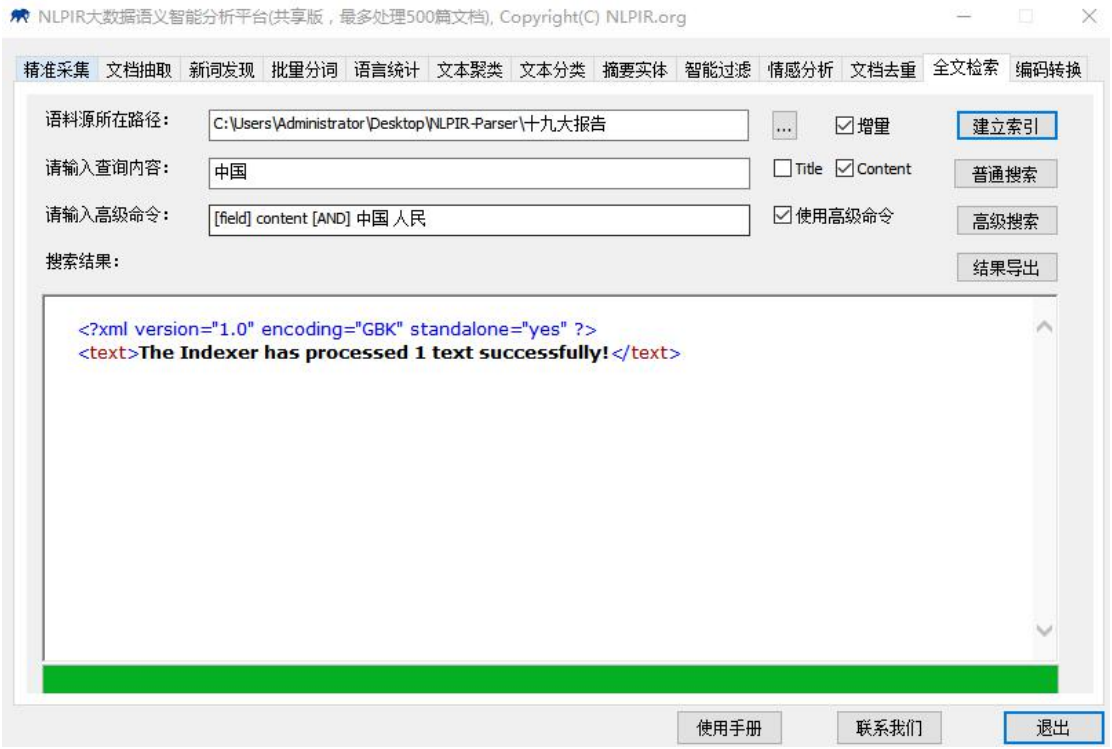


图 3.57 建立索引

➢ 普通检索

Step1：输入查询关键词（中国），选择“Title”（标题查询）与“content”（内容查询），两者可同时选择。

Step2：点击“普通检索”。搜索结果框会呈现查询结果，并配以相似得分。检索结果文件（\NLPIR-Parser\output\搜索结果 JZSearch-result）以网页形式保存。

检索结果包括：文档总量统计、标题、内容与相似得分。

图 3.58 普通检索

Step3：点击“结果导出”，系统将检索目标文档导出\NLPIR-Parser\output\搜索结果\中国，并自动打开文件目录。

图 3.59 结果导出

图 3.60 搜索结果

➢ 高级检索

Step1：点击“使用高级命令”，输入高级命令。例：[field] content [AND] 中国 人民 表示：搜索内容字段中同时包含“中国”和“人民”的文档，采用该语法信息过滤将更有针对性；

Step2：点击“高级检索”，系统将进行高级检索。搜索结果框会呈现查询结果，并配以相似得分。检索结果文件（\NLPIR-Parser\output\搜索结果 JZSearch-result）以网页形式保存。

图 3.61 高级检索

Step3：点击“结果导出”，系统将检索目标文档导出，并自动

打开文件目录。

图 3. 62 结果导出

3.13 编码转换

编码转换功能，自动识别内容的编码，并把编码统一转换为 GBK 编码。目前支持 Unicode/BIG5/UTF-8 等编码自动转换为简体的 GBK，同时将繁体 BIG5 和繁体 GBK 进行繁简转化。

用户首先点击“编码转换”，进入系统编码转换功能模块。

- 转换为 GBK 编码

Step1：选择语料源：\NLPIR-Parser\编码转换测试文本，系统指定输出路径：\NLPIR-Parser\bin-win64\output\GBK。

Step2：点击“转换为 GBK 编码”。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 编码的文件。转换结果提示框将显示转换结果，并将编码转换结果文件夹自动打开，用户可直接查看与使用转换后的文件。

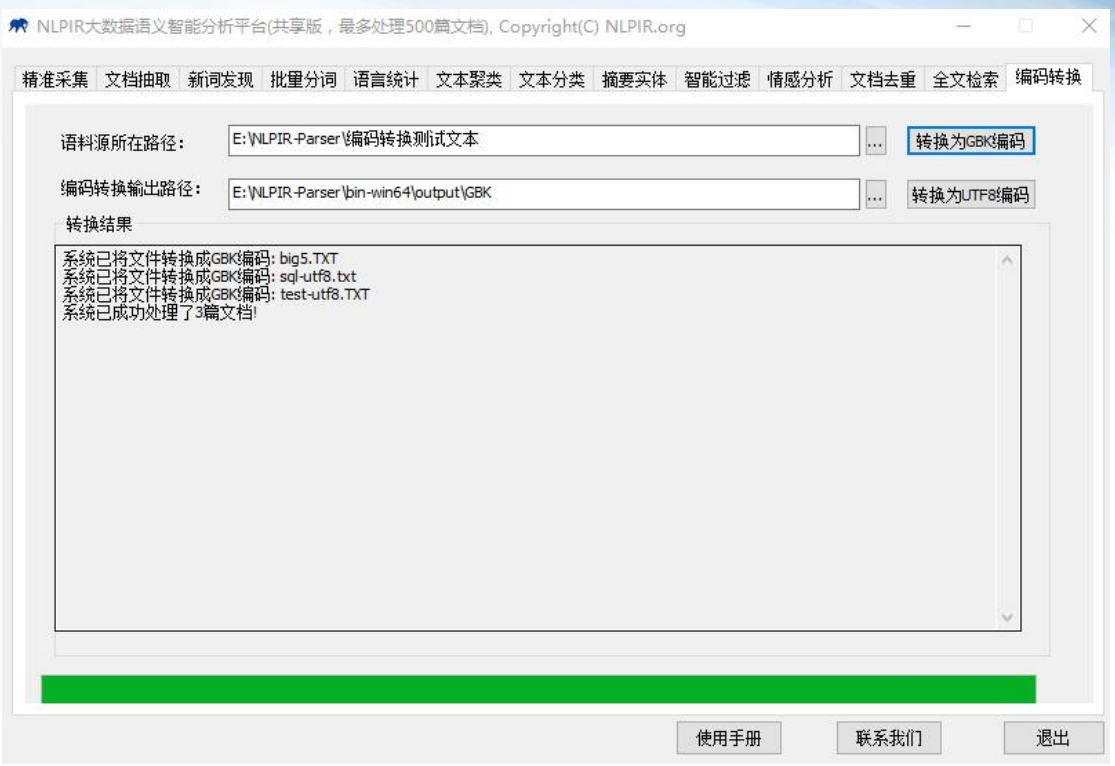


图 3.63 转换为 GBK 编码

图 3.64 转换为 GBK 编码

➢ 转换为 UTF8 编码

Step1：选择语料源：\NLPIR-Parser\编码转换测试文本，系统指定输出路径：\NLPIR-Parser\bin-win64\output\UTF8。

Step2：点击“转换为 UTF8 编码”。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 UTF8 编码的文件。转换结果提示框将显示转换结果，并将编码转换结果文件夹自动打开，用户可直接查看与使用转换后的文件。

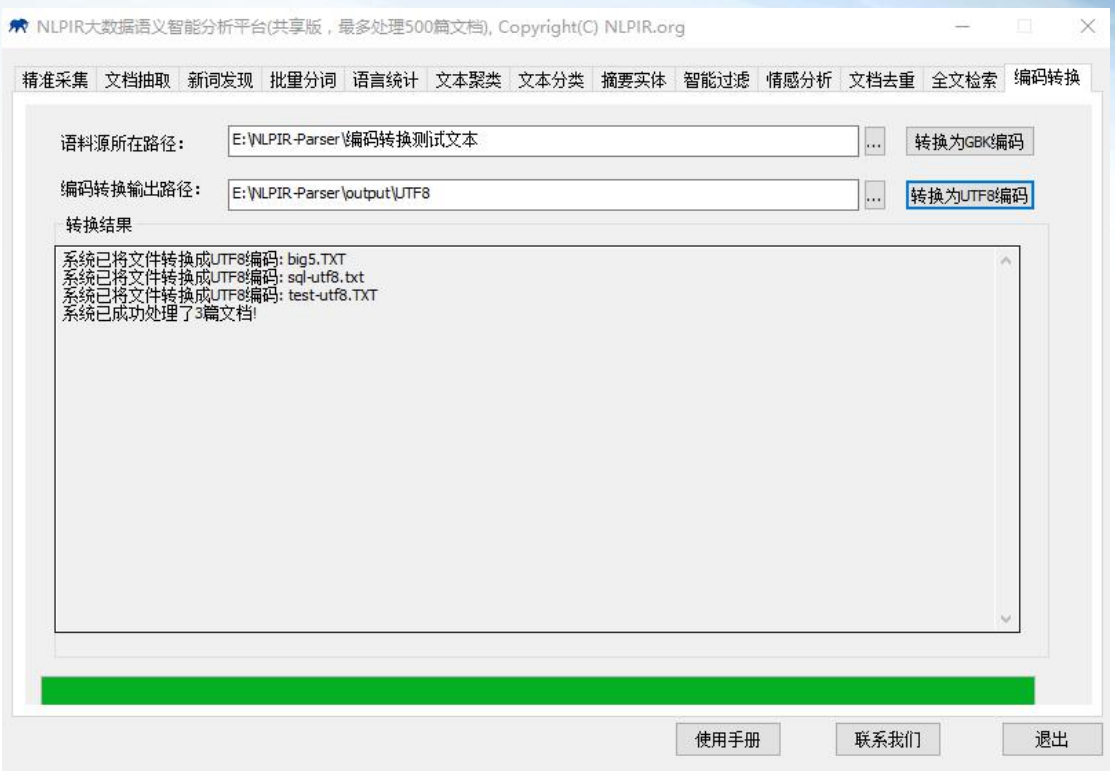


图 3.65 转换为 UTF8 编码

图 3.66 转换为 UTF8 编码

四、应用示范案例

4.1 十九大报告语义智能分析

2017 年 10 月 18 日，中国共产党第十九次全国代表大会在北京隆重召开，习近平代表第十八届中央委员会向大会作报告。这份沉甸甸的报告总结了自十八大以来我国的发展进程，党的引领脚步，人民的生活改变.....以及未来如何开启新时代、谱写新篇章。

如何精准解读这份报告，我们采用自然语言处理工具 nlpir-paser，通过挖掘十九大报告的关键词、概念新词、内容图谱等语义智能处理技术，带你深度感受十九大精神。

➢ 关键词提取

十九大报告全文 3 万余字，本文使用 NLPIR 对十九大报告进行关键词提取，以期揭示十九大报告的核心要点。关键词 top100 结果展示如下：

图 4.1 关键词 top100

由于篇幅所限，本文只展示了部分关键词提取的结果，关键词词云图分析结果比较充分地展示了十九大报告的核心概念。

➢ 词频统计

分析结果显示，词频统计 top10 的关键词分别为：“中国特色社会主义”、“中华民族伟大复兴”、“依法治国”、“全面建成小康社会”、“中国梦”、“人民当家作主”、“美好生活”、“现代化经济体系”、“人民军队”、“小康社会”。这些高频词汇基本概括

了十九大报告中的基础概念。

图 4.2 词频统计

➢ 新词发现

“人类命运共同体”，“新征程”，“现代化经济体系”，“社会主要矛盾转化”，“历史性变革”……

十九大报告中出现的不少新的“关键词”，这些新词展示了新理念、新观点，给予了重大时代课题明确的回答，在实践上作出了新部署。

图 4.3 十九大新词

4.2 文章风格对比：方文山 VS 汪峰

不同人的文章风格不同，汪峰的摇滚歌词给人奔放、热烈的情感激荡，而方文山中国风歌词则会给我们造成委婉、缠绵悱恻的心湖涟漪。这类文章风格主观感受的差别能否经得起科学实验的验证或证明呢？再者，文学、艺术等多个领域都存在文章作品对比与评价的争议，造成了很多不良的影响。通过技术能否为此提供一个评估的新维度或方法呢？我们通过 nlpir-paser 进行语言统计与分析、情感分析与词曲语言广度分析（信息熵）来进行文章风格的对比分析。

- 词频广度分析

通过歌词数目对比，通过工具可以得出以下方文山与汪峰对比：（比率=方文山/汪峰，平均用词=总词数/歌曲数）

表 4.1　方文山和汪峰用词分析

	总词数	歌曲数	平均用词
方文山	8195	200	40.975
汪峰	2270	127	17.874
比率	3.610	1.574	2.292

可以很明显的看出方文山所用词汇数量远远多于汪峰。通过平均用词可以发现方文山比汪峰用用词广度大。每首歌曲方文山是汪峰的用词量的二倍。

- 情感对比分析

将方文山和汪峰的形容词作为情感分析的主要词汇。

图 4. 4 方文山（左）和汪峰（右）的情感词汇词云图

从形容词上统计方文山和汪峰，可以看出汪峰是一种激进的用词，负向很明显“孤独”“破碎”，正向“美丽”“坚强”，这些对生命的感悟的词汇。汪峰多写生命的感悟，同时把摇滚歌手那种想表达的孤单，力量感，表达出来。而方文山的形容词性则以比较温柔的情感词为主“温柔”“美丽”“简单”。这里也能说明两个作词人风格不同，方文山多写爱情和亲情。通过比对能很明显的发现两个作词人词风不同。

- 信息熵分析

信息熵公式：$H(X)=-\sum_{1}^{n}P(X)\log P(X)$。信息熵用来表示作词人用词的广度。用词数量越小，信息熵越小。通过用词信息熵进行加和来比较方文山和汪峰的用词广度。

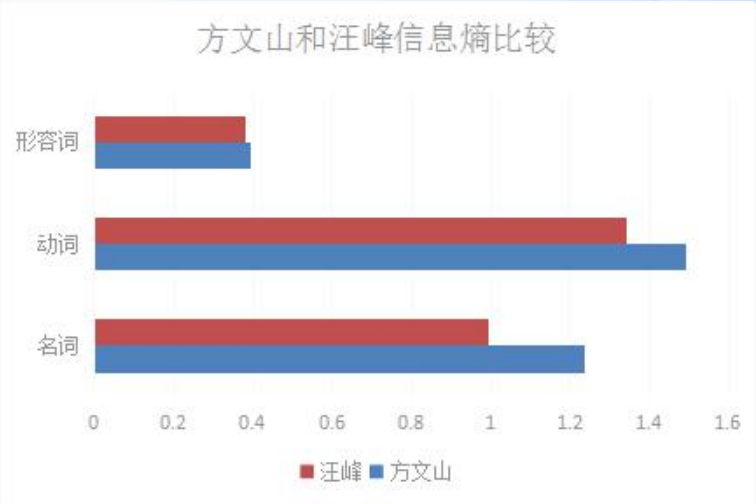


图 4.5 信息熵对比

可以看出汪峰词作在三组词性上的信息熵均小于方文山。同时验证了汪峰的词作中用词信息量较少。可以推理出汪峰词作多重复性词汇，方文山用词量大，广泛。

4.3 《红楼梦》作者前后同一性识别

《红楼梦》前八十回和后四十回到底是不是同一个作者？我们都知道《红楼梦》的作者有两个：曹雪芹写了前八十回，高鹗续写了后四十回。然而，红学上关于《红楼梦》的作者争议一直很大，存在着很多种版本。我们将利用大数据语义智能分析工具 nlpir-paser，通过语言统计、概率计算与文本相似度分析来进行《红楼梦》前后作者同一性判别。

- 虚词统计

每个人的写作都有些小习惯，虽然文章前后说的内容会有差别。但是每个人使用虚词的顺序与数量可能存在着差异。

将《红楼梦》120 回按顺序均分为 3 组，使用 NLPIR 统计出文言

虚词的词频，再对不同组数据之间进行 KL 距离计算。第一组将 120 回按顺序均分为三等份即第 1 回-第 40 回、第 41 回-第 80 回、第 81-第 120 回。这 3 组数据中部分虚词以及该词的概率如表所示：

表 4.2　三组虚词统计分析

	第1回-第40回		第41回-第80回		第81回-第120回	
词	词频	概率	词频	概率	词频	概率
了	5981	0.199712836	7740	0.213299529	6710	0.206786033
的	3854	0.128689729	5156	0.142089454	5269	0.162377885
不	3063	0.102277281	3805	0.104858489	3510	0.108169743
是	2293	0.076566048	2975	0.081985284	3039	0.093654658
一	2202	0.073527448	2750	0.075784716	1953	0.060186755
着	1607	0.053659677	1855	0.051120236	2112	0.065086752
便	1075	0.035895552	1272	0.035053876	1295	0.03990878
在	1026	0.034259383	1089	0.030010748	1253	0.038614441
就	935	0.031220783	1101	0.030341445	817	0.025177972
儿	899	0.030018699	1108	0.030534351	1143	0.035224506
好	786	0.026245492	956	0.026345523	939	0.028937718
之	747	0.024943235	658	0.018133216	243	0.007488675
呢	601	0.020068118	515	0.014192411	719	0.022157848
因	571	0.019066382	724	0.019952049	363	0.011186785
再	395	0.013189529	456	0.012566484	262	0.008074209
可	385	0.012855616	362	0.009976024	254	0.007827668
罢	328	0.010952317	354	0.00975556	407	0.012542759
把	324	0.010818753	364	0.010031141	420	0.012943388
方	266	0.008882062	284	0.007826494	59	0.001818238
往	253	0.008447976	243	0.006696613	140	0.004314463
别	250	0.008347803	314	0.008653237	165	0.005084902
向	212	0.007078937	203	0.00559429	119	0.003667293
亦	171	0.005709897	144	0.003968363	28	8.63E-04
比	160	0.005342594	211	0.005814755	110	0.003389935

➢ KL 距离

KL 距离（相对熵）可以衡量两个随机分布之间的距离，当两个随机分布相同时，它们的相对熵为零，当两个随机分布的差别增大时，它们的相对熵也会增大。所以相对熵（KL 散度）可以用于比较文本的相似度。

从下表中可以观察到第一行中 1-40 与 81-120 的 KL 值是 1-40 与 41-80 的 KL 值的十倍。由于当两个随机分布的差别增大时，它们的相对熵也会增大。所以 1-40 与 81-120 的相似性比 1-40 与 41-80 低。

表 4.3　三组 KL 距离分析

回数 / KL 值 / 回数	1-40	41-80	81-120
1-40	0	0.008	0.082
41-80	0.007	0	0.06
81-120	0.051	0.049	0

可以看出前八十回的各组数据的 KL 值与后四十回的数据的 KL 值有不同程度的差距。后四十回之间的 KL 值比其他组得 KL 值要小，说明后四十回的相似度较高。可以大胆猜测后四十回是出自于另外一个人。

五、联系我们

需要购买 NLPIR 大数据语义智能分析平台正式版本，或者需要使用 NLPIR 各类二次开发包，可以通过以下方式联系到我们：

大数据搜索与挖掘实验室（北京市海量语言信息处理与云计算应用工程技术研究中心）

地址：北京海淀区中关村南大街 5 号 100081

电话：13681251543(商务助手电话)

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)

微博:http://www.weibo.com/drkevinzhang/

微信公众号：大数据千人会

Beijing Engineering Research Center of Massive Language Information Processing and Cloud Computing Application

Beijing Institute of Technology

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

Tel: 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)

Twitter:http://www.weibo.com/drkevinzhang/

Subscriptions: Thousands of Big Data Experts

六、附录

6.1 下载途径

NLPIR-Parser 系统的多种下载途径：

1、GitHub:

https://github.com/NLPIR-team/NLPIR/tree/master/NLPIR-Parser

【有可能国内访问国外网址受限】

2、官方网站下载：

链接：

http://www.nlpir.org/wordpress/wp-content/uploads/2018/12/NLPIR-Parser.zip

打开浏览器，复制下载链接，即可启动下载。

3、也可以百度各软件下载平台，下载 NLPIR-Parser。

访问 NLPIR-Parser 目录即可。

注：用户在 github 上下载 NLPIR-Parser 文件时需要专门的下载工具，建议使用 svn 工具下载文件。

6.2 Github 下载演示

首先，打开 github 上 NLPIR-Parser 文件下载地址，复制该地址；

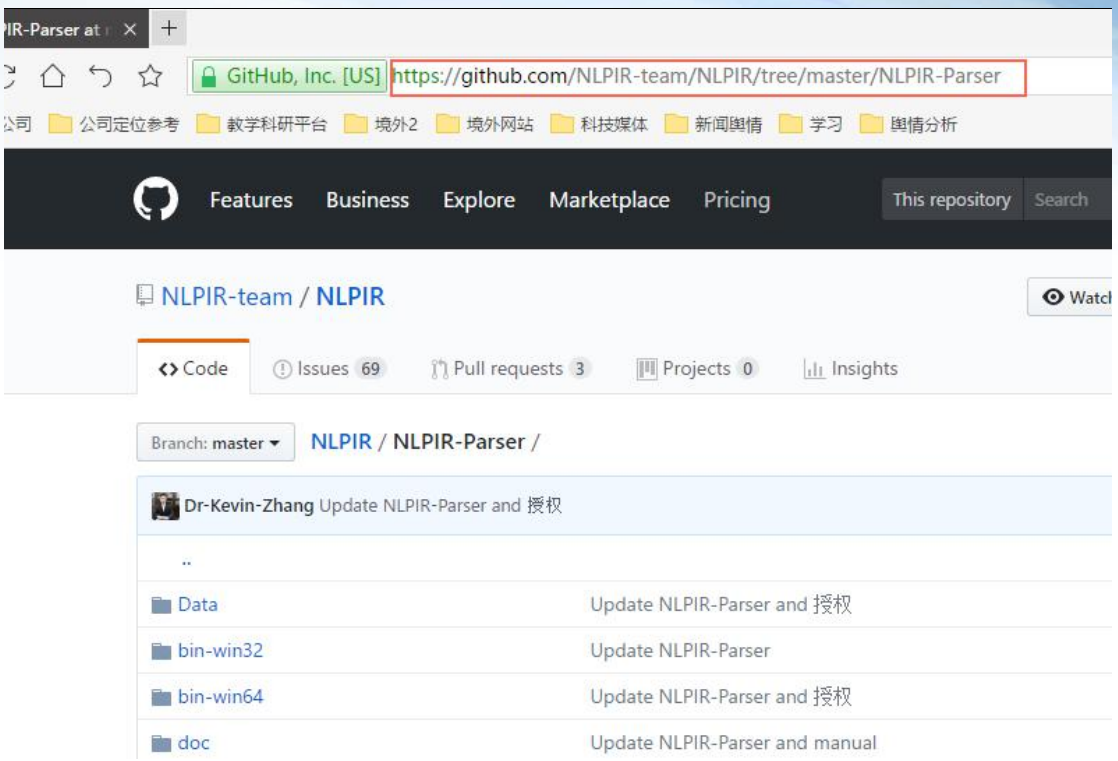


图 6.1 github 网址

然后，点击鼠标右键 SVN Checkout，弹出以下窗口，文件下载地址已经自动复制。

图 6.2 右键“svn checkout”

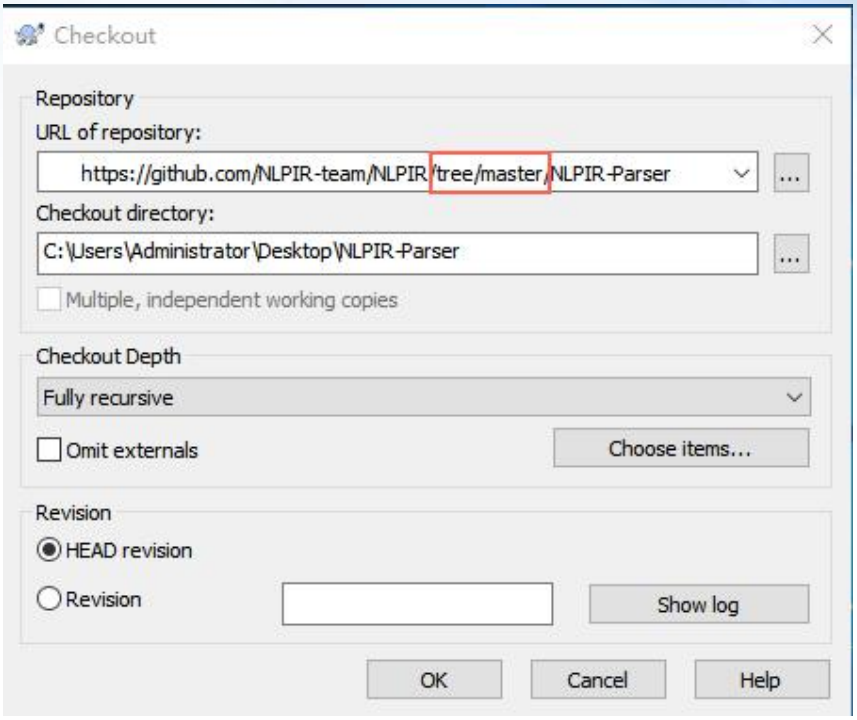


图 6.3 自动复制网址

最后，将地址中的“/tree/master/”修改为“/trunk/”，选择文件存放地址（桌面 desktop 或其他地址），点击”ok”，文件下载启动，下载完毕后点击“ok”，文件下载完毕。

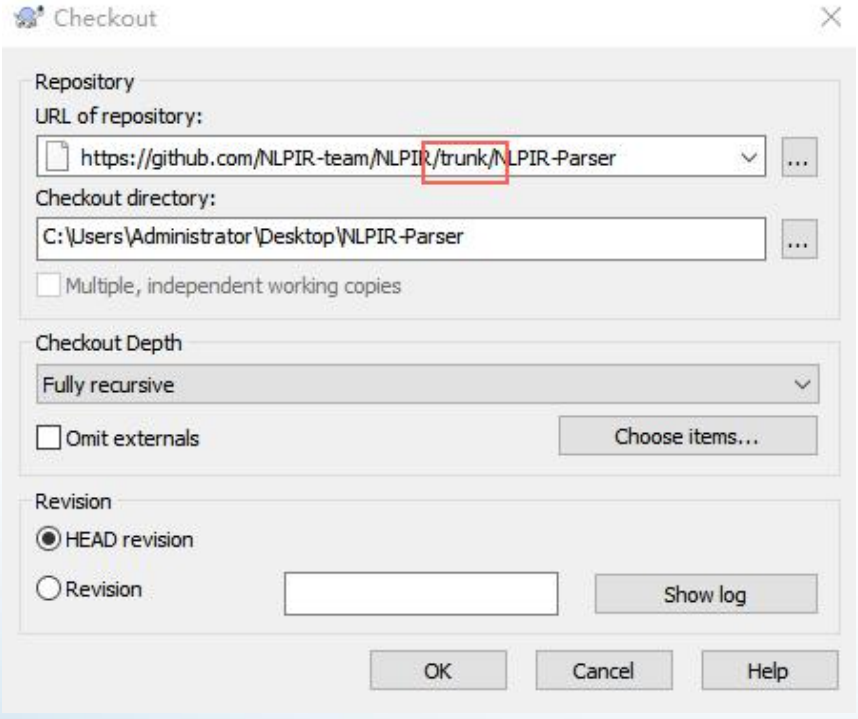


图 6.4 修改地址

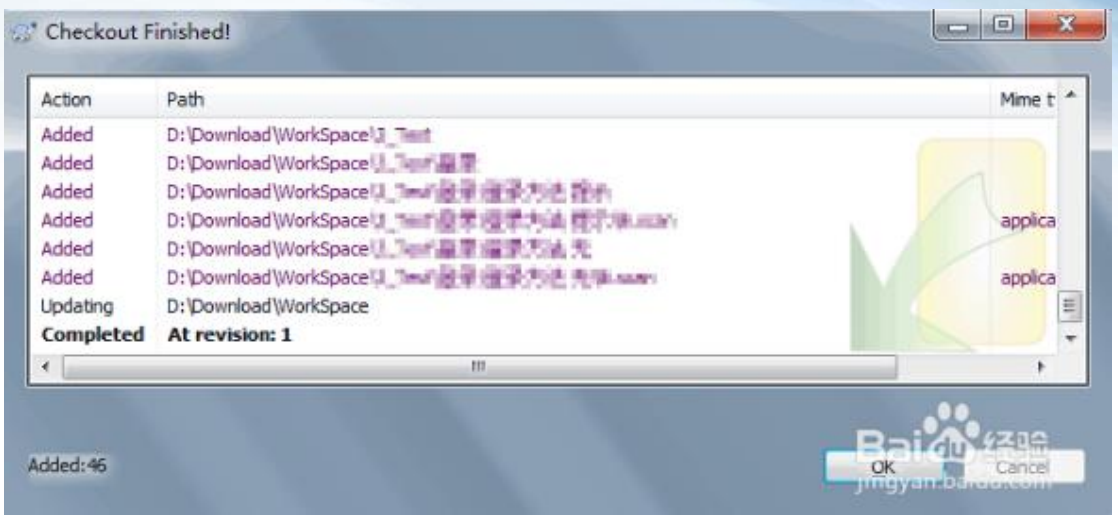

Checkout Finished!
Action
Path
Mime t
Added
Added
Added
Added
Added
Added
Updating
Completed
D:\Download\WorkSpace
At revision: 1
applica
applica
Added: 46
OK
Cancel

图 6.5 下载成功