NLPIR-Parser：大数据语义智能分析平台*

北京理工大学 张华平 商建云

提要：随着社交网络等新型网络的迅猛发展，文本大数据呈几何级数增长，语料库的加工处理一般都是由文科背景的研究人员完成，因此，急需快速简便的大数据内容批处理平台。NLPIR-Parser大数据语义智能分析平台历时20余年的积累，融合了网络数据采集、自然语言处理、文本挖掘与文本检索等核心技术。平台为一般用户提供了本地化部署的客户端实现语义智能分析的全链条一站式服务，也为软件工程师提供了二次开发接口。NLPIR-Parser平台包含精准采集，文档格式转换、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十三项独立功能，涵盖了从数据的采集预处理、自然语言处理到文本挖掘、信息检索再到可视化呈现、结果导出等全链条各个环节的语义分析工具，服务了全球40万家机构用户和百余家高校科研院所，为自然语言的研究者与工程应用提供了便利的技术支持。

关键词：语义智能分析、汉语分词、新词发现、全文检索、语料库处理

1. 引言

在大数据背景下，2017年7月8日，国务院印发《新一代人工智能发展规划》，明确了我国发展人工智能的战略目标，到2030年，人工智能核心产业规模超过1万亿元，带动相关产业规模超过10万亿元。人工智能已经成为现代科学皇冠上的明珠，而自然语言处理直接面对数据中的语义内容，号称是“人工智能皇冠上的明珠”,直接决定大数据智能的广度与深度。自然语言处理是计算机对自然语言所包含的字形、读音和含义等信息进行处理，包括对字、词、句和篇章的输入输出、识别分析、理解生成等操作和加工，是当前人工智能研究的核心课题之一，自然语言处理的关键是让计算机“理解”自然语言。

在语料库加工处理过程中，一般都是大量的人工标引，急需快捷简便的自然语言处理工具。但对于一般研究人员来说，具体操作过程中有如下挑战：

*本课题得到国家自然科学基金（No.61772075、2018-U11636123）、“两高一部”课题（2018YFC0832304）资助。

1.1 需要技术人员参与开发，文科背景的研究人员学习代价过高

对语料处理的项目，在时间要求不太紧急时，让有知识背景的人做人工标记，如档案局历史材料、专利局的专利申请材料等，但成本高、耗时长，标记后的内容处理，如聚类、分类、可视化等人工无法完成，需要计算机软件来处理，虽然可以用现成的商业或开源工具，效果也不一定好，对工具的使用又有学习成本，尤其对于文字处理需求多的文科类人员比较困难。有些项目时间要求紧急，如网上应急事件的处理，再用人工逐一筛选，既不全面也不能满足快速应急处理的需要。

1.2 待处理的语料库知识资源存在数据泄漏的隐患

大部分研究者的语料库都是耗费了大量人力物力收集整理的，价值密度极高，甚至是毕生的心血积累。而目前自然语言处理的机构大部分提供的都是自然语言处理云服务平台，要求使用者上传待处理的语料库，如腾讯NLP云服务、百度NLP云服务。云端存储的数据资源脱离了上传者后，数据确权上没有法律保障，存在数据泄露并被窃取滥用的巨大隐患，导致大部分使用者望而却步。

1.3 大部分工具功能单一，缺乏一站式全链条的语义分析工具

目前已经有大量的研究者分别对自然语言处理中的各个关键点上问题进行研究开发出了一些开源的工具，有的只是单一功能，有的具有多个功能，但不是全链条。如urllib2、Scrapy、Pyspider等提供信息抓取工具；jieba提供分词工具；SnowNLP提供分词、情感分析、文本分类、转换成拼音、繁简转换、文本关键词和文本摘要提取、计算文档词频和文本相似度计算等工具；sklearn提供分类、聚类、回归、预处理、模型选择等工具；HanNLP提供中文分词，命名实体识别，关键词提取，自动摘要，短语提取，拼音转换，简繁转换，文本推荐，依存句法分析工具，但只有在java上可以用，而且配置、安装复杂；哈尔滨工业大学语言技术平台LTP提供中文分词、词性标注、命名实体识别、依存句法分析、语义角色标注等工具，但需要根据API参数构造HTTP请求在线获得分析结果；

针对众多研究者对自然语言处理的迫切需求与实际挑战，NLPIR-Parser历时20余年，为一般用户提供了本地化部署的客户端实现语义智能分析的全链条一站式服务，也为软件工程师提供了二次开发接口。NLPIR-Parser平台包含精准采集，文档格式转换、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十三项独立功能，涵盖了从数据的采集预处理、自然语言处理到文本挖掘、信息检索再到可视化呈现、结果导出等全链条各个环节的语义分析工具。

2. NLPIR-Parser的总体架构

NLPIR大数据语义智能分析平台是一个全链条的分析工具，完全本地化部署，不上传用户数据，安全可靠。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、云服务以及二次开发接口，包含了大数据背景下有关语义分析的各个环节的工具，无论对没有任何编程背景但要大量处理语言、媒体信息的文科生辅助处理分析，还是对需要二次开发才能完成特定领域的信息服务都可以满足要求。平台先后历时20年，融入了20年的科研成果。服务了全球40万家机构用户和100余家高校用户，免费给研究人员从事研究工作。

语义智能分析的全链条指的是从语料数据的采集预处理，经过自然语言处理到文本挖掘，信息检索再到可视化呈现和导出以便适合于不同人员的使用需求的全部处理过程。数据收集和预处理中包括了通过主题采集和站点采集从互联网上爬取信息和处理本地上传或录入的信息，同时还提供了不同文档格式转换和编码转换的工具；自然语言处理部分可以进行批量分词、新词发现和主题抽取和语言统计；文本挖掘部分可以进行文本分类、文本聚类、摘要实体生成、智能过滤、情感分析、文档去重；信息检索部分可以进行模糊查询快速全文检索，附带还有文档去重的工具；可视化呈现部分可以画出各种用户喜欢的信息表示图案，如词云图等；导出部分贯穿在各个功能当中，将输出结果导出，用户可以采用导出的内容写入分析报告当中。对于有开发背景的还可以通过API进行二次开发满足特定需要，自动生成分析报告。

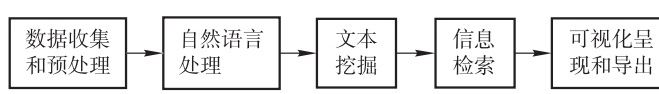


图1　NLPIR全链条大数据语义智能分析平台

具体的功能在第三大部分有详细的描述并给出了实例。

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。

3. NLPIR-Parser功能介绍

3.1　数据收集和预处理

3.1.1. 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，尤其是境外信息与情报的挖掘。用户可自定义采集模式、采集时间、采集区域、采集存储、采集层。采集完成以后，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20190301，指的是2019年3月1日的境内新闻。

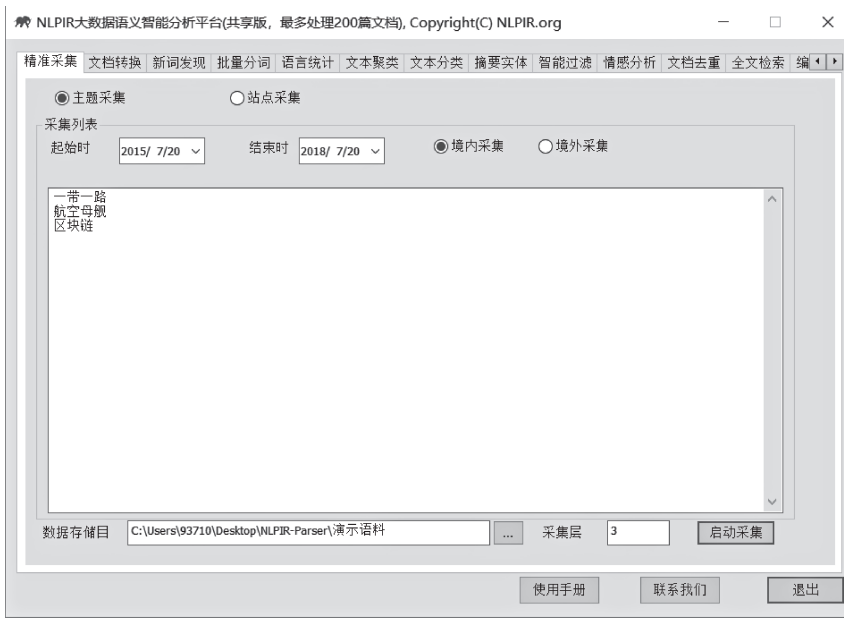
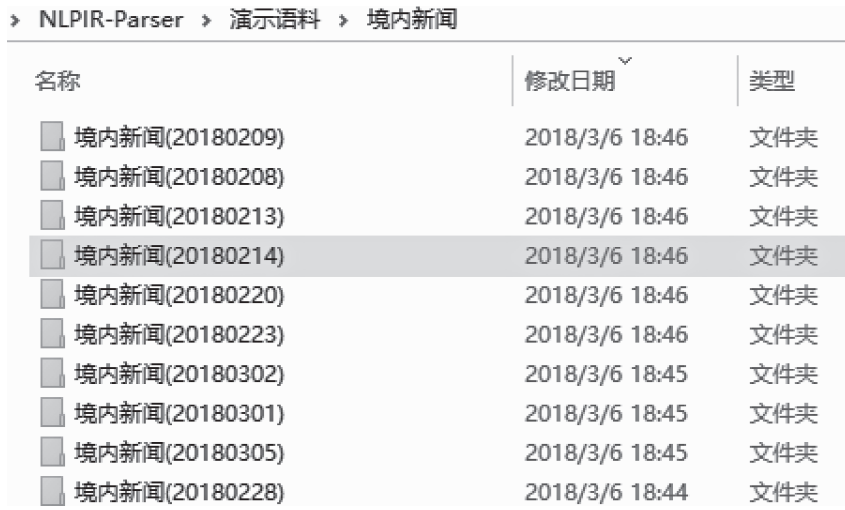


图2　NLPIR大数据语义智能分析平台客户端

（1）主题采集

采集模式选择了“主题采集”，如图2实例所示，按照给定的关键词或主题词进行信息采集时，输入关键词“一带一路”、“航空母舰”与“区块链”等三个主题；采集时间区域（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，这里选取的是2015年的7月20日到2018年的7月20日）；采集区域选择了“境内采集”（如果选择境外采集，需要启动翻墙措施方可使用）；采集层3层。获取主题相关的主流新闻报道、BBS与博客等内容。

采集完成以后，用户可查看采集结果，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。



› NLPIR-Parser › 演示语料 › 境内新闻

名称	修改日期	类型
境内新闻(20180209)	2018/3/6 18:46	文件夹
境内新闻(20180208)	2018/3/6 18:46	文件夹
境内新闻(20180213)	2018/3/6 18:46	文件夹
境内新闻(20180214)	2018/3/6 18:46	文件夹
境内新闻(20180220)	2018/3/6 18:46	文件夹
境内新闻(20180223)	2018/3/6 18:46	文件夹
境内新闻(20180302)	2018/3/6 18:45	文件夹
境内新闻(20180301)	2018/3/6 18:45	文件夹
境内新闻(20180305)	2018/3/6 18:45	文件夹
境内新闻(20180228)	2018/3/6 18:44	文件夹

图3　采集结果文件

（2）站点采集

采集模式选择了“站点采集”，输入站点地址，http://news.sina.com.cn、http://www.nlpir.org、http://www.bit.edu.cn；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，结果如下。

.PIR-Parser › 演示语料 › 站点采集

名称	修改日期	类型	大小
“高铁一姐”曾被多人诈骗 她挖的坑一直...	2018/7/20 9:55	XML 文档	7 KB
“货拉拉”当网约车载客？平台-司机载客...	2018/7/20 9:55	XML 文档	10 KB
“货拉拉”载客视频网上热传 发生事故谁...	2018/7/20 9:55	XML 文档	8 KB
“特普会”发言被骂惨 特朗普急回应：昨...	2018/7/20 9:55	XML 文档	3 KB
“特普会”刚结束 美国就拘捕一名俄女特...	2018/7/20 9:55	XML 文档	2 KB
《九州》拍网剧，“天空城”里一片新面孔...	2018/7/20 9:55	XML 文档	5 KB
2年前交1.5亿彩礼又退婚 “新郎”起诉：...	2018/7/20 9:55	XML 文档	13 KB
3万6就能买到京牌和车是真的？如何参...	2018/7/20 9:55	XML 文档	8 KB
4人集资10.8亿被判刑：千余人参与 实际...	2018/7/20 9:55	XML 文档	3 KB
7岁女儿目睹爸爸救起落水母子 自制奖状...	2018/7/20 9:55	XML 文档	4 KB
14岁女孩从瀑布高处跳水玩耍 被急流卷...	2018/7/20 9:55	XML 文档	2 KB
16岁上大学的他将成福建最年轻市长(图-...	2018/7/20 9:55	XML 文档	6 KB
24岁小伙因还不清高利贷喝农药自杀 抢...	2018/7/20 9:55	XML 文档	3 KB
31名老赖子女就读高收费私立学校 法院...	2018/7/20 9:55	XML 文档	7 KB
31岁富太称能拿紧俏房源 亲友老公都被...	2018/7/20 9:55	XML 文档	7 KB
72岁英国男资助非洲贫困家庭 性虐当地...	2018/7/20 9:55	XML 文档	4 KB

图4　站点采集结果文件

3.1.2　文档格式转换

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。通过文档转换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。

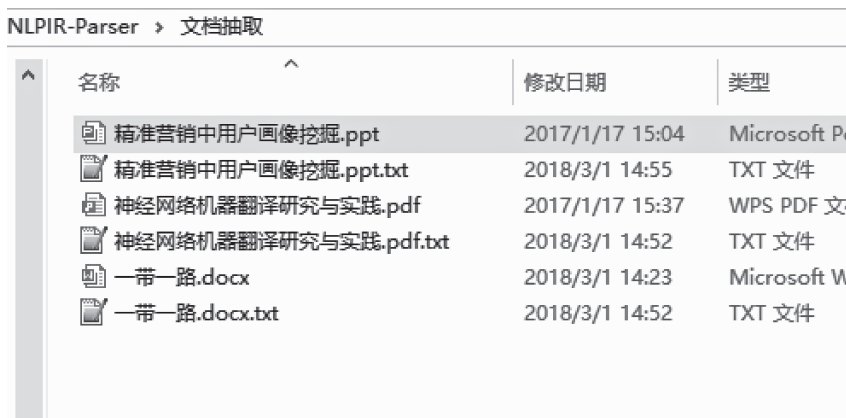


图5　文档转换结果文件

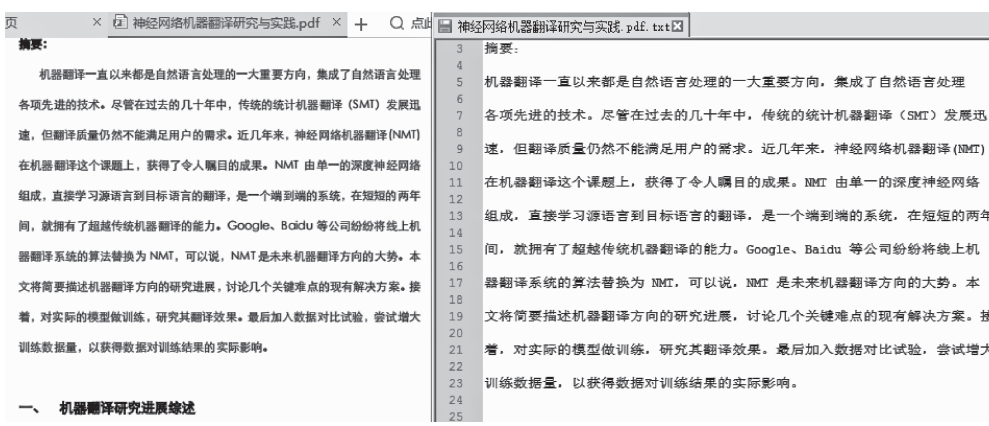


图6　文档转换效果对比

3.1.3　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。

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或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。

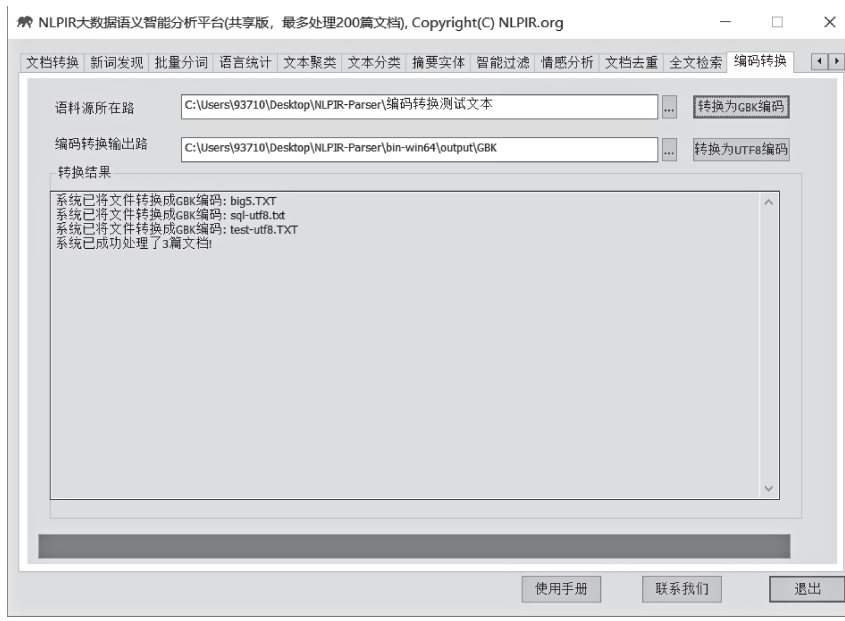


图7　转换为GBK编码

3.2　自然语言处理

3.2.1　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。词频统计及翻译分析结果有四个Excel输出文件。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：$H(X)=-\sum_{1}^{n}P(X)\log P(X)$。

（1）按词频排序的统计结果文件

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。

1	总词数为：3139，所有词的平均频率为：4.355527					
2	词语	词性	词频	一元概率	信息熵	译文
3	的	HLM	625	0.045714	0.141043	target; bull's-eye 有～放矢 shoot the arrow at the target; have a definite object in
4	党	n	195	0.014263	0.060618	①（政党） political party; party ②（指中国共产党） the Party (the Communist Party
5	人民	n	151	0.011044	0.049764	the people; popular (adj.) 世界各国～ peoples of the world ～之间的联系和交流 people
6	是	HLM	145	0.010606	0.048217	①（对；正确） correct; right ②（表示答应） yes; right ～，我就来。 Yes, I'm coming
7	建设	vn	144	0.010532	0.047957	build; construct; construction (n.) 社会主义～ socialist construction ～有中国特色的
8	坚持	v	131	0.009582	0.044535	persist in; persevere in; uphold; insist on; stick to; adhere to ～原则 adhere to pr
9	国家	n	105	0.00768	0.037395	country; state; nation 发展中～ developing countries 中等发达～ moderately developed
10	发展	v	101	0.007387	0.036257	①（变化）develop; expand; grow; development(n.) ～生产力 development of production
11	在	HLM	94	0.006875	0.034238	①（存在；生存） exist; be living ②（表示位置） at 在120 公里处 at 120 kilometers
12	社会	n	93	0.006802	0.033947	society; social (adj.) 工业～ industrial society 农业～ agricultural society 社会主
13	新	a	92	0.006729	0.033654	①（跟“老”或“旧”相对） new; fresh; up-to-date ～发明 a new invention ～技术 new
14	发展	vn	91	0.006656	0.033361	①（变化）develop; expand; grow; development(n.) ～生产力 development of production
15	政治	n	90	0.006583	0.033067	politics; political affairs
16	要	v	90	0.006583	0.033067	①（重要） important; essential ～事 an important matter ②（希望得到） want; ask fo
17	制度	n	89	0.00651	0.032773	①（规章） rules; regulations 税收～ tax rules and regulations ②（体系） system; in
18	推进	vi	81	0.005925	0.030385	①（推动前进） push on; carry forward; advance; give impetus to ～国民经济信息化 try
19	中国	ns	78	0.005705	0.029475	China; Chinese (adj.)
20	体系	n	77	0.005632	0.02917	system; setup 经济～ economic system 思想～ ideological system
21	实现	v	74	0.005413	0.028248	realize; achieve; bring about ～工业化和经济的社会化、市场化、现代化 accomplish indu

图8　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。

（2）按字典排序的词频统计文件

输出到一个名为FreqSortByWord的文件，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、一元概率与信息熵。

（3）Bigrams输出文件

输出到一个名为Bigrams的文件，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84词，频率为0.430769，其信息熵值为0.031287。

（4）文件统计信息输出文件

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。

3.2.2 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。

目前多数的分词算法都采用规则和统计相结合的方法，这样做的目的是为了降低统计对语料库的依赖性，可以将已有的词法信息进行充分利用，同时还能弥补规则方法的不足。现在经常使用方法是利用词典进行初次切分，得出切分结果后，使用其他的概率统计方法和简单规则消歧进行未登录词的识别。NLPIR分词法（Chen *et al*. 2014）利用词典匹配进行初词切分，得到词切分图后，利用词频信息求词图N条最短路径的N最短路径法。

图9 分词结果文件

3.2.3　新词发现

新词发现（张华平、商建云2017）模块包括新词提取与关键词提取两个功能。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，无须再次设计美化。

新词发现能从文本中挖掘出具有内涵的新词、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，导入分词词典可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。

（1）新词提取

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。

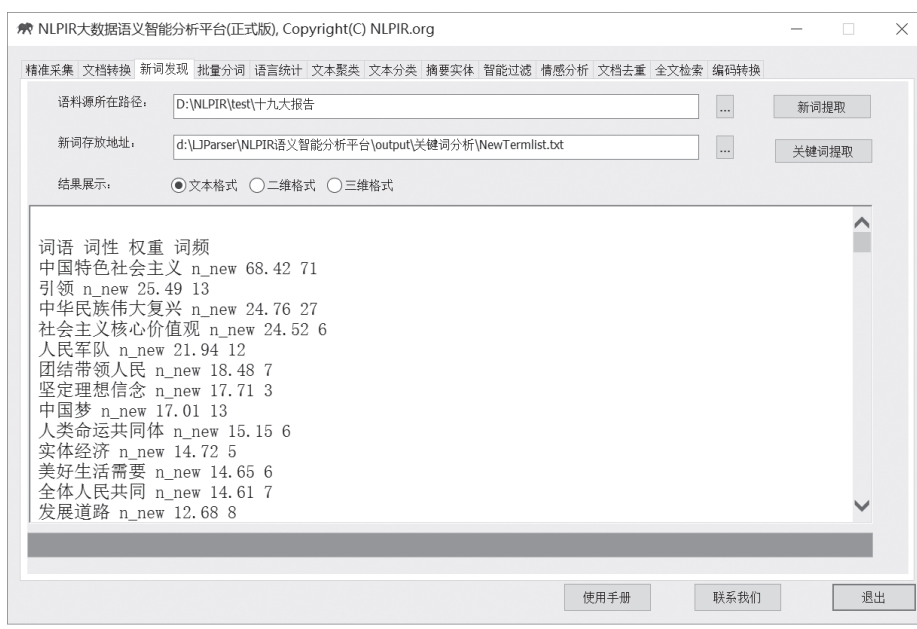


图10　新词提取

（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。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。

图11　关键词提取

3.3　文本挖掘

3.3.1　文本分类

一个文本分类问题就是将一篇文档归入预先定义的几个类别中的一个或几个，而文本的自动分类则是使用计算机程序来实现这样的分类。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。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%。

文本分类（赵连伟等2014）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“中国建筑”类别，可定义该类别的规则：“长城；牌坊；园林；寺院；钟；塔；庙宇；亭

台楼阁；井；石狮；民宅；秦砖汉瓦；兵马俑；故宫；紫禁城；颐和园；布达拉宫；平遥古城；乔家大院；苏州园林；杭州园林；徽派建筑；十里长亭；长城；天坛;鸟巢；水立方”，系统会根据文本中出现的特征词语判定文本类别为：中国建筑。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。

通过“专家规则分类过滤”、“机器学习分类过滤”，分类结果会呈现在结果提示框中。

图12　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186,964，共有1,000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。

3.3.2　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。

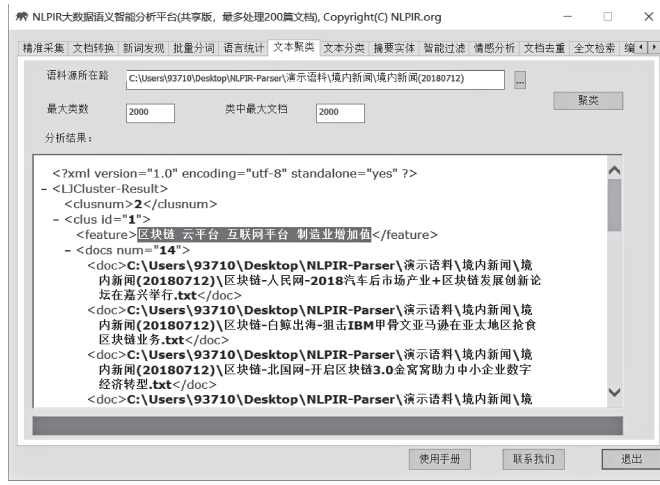


图13　聚类

图14　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。

3.3.3　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。

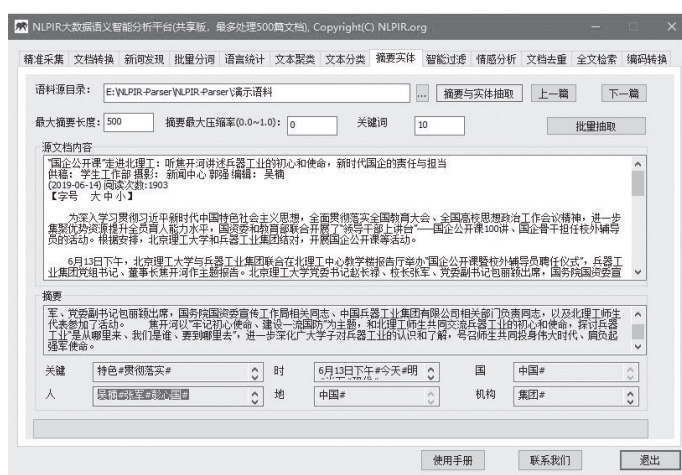


图15　摘要与实体抽取

北理工校内新闻的分析结果如下：

摘要（250字）“国企公开课”走进北理工：听焦开河讲述兵器工业的初心和使命，新时代国企的责任与担当，北京理工大学党委书记赵长禄、校长张军、党委副书记包丽颖出席，国务院国资委宣传工作局相关同志、中国兵器工业集团有限公司相关部门负责同志，以及北理工师生代表参加了活动。

关键词：北理工#北理工师生#兵器工业#兵器工业集团#焦开河#国企公开课#包丽颖#北京理工大学#中国特色#贯彻落实#

人物：包丽颖#习近平#赵长禄#郭强#吴楠#张军#彭心国#

时间：6月13日下午#今天#明#当下#现代#

国家：中国#

机构：#党中央#中国共产党#教育部#中国兵器工业集团#

3.3.4　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。

系统已内置约10类近4万关键词，用户仍可根据需求用“导入关键词”添加个人的关键词；用“批量扫描”，系统进行不良信息过滤；还可以用“打开文件”或者直接将扫描文本粘贴至文本框中进行输入扫描

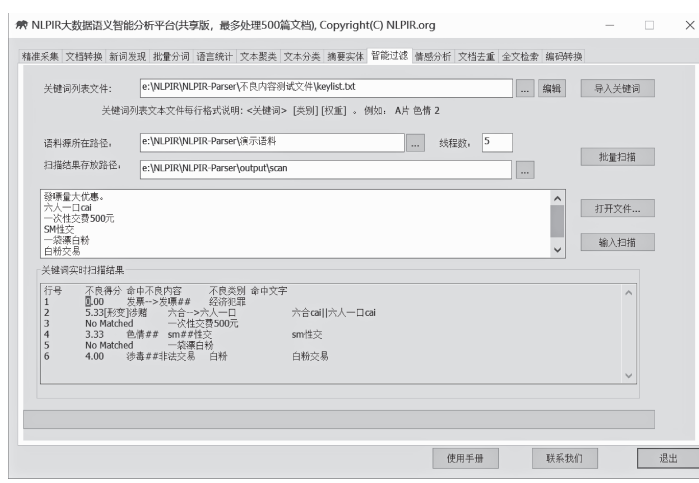


图16　输入扫描

3.3.5　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。

可以单个对象分析和批量对象分析来进行情感分析。

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。

情感分析	区块链													
文档总数	200	负面总数	2	负面占比	1.00%	正面总数	12	正面占比	6.00%					
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							
银保监会	百度新闻搜索-网易	2018/7/8 15:37	-2	1	-3	http://mo	区块链-财联社-银保监会国际部负责人范文仲：不要神化区块链.txt							
区块链板	百度新闻搜索-网易	2018/7/10 13:49	-1	0	-1	http://mo	区块链-上海证券报·中国证券网-区块链板块拉升宜亚国际涨停.txt							
相关传销	百度新闻搜索-网易	2018/7/10 10:14	1	77	-76	http://na	区块链-人民网-相关传销平台已超3000家“区块链”竞成“上亿大坑”.t							
蚂蚁金服	百度新闻搜索-新浪	2018/7/8 15:03	6	10	-4	http://f	区块链-新浪财经-蚂蚁金服韩歆毅：ICO不是真正的区块链金融服务.txt							
【三言】	百度新闻搜索-新浪	2018/7/6 8:31	11	15	-4	http://f	区块链-三言财经-【三言】中国区块链50城之长沙：将设30亿产业基金企							
世界杯点	百度新闻搜索-新浪	2018/7/7 2:04	17	57	-40	http://f	区块链-证券日报-世界杯点燃区块链预测“引线”宣称“只赢不输”难避							
快讯：区	百度新闻搜索-新浪	2018/7/10 13:29	18	19	-1	http://f	区块链-新浪财经-快讯：区块链概念股集体走强宜亚国际涨停.txt							
区块链投	百度新闻搜索-网易	2018/7/5 8:32	20	38	-18	http://mo	区块链-新京报-区块链投资切勿盲目跟风“网红”的诱惑.txt							
合茂数字	百度新闻搜索-腾讯	2018/7/2 14:59	28	31	-3	http://b	一带一路-北国网-合茂数字发起成立“一带一路数字经济联盟”.txt							
梅地卡尔	百度新闻搜索-网易	2018/6/29 19:50	38	52	-14	http://mo	区块链-证券日报-梅地卡尔利用区块链技术发力医疗数据交换系统.txt							
区块链+	百度新闻搜索-网易	2018/7/6 9:21	39	55	-16	http://mo	区块链-钛媒体-区块链+人工智能如何重新定义世界黑科技？.txt							
区块链跨	百度新闻搜索-新浪	2018/7/5 9:18	41	66	-25	http://f	区块链-每日经济新闻-区块链跨国“围城”：中韩两国都希望寻求外部空							
迅雷发布	百度新闻搜索-新浪	2018/7/9 18:18	54	86	-32	http://f	区块链-21世纪经济报道-迅雷发布区块链文件协议TCFS陈磊：技术会发生							
央行数字	百度新闻搜索-网易	2018/7/1 8:13	92	113	-21	http://mo	区块链-21世纪经济报道-央行数字货币研究所蒋国庆：法定数字货币与区							

图17　单个对象（区块链）的情感分析结果

对象：区块链，情感得分：38，正面得分：52，负面得分：-14

“多对象分析”，系统开始对多个对象进行情感分析。

情感分析	C:\Users\93710\Desktop\NLPIR-Parser\stConduct.xml													
文档总数	200	负面总数	4	负面占比	2.00%	正面总数	32	正面占比	16.00%					
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							
解构美国	百度新闻	2012/9/30 7:58	-4	1	-5	https://i	航空母舰-新华网-解构美国航母战斗群.txt							
专访华侨	百度新闻	2011/11/20 17:38	-1	3	-4	https://i	航空母舰-新华网-专访华侨城有关负责人：打造中国文化产业“航空母舰”.txt							
我国第二	百度新闻	2017/4/27 2:51	-1	1	-2	http://ne	航空母舰-信息时报-我国第二艘航空母舰昨日下水.txt							
昨天航母	百度新闻	2017/4/27 22:05	-1	0	-1	http://ne	航空母舰-每日经济新闻-昨天航母刚刚下水 又一“国之重器”将刺破苍穹.txt							
日本航母	百度新闻	2017/4/27 9:05	0	3	-3	http://bi	航空母舰-宋鸿兵-日本航母命名显狼子野心！国产航母001A如何命名？.txt							
韩美军费	人民网-人	2018/4/14 5:35	0	0	0	http://ww	航空母舰-人民网一人民日报-韩美军费谈判分歧加大.txt							
韩美军费	人民网-人	2018/4/16 9:02	0	0	0	http://k	航空母舰-人民网-人民日报-韩美军费谈判分歧加大.txt							
乌克兰媒	百度新闻	2011/7/30 12:52	1	1	0	https://i	航空母舰-环球网-乌克兰媒体称中国正在建造两艘国产航空母舰.txt							
世界杯点	百度新闻	2018/7/7 2:04	1	1	0	http://f	区块链-证券日报-世界杯点燃区块链预测“引线”宣称“只赢不输”难避博彩嫌							
美容连锁	百度新闻	2013/11/27 16:49	1	1	0	http://sl	航空母舰-中华网-美容连锁行业航空母舰克丽缇娜成功挂牌上市.txt							
日本企业	百度新闻	2018/6/20 9:16	1	1	0	http://mo	一带一路-环球时报-环球网-日本企业正在借助“一带一路”倡议寻找中欧商机.							
雪人保安	百度新闻	2013/1/6 7:52	2	2	0	http://z	航空母舰-浙江在线-雪人保安要是给我很多雪，就能堆个航空母舰.txt							
韩媒集体	百度新闻	2017/4/19 9:28	2	2	0	http://ne	航空母舰-综合-韩媒集体乌龙 称美国三艘航母将抵朝鲜遭美方打脸.txt							
重庆：融	百度新闻	2018/7/6 10:41	2	2	0	http://co	一带一路-人民网-重庆：融入’一带一路’建设打造内陆开放高地.txt							
恒天然受	百度新闻	2018/7/3 10:50	2	3	-1	http://t	一带一路-中国网-恒天然受邀加入“一带一路”奶业联盟发力中国鲜奶市场.txt							
“一带一	百度新闻	2018/7/9 15:19	2	2	0	https://i	一带一路-新华网-“一带一路”助中阿构建命运共同体.txt							
韩美军费	人民网-人	2018/7/3 4:55	3	3	0	http://ww	航空母舰-人民网一人民日报-韩美军费分担谈判无果而终.txt							

图18　情感批量分析结果

3.3.6　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。

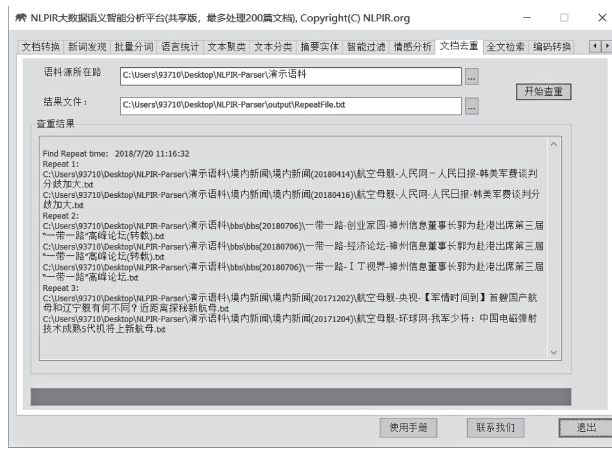


图19　文档去重

3.4　文本检索

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

检索结果包括：文档总量统计、标题、内容与相似得分。

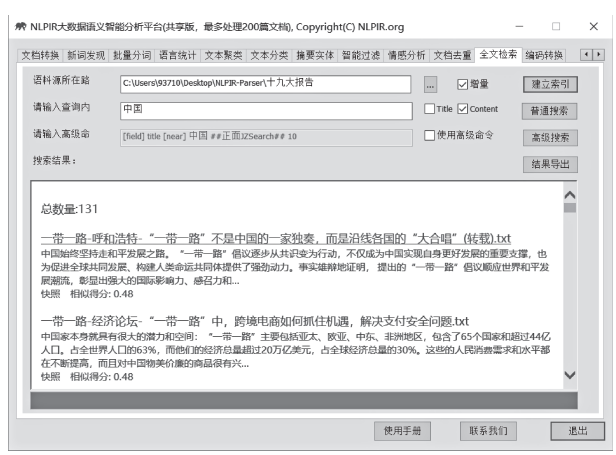


图20　普通检索

图21　搜索结果

输入高级命令。例如：[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。

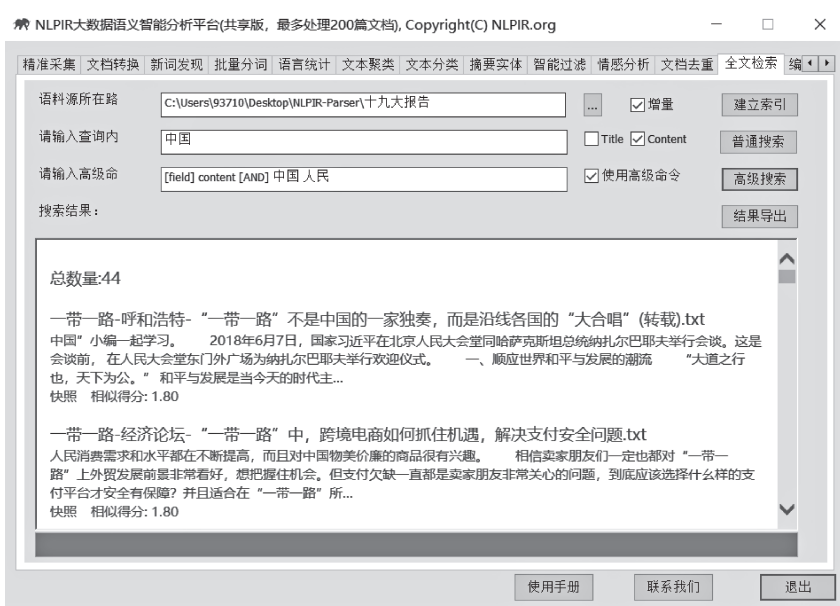


图22　高级检索

3.5　可视化展示

“结果展示”选的是二维格式：top42 词汇的词云形式展示效果如图23所示。

图23　二维格式

“结果展示”选的是三维格式：top20词汇的三维动态展示，简洁美观。

图24　三维格式

4. 结语

NLPIR大数据语义智能分析平台为语义分析提供数据和技术支持，在大数据背景下，可以满足常见的需求，支持用户专业词典与微博分析、支持多种编码、多种操作系统、多种开发语言与平台。一方面为语言处理，提供友好、实用的工具，另一方面为软件开发人员和研究人员提供二次开发的接口满足特定的空间信息处理和应用程序的需要；同时也为研究者提供统计数据和实例支撑。

此平台是不断完善的，最新客户端的二次接口定期会发布白皮书，不断把新的研究成果融入平台以满足各种不同的需要，并提供更大规模的词典库以及更多的语料。最新的在线演示、客户端下载以及用户手册均可以在http://www.nlpir.org/站点获得，欢迎更多的研究者和大数据分析者使用并提出宝贵意见。

参考文献

Chen, X., J. Shang, Y. Zhao & X. Shi. 2014. A comparison of part-of-speech tag sets for transition-based dependency parsing [A]. In *The Second International Conference on Information Technology and Computer Application Engineering* [C]. 191-194.

张华平、商建云，2017，面向社会媒体的开放领域新词发现 [J],《中文信息学报》(3)：115-121。

赵连伟、史学文、张华平、商建云，2014，文本分类字词特征对比研究 [A]，载《第三届全国社会媒体处理会议论文集》[C]。120-123。

通信地址： 100081 北京市北京理工大学计算机学院