# 智分析使用手册

©码有引力 (重庆) 科技有限公司

最新版请前往智分析官网—产品手册查看: <u>https://saz.codeghub.com</u>

# 目录

智分析使用手册 目录 前言 快速入门 四步开启您的智分析之旅 结果调试 功能介绍 项目/词库管理 项目/词库分享(协同分析) 开始分析 数据灵感 计数项分析 计数项条形图 计数项扇形图 计数项词云图 计数项趋势分析折线图 计数项词频堆积图 计数项TF-IDF条形图 词频统计图 TF-IDF统计图 词云图 发文趋势图 词频堆积图 TF-IDF趋势变化图 新词发现 关键词提取 关键词提取 绘制条形图 绘制词云图 绘制堆积图 绘制折线图 计数项分析 单词图形 多词-堆积图 多词-多指标柱形图 多词-条形图 词关联性分析 一对多分析 数值分析 关系图 距离条形图 共现概率图 多对多分析 共现矩阵 相似矩阵 共现多维度图 相似多维度图 词对分析 词定位 主题分析 主题聚类 设置主题数

编辑主题词 主题归纳 主题树图 主题河流图 主题趋势图 主题模型 社区划分 词库编辑 分析词库 停用词库 同义词库 查看文件, 重新分析 进阶使用 基础统计 计数项 时间项 多维分析 分词与词库

分词 分析词库 停用词库 同义词库 新词发现/短语抽取 词语分析 关键词分析 词频统计 TF-IDF 词关联性分析 共现 词向量 词距离/相似度 矩阵分析 MDS降维 K-means聚类 主题挖掘 主题聚类 主题模型 社区划分 常见问题

# 前言

智分析是一款基于自然语言处理技术的文本分析工具,同时通过可视化技术,大大减小了使用自然语言处理技术的门槛。

智分析主打傻瓜式、自动化,最大限度减少人工成本,完全不懂技术的用户也可以无障碍使用。

同时智分析秉着 "人工+智能=越用越智能" 的理念,使人工仅专注于领域内知识的刨析,相关技术性的东西可以完全交给智分析来处理,且人工也会使智分析更加的智能,分析得到更准确的结果

# 快速入门

# 四步开启您的智分析之旅





©码有引力(重庆)科技有限公司渝ICP备19011279号

2. 点击"+"添加项目,输入项目名称并上传数据文件(数据文件暂支持excel、txt文件),点击新建项目完成项目创建

成的項目     我的词单	
新建项目       ×         习近平讲话         习近平讲话 xlsx       演覧         資産項目词库       选择已有词库	
新建项目       ×         习近平讲话         习近平讲话 xisx       激沈         ● 新建项目词库       送择已有词库	
习近平讲话         フ近平讲话 xlsx       演览         ⑥ 新建项目词库       选择已有词库	
フ近平讲话 xisx 対応 新建项目词库 选择已有词库	
新建项目词库选择已有词库	
新建项目	

## 3. 创建项目成功后点击您所创建的项目,进入数据浏览页面



4. 点击"开始分析"按钮, 稍等片刻, 即可开始您的智分析之旅~



# 结果调试

可从数据灵感结果中看到,有的词语并不是我们想要的,比如:"我们"、"年月日"等;有的词语比较简单,并未表达出准确意思,比如:"问题"、"坚持"等。怎么让结果更准确呢?

1. 进入新词发现, 筛选出新词短语, 并右键添加到词库中



2. 筛选添加到词库完毕后,进入关键词分析,查看排名靠前的词语中有没有我们不想要的,如果有,则右键添加到停用词库中





@码有리力 (重庆) 科技有關公司 渝ICP条19011279号

3. 添加完毕后,点击<查看文件>,再点击重新计算,即可进行重新分词

灵感之智分析				
① 当前项目: 习	近平讲话			
序号	id	和示意的	时间	内容
1	1	习近平:把权力关进制度的笼子里	2013/1/21	中共中央总书记、中共中央军委主席习近平 2 2 日在北京
2	2	习近平总书记在十八届中央纪委第二次全会上重要讲话	2013/1/22	我们党员干部队伍的主流始终是好的。同时,我们也要清
3	3	中共中央政治局召开会议习近平主持	2013/1/27	中共中央政治局召开会议研究部署加强新形势下党员发展
4	4	全面贯彻潜实党的十八大精神要突出抓好六个方面工作	2013/1/3	第五,全面推进党的建设新的伟大工程。新形势下,我们
5	5	习近平:把"枫桥经验"坚持好、发展好把党的群众路线坚	2013/10/11	人民网杭州10月11日电 (记者王比学) 中共中央总书记、
6	6	习近平主持中共中央政治局会议讨论拟提请十八届三中全	2013/10/29	中共中央政治局10月29日召开会议,讨论十八届二中全会
7	7	习近平:不断提高军队党的建设科学化水平为实现强军目	2013/11/5	习近平在接见全军党的建设工作会议代表时强调不断提高
8	8	(党规)建立健全惩治和预防腐败体系2013—2017年工作	2013/12/25	为深入贯彻落实党的十八大和十八届三中全会精神,加强
9	9	习近平:推动全党学习和掌握历史唯物主义	2013/12/3	习近平在中共中央政治局第十一次集体学习时强调推动全
10	10	中央政治局召开会议决定成立中央全面深化改革领导小组	2013/12/30	中共中央政治局召开会议决定成立中央全面深化改革领导
11	11	习近平听取河北省委党的群众路线教育实践活动总体情况	2013/12/8	习近平在听取河北省委党的群众路线教育实践活动总体情
12	12	(党规)关于加强新形势下发展党员和党员管理工作的意见	2013/2/24	为深入贯彻潜实党的十八大精神,保持党的先进性和纯洁
13	13	习近平在中央党校建校80周年庆祝大会暨2013年春季学期	2013/3/2	同志们:今天,我们在这里集会,庆祝中央党校建校80周
14	14	中共中央政治局召开会议研究部署在全党深入开展党的群	2013/4/19	中共中央政治局4月19日召开会议,决定从今年下半年
15	15	习近平在党的群众路线教育实践活动工作会议上的讲话	2013/6/18	围绕保持党的先进性和纯洁性,在全党深入开展以为民务
		· · · · · · · · · · · · · · · · · · ·	共11页 <b>跳转</b>	点击进行重新分词
		进入项目重新	い いちょう いちょう いちょう いちょう いちょう しんしょう いちょう しんしょう しんしょ しんしょ	

@码有引力 (董庄) 利持有限公司 淪ICP&19011279月



4. 反复进行以上结果调试步骤,直到结果满意为止



# 功能介绍

# 项目/词库管理

在<我的项目/词库>页面可进行项目查看、创建或删除操作,同时, <项目/词库>支持分享,分享后,别人即可获得您的分析结果,还能与您 进行协同分析。

<我的词库>页面不能进行词库创建, 词库的创建是在创建项目时勾选新建项目词库创建的

# 项目/词库 分享(协同分析)

只有<项目/词库>创建者才能进行<项目/词库>分享操作

1. 点击<项目/词库>分享按钮



2. 输入对方的用户名,编辑分享权限,权限分为"可读"和"可编辑"

项目权限解释:

- 可读:对方只能浏览您的项目结果,不能进行新词发现调参计算、词关联性分析初始化、主题分析重置、重新分词计算等改 变您计算结果的操作。一般用于项目结果分享
- 可编辑:对方拥有可读权限的同时,可以进行新词发现调参计算、词关联性分析初始化、主题分析重置、重新分词计算等操作。一般用于项目协同分析

词库权限解释:

可读:对方有读取您的词库权限,并可用您的词库进行项目分词计算,但不能进行词库编辑。一般用于词库结果分享,借用
 可编辑:对方拥有可读权限的同时,可以进行词库编辑。一般用于词库协同共建

数据灵感之智分析							
۵		我的项目	我的词库				
	<b>习近平讲话</b> <b>基本信息</b> 创建者: smartanaly.	Z ā					
+	分享项目				×		
		共享用户名					
	[	īj	读	$\checkmark$			
	I		分享				
				708			
		())))))))))))))))))))))))))))))))	VERV/201000020011				

#### 3. 点击查看成员信息可查看<项目/词库>成员

数据灵感之智分析			() ()
Ô	我伯	り项目 我的词库	
	J近年明時       「         ●       基本信息         回想所領:       2020-12-24 20:50:03         通常頂領:       2020-12-24 20:50:03         进作取領:       ごを用いる         通常の目前       通信の目前         ●       日日の目前         ○       日日の目         ○       日日の日         ○       日日         ○       日日         ○       日日         ○       日日         ○       日日         ○       日日         ○       日         ○       日         ○       日 <th>3 章 点击查看成员信息</th> <th></th>	3 章 点击查看成员信息	
	©码有引力	(重庆) 科技有限公司 渝ICP备19011279号	

数据灵感之智分析				
۵	我的项目	我的词库		
	Л近平讲话 🚺 🗇			
	项目成员		×	
	用户名	权限	操作	
	smartanalyze	可转让	删除	
	alextan	可编辑	删除	

@码有引力 (重庆) 科技有限公司 淪ICP备19011279号

## 同时在<项目/词库>成员页面可进行项目成员删除操作(取消分享)

# 开始分析

进入项目后,点击开始分析,智分析便能帮您进行全自动分析,在这里,智分析会自动识别您所上传文件的内容,其中最重要的是内 容列、时间列、计数项列(分类列),然后自动提取内容列里所有单元格的内容进行分词计算,同时智分析会通过算法匹配可能能结 合分析的列(项)进行多维度分析计算

序号	id	标题	时间	内容
1	1	习近平:把权力关进制度的笼子里	2013/1/21	中共中央总书记、中共中央军委主席习近平22日在北京
2	2	习近平总书记在十八届中央纪委第二次全会上重要讲话	2013/1/22	我们党员干部队伍的主流始终是好的。同时,我们也要清
3	3	中共中央政治局召开会议习近平主持	2013/1/27	中共中央政治局召开会议研究部署加强新形势下党员发展
4	4	全面贯彻落实党的十八大精神要突出抓好六个方面工作	2013/1/3	第五,全面推进党的建设新的伟大工程。新形势下,我们
5	5	习近平:把"枫桥经验"坚持好、发展好把党的群众路线坚	2013/10/11	人民网杭州10月11日电 (记者王比学) 中共中央总书记、
6	6	习近平主持中共中央政治局会议讨论拟提请十八届三中全	2013/10/29	中共中央政治局10月29日召开会议,讨论十八届二中全会
7	7	习近平:不断提高军队党的建设科学化水平为实现强军目	2013/11/5	习近平在接见全军党的建设工作会议代表时强调不断提高
8	8	(党规) 建立健全惩治和预防腐败体系2013—2017年工作	2013/12/25	为深入贯彻落实党的十八大和十八届三中全会精神,加强
9	9	习近平: 推动全党学习和掌握历史唯物主义	2013/12/3	习近平在中共中央政治局第十一次集体学习时强调推动全
10	10	中央政治局召开会议决定成立中央全面深化改革领导小组	2013/12/30	中共中央政治局召开会议决定成立中央全面深化改革领导
11	11	习近平听取河北省委党的群众路线教育实践活动总体情况	2013/12/8	习近平在听取河北省委党的群众路线教育实践活动总体情
12	12	(党规)关于加强新形势下发展党员和党员管理工作的意见	2013/2/24	为深入贯彻落实党的十八大精神,保持党的先进性和纯洁
13	13	习近平在中央党校建校80周年庆祝大会暨2013年春季学期	2013/3/2	同志们:今天,我们在这里集会,庆祝中央党校建校80周
14	14	中共中央政治局召开会议研究部署在全党深入开展党的群	2013/4/19	中共中央政治局4月19日召开会议,决定从今年下半年
15	15	习近平在党的群众路线教育实践活动工作会议上的讲话	2013/6/18	围绕保持党的先进性和纯洁性,在全党深入开展以为民务
		首页上一页下一页尾页第一	共11页 跳转	

# 数据灵感

数据灵感旨在为用户提供一键智能化分析的功能,顾名思义,一键为用户带来数据的灵感,根据所上传的数据,呈现出相应的图,所 有图形支持直接保存到本地。



# 计数项分析

计数项指的是excel表格中可以计数的一项,或者可以理解为能分类的一项,比如下列表格中的"态度"可分为:积极、消极、中性。

智分析会根据表格内容自动分析出可能作为"计数项"的一列,自动呈现出计数项图,如果没有找到计数项,则不会呈现。

如果系统判断有误,或者想对其他项进行计数,也可以在图形中手动选择计数项。

部分图形计数项和计数类支持用户手动选择

	Tan Alex TA	• - • ×
文件 开始 插入 页面布局 公式 数据 审阅 视图 帮助 ACROBAT		☆ 共享 □ 批注
●       ●	∑         自动求和         ▲         ○           順除         偕式         填充→         排序和筛选 查找和选择           ◇         消除◇         ○	分析     敏感       数据     性 ~
剪贴板 15 字体 15 对齐方式 15 数字 15 样式 单近	元格编辑	分析 敏感度 へ
02 • : × ✓ fr		*



#### 计数项条形图

当数据中含有计数项时会呈现此图, 该图会对计数项进行数量统计, 横坐标代表不同的类别, 能直观展示出不同类别的数量。



#### 计数项扇形图

当数据中含有计数项时会呈现此图, 该图对计数项进行了数量统计, 图中不同颜色代表计数项中不同的类别, 可整体展现不同类别所 占的数量比例。



## 计数项词云图

当数据中含有计数项时会呈现此图,可对计数项中某一类别中的内容进行词频统计,绘制成词云图。该图中词频越大,词的大小就越大,能够提炼出某一类别中的关键词语。

词频:即词语在全文中所出现的次数

数据灵感之智分析		©
○ 当前项目: 计数项测试	数据灵感 新词发现 关键词提取 词关联性分析 主题分析 词库编辑	③ 查看文件
数据灵感		
计数项分析 计数项 请选择	└ 対数类 消极	返回
		C 7



#### 计数项趋势分析折线图

当数据中同时拥有时间项、计数项时会呈现此图,将计数项与时间项结合分析,按时间统计某一类别的数量,绘制成折线图,不同颜 色代表不同类别,能表示出不同类别的趋势信息。

智分析会自动识别判断时间项,并选择合适的时间间隔,时间间隔也可以手动选择。



#### 计数项词频堆积图

当数据中包含计数项时会展现此图, 该图将计数项和词频结合分析, 统计某一类别中某个词语的词频, 绘制成堆积图, 不同颜色代表 不同类别,能表现出词语在各类别中所占的数量与比例,默认取词频排名前10的词语。



#### 计数项TF-IDF条形图

当数据中包含计数项时会展现此图, 该图将计数项和TF-IDF信息结合分析, 统计某一类别中某个词语的TF-IDF值, 绘制成条形图, 能表 现出各类别中的关键词语以及词语在类别中的关键程度,默认取TF-IDF排名前10的词语。

数据灵感之智分析			©   U
△ 当前项目: 计数项测试	数据灵感 新词发现 关键词提取	词关联性分析 主题分析 词库编辑	③ 查看文件
数据灵感			
计数项分析 计数项 态度	✓ 计数类 消极 ✓		返回
1		数量	



# 词频统计图



### 统计全文词频信息,取排名前十的词语,横坐标为词语,纵坐标为词频,能表现出全文中排名前十的关键词词频信息

# TF-IDF统计图

统计全文TF-IDF信息,取排名前十的词语,横坐标为TFIDF值,纵坐标为词语,能表现出全文中排名前十的关键词TF-IDF信息



# 词云图

统计全文的词频信息, 取排名前200的词语, 绘制成词云图, 能展现出全文中的关键词信息

数据灵感之智分析			$^{\odot}$   $^{\odot}$
☆ 当前项目: 习近平讲话	数据灵感 新词发现 关键词提取	词关联性分析 主题分析 词库编辑	③ 直看文件
数据灵感			
词云图			返回
斗争			
作用 做到		「「「「」」、「」、「」、「」、「」、「」、「」、「」、「」、「」、「」、「」、	教育改革体系
创新坚定	於流反肢		



@码有引力 (重庆) 科技有限公司 渝ICP条19011279号

# 发文趋势图

当数据中含有时间项时呈现此图,按时间统计文章数量,能表现出发文数量趋势



# 词频堆积图

当数据中含有时间项时呈现此图,按时间统计词频信息,绘制成堆积图,不同颜色代表不同的关键词,能表选出各个时间段里的词频 信息。



# TF-IDF趋势变化图

当数据中含有时间项时呈现此图,按时间统计TF-IDF信息,绘制成折线图,不同颜色代表不同的关键词,能表现出关键词TFIDF随时间 变化的趋势。

数据灵感之智分析		© _
○ 当前项目: 习近平讲话	数据灵感 新词发现 关键词提取 词关联性分析 主题分析 词库编辑	<ul> <li>查看文件</li> </ul>
数据灵感		
TFIDF趋势变化图    时间间隔	日月年	返回
	党内	
2.5 -		
2		
1.5 -		



©码有引力 (重庆) 科技有限公司 淪ICP备19011279号

# 新词发现

新词发现又叫做短语抽取,是通过左右熵、互信息算法自动计算出有可能组成短语的新词,计算出的新词一般是通过两两词语组成, 其设计目的在于帮助用户减小词库建设的人工成本,用户可直接根据智分析自动计算得到的结果进行人工筛选后添加到词库,以此来 提高智分析的分词准确性。

最小互信息:即互信息小于此设定值的词语将为被排除,建议取值范围5-12,可根据数据量自行调节

权重比:指互信息与左右熵的权重比例,关于互信息和左右熵的解释详见<进阶使用-分词与词库-新词发现>

新词指数:一般来说新词值数越高,越可能是一个短语(新词)

共现次数:组成短语的两个词语共同出现在一句话中的次数

相似度:组成短语的两个词语的关联性,关于其理论知识详见<进阶使用-词语分析-词关联性分析-相似度>

扇形图中的比例:

本功能中的扇形图里比例的计算公式为:

X/(Count A+Count B+Count AB)

因此,本图中各个颜色的比例越相近,越可能是一个新词短语



# 关键词提取

在关键词提取功能里,用户可根据词频或TF-IDF值排序,查看所有关键词,并勾选自己想要的关键词进行绘图或导出为表格文件,同时 关键词支持搜索功能。

# 关键词提取

关键词提取默认根据词频绘制词云图,用户也可把词频切换为TF-IDF,同时还可根据用户所勾选的词语绘制条形图、堆积图、折线图, 所有图形均可下载到本地,同时支持把勾选的关键词信息导出为表格文件。

由于图形最大显示数量原因,支持勾选的词语数量为2-20个

导出表格时不限制词语勾选数量

#### 数据灵感之智分析 <u></u> 数据灵感 新词发现 关键词提取 词关联性分析 主题分析 词库编辑 查看文件 关键词提取 计数项统计 可切换为TFIDF 数据可视化 关键词 Q 搜索词语 词语 词频 C 1 ≙ 重庆 3343 开学 2872 ഥ 开学时间 810 希望工作在家 长江上游地区 $\sim$ 快点科技中心中国 准备时候已经 讨论 662 ≡⊠ 阅读 659 <sup>重庆市</sup>城市 什么时候开学 学校 516 确诊病例 明确开学时间 跳底 继续 大学 大学 321 学院 高三 不想 <sup>综合</sup> 讨论 学院 262 疫情 清零 有缘人 二级 国际 疫情 229 11 国家 新鮮事 多地 学校 ■ día 城市 228 应急响应调整 点击可放大 本地热门溶讯 213 二级



#### 绘制条形图

任意勾选2-20个词语,点击绘制条形图图标,即可绘制出图形。



## 绘制词云图

词云图无需勾选词语,默认选取当前页前50个词语进行词云图绘制,点击绘制词云图图标即可绘制图形。



# 绘制堆积图

任意勾选2-20个词语,点击绘制堆积图图标,即可绘制出图形。

注意:绘制该图的数据中必须包含时间字段。



@码有引力 (重庆) 科技有限公司 渝ICP&19011279号

#### 绘制折线图

任意勾选2-20个词语, 点击绘折线图图标, 即可绘制出图形。

注意: 绘制该图的数据中必须包含时间字段。



# 计数项分析

该功能支持计数项和关键词的结合分析,用户可手动选择计数项和计数类,对某个类别中的关键词进行更深入的分析,同时也支持用 户手动勾选词语进行图形绘制与表格导出功能。



# 单词图形

单词指单个词语,该图形由条形图和饼图组成,主要用于展现所选择的词语在不同类别中的数量/TFIDF和占比。



切换词语后右边图形也会随之切换为该词结果。

#### 多词-堆积图

任意勾选2-20个词语,点击绘制堆积图图标,即可绘制出图形。和关键词提取中的堆积图不同的是,该图中横坐标代表不同词语,不同颜色代表不同的类别(计数类)。



## 多词-多指标柱形图

任意勾选2-20个词语,点击绘制多指标柱形图图标,即可绘制出图形。该图横坐标表示不同类别,不同颜色的柱子代表不同词语,纵 坐标表示词频或TFIDF。



### 多词-条形图

任意勾选2-20个词语,点击绘制条形图图标,即可绘制出图形。该图横坐标表示词频或TFIDF,纵坐标表示不同词语,切换不同类别可 展现不同类别中的关键词。



# 词关联性分析

词关联性分析的设计目的主要是挖掘词与词之间的关联性,同时根据选择不同的词语,亦或者是主题词,可深入挖掘主题与主题之间 的关联性。

初始化:选择需要进行词关联性分析的词语,及距离对象和共现标准

距离对象:计算词向量(word2vec)时的不同算法

词语:对应CBOW模型,可以通俗理解为计算相似度是比对词语本身的相似度

上下文:对应Skip-gram模型,可以通俗理解为计算相似度是比对词语上下文的相似度

关于其理论知识详见<进阶使用-词语分析-词关联性分析-词向量>

共现标准:统计共现次数的衡量标准

以文章:两个或两个以上词语共同出现在同一篇文章中,则计为共同出现,并统计其次数

以句子:两个或两个以上词语共同出现在一句话中,则计为共现,并统计其次数



# 一对多分析

一对多分析用于以一个词为中心,分析它与其他词之间的关系

#### 数值分析

点击<查看数值>图标按钮可具体查看与中心词关联性最强的前十个词语的关系数值,其中包括单独出现的次数、共现次数、共现比例。

NK abbra	50. 100							*) 本手寸//
」 当則以	(日: 123	致情灭感	新闻风观 大键问提取	山大坂住方竹 土地	型257/TT 101/年9月9月			し」道有文件
一对多共	现分析 多对多共现分析	词对分析 词定位						
	占击可杳看关	系数值						
词频表		ノーを見ていていた。	所选择的词语			×		
	词语	词频	词语 干部 A出现次数	: 1737				
5	工作	2062	词B	B出现的次数	AB同时出现的次数	AB同时出现所占比例		林島山 🗰
5	政治	1930	做到	339	313	92.33%		做到 🔳
	干部	1737	各级	506	460	90.91%		领导 🔳
	(新日)	4540	自觉	356	320	89.89%		各级 🔳
<u>-0</u>	<b>秋</b> 寺	1513	特别	344	308	89.53%		目元 ■
	发展	1498	纪律	521	459	88.10%		主要 🔳
-0	问题	1486	坚决	382	332	86.91%		● 5自
<u>-0</u>	坚持	1477	从严治党	548	476	86.86%	b定	作为
ā	建设	1394	全党	366	308	84.15%	坚持	- 初天 ■
5	我们	1363	管理	333	278	83.48%		党章 🗧
10	中国	1343	维护	349	291	83.38%		必须 🗖
	党内	1319						责任
		1010						安水 📕
-01		1121				不能自觉		管理 🛢
1 <u>a</u>	全面	1089						党员 📒
-0	社会主义	1083						就是
	力III吊	1008						至時 🧧

#### 关系图

关系图中,展示的是词与词之间的距离,即距离越近,词之间的关联性越强,该图默认显示与所选词距离最近的前20个词语,不同颜 色代表不同的词语,长短代表距离的长短,能直观展现出所选词与其他词之间的距离关系。

@四右引力 (重庄) 利技右限公司 婨

点击左边表格中不同的词语可切换中心词

关于距离的理论知识详见<进阶使用-词语分析-词关联性分析-词距离>



#### 距离条形图

距离条形图能直观显示出词之间的距离数值

点击左边表格中不同的词语可切换中心词

关于距离的理论知识详见<进阶使用-词语分析-词关联性分析-词距离>



#### ©码有引力 (重庆) 科技有限公司 淪ICP条19011279号

#### 共现概率图

共现概率图能直观显示出词之间的共现概率,注意:和词之间的距离不一样,共现概率越高,词之间的关联性越强,即代表两个词经 常出现在同一句话或同一篇文章里,至于是同一句话还是同一篇文章,根据词关联性分析初始化(选词)时勾选的共现标准决定。

点击左边表格中不同的词语可切换中心词

关于共现概率的理论知识详见<进阶使用-词语分析-词关联性分析-共现>

数据灵感之智分析		© _
☆ 当前项目: 123 数据灵	a 新词发现 关键词提取 词关联性分析 主题分析 词库编辑	查看文件
一对多共现分析 多对多共现分析 词对分析 词定位		
词频表	数据可视化	
词语 词频		



# 多对多分析

多对多分析用于整体上分析词与词之间的关联性

## 共现矩阵

进入多对多分析,点击<显示表格>按钮可查看共现矩阵,共现矩阵横纵第一项都为所选择的词语,中间数值为两两的共同出现的次数,数值越大,代表关联性越强。

关于共现的理论知识详见<进阶使用-词语分析-词关联性分析-共现>

灵感之智分析												0
介 当前项目: 123	数据灵感 新词发	现 关键词提取	词关联性分析	主题分析	词库编辑						<u>)</u> 查看3	之件
一对多共现分析 多对多共现分析	词对分析 词定位											
词频表	Q 搜索词语	矩阵分析	++\=\cept		按住shiff逐	动滚轮可左右和	多动奋君数据				Wester	-T40 //+
词语	词频	×	共间知阵	~							EALXS	PJ 49846
工作	2062	序号	工作	政治	干部	领导	发展	问题	坚持	建设	我们	中国
74724	1020	工作	2062	876	978	1027	738	982	991	895	714	64
成百	1930	政治	876	1930	876	912	585	921	1005	811	758	55
干部	1737	干部	978	876	1737	1120	407	942	800	710	688	39
领导	1513	领导	1027	912	1120	1513	540	913	944	811	672	51
发展	1498	发展	738	585	407	540	1498	586	824	776	728	89
问题	1486	回题	982	921	942	913	586	1486	857	780	831	48
1*3.8 <u>m</u>	1400	坚持	991	1005	800	944	824	857	1477	1002	808	80
坚持	1477	建设	895	811	710	811	776	780	1002	1394	725	73
建设	1394	3211	714	758	688	672	728	831	808	725	1363	69
我们	1363	中国	645	552	393	517	895	483	801	732	694	134
中国	1343	<u><u></u><u></u><u></u><u></u><u></u><u></u><u></u><u></u><u></u><u></u><u></u><u></u><u></u><u></u><u></u><u></u><u></u><u></u><u></u></u>	551	923	680	695	269	652	576	435	429	25
	1010	~雨	705	701	472	710	647	714	775	702	622	<u>50</u>
元内	1319	社会主义	605	553	397	530	818	/14	759	701	584	95
人民	1121	X ± × ± r	830	750	680	756	524	755	816	776	579	47
全面	1089	正にして	750	763	695	760	562	714	808	649	633	50
社会主义	1083	制度	666	673	584	659	507	648	727	685	543	48
103日	1008	监督	478	588	567	518	214	589	484	441	335	22
RECEIPT	•	<u></u> (日纪	755	567	600	707	364	501	577	468	328	33

#### @ 忍者引力 (重庆) 科技有限公司 淪ICP & 19011279 是

## 相似矩阵

进入多对多分析,点击<显示表格>按钮,再点击表格切换为相似矩阵,可查看相似矩阵,相似矩阵横纵第一项都为所选择的词语,中间数值为两两的相似度,相似度越大,代表关联性越强。

关于相似度的理论知识详见<进阶使用-词语分析-词关联性分析-相似度>

数据灵感	感之智分析													?
Û	当前项目: 123	数据灵感 新词发	远现	关键词提取	词关联性分析	主题分析	词库编辑						<u>)</u> 查看?	之件
	一对多共现分析 多对多共现分析	词对分析 词定位												
	词频表	○ 搜索司再		矩阵分析										
	1111年	词類	K.	表格	相似矩阵	$\sim$	按住shift溪	發动滾轮可左右種	多动查看数据				数据	可视化
	 工作	2062		序号	工作	政治	干部	领导	发展	问题	坚持	建设	我们	中国
		4000		工作	1.0000	0.2488	0.2274	0.3523	0.1844	0.2429	0.3195	0.3333	-0.0652	0.15
	政)音	1930		政治	0.2488	1.0000	0.2184	0.3741	0.0998	0.3289	0.4150	0.4255	0.3587	30.0
	干部	1737		干部	0.2274	0.2184	1.0000	0.6929	-0.1706	0.1524	0.3525	0.0106	0.0230	-0.0
	领导	1513		领导	0.3523	0.3741	0.6929	1.0000	-0.0635	0.0118	0.4974	0.1969	0.0664	-0.0*
	发展	1498		发展	0.1844	0.0998	-0.1706	-0.0635	1.0000	0.0656	0.3402	0.3774	0.4001	0.58
	(D)55	1496		问题	0.2429	0.3289	0.1524	0.0118	0.0656	1.0000	0.1133	0.1412	0.2356	-0.12
	1972	1460		坚持	0.3195	0.4150	0.3525	0.4974	0.3402	0.1133	1.0000	0.3858	0.3925	0.33
	坚持	1477		建设	0.3333	0.4255	0.0106	0.1969	0.3774	0.1412	0.3858	1.0000	0.2434	0.23
	建设	1394		我们	-0.0652	0.3587	0.0230	0.0664	0.4001	0.2356	0.3925	0.2434	1.0000	0.57
	我们	1363		中国	0.1551	0.0854	-0.0257	-0.0108	0.5807	-0.1239	0.3365	0.2317	0.5707	1.00
	山国	1242		党内	0.3316	0.7294	0.0852	0.2852	0.0212	0.3281	0.3503	0.3003	0.1528	-0.04
		1343		人氏	0.1703	0.0168	0.1026	0.1316	0.5361	-0.0543	0.5218	0.1061	0.4668	0.49
	党内	1319		全面	0.1822	0.3070	0.0311	0.1788	0.3506	0.0725	0.3380	0.5100	0.2162	0.22
	人民	1121		社会主义	0.1470	0.3194	0.0934	0.1897	0.4737	-0.0431	0.5322	0.4536	0.5146	0.82
	全面	1089		705頭	0.6495	0.4422	0.3188	0.4330	0.1481	0.1420	0.4865	0.6752	-0.0212	0.02
	社会主义	1093		必须	0.3578	0.5197	0.4256	0.5973	0.1212	0.2170	0.7488	0.3826	0.4887	0.18
	TTATX.	1003		制度	0.2626	0.2281	0.1875	0.2224	0.1925	-0.0263	0.5168	0.3908	-0.0873	0.20
	力以通	1008		血管	0.4522	0.3555	0.3488	0.4559	-0.0544	0.1749	0.4189	0.2986	0.0277	-0.0

◎码有引力 (重庆) 科技有限公司 淪ICP条19011279号

#### 共现多维度图

共现多维度图是由共现矩阵通过MDS降维,并通过K-means聚类得到,图中还加入了词频信息。

即其表示意义为:圆越大的词,词频越高,图上距离越近表示两个词关系越近,不同颜色表示不同类别。

关于MDS的理论知识详见<进阶使用-词语分析-词关联性分析-MDS降维>

关于K-means聚类的理论知识详见<进阶使用-词语分析-词关联性分析K-means聚类>



## 相似多维度图

和共现多维度图类似,只不过相似多维度图是用相似矩阵通过MDS降维,并通过K-means聚类得到,图中还加入了词频信息。

即其表示意义为:圆越大的词,词频越高,图上距离越近表示两个词关系越近,不同颜色表示不同类别。

关于MDS的理论知识详见<进阶使用-词语分析-词关联性分析-MDS降维>

关于K-means聚类的理论知识详见<进阶使用-词语分析-词关联性分析K-means聚类>



# 词对分析

词又	讨分析用于分析一个词双	付 (即两个词)	与其他	加之间	的关系,	可进一步说明两个词之间的关联性。		
	1. 该功能需要用户说	选择两个词语,	即词a和	和词b				
数据	灵感之智分析							®
	① 当前项目: 123	数据灵感	新词发现	关键词提取	词关联性分析	主题分析 词库编辑	查看文件	
	一对多共现分析 多对多共现分析	词对分析 词定位						
点击试		Q 搜索问语 词频 2062	<b>%</b>	<b>词语信息</b> <sub>词a</sub> 数据可视化	词b	重选	分析词对	

	2002		
③ ⑤ 政治	1930		
<ol> <li>(a) (b) 干部</li> </ol>	1737		
③ ⑤ 领导	1513		
③ ⑤ 发展	1498		
(3) (5) 问题	1486		
③ ⑤ 坚持	1477		
③ ⑤ 建设	1394	$\checkmark$	
(3) (b) 我们	1363		
(a) (b) 中国	1343		
③ ⑤ 党内	1319		
(3) (b) 人民	1121		
(a) (b) 全面	1089		
③ ⑤ 社会主义	1083		
─────────────────────────────────────	拿刃词 D		

2. 下图分别显示出"工作"单独出现的次数、"加强"单独出现的次数以及两个词共同出现的次数。然后点击分析词对



©码有引力 (重庆) 科技有限公司 渝ICP备19011279号

3. 下图展示了该词对与其他词之间的共现关系,即下图中

红色柱子代表左边的词与"工作"单独出现的次数 深绿色代表左边的词与"加强"单独出现的次数 浅绿色代表三个词共同出现的次数



4. 点击表格分析可显示所选词对和其他词之间的共现数值



5. 点击<词A与词B的共现数值>可查看该词对的共现数值



# 词定位

词定位用于查看词语在原文中的具体位置与上下文关系

#### ? 数据灵感之智分析 查看文件 数据灵感 新词发现 关键词提取 词关联性分析 主题分析 词库编辑 介 当前项目: 123 一对多共现分析 多对多共现分析 词对分析 词定位 定位结果 词频表 Q 搜索词语 词语 词频 \* 上文 下文 ID 关键词 工作 2062 为我们这一届中央领导集体的 工作 指明了方向。中央已经发出关于认真学习宣传 中也要牢记初级阶段。党在社会主义初级阶段。 政治 1930 不仅在经济建设中要始终立足初级阶段,而且. 工作 2 工作 的谋划和部署都是遵循和体现这些基本要求的。 党的十八大提出的基本要求,是对当前我国经. 干部 1737 密切党群、干群关系,保持同人民群众的血肉... 工作 新特点新要求, 领导 1513 深入做好组织群众、宣传群众、教育群众、服.. 工作 , 虚心向群众学习, 诚心接受群众监督, 始终. 5 发展 1498 大量事实告诉我们,腐败问题越演越烈,最终... 人员的教育和约束,决不允许以权谋私,决不。 工作 问题 1486 ,全面提高党的建设科学化水平。 要深刻理解把生态文明建设纳入中国特色社会... 工作 7 坚持 1477 工作 也是全党同志的应尽义务和庄严责任, 对强... 8 是加强党的建设的一项基础性经常性 来抓,通过日常学习、专题培训等形式,组织...点击切换位置 建设 1394 9 党章对党的性质、宗旨、指导思想、奋斗纲领... 工作 我们 点击切换词语 1363 10 要把检查学习和遵守党章情况作为组织生活会... 工作 、党内活动、党的建设的根本依据,把党章各... 中国 1343 首页 上一页 下一页 尾页 第 1 /共237页 别转 选段来自 文章ID: 100 党内 1319 党的十八大提出的基本要求,是对当前我国经济社会发展中存在的突出问题。改革攻坚和加快转变经济发展方式面临的進点问题,干部群众普遍关注 的热点问题的积极回应,是对我国进入全面建成小康社会决定性阶段改革发展稳定,内政外交国防、治党治国治军的正确指引。这些基本要求,既涉及 生产力和生产关系、又涉及经济基础和上层建筑,既涉及中国特色社会主义伟大事业、又涉及党的建设新的伟大工程,同时还涉及统筹国内国际两个大 人民 1121 全面 1089 局。党的十八大对各项 <u>工作</u> 的谋切和部署都是遺藥和体现这些基本要求的,抓住了这些基本要求,就就要好凝聚力量、灾坚死难,继续推动科学发展、 促进社会和谐,继续改善人民生活、增进人民福祉,完成时代赋予的光荣而艰巨的任务。第五,深刻领会确保党始终成为中国特色社会主义事业的坚强 社会主义 1083 领导核心。 加强 1008 ©码有引力 (重庆) 科技有限公司 渝ICP备19011279号

# 主题分析

用于分析数据中的主题

# 主题聚类

采用K-means+TFIDF进行主题聚类,用主题词的形式来表现主题,并结合时间进行主题趋势分析

## 设置主题数

由于图形限制, 主题数可支持2-10个主题

当数据中含有时间项时,可编辑时间间隔

主题设置 主题词表 主题可视化

数据灵感之智分析

△ 当前项目: 123 数据灵感 新词发现 关键词提取 词关联性分析 主题分析 词库编辑

?

👌 查看文件



## 编辑主题词

主题词默认选择每个主题里TF-IDF排名前10的词语,当其中有你不想要的词语时,可点击"X"按钮删除,删除后,右键选中左边词表中任一词语,可替换成您选择的词语。

♪ 当前项目: 123	数据灵感 新词发现	观 关键词提取 词关联性分析 主题分析 词库编辑		う 查看文件
主题设置  主题词表  主题可视化 			点击切换主题	
当前主题: 主题1		<sup>主题调库</sup> 点击可删除		24 IV-
词语	TFIDF			□ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○
发展党员	0.877	主题1 主题2	主题3 主题4	主題5
◎ 田丁 //=	0.607	1 发展党员 1 会议	1 巡视 1 警告	1 党内
自·理工1F	0.007	2 管理工作 2 习近平	2 论述 X 2 设立	× 2 讲话 ×
教育	0.560	3 教育 3 宣传	3 摘編 3 留党察者	3 年月日
<b>告</b> 李	0.540		4 元方     4 円页       5 活动     5 信节严重	4 生活
元早	0.519		6 我们 X 6 处分	X 6 全会 X
发展	0.482	7 党委 7 党章	7 反腐败 7 撤销	7 生态
		8 廖俊波 × 8 人民 ×	8 监督 X 8 人选	X 8 换届 X
7015年	0.471	9 应当 9 法治	9 年月日 9 责任者	9 严肃
党委	0.466	10 党员 (10 时代) (10 日代) (10 H) (10	10 纪委 X 10 开除党籍	10 集中统一 · · · · · · · · · · · · · · · · · · ·
ppp///th	0.424			重置  下一步
18-15-10	0.424			
应当	0.419			
党员	0.414			
入党	0.414			
两学	0.400			
预备党员	0.394	右键点击添加到主题词		
同志	0.374 💌			

# 主题归纳

编辑完成后点击下一步,显示您所选择的主题词,用户可根据主题词归纳,编辑主题名,然后再点击可视化可进行图形绘制

		数据灵感	新词发现	关键词提取 词关联性分	析主题分析	词库编辑				查看文件	
主题设置主题词表	ē 主题可视化 -										
					主题树						
主题	主題词1	主题词3	主题词5	主题词7	主题词9	主題词11	主题词13	主题词15	主题词17	主题词19	
党员教育	发展党员	管理工作	教育	党章	发展	力口引虽	党委	廖俊波	应当	党员	
主题2	会议	习近平	宣传	县委书记	社会主义	中国	党章	人民	法治	时代	
主题3	巡视	论述	摘编	党务	活动	我们	反腐败	监督	年月日	纪委	
主题4	警告	设立	留党察看	问责	情节严重	处分	撤销	人选	责任者	开除党籍	
				11. 1007	ANY 10	~~	# *	摘屋	222 (89)	售由统	
主题5	党内	讲话	年月日	生活	第只	王云	TTT	1940/881	/ <i>P</i> N	340 T 970	
<sup>主願5</sup> 人	<sup>党内</sup> 古可进行编	<sup>жа</sup>	年月日	生活	朝贝	王云	±73	190(88	, 17 L	ACT 700-	
<sup>主题5</sup> 点	<sup>兆内</sup>	<sub>на</sub>	年月日	生活	可视化	王帝	<u> </u>	150/00	1 m	944 T-90	
<sup>3酸±</sup>	地	<sub>讲活</sub>	年月日	生活	可视化	王帝	<u> </u>	1550183	1 m	944 T-96	

# 主题树图



©码有引力 (重庆) 科技有限公司 渝ICP条19011279号

## 主题河流图

不同颜色代表不同主题,结合时间分析主题,看各时间段的主题分布、占比与发展趋势



### 主题趋势图

该图中不同颜色代表不同主题;一个小方块代表一个主题词,方块的高度代表该词的TF-IDF数值,方块越高TF-IDF越大;图中同种色块前后时间段有曲线连接(贝塞尔曲线),代表在两个时间段都出现了该主题,即说明该主题有发展趋势,色块越高,主题占比越大。



# 主题模型

即将推出, 敬请期待

# 社区划分

即将推出, 敬请期待

# 词库编辑

词库编辑页面可查看词库中的词语以及词库编辑日志,同时可以编辑词库中的词语,分析词库和停用词库支持导入功能。

# 分析词库

分析词库中的词语是智分析自动识别不出而又要用到的词语或短语,添加到词库的词语,点击查看文件后再点击重新计算(重新分

词) 生效。

分析词库支持导入功能,暂只支持TXT文件,且文件内不同词语需用**回车**隔开。

当前项目:计数项测试 数据灵感 新词发现 关键词提取 词关联性分析 雪分析词库 停用词库 同义词库 日志 输入要添加到	<sup>上輕分析</sup> 词 <sup>互编辑</sup> 月分析词库中的词语,	再点击回车,	即可添加	) 查看文件
Q 搜索词语 + 输入添加词语	编辑记录	点击可导入	、TXT文件 ———	词库导入
什么时候开学 啥时候开学 什么时候才能开学 分批开学 不想开学不想开学	时间	操作	记录	执行人
1171%和血口 初加加加及活者 麻辣香口鸡 月月四同节 页加百同二 每日一问 变异病毒 高三年级 微博视频 讨论阅读 明确开学时间	2020-12-28 14:47:54	添加	什么时候开学	alextan
应急响应调整 多地明确 本地热门资讯 长江上游地区 山城雲都	2020-12-28 14:47:50	添加	啥时候开学	alextan
新型冠状病毒 巡查执法 热门资讯 本地热门 确诊病例 网页链接	2020-12-28 14:47:47	添加	什么时候才能开学	alextan
下载使用 提出问题 网警巡查 啊啊啊啊啊啊啊 电子凭证 央视新闻	2020-12-28 14:46:57	添加	分批开学	alextan
肺炎疫情 高校应届 初三年级 国家中心 经济中心 应届毕业生	2020-12-28 14:45:59	添加	不想开学不想开学	alextan
重庆市教委 疫情防控 新冠肺炎 应急响应 开学时间 不想开学	2020-12-28 14:45:50	添加	红外线体温计	alextan
不想上网	2020-12-28 14:45:48	添加	新冠肺炎患者	alextan
	2020-12-28 14:45:34	添加	麻辣香锅	alextan
	2020-12-28 14:45:31	添加	青海高中	alextan
	2020-12-28 14:45:29	添加	贵州省高三	alextan
	2020-12-28 14:45:27	添加	每日一问	alextan
	2020-12-28 14:45:24	添加	变异病毒	alextan
	2020-12-28 14·45·18 首	· · · · · · · · · · · · · · · · · · ·	高三 <u>年级</u> 瓦 第 <mark>1</mark> /共1页 <mark>跳转</mark>	alextan

# 示例TXT文件:

🥘 word.txt - 记事本				- 0	$\times$
文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H)					
政党协商					^
政治制度					
新时代					
作风建设					
作风不纯					
左翼文化					
道会党纪					
4/4/8/注 ////////////////////////////////					
利和時後					
ロンルに同					
日75年初 白75年初					
日次中印					
中国特色社会工業取得の時代に					
中国特色社会主义制度的最大优势					
부명하면 전 국 도 것이 만경 전 문 문 문 문 문 문 문 문 문 문 문 문 문 문 문 문 문 문					
中国头广况最佳 					$\sim$
<					>
	第1行,第1列	100% W	indows (CRLF)	UTF-8	

# 停用词库

停用词库中的词语是需要移除的词语,以免这些词语影响分析结果。停用词库也和分词词库一样支持导入功能,且暂也只支持TXT文件。

》 当前项目: 计数项测试 数据灵感 新词发现 关键词提取 词关联性分析 主题分析	词库编辑		5 查查	較件
分析词库 停用词库 同义词库 日志		Ŗ	点击可进行TXT文	件导入
Q 搜索词语 + 给入添加词语	编辑记录			词库导入
通知 还有 一个 可以 没有 我们 简称 真的 什么	时间	操作	记录	执行人
	2020-12-30 21:25:23	删除	2	alextan
	2020-12-28 14:41:57	添加	通知	alextan
输入词语,回车添加	2020-12-28 14:41:55	添加	还有	alextan
	2020-12-28 14:41:50	添加	<b>一</b> 个	alextan
	2020-12-28 14:41:46	添加	可以	alextan
	2020-12-28 14:41:45	添加	没有	alextan
	2020-12-28 14:41:43	添加	我们	alextan
	2020-12-28 14:41:38	添加	简称	alextan
	2020-12-28 14:41:31	添加	真的	alextan
	2020-12-28 14:41:28	添加	什么	alextan
	2020-11-05 08:14:37	添加	2	alextan

@ 码 右 引 力 ( 重 庆 ) 科 技 右 調 公 司 淪 ICP 各 19011279 문

# 同义词库

同义词库中的词语是需要进行合并分析的词语,分为中心词和同义词,在智分析自动分析过程中,会把一个中心词下的所有同义词都 替换成中心词,以此完成合并分析。其操作过程如下图:

数据灵感之智分析					0
○ 当前项目: 计数项测试	数据灵感 新词发现 关键词提取 词关联性分析	主题分析 词库编辑		③查	<b>活</b> 文件
分析词库 停用词库 同义词库 日志		1. 输入添加中心词,	并按回车健		
· · · · · · · · · · · · · · · · · · ·		时间	操作	记录	执行人
		2020-12-30 21:28:10	添加	重庆:雾都	alextan
		2020-12-30 21:27:48	添加	重庆:山城	alextan
2. 点击中心词		2020-12-30 21:27:46	添加	重庆:重庆市	alextan
同义词 雾都 山城 重庆市	—— 输入添加问证	3. 输入同义词,并按	<b>回车健</b> □ 上─页 下─页 尾页 策[	<b>1</b> /共1页 <mark>姚转</mark>	

◎福有引力 (重庆) 科技有限公司 渝|

如果需要查看某个中心词的同义词,点击该中心词即可。

# 查看文件, 重新分析

在使用任何分析功能时,都可以点击右上角的<查看文件>按钮查看数据。



@码有引力 (重庆) 科技有限公司 淪ICP条19011279号

在查看文件页面,可点击<重新计算>按钮进行重新分词计算,使词库中的词语生效。

#### 数据灵感之智分析

序号	id	时间	作者	正文	网站名称	pandaweburl	态度
1	1	2020-02-19T19:45:00.00000000	我听fm	疫情今日辟谣今天是2月19日,	新浪微博	https://m.weibo.cn/status/luX1N	中性
2	2	2020-02-19T21:22:00.000000000	华龙网	重庆辟谣【2020.2.19 谣言,别	新浪微博	https://m.weibo.cn/status/luXF0	中性
3	3	2020-02-19T21:36:00.000000000	重庆人社	重庆辟谣【2020.2.19】谣言,	新浪微博	https://m.weibo.cn/status/luXK	中性
4	4	2020-02-19T21:40:00.000000000	沙坪坝微政务	微博辟谣重庆辟谣【谣言,别信	新浪微博	https://m.weibo.cn/status/luXM	中性
5	5	2020-02-19T21:41:00.000000000	cqdk全媒体	重庆辟谣【2020.2.19   谣言,	新浪微博	https://m.weibo.cn/status/luXM	中性
6	6	2020-02-19T21:43:00.000000000	重庆日报	重庆辟谣【2020.2.19   谣言,	新浪微博	https://m.weibo.cn/status/luXN	中性
7	7	2020-02-19T21:49:00.000000000	重庆市江北区委外宣办	【2020.2.19】谣言,别信!—…	新浪微博	https://m.weibo.cn/status/luXPS	中性
8	8	2020-02-19T21:52:00.000000000	微播南川	重庆辟谣【2020.2.19    谣言,	新浪微博	https://m.weibo.cn/status/luXR5	中性
9	9	2020-02-19T22:05:00.000000000	重庆共青团	疫情辟谣【这些都是谣言,别信	新浪微博	https://m.weibo.cn/status/luXW	中性
10	10	2020-02-20T08:51:00.000000000	巴南政务	重庆辟谣【谣言,别信! ——…	新浪微博	https://m.weibo.cn/status/lv2aB	中性
11	11	2020-02-20T08:54:00.000000000	巴南广播电视台	早安巴南[超话]重庆辟谣【谣言	新浪微博	https://m.weibo.cn/status/lv2bU	中性
12	12	2020-02-20T09:48:00.000000000	两江新区发布	重庆辟谣【2020.2.19   谣言,	新浪微博	https://m.weibo.cn/status/lv2xQ	中性
13	13	2020-02-20T11:30:00.000000000	重庆检察	重庆身边事【谣言,别信! ——…	新浪微博	https://m.weibo.cn/status/lv3drx	中性
14	14	2020-02-20T14:34:00.000000000	重庆大渡口检察院	重庆身边事【谣言,别信! ——…	新浪微博	https://m.weibo.cn/status/lv4pQ	中性
15	15	2020-02-20T16:11:00.000000000	微博辟谣	微博辟谣抗击新型肺炎第一线【	新浪微博	https://m.weibo.cn/status/lv53m	中性
			首页 上一	页 下一页 尾页 第 1     /供126页       项目     重新计算	跳转	— 点击可进行重新分	词计算

# 进阶使用

本章节内容建议用户根据自己的数学基础酌情阅读

# 基础统计

基础统计指不包括文本分析的统计,只是对数据进行一些基础分析,其中主要包括利用时间项、计数项等结合分析。

# 计数项

计数项是指可以计数的一项,其命名来源于excel,换句话说,是可以标注该条(行)数据的类别的一项,比如说:作者、来源、期刊 等,即计数项为作者时,可以把数据按作者归类进行统计,例如统计《红楼梦》前120回的作者数据:



如图所示,《红楼梦》前120回中,其中有40回的作者是高鹗,80回的作者是曹雪芹。

智分析能自动判断出数据中是否包含计数项,如果有,则会自动出现计数项相关功能;如果数据中含有多个计数项,智分析会默认使 用第一个计数项,用户可手动选择计数项,选择不同的计数项,可以实现对数据按不同的标准进行归类统计。

计数项最多支持30个类别。

# 时间项

时间项指数据中全为时间的一列,智分析能通过时间,对数据进行统计分析,如图所示:



智分析按年统计了数据里的文章数量,智分析能自动判断出数据中是否包含时间项,如果包含,会自动选择一个合适的时间间隔进行 数据统计,同时也支持用户手动选择时间间隔。

注意:时间项中必须保证每个

# 多维分析

多维分析是指多个项结合分析,例如计数项和时间项可结合得到下图:



该图设置计数项为态度,时间间隔为1日,分别统计了态度中积极、消极、中性在每日中各有多少数量,并以折线图的形式展示,展现 出了不同态度每日的趋势情况。

当然, 计数项、时间项也可以结合文本进行分析, 比如, 可结合计数项和关键词统计得到如下词云图:



图中展示了消极态度的词云,用户也可手动切换为其他计数类。

# 分词与词库

分词与词库是智分析中最核心的一个功能,分词是文本分析的基础,进行了分词后,才能进行一系列的数学处理;而词库是"人工+智能 =越用越智能"的基础,即词库是人工的结晶,有了专业的词库,分出来的词语才能更加地准确,而后的数学分析才能分析出更精确的结果。

**设计来源:**由于汉语文化博大精深,目前中文分词技术在全世界范围内也没有得到良好的解决,属于中文自然语言处理技术的一大难题,也是最重要的难题。中文不像英文,英文每个单词是用空格隔开的,一个单词就能有具体的意思,即便是单词与单词组合形成的短语,大部分也与原单词含义关联性较强,甚至是含义相近。而中文不同,中文是字和字形成词,不同字组成的词含义可能天差地别,因此,中文自然语言处理技术一般以词为单位进行分析;而在中文文本中,词之间并没有任何符号隔开,因此就需要分词。

而在中文词语的发展中,随着时间的推移,随时可能出现新的词语,例如:一带一路、耗子尾汁这类词语,且不同专业有着不同专业 的词语,随着时间也会出现大量新词,像这类词,算法是很难自动把这些词分出来的。因此就需要借助(人工)词库,来帮助计算机 进行自动分词,让计算机也能理解博大精深的汉语。

目前在自然语言处理领域,常用的分词方法主要有:基于词典分词、词典和统计结合的分词和基于统计模型的分词。基于词典的分词 虽然过度依赖词典和规则库,对歧义词和未登录词的识别能力低,但由于其速度快、效率高,仍在一些特定语境环境仍被"智分析"采 用。在统计模型中,CRF(条件随机场模型),HMM(隐马尔可夫模型),MEMM(最大熵隐马尔可夫模型)都常用来做序列标注的建模, 像分词、词性标注,以及命名实体标注,这三种模型应用在中文分词中都已经取得了不错的效果。由于HMM模型的输出独立性假设, 导致该模型不能考虑上下文的特征,限制了特征的选择;MEMM模型则解决了HMM模型的问题,可以任意选择特征,但由于其在每一 节点都要进行归一化,所以只能找到局部的最优解,同时也带来了标记偏见问题,即凡是训练语料中未出现的情况全都忽略掉;CRF模 型则很好的解决了局部最优解的问题。CRF模型并不在每一个节点进行归一化,而是对所有特征进行全局归一化,因此可以求得全局的 最优解。

"智分析"在运用CRF分词时,结合了人文社会科学领域的多位专家经验,通过人工配置的特征函数模板,有针对性的对人文社科全领域 大量语料库,提取局部上下文特征,从而学习到特征权重。"智分析"在"特征工程"构建上,有大量人文社科领域专家参与,使得在分词 (特别是规范性文本)的准确性上,表现出优秀的能力。对于互联网短文本等非规范性文本,"智分析"结合了神经网络最新成果,在 "特征学习"上运用了RNN、LSTM、BiLSTM等多种神经网络模型。但由于RNN类模型的效果,一是依赖于语料库的规模,二是算法本身 的效率问题,"智分析"在处理跨领域全文本的分词问题时,以平台化的方式提供该功能。

分词是文本分析的起始点,即几乎所有文本分析都离不开分词,它是最基础也是最重要的功能。智分析中,对分词算法进行了大量优化,不管从速度上还是从精确度上,都在行业处于领先水平。

智分析分词是基于概率算法及词库的,此部分内容读者只需重点关注词库即可。在词库内容为空时,智分析会根据分词算法自动分出 词语,此时基本上能包含汉语中经常出现的词语,当分词结果不准确时,可通过添加词语到词库,再重新计算(分词),来调试结 果。

# 分析词库

分析词库中的词语是分词时要切分出的词语(分析时要用到的词语或短语),分析词库里一般添加的是通过智分析分词无法分出来的 词语,例如像:一带一路、耗子为汁这类词语。

# 停用词库

一般在中文文本分析中,出现次数很高的词语一般是分析里不需要的词,而这类词会大大影响分析结果,比如"我们"、"这个"、"怎 样"、"怎么"等,一般是一些代词、介词等。而想去掉这些词语怎么办?可以把这些词语添加到停用词库中,然后再进行重新分词计 算,即可去掉停用词库里的词语。

### 同义词库

在中文文本分析中,有的词含义相同,即同义词。在智分析中,可以把同义词添加到同义词库中进行合并分析,在智分析的自动分析 过程中会把所有同义词合并成同一个词。

#### 新词发现/短语抽取

"智分析"综合了互信息和左右熵,来对中文文本进行新词发现。互信息(Mutual Information)是信息论里一种有用的信息度量,它可以 看成是一个随机变量中包含的关于另一个随机变量的信息量,或者说是一个随机变量由于已知另一个随机变量而减少的不肯定性。对 于词而言,互信息越大,越能说明两个或多个词经常一起出现,这意味着两个或多个词的凝固程度越大,这两个或多个词组成一个新 词的可能性也就越大。

"智分析"用信息熵来衡量一个文本片段的左邻字集合和右邻字集合有多随机。计算一对词之间的左熵和右熵,熵越大,越说明是一个 新词。因为熵表示不确定性,所以熵越大,不确定越大,也就是这对词左右搭配越丰富,越多选择。

我们已经对大量人文社会科学语料库进行了分类的互信息搭配和信息熵统计。"智分析"包含了大量领域词库。通过"智分析"建议的最低 互信息(PMI\_LIMIT),互信息权值、左右熵权值,能够得到待研究文本精确的短信抽取(新词发现)结果。通过使用此方法,可对专 业领域文章进行分析,自动发现、提取专业领域新词。

互信息

摘自百度百科,互信息(Mutual Information)是<u>信息论</u>里一种有用的信息度量,它可以看成是一个<u>随机变量</u>中包含的关于另一个 随机变量的信息量,或者说是一个随机变量由于已知另一个随机变量而减少的不肯定性。其计算公式为:

$$I\left(X;Y
ight) = \sum_{x \in X} \sum_{y \in Y} p\left(x,y
ight) log rac{p\left(x,y
ight)}{p\left(x
ight) p\left(y
ight)}$$

在文本分析中,互信息能说明词语的凝固程度。例如:在公式中,P(x)可看作词X出现的概率,P(y)可看作词Y出现的概率,P(x,y)可理解为词X和词Y一起出现的实际概率,而P(x)P(y)则能表示词X和词Y一起出现的理论概率,因此上述公式可简单理解为两个词一起出现的实际概率/理论概率。

左右熵

熵是一种表示信息量的指标,熵越高就意味着信息含量越大,不确定性越高,越难以预测。通常对于一个随机变量 X,它的熵可以 被表示成:

$$H(X) = -\sum_{x \in X} p(x) \log_2 p(x)$$

p(x) 表示的是事件 x 出现的概率,在文本分析中就是一个词出现的概率。我们用信息熵来衡量一个文本片段的左邻字集合和右邻 字集合有多随机,一个文本片段的左右邻接字越随机,即这个文本片段越可能是一个新词。

• 关于左右熵和互信息的更多知识请参阅<u>互联网时代的社会语言学:基于SNS的文本数据挖掘</u>

# 词语分析

在汉语中,大多数词语就能表达一个独立的含义。因此,一般在中文文本分析中,是以词语为基础进行分析的,而智分析也是如此。 智分析里的词语分析目前主要分为关键词分析和词关联性分析。

# 关键词分析

关键词能一定程度上表示文本里的重点,智分析支持按词频和TF-IDF两种指标来统计关键词。

#### 词频统计

词频统计是统计一个词语在文本中所出现的次数。注意:对于某个四字词语包含了某个两次词语或三字词语时,智分析是分别进行词 频统计的。例如:"领导小组"和"领导"这两个词语,智分析会分别统计词频,即"领导"所出现的次数里不包括"领导小组"中"领导"所出 现的次数。

#### TF-IDF

TF-IDF算法已经被证明是一个高效的关键词提取算法。TF-IDF是建立在这样一个假设之上的:对区别文档最有意义的词语应该是那些在 文档中出现频率高,而在该文档所属领域语料库文档中出现频率低的词语。另外,考虑到单词区分不同类别的能力,TF-IDF算法认为一 个单词出现的文本频数越小,它区别不同类别文本的能力就越大。TF-IDF实际上是:TF\*IDF,TF词频(Term Frequency),IDF逆向文 件频率(Inverse Document Frequency)。实际情况是如果一个词在一个类别的文档中频繁出现,说明该词能够比较好代表该类别的文 本特征,这样的词应该赋予较高的权重,并用来作为该类别文本的特征词。但按照IDF的定义,如果在语料库中出现的频次高,则反应 不出来该词的重要性,这就是IDF的不足之处。另外,对于IDF来说,它本身是一种抑制噪声的权重,倾向于文本中频率小的词,使得 TF-IDF算法的精度存在问题。另外,像词的位置信息如标题、首尾等含有较重要的信息,也未在算法中得到体现。

'实事上,"智分析"在分词阶段所构建的"特征工程",在很大程度上被用于关键词提取阶段。"智分析"结合了TF-IDF效率高的优点,同时 又结合了浅层语义分析。在TF-IDF候选关键词语义不明确时,又结合运用了主题模型的方法,在"特征工程"基础上,取得了不错的关键 词提取效果。

#### 词关联性分析

'在"智分析"中,词关联分析是领域专家甚至用户自己构建"特征工程"非常重要的功能。词关联分析是在分词、关键词提取和短语抽取的 基础上,由用户根据单个词频、共现词频、TFIDF权值等进行综合分析,得出关键词或者短语之间的逻辑关系,从而构建起具有待研究 领域的"特征工程"。在此基础上,进一步增加后续文本语义分析的精确性。

由于词是支撑主题的最小语义元素,因此挖掘出词与词之间的关系,是进行语义挖掘的基本手段之一。"智分析"中,我们用词的关系, 粗略代表了主题与主题之间的关系。其中,我们对词与词的关系的界定,包含了(不限于)对如下关系的分析:

- 主从关系
- 并列关系
- 层次关系
- 递进关系
- 先后关系
- 因果关系
- 对比关系

#### 共现

,共现指两个或多个词语共同出现,其界定单位可以是一句话或一篇文章。共现次数能一定程度表现词语之间的关联程度,即如果两个 ·词经常一起出现,则表示这两个词语的关联程度较强,至于其是否有着特殊含义,则需要通过人工解读。

#### 词向量

词向量是词的一种数学表示,词向量分很多种,智分析中的词向量是通过word2vec算法计算得到。得到词向量后可通过数学运算计算 词与词之间的相似度,同时也能计算词与词之间的距离,关于word2vec的理解大家可以参考<u>秒懂词向量Word2vec的本质</u>。

#### 词距离/相似度

词距离/相似度是表示词与词之间的关联程度的另一种形式。相似度越大,词距离越小,即代表词语之间关联程度越大。

在智分析中,是通过余弦相似度来计算词与词之间的相似度的,其公式如下所示:

$$ext{similarity} = \cos( heta) = rac{A \cdot B}{\|A\| \|B\|} = rac{\sum\limits_{i=1}^n A_i imes B_i}{\sqrt{\sum\limits_{i=1}^n (A_i)^2} imes \sqrt{\sum\limits_{i=1}^n (B_i)^2}},$$

向量间的距离公式有很多,比如说欧式距离、余弦距离、曼哈顿距离、切比雪夫距离等等,而在智分析中,使用的距离公式是余弦距 离,即余弦距离=1-余弦相似度。

#### 矩阵分析

矩阵(Matrix)是一个按照长方阵列排列的<u>复数</u>或<u>实数</u>集合,是高等代数中的常见工具。而在智分析中,主要用到的是共现矩阵和相 似矩阵,其中横纵里的第一项都为用户所选择的词,内容为词与词的共现次数或者相似度,可以整体上分析词与词之间的关联性。

#### MDS降维

MDS(多维尺度变换)算法解决的问题是:当n个对象之间的相似性给定,确定这些对象在低维空间中的表示,并使其尽可能与原先的相似 性大致匹配。高维空间中每一个点代表一个对象,因此点与点之间的距离和对象之间的相似度高度相关。可以这么理解,两个相似的 对象在高维空间中由两个距离相近的点所表示,两个不相似的对象在高维空间中由两个距离比较远的点表示。经典MDS算法流程如 ጉ፡

(1)计算原始空间中数据点的距离矩阵。

(2)计算内积矩阵 **B**。

(3)对矩阵B进行特征值分解,获得特征值矩阵 👖 和特征向量矩阵 🗸 。

(4)取特征值矩阵最大的前Z项及其对应的特征向量 $Z = V_z A_z^{1/2}$ 。

因此, MDS能较好地对共现矩阵和相似矩阵进行降维处理, 使词语分布在同一平面中, 距离越近的词语则关联性越强。

#### K-means聚类

k均值聚类算法(k-means)是一种迭代求解的聚类分析算法,其步骤是随机选取K个对象作为初始的聚类中心,然后计算每个对象与 各个种子聚类中心之间的距离,把每个对象分配给距离它最近的聚类中心。聚类中心以及分配给它们的对象就代表一个类别。每分配 一个样本,聚类的中心会根据聚类中现有的对象被重新计算。这个过程将不断重复直到满足某个终止条件。终止条件可以是没有(或 最小数目)对象被重新分配给不同的类别,没有(或最小数目)聚类中心再发生变化,误差平方和局部最小即达到算法收敛。

# 主题挖掘

主题挖掘功能用于挖掘文本中的主题,对文本内容进行一个主题性的概况,目前智分析支持无监督的主题聚类算法,半监督和全监督 的主题算法正在研发中,敬请期待。

## 主题聚类

针对时间序列文本,"智分析"通过Kmeans等算法实现主题聚类,其具体流程如下:

1.首先,把所有文本通过TF-IDF算法进行文本特征向量化,得到文本向量;

2.其次,运用Kmeans等算法对各个向量进行聚类,得到各个主题类别关键词;

3.再次,对各个主题类别关键词,分别进行TF-IDF计算提炼出各个类别的主题关键词;

4.最后,再对原文本按照不同时间段进行划分,划分后再次进行1、2和3步骤,可得到各个时间段内每个主题的代表词。

其中, k均值聚类算法 (k-means) 是一种迭代求解的聚类分析算法,其步骤是随机选取K个对象作为初始的聚类中心,然后计算每个 对象与各个种子聚类中心之间的距离,把每个对象分配给距离它最近的聚类中心。聚类中心以及分配给它们的对象就代表一个类别。 每分配一个样本,聚类的中心会根据聚类中现有的对象被重新计算。这个过程将不断重复直到满足某个终止条件。终止条件可以是没 有(或最小数目)对象被重新分配给不同的类别,没有(或最小数目)聚类中心再发生变化,误差平方和局部最小即达到算法收敛。

作为"智分析"文本主题挖掘的第一步,在TF-IDF基础上使用聚类算法对文本进行分类,能够快速有效确定待分析文本的类别数量及主题 特征。在文档类别数量不是特别大、且单主题(无多重语义的文档)文档集合,能达到比较好的文本主题聚类效果。而对于多主题文 档构成的文档集合,则适合用LDA主题模型或者带监督学习的LDA模型进行主题挖掘。

# 主题模型

即将推出, 敬请期待

# 社区划分

即将推出, 敬请期待

# 常见问题

### 1. 我该如何准备要分析的数据?

- 对于文章数据而言,如果想结合计数项和时间项进行分类和趋势分析,可把时间、内容、计数项字段复制到excel表格中 (表格中要有时间、内容等标题字段),然后直接导入智分析进行分析
- 针对线上数据而言,如微博、豆瓣、知乎、今日头条等,可通过爬虫等技术手段采集数据,如果不会技术,可联系我们提供 技术服务
- 针对线下数据,如书、报纸等数据,可通过OCR等电子化手段变成电子数据,整理成excel后便可导入智分析进行分析,同时也可以联系我们提供电子化服务

# 2. 智分析支持什么数据? 支持什么文件格式?

智分析支持任何中文文本数据,不管是知网论文,还是微博头条,都能导入智分析开启文本分析世界的大门

智分析目前支持Excel、TXT文件格式的数据,注意:TXT文件只支持文本分析,不支持计数项和趋势分析

### 3. 为什么我导入数据后没找到计数项功能?

这是因为系统未识别出计数项,智分析会自动识别数据中的计数项,计数项里的类别必须大于等于2歌,且不能超过30个,否则 无法识别出计数项。

### 4. 为什么我导入数据后没有趋势分析功能?

这是因为系统未识别出时间项,时间项最好为:年/月/日的格式,且要求每条数据里都得有时间,即不能某一条数据里的时间为 空或者其他内容(可以使用excel的筛选功能查看),检查每条数据都有时间后,智分析便能自动识别出时间。

### 5. 为什么我导入数据后,数据灵感只有三个图,而别人的有十多个?

当您导入的数据中既没有时间字段,又没有可以计数的字段(计数项)时(例如导入TXT文件),就只能进行文本分析,无法进行其他分析,因此数据灵感中就只有三个图。

6. 我把词语已经添加到词库中了,如何让它们生效?



©码有引力 (重庆) 科技有限公司 渝ICP备19011279号

#### 可点击右上角的<查看文件>按钮,然后再点击重新计算,即可让词库中的词语生效。