数据挖掘-分词

1. 功能概述

分词，是按照特定的规范将汉语中连续的字序列切分为合理的词序列的过程；换句话说，它能够自动识别出文本的词，在词库中进行搜索匹配，根据匹配的结果在词间加入边界标记符(如“/ ”等)，分隔出各个词组。

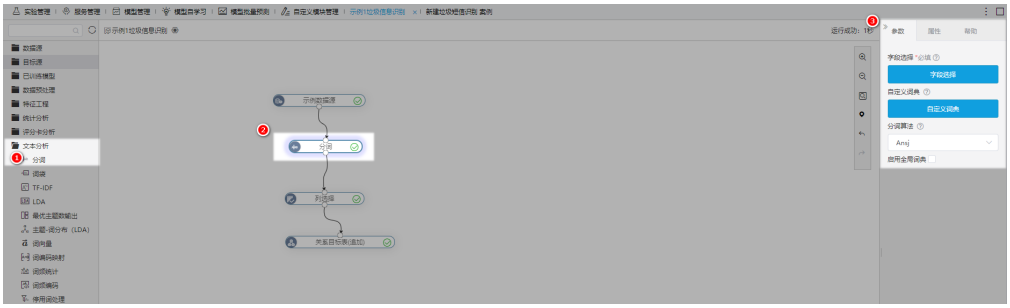
我们为什么需要用分词？

- 分词可帮助用户对海量信息的整理更准确更合理，提高信息处理效率，常用于做文本挖掘分析。
- 分词工作看似细微，但它对后续文本挖掘的关键作用是不容忽视的。若是分词效果不佳，即使后续算法优秀也无法实现理想的效果。

2. 使用说明

2.1 操作流程

从左侧资源树的 **文本分析** 中拖拽 **分词** 节点到画布中，选择数据输入，配置分词的参数（词典、算法等），最后再按需输出数据，具体操作步骤请浏览以下内容。



2.2 输入数据

本文采用**示例数据源**“深圳企业信息”作为输入。若是想使用其他数据来源，操作详情可参考输入数据

2.3 配置参数

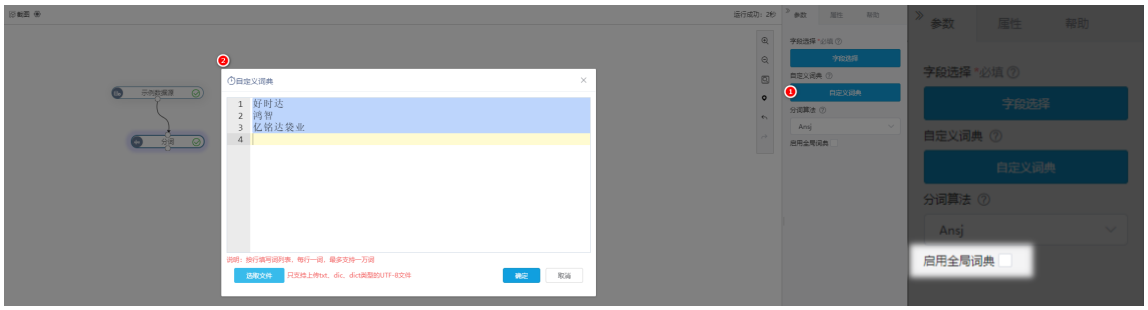
2.3.1 字段选择

字段选择，支持用户选择需要分词的字段（仅限**文本类型**的数据）。

2.3.2 词典

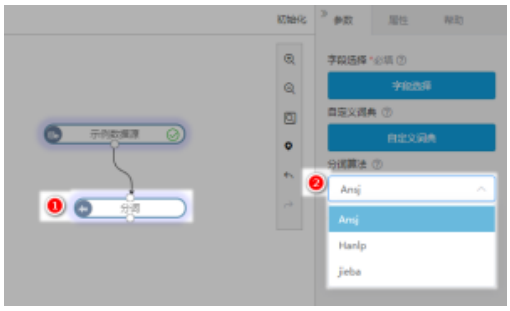
词典，用来满足用户根据不同专业领域制定不同的分词规范的需求，从而提高分词准确率。

名称	功能说明	生效范围	配置方式
自定义词典	用户可以添加业务用词、新词（未被系统自带词典收录的），作为重新组合词序列的规范。	只为**当前的分词节点**使用。	用户可通过两种方式输入自定义词： • 手动输入，每行填写一个自定义词； • 上传文件，只支持上传txt、dic、dict类型的UTF-8的文件，文件中每行填写一个自定义词且不超过一万行。
全局词典		在系统中所有实验中使用。	配置全局词典请参考引擎设置。



2.3.3 分词算法

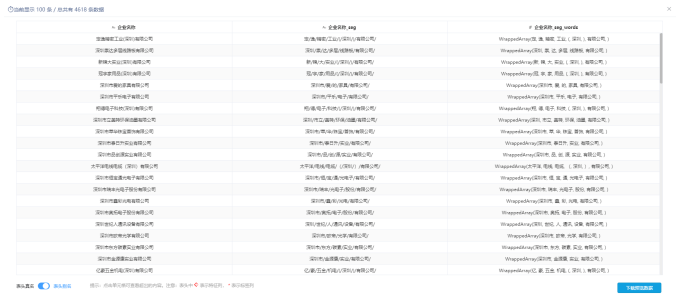
由于不同分词算法各有优劣，产品提供的分词算法可以满足用户根据不同的需求选用不同的算法。



算法名称	算法说明	应用场景
Ansj	可直接根据系统词库分出人名、机构等信息。但是多单词英文姓名无法分出。	适用于不使用自定义词典的场景。 （配置Ansj分词算法时，系统会优先使用自带的词典的词组，无论用户是否适用自定义词典。）
Hanlp	可分出多单词的英文名称。 但是，以文件添加自定义词典速度略慢。	自定义词典数据可包含空格。
Jieba	自定义分词方便。 以文件添加自定义词典比Hanlp 速度快。	

2.4 输出数据

如下图所示，分词是将 **企业名称** 进行分词，企业名称_seg为分词后的字符串型结果，企业名称_seg_words为分词后的WrappedArray类型结果。



注意事项

通过分词会输出array数据类型的字段列 "***seg_words" ，考虑到目前大部分目标数据库没有与之匹配的数据类型，因此，建议先通过 **列选择** 过滤掉array类型字段，或者先通过 **元数据编辑** 更改array数据类型为字符串，再导出到目标源。

3. 应用案例

请参考产品内置案例：“某政府单位疫情期间网民情绪识别” 和 “垃圾短信识别”。