

数据挖掘-停用词处理

概述

停用词处理是由于有些词频很高，但对文章却没有太多意义的语气词和助词等等，比如这、我、你们、吗等等，对这些词进行去掉处理以便能够更好的分析语义。常用于关键词提取分析。

- 概述
- 输入/输出
- 参数设置
- 示例

输入/输出

输入	一个输入端口，用于需要过滤停用词的字段。
输出	一个输出端口，用于输出去掉停用词之后的结果。

参数设置

参数名称	说明	备注
字段选择	选择需要去除停用词的列	必填
停用词列表	在对话框填写停用词，每行填写一个停用词	必填

示例

使用“深圳企业信息”示例数据，选择分词后去除停用词的列(企业名称_seg_words)，填写停用词，输出结果如下图所示：

当前显示 100 条 / 总共有 4618 条数据 提示:点击单元格可查看超出的内容

	# 企业名称_seg_words	# 企业名称_seg_words_filtered
公司/	WrappedArray(远, 特, 信, 电子, (, 深圳,), 有限公司,)	WrappedArray(远, 特, 信, 电子, 深圳, 有限公司)
司/	WrappedArray(深圳市, 万德福, 珠宝, 有限公司,)	WrappedArray(深圳市, 万德福, 珠宝, 有限公司)
公司/	WrappedArray(联, 致, 互, 应, 科技, (, 深圳,), 有限公司,)	WrappedArray(联, 致, 互, 应, 科技, 深圳, 有限公司)
限公司/	WrappedArray(百, 富, 计算机, 技术, (, 深圳,), 有限公司,)	WrappedArray(百, 富, 计算机, 技术, 深圳, 有限公司)
公司/	WrappedArray(深圳市, 西, 龙同辉, 技术, 有限公司,)	WrappedArray(深圳市, 西, 龙同辉, 技术, 有限公司)
司/	WrappedArray(深圳市, 凯, 得, 克, 科技, 有限公司,)	WrappedArray(深圳市, 凯, 得, 克, 科技, 有限公司)
司/	WrappedArray(冠, 德, 塑胶, (, 深圳,), 有限公司,)	WrappedArray(冠, 德, 塑胶, 深圳, 有限公司)
司/	WrappedArray(深圳市, 富魏盛, 科技, 有限公司,)	WrappedArray(深圳市, 富魏盛, 科技, 有限公司)
司/	WrappedArray(深圳市, 精, 腾, 电子, 有限公司,)	WrappedArray(深圳市, 精, 腾, 电子, 有限公司)
司/	WrappedArray(深圳市, 先, 德, 通, 科技, 有限公司,)	WrappedArray(深圳市, 先, 德, 通, 科技, 有限公司)
公司/	WrappedArray(深圳, 大森, 全息, 包装, 材料, 有限公司,)	WrappedArray(深圳, 大森, 全息, 包装, 材料, 有限公司)

注意: 表头中*表示特征列, *表示标签列

表头真名 表头别名