数据挖掘-TF-IDF

概述

一种统计方法，TF意思是词频，IDF意思是逆文本频率指数，用以评估一字词对于一个文件集或一个语料库中的其中一份文件的重要程度。字词的重要性随着它在文件中出现的次数成正比增加，但同时会随着它在语料库中出现的频率成反比下降。

TF-IDF值越高，说明该词越重要。

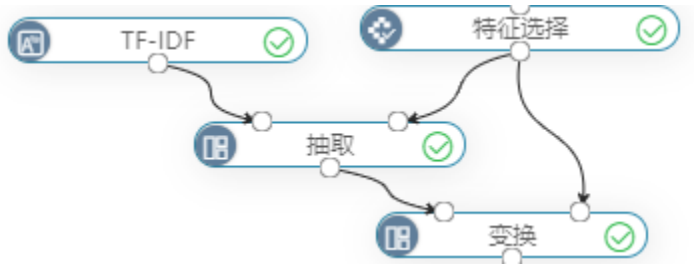
输入/输出

输入	没有输入端口
输出	一个输出端口，与抽取、变换节点组合使用。

参数设置

参数名称	说明	备注
特征项数	输入的数值n，代表算法最终会筛选出TF-IDF值最高的n个词	必填

示例



效果

使用“垃圾短信识别”数据，选择分词后的文本列_c2_seg_words_filtered，设置的特征项数为10，进行统计TF-IDF，输出结果如下图：