数据挖掘 - 相似集计算(LSH)

> 该功能为V10.5版本功能。

概述

使用训练好的LSH模型，对两份数据中的向量进行相似度匹配，把相似度距离低于预设阈值的组合输出到结果。

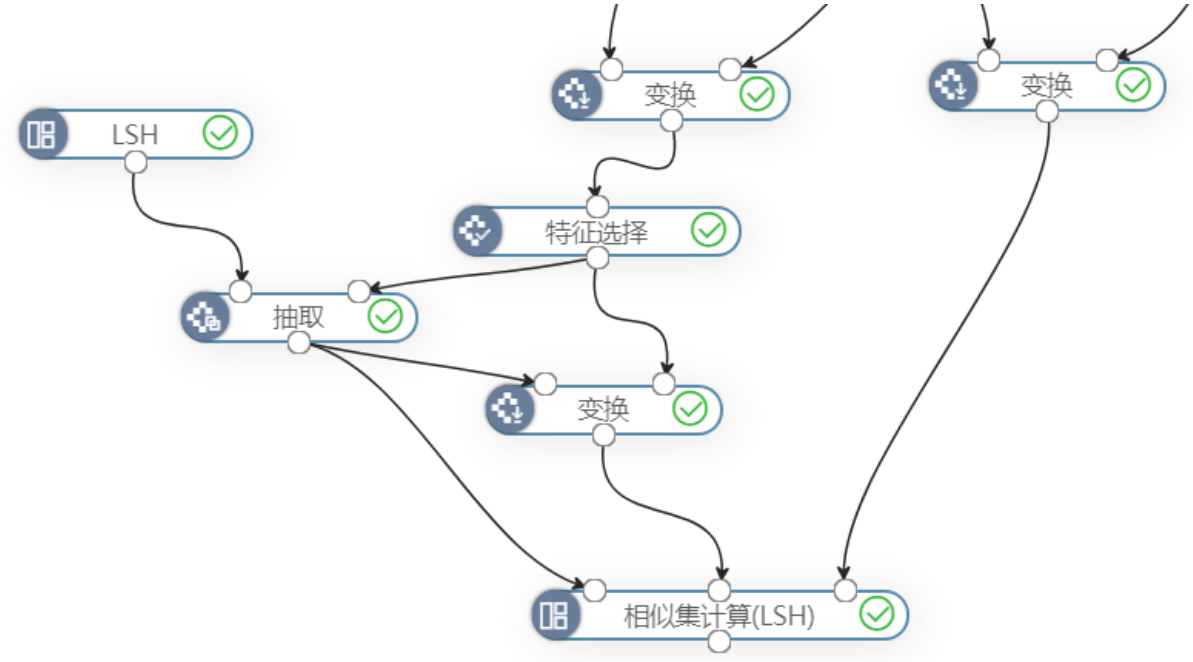
输入/输出

输入	三个输入端口，输入1接收训练好的LSH模型，输入2和3接收要进行匹配的数据。
输出	一个输出端口，用于输出匹配后的结果。

参数设置

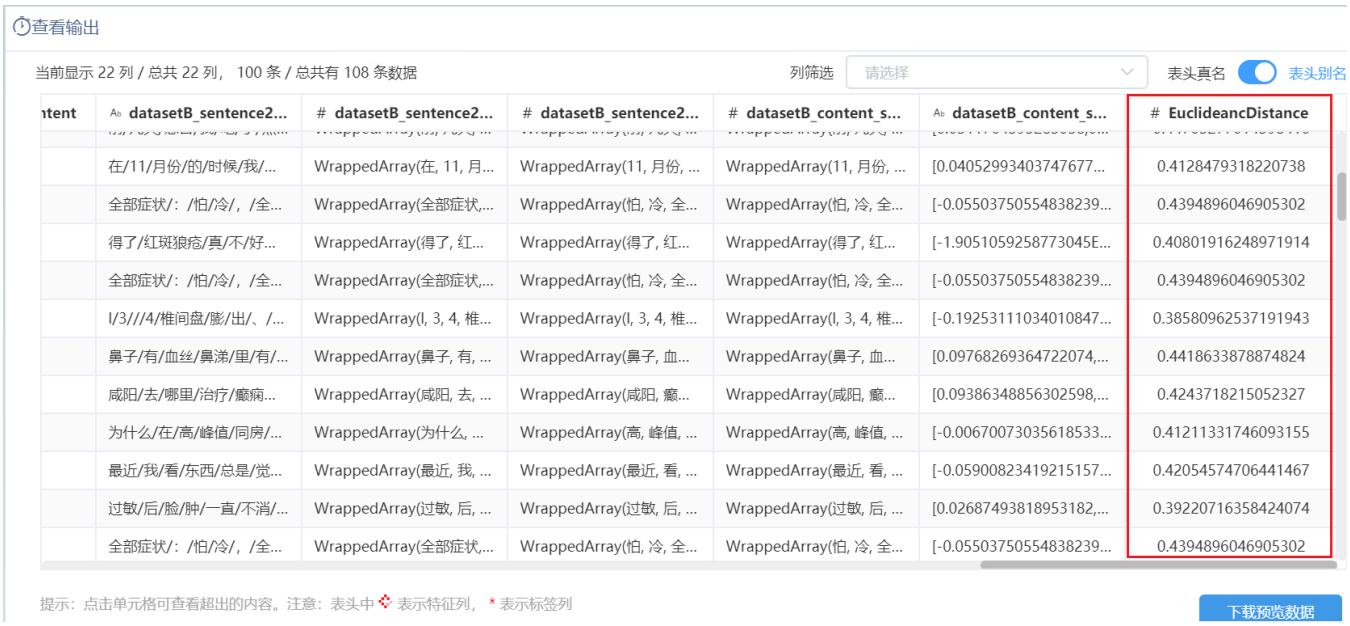
参数名称	说明	备注
相似阈值	设置相似度距离阈值	距离低于阈值的组合才会被输出

示例



效果

分别接入LSH模型，数据1和数据2，其中数据1和数据2都已使用词向量模型对其文本进行转换。把相似阈值设为0.45，输出结果中返回了两份数据中所有相似度距离小于该阈值的组合，其中数据1中的列会被标记为datasetA，数据2中的列会被标记为datasetB，如下图：



注意事项

数据1和数据2中必须包含训练LSH模型时使用的列名。如下图，实验图中特征选择节点使用了 content_seg_words_filtered_wordToVec 列，进行LSH模型的训练，那么在相似集计算节点，会对两份数据中的对应列作相似度匹配。