数据挖掘 - LSH

> 该功能为V10.5版本功能。

概述

LSH（局部敏感哈希）是一种哈希算法，用于对高维数据进行快速最近邻查找。LSH把两个高相似度的数据以较高的概率映射成同一个哈希值，把两个相似度很低的数据以较低的概率映射成同一个哈希值。利用哈希过后的数据进行最近邻查找，能提高查找效率，减少耗时。

对于数据向量的相似度距离，LSH节点提供了两种距离度量：欧式距离和杰卡德距离。其中，欧式距离适用于绝大多数数据向量，而杰卡德距离适用于由0和1组成的向量（如，00101，10011等，非0的数值都会被视为1）。在文本分析问题中，可先使用词向量或TF-IDF把文本转换为数值型向量，再选用欧氏距离的LSH对向量进行哈希，哈希后的向量可用于相似度匹配。

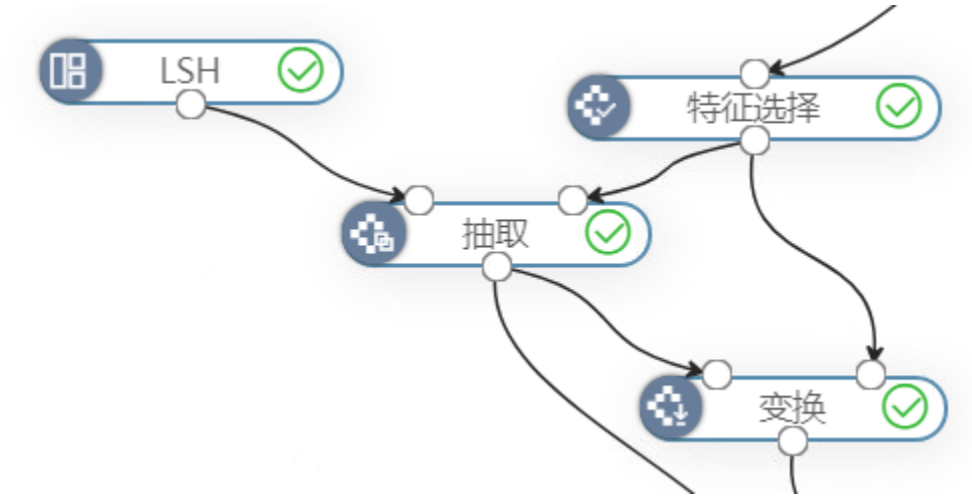
输入/输出

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
相似度计算方法	相似度距离度量	欧式距离和杰卡德距离
哈希存储桶的长度	每个哈希表内的哈希桶的长度，长度更长能降低假阴率	只适用于欧氏距离
哈希表数量	哈希表的数量	哈希后向量的长度

示例



效果

使用词向量算法把文本转换为向量后，选择该向量列。使用LSH算法，设置哈希表数量为4，输出结果如下图：

查看输出

当前显示 11 列 / 总共 11 列， 100 条 / 总共有 843 条数据　　列筛选 请选择　　表头真名　表头别名

2_seg	# sentence2_seg_wor...	# sentence2_seg_wor...	# content_seg_words...	content_seg_words...	# content_seg_words_filtered_wordToVecLSH
头/很/晕/...	WrappedArray(跑, 完, 步...	WrappedArray(跑, 完, 步...	WrappedArray(跑, 完, 步...	[-0.01297746459022164...	WrappedArray([-1.0], [0.0], [-1.0], [-1.0])
脚气/传...	WrappedArray(医生, 阿...	WrappedArray(阿姨, 脚...	WrappedArray(阿姨, 脚...	[0.04150690188010533...	WrappedArray([0.0], [0.0], [0.0], [-1.0])
/孕/检/...	WrappedArray(问, 一下, ...	WrappedArray(问, 一下, ...	WrappedArray(问, 一下, ...	[-0.03211454232223332...	WrappedArray([0.0], [-1.0], [0.0], [-1.0])
么/口/服...	WrappedArray(不, 想, 打...	WrappedArray(打针, 口, ...	WrappedArray(打针, 口, ...	[0.28031821353361014,...	WrappedArray([0.0], [0.0], [-1.0], [0.0])
月/发/2...	WrappedArray(时不时, ...	WrappedArray(时不时, ...	WrappedArray(时不时, ...	[0.13035690092614718,...	WrappedArray([-1.0], [-1.0], [-1.0], [-1.0])
头痛/作/...	WrappedArray(我, 爱人, ...	WrappedArray(爱人, 经...	WrappedArray(爱人, 经...	[-0.09482388012111187...	WrappedArray([-1.0], [-1.0], [0.0], [-1.0])
了/，/身...	WrappedArray(现在, 孩...	WrappedArray(一周, 身...	WrappedArray(一周, 身...	[0.20928237922489645,...	WrappedArray([0.0], [0.0], [-1.0], [0.0])
么/好/的/...	WrappedArray(未婚, 痛...	WrappedArray(未婚, 痛...	WrappedArray(未婚, 痛...	[-0.02439153641462326...	WrappedArray([0.0], [-1.0], [0.0], [-1.0])
孕/育/而/...	WrappedArray(第一胎, ...	WrappedArray(第一胎, ...	WrappedArray(第一胎, ...	[0.03263730038980059,...	WrappedArray([0.0], [-1.0], [0.0], [0.0])
法/哪/种...	WrappedArray(脑, 摊, 康...	WrappedArray(脑, 摊, 康...	WrappedArray(脑, 摊, 康...	[0.090272073041309,0....	WrappedArray([0.0], [0.0], [-1.0], [-1.0])
是/怎么/...	WrappedArray(小腿, 经...	WrappedArray(小腿, 经...	WrappedArray(小腿, 经...	[-0.00997659439841906...	WrappedArray([0.0], [0.0], [-1.0], [0.0])

提示：点击单元格可查看超出的内容。注意：表头中 表示特征列， * 表示标签列