

数据挖掘 - 相似集计算(LSH)

该功能为V10.5版本功能。

- 概述
- 输入/输出
- 参数设置
- 示例
- 注意事项

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。

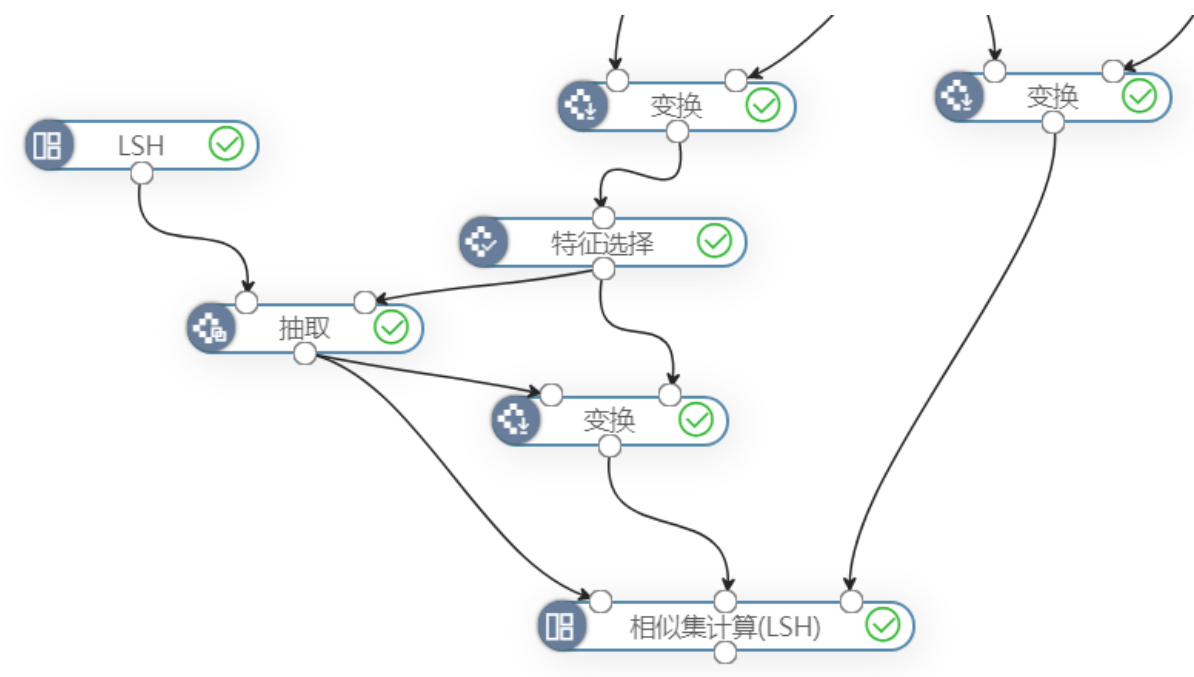
输入/输出

| | |
|----|--|
| 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：

查看输出

当前显示 22 列 / 总共 22 列, 100 条 / 总共有 108 条数据

列筛选

请选择

表头真名

表头别名

| intent | Ab datasetB_sentence2... | # datasetB_sentence2... | # datasetB_sentence2... | # datasetB_content_s... | Ab datasetB_content_s... | # EuclideanDistance |
|--------|--------------------------|----------------------------|----------------------------|----------------------------|---------------------------|---------------------|
| | 在/11/月份/的时候/我/... | WrappedArray(在, 11, 月... | WrappedArray(11, 月份, ... | WrappedArray(11, 月份, ... | [0.04052993403747677... | 0.4128479318220738 |
| | 全部症状:/怕/冷/, /全... | WrappedArray(全部症状,... | WrappedArray(怕, 冷, 全... | WrappedArray(怕, 冷, 全... | [-0.05503750554838239... | 0.4394896046905302 |
| | 得了/红斑狼疮/真/不/好... | WrappedArray(得了, 红... | WrappedArray(得了, 红... | WrappedArray(得了, 红... | [-1.9051059258773045E... | 0.40801916248971914 |
| | 全部症状:/怕/冷/, /全... | WrappedArray(全部症状,... | WrappedArray(怕, 冷, 全... | WrappedArray(怕, 冷, 全... | [-0.05503750554838239... | 0.4394896046905302 |
| | l/3///4/椎间盘/膨/出/ /... | WrappedArray(l, 3, 4, 椎... | WrappedArray(l, 3, 4, 椎... | WrappedArray(l, 3, 4, 椎... | [-0.19253111034010847... | 0.38580962537191943 |
| | 鼻子/有/血丝/鼻涕/有/有/... | WrappedArray(鼻子, 有, ... | WrappedArray(鼻子, 血... | WrappedArray(鼻子, 血... | [0.09768269364722074,... | 0.4418633878874824 |
| | 咸阳/去/哪里/治疗/癫痫... | WrappedArray(咸阳, 去, ... | WrappedArray(咸阳, 癫... | WrappedArray(咸阳, 癫... | [0.09386348856302598,... | 0.4243718215052327 |
| | 为什么/在/高/峰值/同房/... | WrappedArray(为什么, ... | WrappedArray(高, 峰值, ... | WrappedArray(高, 峰值, ... | [-0.00670073035618533,... | 0.41211331746093155 |
| | 最近/我/看/东西/总是/觉... | WrappedArray(最近, 我, ... | WrappedArray(最近, 看, ... | WrappedArray(最近, 看, ... | [-0.05900823419215157,... | 0.42054574706441467 |
| | 过敏/后/脸/肿/一直/不消/... | WrappedArray(过敏, 后, ... | WrappedArray(过敏, 后, ... | WrappedArray(过敏, 后, ... | [0.02687493818953182,... | 0.39220716358424074 |
| | 全部症状:/怕/冷/, /全... | WrappedArray(全部症状,... | WrappedArray(怕, 冷, 全... | WrappedArray(怕, 冷, 全... | [-0.05503750554838239,... | 0.4394896046905302 |

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

下载预览数据

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。

选择特征列

×

源数据列表

0/9

请输入搜索内容

- ☐ #question_id
- ☐ Ab content
- ☐ #question_id2
- ☐ #ans_id
- ☐ Ab content2
- ☐ Ab sentence2_seg

☒ 全部 ☐ 字符 ☐ 数字

到右边 >

< 到左边

已选字段列表

0/1

请输入搜索内容

Ab content_seg_words_filtered_wordToVec

确定

取消