数据挖掘-词向量

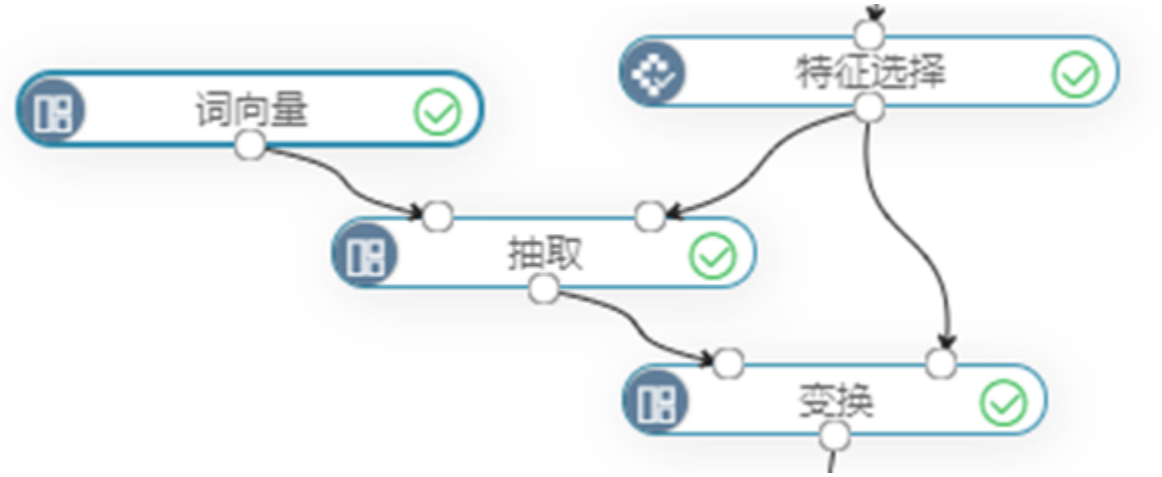
概述

词向量是表示文档的单词序列，通过训练Word2vec模型，将词语转化为向量。该模型将每个单词映射到一个唯一的固定大小向量。Word2Vec模型通过文档中所有单词的平均值将每个文档转换为一个向量；然后可以将该向量用作预测、文档相似性计算的特征。

参数设置

参数名称	说明
生成向量的数量	词向量的维度，默认值为50
词频	默认值为2，词频大于该值的词才能入选词典

示例



效果

使用“垃圾短信识别”示例数据，词向量的参数生成向量数量为50.词频为2，特征选择后，输出结果如下：

# target	text	_c2_seg	_c2_seg_words	_c2_seg_words_filtered	_c2_seg_words_filtered_...
0	23年从盐城拉回来的麻麻的嫁妆	23年/从/盐城/拉/回来/的/麻...	WrappedArray(23年, 从, 盐城...	WrappedArray(23年, 盐城, 拉...	[0.08159597932050625,-0.1...
0	乌兰察布丰镇市法院成立爱心...	乌兰察/布丰镇/市/法院/成立/...	WrappedArray(乌兰察, 布丰...	WrappedArray(乌兰察, 布丰...	[0.1481510316953063,-0.04...
0	有效服务和保障？“一带一路”...	有效/服务/和/保障/?/“/一/带...	WrappedArray(有效, 服务, 和...	WrappedArray(有效, 服务, 保...	[0.11429185722954571,-0.1...
0	predictionio去不了了	predictionio/去/不/了/了	WrappedArray(predictionio,...	WrappedArray(predictionio,...	[0.011170446872711182,-0....
0	为什么一层楼那么多个Ludovi...	为什么/一层/楼/那么/多个/lu...	WrappedArray(为什么, 一层, ...	WrappedArray(一层, 楼, 多个...	[-0.040491219889372584,-0...
0	南京下暴雨我在四川享受完美...	南京/下/暴雨/我/在/四川/享受...	WrappedArray(南京, 下, 暴雨...	WrappedArray(南京, 下, 暴雨...	[0.14469380144562038,-0.0...
0	保卫他们强奸你全家的权利	保卫/他们/强奸/你/全家/的/权...	WrappedArray(保卫, 他们, 强...	WrappedArray(保卫, 强奸, 全...	[0.14923215098679066,-0.0...
0	美职篮自由球员谈判期x日开启	美/职/篮/自由/球员/谈判/期/x...	WrappedArray(美, 职, 篮, 自...	WrappedArray(美, 职, 篮, 自...	[0.10647950656712056,0.00...
0	晶体全净无暇颜色漂亮xxmm	晶体/全/净/无暇/颜色/漂亮/xx...	WrappedArray(晶体, 全, 净, ...	WrappedArray(晶体, 全, 净, ...	[-0.09716353844851255,-0....
0	最好的距离是30到70公里之间	最/好/的/距离/是/30/到/70公...	WrappedArray(最, 好, 的, 距...	WrappedArray(最, 好, 距离, 3...	[0.033778250217437744,-0....
0	并以服务的形式提供Windows	并/以/服务/的/形式/提供/win...	WrappedArray(并, 以, 服务, ...	WrappedArray(服务, 形式, 提...	[0.07316478993743658,-0.2...
0	是大姑娘了要一点一点成熟起...	是/大姑娘/了/要/一点一点/成...	WrappedArray(是, 大姑娘, 了...	WrappedArray(大姑娘, 一点...	[0.1155399956740439,-0.15...