EMBDI工作

1. 目前进度

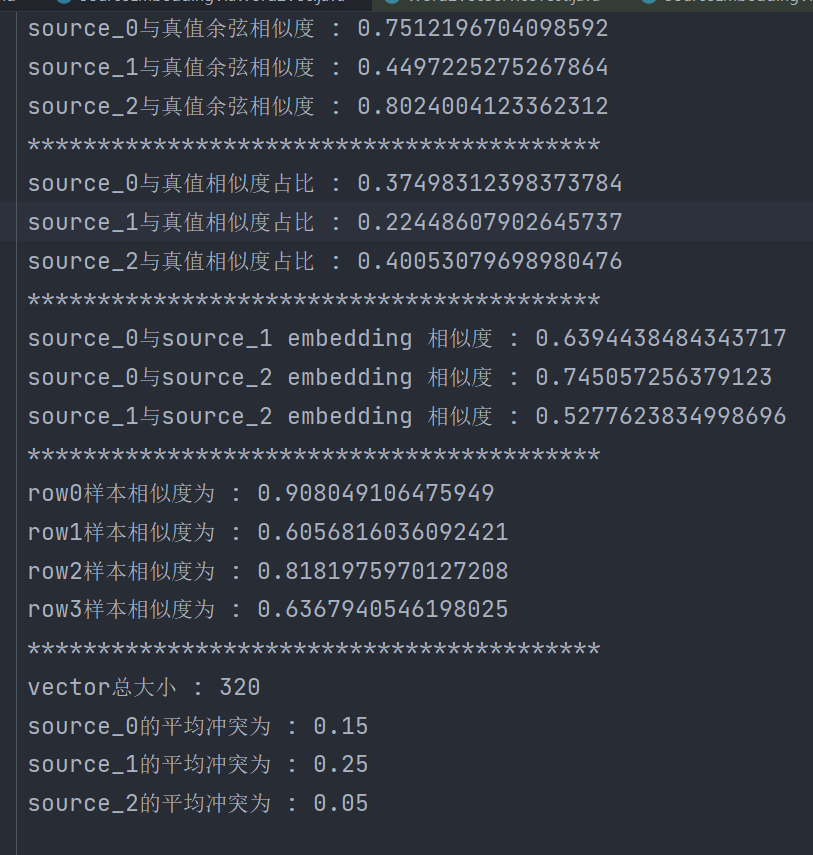
完成了三分图，五分图的建立，支持多数据源的输入，进行embedding，计算相似度等操作。实现了样本相似度的计算，数据源独立embedding的计算，数据源相似度的计算，数据源和真值之间相似度计算，以及数据源的平均冲突的计算。详细过程可以查看Github：<https://github.com/windblow32/ICDE2021>，对于gugu项目内攥写了说明性的README，可以查看到图的结构，相关文件信息等。

1. 实验说明

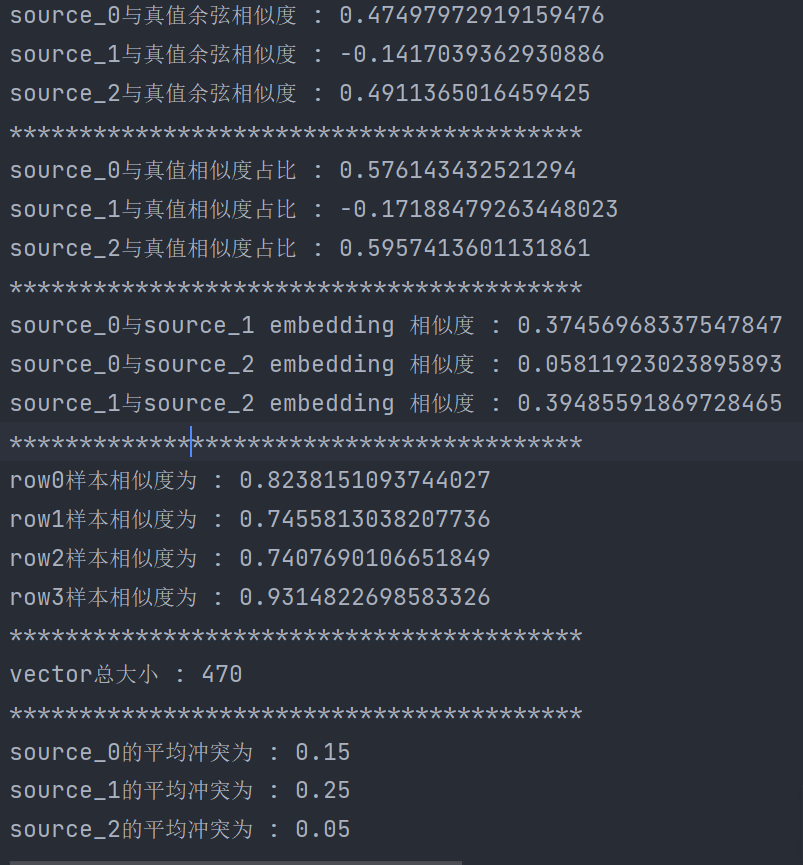
Github中data为实验所用的数据，其中xslx是未经处理的格式，存储在带有origin的文件夹中，CSV结尾是处理后用于测试的vsc文件，utf-8编码，逗号分隔。其中，sourceCSV有truth，即真值文件，进行了完整的训练测试，详细可查看github中log文件夹，存放了训练输出数据。SourceFiveEMBDI存放了五分图的输出结果，SourceThreeEMBDI存放了三分图的结果。

实验效果：

下图是三分图实验效果

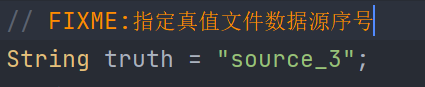


下图是五分图实验效果



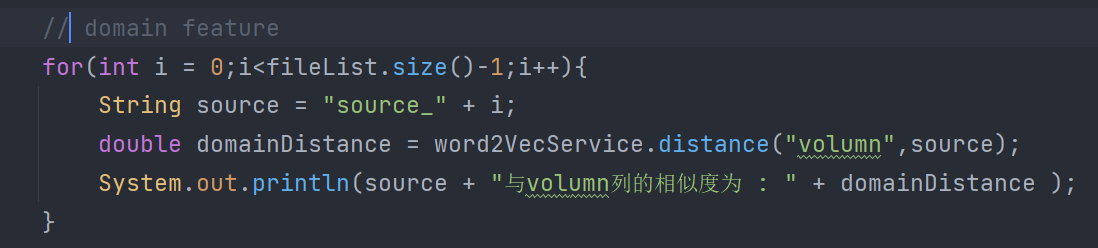
此处采取了三个数据源进行测试，三分图对于平均冲突较小的数据源，显示出了较大的相似度。其中，source1相似度出现了负值（？

EMDBI的使用已经得到了很大优化，只需要在Test中的相关测试文件如SourceEmbeddingViaWord2VecTest.java更改数据源路径，以及指定真值文件在数据源中排序序号（注意，务必使得真值排在数据源得最后一个）即可。



上图中，共4个数据源，真值是最后一个

Domain feature考量：对特定的属性，和source得embdi进行比较分析，这个是基于stock数据集的，如下所示：



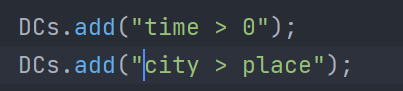
这里的F列，对应网站中的volumn（是嘛？），网站上给出的列数比叶姐处理的少两列，手头的处理后的第一列应该是数据源序号，按照观察，应该是volumn列冲突最多。

存在问题：1. Stock的真值文件，不咋会用，列数少一列，并且前几列的规律和处理后的应该不太一样。需要合适的truth文件。2. 数据源embedding后相似度是负值，存在什么问题。这里只通过接口，然后除了模长乘积，按理来说不会这样。

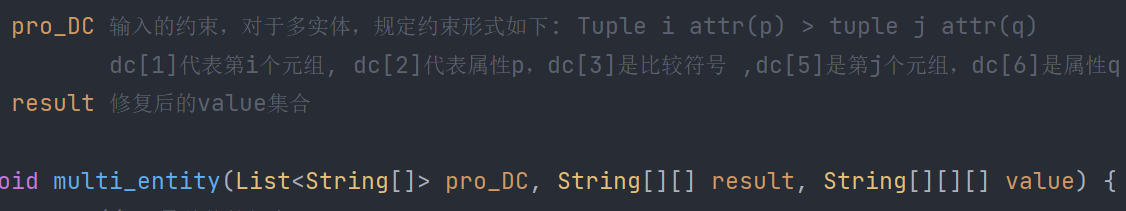
CTD

1. 项目现状

目前CTD已经可以和EMBDI进行衔接了，通过distance函数，访问我们的EMBDI中的model，并且调用model的diatance方法，获取对应的距离。采取的收敛准则是查看每次迭代后的error，这里error用result矩阵的变化来衡量。变化小，error就小。迭代停止条件是error小于1。支持多数据源输入，多约束输入。Entity方面，single entity支持如下输入：



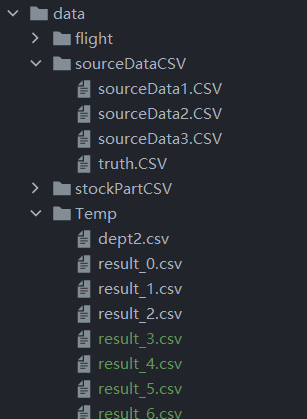
Multi entity需要用tuple前缀进行输入，如下所示（可以查看对应函数规约使用）



衔接符号这里仅是> 为例，支持了常见的 >, <, >=, <=, =, !=这些。

1. 与EMBDI的衔接

我们将EMBDI中模型的软件包复制过来供调用。在CTD每次更新真值矩阵完毕，准备更新权重之前，我们把现有的真值矩阵写入文件（csv格式）存放在data/Temp/result中，其中文件名字就是result序号



可以查看到每轮的真值矩阵更新情况。EMBDI的train就获取到这个文件目录，从中读取文件，用于训练，同时也要读取原数据，生成EMBDI计算距离。

1. 优化方向及实验相关数据

EMBDI训练耗费了最多的时间，因此必须把train这个过程从循环中分离出来，但也要保证每次更新真值后，embedding都重新训练。文件io操作耗时并不长，可以忍受，约为4-5秒。

目前，在三个数据源，每个数据源4个属性的情况下，需要20分钟，7轮训练，达到了要求，并且输出了weight，也就是对应的数据源权重。

详细文件可查看github获取