# 联合多维特征和学术网络嵌入的学者姓名消歧

实验流程（详细版）：

PS：

实验环境：

全局表示阶段使用得是python3.6的环境以及一系列框架。

其余阶段使用得是python3.9的环境以及一系列框架。

代码以及数据存放规则：

代码在src文件夹中

数据在datas文件夹中

两个文件夹内容一一对应

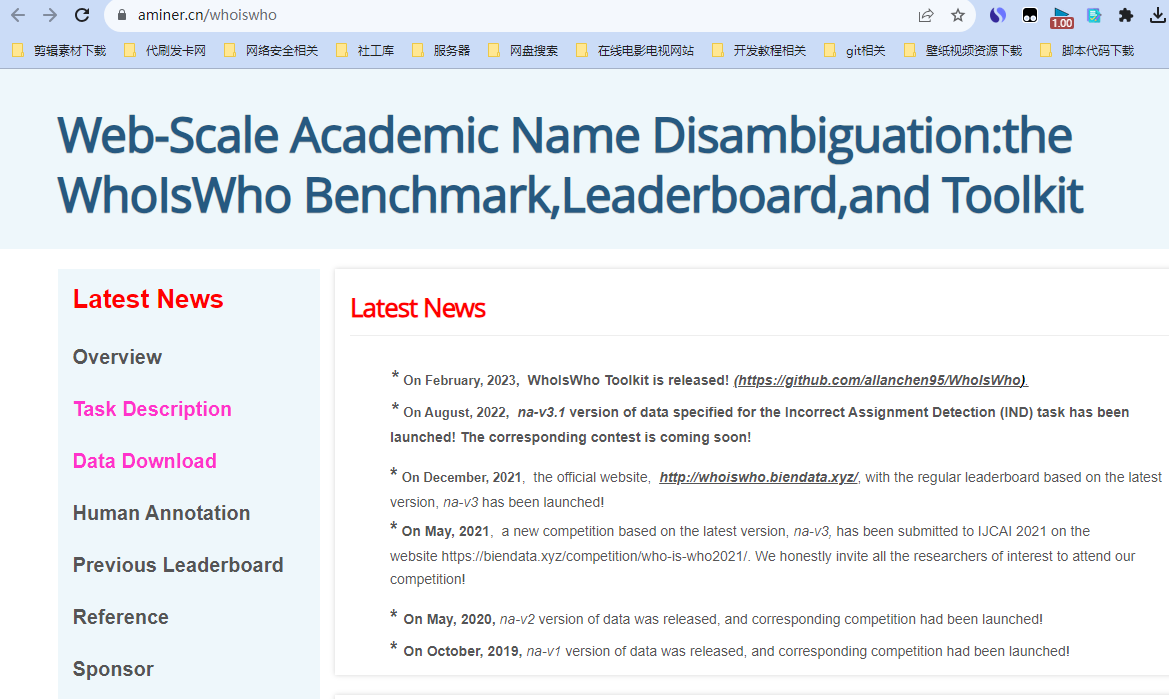
\*代表测试集test or训练集train。

所有规则在所有模块通用。

### 一，数据准备

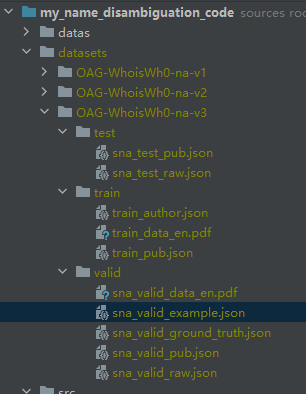
#### 下载数据

从https://www.aminer.cn/whoiswho 下载whoiswho 数据集。



#### 存放位置

将解压的数据放到datasets文件夹中。

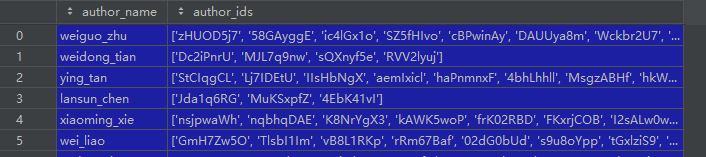


### 二，数据预处理

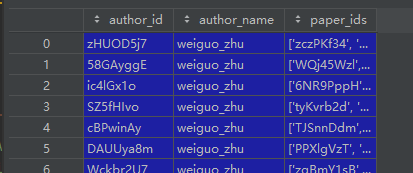
对应/pre\_data文件夹

#### 运行pickle\_data.py，将数据提取到pandas中：

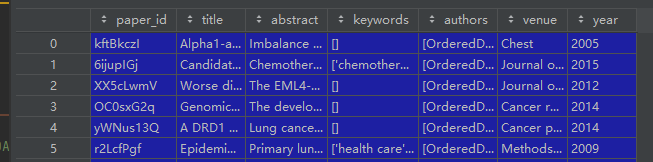
1. 作者姓名-作者ids 数据，存放在\*\_author\_name\_ids.pkl。



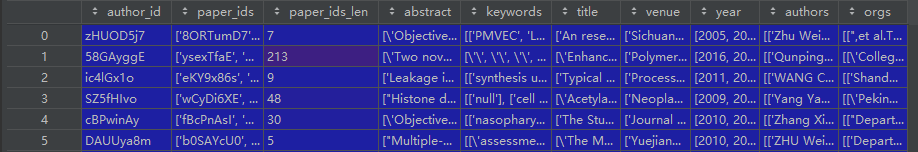
1. 作者id-论文id数据，存放在\*\_author\_name\_paper\_ids.pkl。



1. 论文信息，存放在\*\_pub\_info.pkl。



1. 作者-论文信息，存放在author\_pub\_detail.pkl。



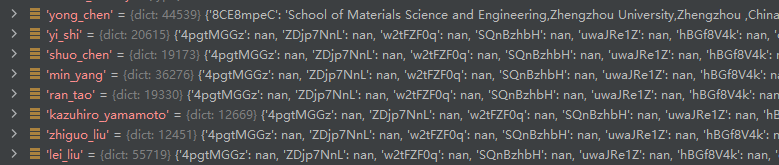
#### 运行repare\_data\_name.py进行合作者姓名与待消除歧义姓名的处理。

处理结果存放在 author\_name\_map.pkl，这一步将不同写法得姓名进行统一识别。

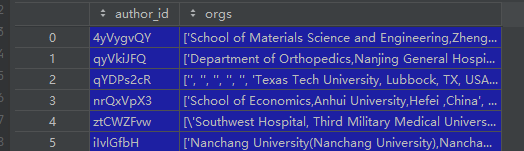


#### 运行repare\_data\_orgs.py 对机构数据进行处理

1. 得到作者-机构映射表，存放在 author\_org\_map.pkl



1. 以及得到作者id-机构映射表，存放在author\_id\_org\_map.pkl



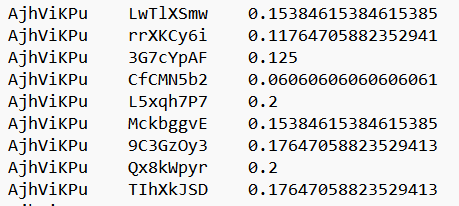
#### 4,运行generate\_data.py将所有特征数据进行详细处理

这一步包括去除停用词，标点，乱码词等脏数据，大小写转换，中文转拼音，句子转词等一系列处理，得到待向量化得数据。



#### generate\_adj.py生成相似度图

图边进行加权，具体为两篇论文的对应特征维度的相似度的值。

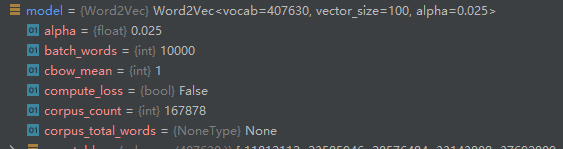


### 初始表示学习阶段

对应/pre\_embedding 文件夹

#### train\_word2vec.py训练word2vec

训练好的特征存放在word2vec\_xl\_100.model ：



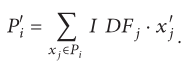
#### get\_pre\_embedding.py生成word2vec+idf初始特征表示

1. 将语义特征输入到 Word2Vec 模型，得到每个单词的低维向量表示x’。
2. 计算每个词出现的的 IDF 值。

IMG_256

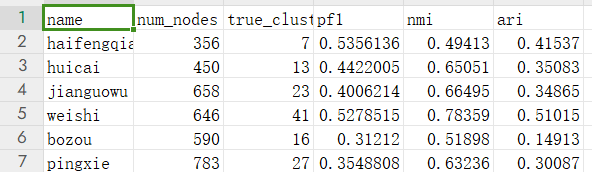
(3)将语义特征中每个词的词向量乘以对应的IDF值，并将所有乘积相加得到论文的统一嵌入向量。

IMG_256



（4）对所有特征计算初始向量表示后，存放在 \*\_pub\_pre\_embedding.pkl



1. 这一部分得消歧结果存放在out文件夹中

### 全局表示学习阶段

对应/global\_embedding 文件夹

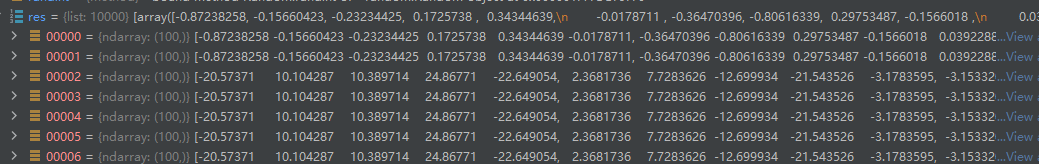
#### valid\_pkl2npy.py 将初始特征pandas数据转为numpy。

因为全局表示学习为python36，而其他模块为pytohn39，pandas数据不兼容，必须使用nunpy进行衔接。

转换后的数据存放在 \*\_pub\_pre\_embedding\_npy.npy

#### get\_data\_for\_global\_embedding.py生成正负样本对

用于进行后续的全局表示学习模型得训练。结果存放在 特征维度/\*-triplets/文件夹中。



#### global\_embedding.py进行全局表示学习

将测试集数据通过训练后的模型，得到全局嵌入表示，存放在global\_embedding\_npy.npy，中。

#### global\_npy2pkl.py 将全局表示学习数据转换回pandas数据。

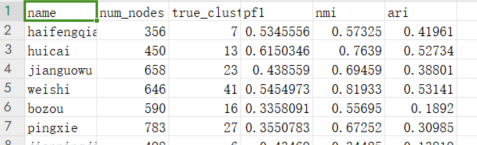
因为全局表示学习为python36，而其他模块为pytohn39，后续必须在pandas数据上进行，必须将nunpy转换回pandas数据。

转换后的数据存放在 \*\_global\_embedding.pkl。

全局表示阶段得结果为：



这一部分得消歧结果存放在out文件夹中

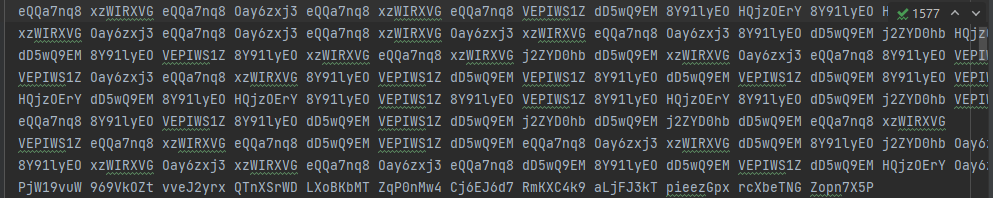


### 异构图表示阶段

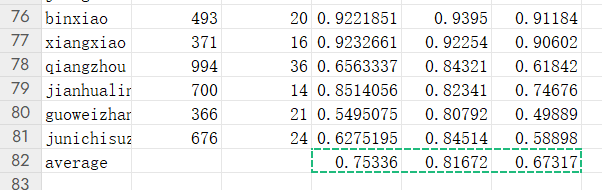
对应/hin\_embedding 文件夹

#### 1,meta2vec.py进行异构图表示学习

（1）每个姓名构成得异构图得随机游走路径放到 姓名/rand\_walks.txt 中



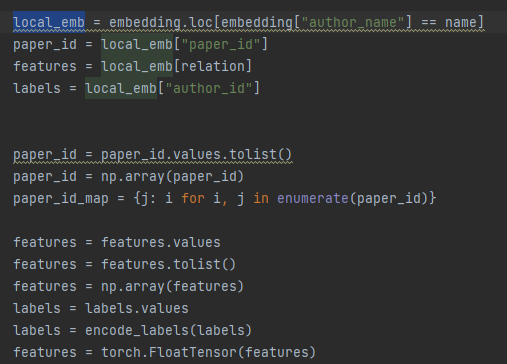
（2）所有异构图结果放在hin\_embedding.pkl

1. 这一部分得消歧结果存放在out文件夹中

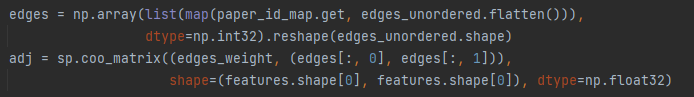
### 融合表示学习阶段

对应/local\_embedding 文件夹

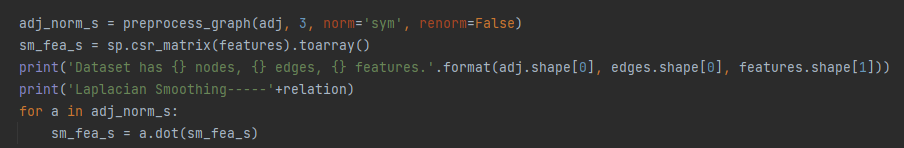
1. 加载全局表示数据



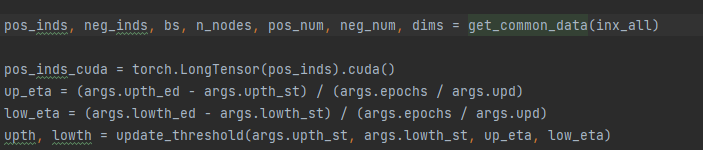
2，加载加权图



3，拉普拉斯滤波



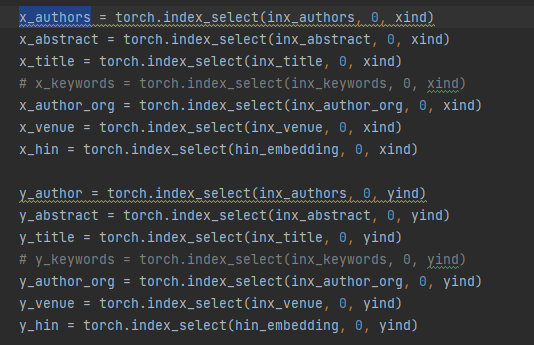
4，第一次动态构建数据集（使用所有特征内积）



1. 样本对抽样

IMG_256

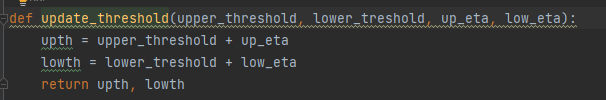
1. 所有特征得到对应得正负样本对。



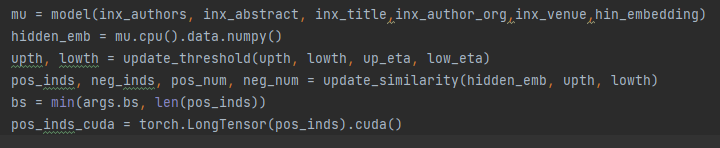
1. 使用交叉熵损失训练模型。



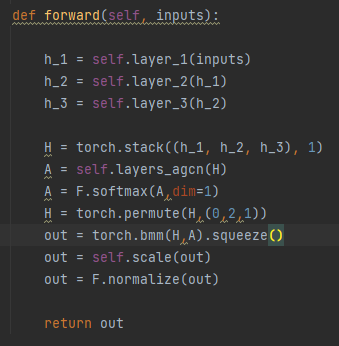
1. 阈值更新



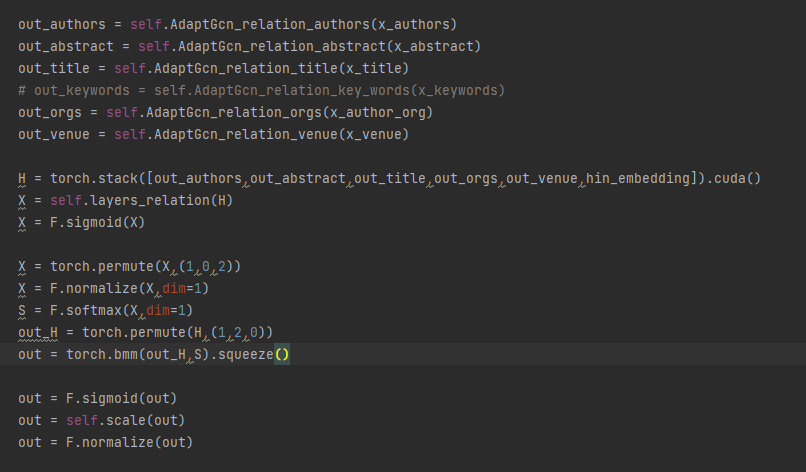
1. 更新相似度，生成正负样本对，动态构建下一个数据集



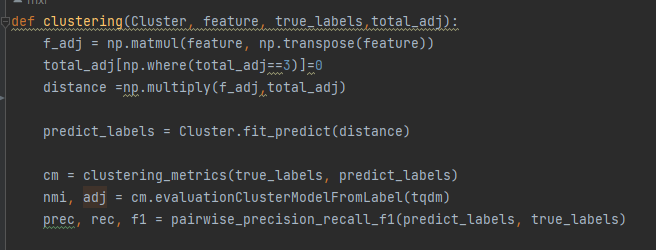
1. 自适应SGC模型



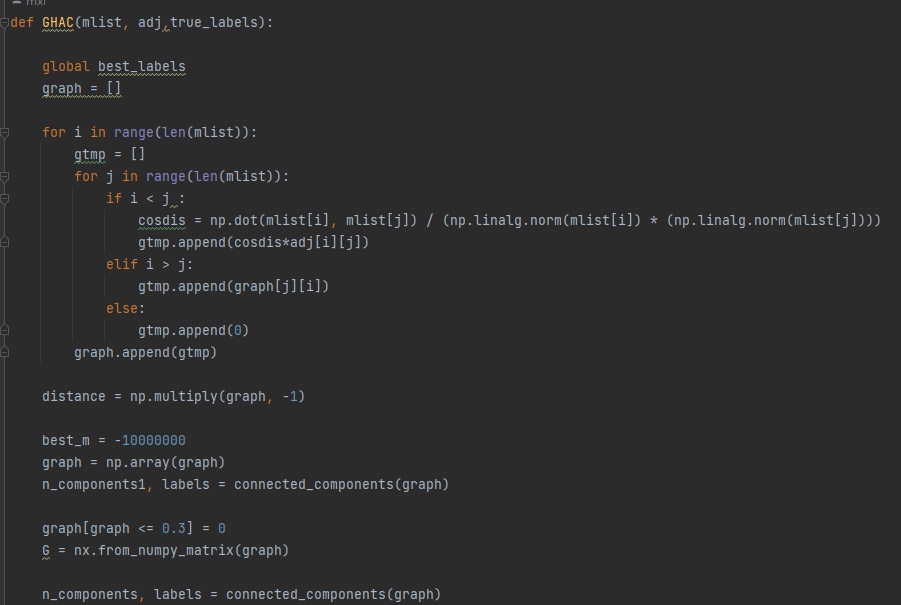
1. 多维关系加权融合（包括了异构图得embedding）



12,基于规则构造加权图，进行谱聚类。

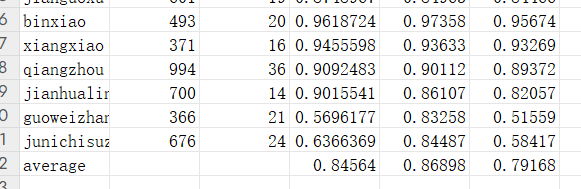


13，使用GHAC进行在不需要真实簇个数情况下进行聚类。





这一部分得消歧结果（即最终得姓名消歧结果）存放在out文件夹中



### 单独聚类评估

1. evaltion\_embedding.py对各个阶段的embedding进行谱聚类评估。

可以单独对各个模块得embeding进行评估，结果放在各个模块对应得out文件夹中。