241006

统计共现关系：对每个实体建立一个边框，统计重叠率，考虑10类实体，一张图就是10个节点，每个节点算一个实体类，若某类实体重叠某类实体超过某个阈值，则建立一条边，1（该实体类，剩余不重叠的数量）+10（该实体类和其它实体类重叠的数量，自身类与自身类为0）组成，即该节点对于所有其它实体的重叠数量和该实体类不重叠的实体数量，即考虑了数量又考虑了位置

节点定义：将10类实体类型中的每一类视为一个节点，因此图中共有10个节点。

边的建立：

- **边界框计算**：对每个实体建立一个边界框（先简化处理，获得上、下、左、右四个坐标）。
  - 对于：POLYLINE，ELLIPS，我们的数据集没有这两种，SPLINE只有极少数有1个，只考虑下述10种，（需要更精细的处理）
    - INSERT ： 遍历块中的实体，计算其边界框
    - LINE：将起点和终点作为边界框的两个极端点
    - TEXT ： 文本的插入位置 `(x, y)` ，文本的高度和宽度，默认是横向或竖向，如果是斜的，就不精确
    - MTEXT
    - HATCH ： 遍历填充实体的边界路径，取边界路径的关键点
    - LWPOLYLINE ： 将所有顶点添加到 `points` 列表中。通过计算 `points` 中的最小和最大坐标，得到边界框。
    - LEADER ： 获取引线的顶点
    - CIRCLE：圆的边界框是其外接正方形
    - DIMENSION： （计算重叠不够精确）收集标注的关键点：
      - `def_point` ： 定义点，通常是标注的基准点。
      - `text_midpoint` ： 标注文字的中点。
      - `dim_line_point` ： 标注线的位置。
    - ARC：采样圆弧上的点，通过 `points` 列表中的点，计算最小和最大坐标，得到边界框
- **重叠率计算**：统计不同实体类型之间的边界框重叠率。

原理：

- 每行表示一种实体类型的特征向量。

- 自身类型对应的值为 **不与任何其他实体类型重叠的实体数量。**
- 其他类型对应的值为 **与其他实体类型发生重叠的实体数量。**
- 使用**空间索引**，**R 树（R-tree）**，来减少需要进行重叠检查的实体对数量。

实体类型	ARC	TEXT	MTEXT	LWPOLYLINE	INSERT	DIMENSION	LEADER	CIRCLE	HATCH	LINE
ARC	309	0	0	204	1	19	0	2	252	564
TEXT	0	9	0	0	0	0	0	0	0	0
MTEXT	0	0	4	0	0	0	0	0	0	0
LWPOLYLINE	8	0	0	70	4	27	4	4	149	1
INSERT	1	0	0	1	0	1	0	0	1	1
DIMENSION	7	0	0	18	1	4	1	6	14	6
LEADER	0	0	0	2	0	1	0	2	1	0
CIRCLE	1	0	0	1	0	1	2	0	2	1
HATCH	4	0	0	10	5	8	1	4	2	9
LINE	528	0	0	2	1	9	0	1	34	1730

- **边的建立**：基础就是如果两个实体类型之间的重叠数量大于零，则在它们之间添加一条无权边。
- **归一化**：实现：有少数只有一两个实体重叠，则归一化为0，这样就无需在这两类之间建立边，两种方式：
  - 对于特征矩阵的每一行，将自身类对应的自身类放在第一列，后续列为固定顺序的其它实体类，对第一列的所有行进行列归一化，对后续列的每一行进行行归一化，这样就造成了每一行，会有一个为0
  - 归一化的特征矩阵：

实体类型	SELF	ARC	TEXT	MTEXT	LWPOLYLINE	INSERT	DIMENSION	LEADER	CIRCLE	HATCH	LINE
ARC	0.1751	0.0	0.0	0.0	0.2162	0.0	0.3784	0.0	0.0	0.0811	0.3243		
TEXT	0.035	0.0	0.0	0.0	0.0303	0.0	0.8182	0.0	0.0	0.0606	0.0909		
MTEXT	0.0078	0	0	0	0	0	0	0	0	0	0		
LWPOLYLINE		0.0156	0.3137	0.0196	0.0	0.0	0.0	0.5686	0.0	0.0	0.0196	0.0784	
INSERT	0.0	0	0	0	0	0	0	0	0	0	0		
DIMENSION		0.1673	0.1407	0.1357	0.0	0.1457	0.0	0.0	0.005	0.0	0.3166	0.2563	
LEADER	0.0233	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0		
CIRCLE	0.0	0	0	0	0	0	0	0	0	0	0		
HATCH	0.0233	0.0732	0.0244	0.0	0.0122	0.0	0.7683	0.0	0.0	0.0	0.122		
LINE	0.5525	0.2609	0.0326	0.0	0.0435	0.0	0.5543	0.0	0.0	0.1087	0.0		

  - 对每一行直接进行归一化，这样就使得重叠的和不重叠的一起归一化，这样就没有利用起不同实体类，数量不同的信息

- 归一化的特征矩阵（每一行对应一个实体类型，列顺序为 ENTITY_TYPES）：

ARC	TEXT	MTEXT	LWPOLYLINE		INSERT	DIMENSION		LEADER	CIRCLE	HATCH	LINE
0.1268	0.0	0.0	0.2254	0.0	0.2817	0.0	0.0	0.0845	0.2817		
0.0	0.3077	0.0	0.0769	0.0	0.3846	0.0	0.0	0.0769	0.1538		
0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0		
0.125	0.125	0.0	0.25	0.0	0.25	0.0	0.0	0.125	0.125		
0	0	0	0	0	0	0	0	0	0		
0.1221	0.1374	0.0	0.2137	0.0	0.0382	0.0076	0.0	0.2748	0.2061		
0.0	0.0	0.0	0.0	0.0	0.1667	0.8333	0.0	0.0	0.0		
0	0	0	0	0	0	0	0	0	0		
0.0833	0.1667	0.0	0.0833	0.0	0.3333	0.0	0.0	0.1667	0.1667		
0.1203	0.019	0.0	0.0253	0.0	0.1076	0.0	0.0	0.057	0.6709		

PS C:\artn\anaconda\ascend>