

利用 FME 进行 POI 名称相似检查并分组

作者：且行且珍惜

指导老师：刘畸生

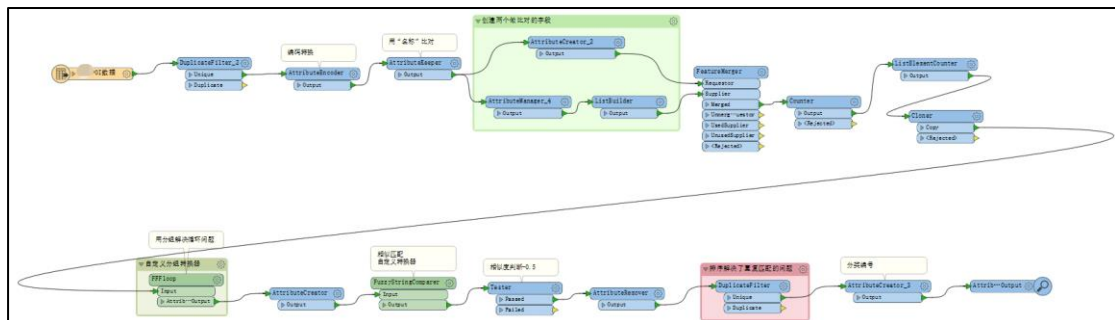
1 目的

做项目时，有时我们会接触到不同来源的数据，数据属性字段中的“数据名称”或“数据地址”值可能会有重复或相似，面对此种情况，我们可以利用 FME 软件进行数据的统一分类，像名称相同的或名称相似度高(大于 0.5)及以上的，可以做一个统一编号，方便后续对数据进行处理，提高工作效率。

2 FME 处理流程

2.1 整体流程

从下图可以看到，处理的流程不算复杂，首先创建一个存储所有需要比对的属性的列表，然后用 ListIndexer 依次将列表中的属性暴露出来，与原属性进行比对，在属性暴露的时候采用了克隆加分组的方式，这样避免了用循环。然后用 FuzzyStringComparer 自定义转换器进行打分，最后把分数大于 0.5 的记录过滤出来，并把重复匹配的记录过滤掉。



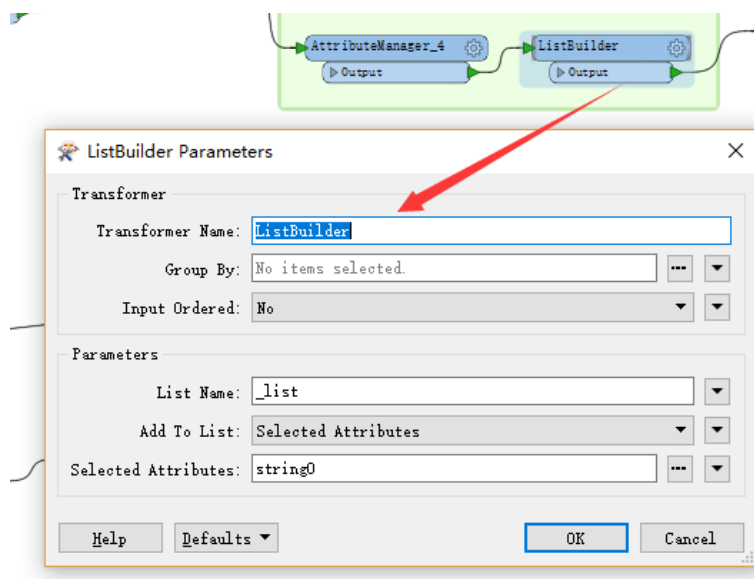
2.2 处理步骤

2.2.1 数据预处理

数据预处理包括对需要匹配的字段去除重复、编码转换、属性字段整理等。

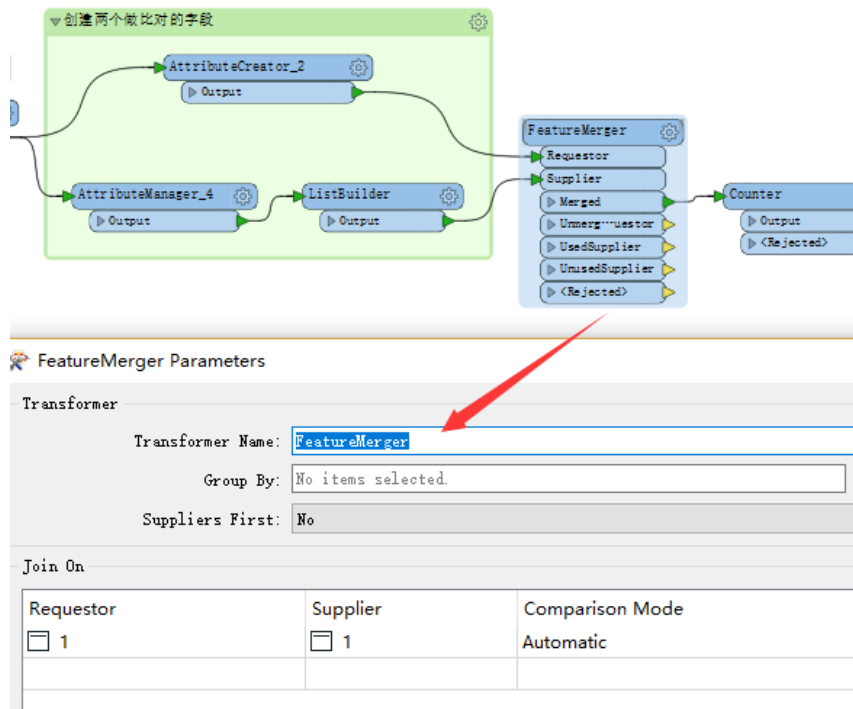
2.2.2 创建两个需要匹配的字段

首先用 AttributeCreator 转换器 ,创建一个新 的字段 “string” ,属性值用 “名称” , 然后创建第二个字段 “string0” , 属性值也用 “名称” , 对第二个字段用 ListBuilder 创建一个列表。



2.2.3 利用 FeatureMerger 进行属性合并

利用 FeatureMerger 进行属性合并 , 需要注意的是 , 在设置时 , requestor 和 supplier 都设置为 1 , 即可进行属性的合并。

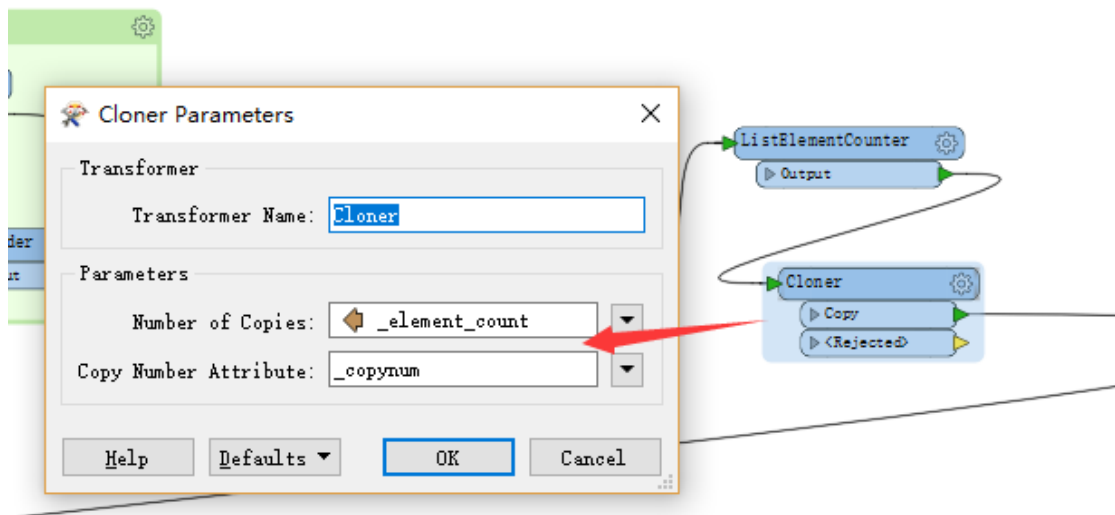
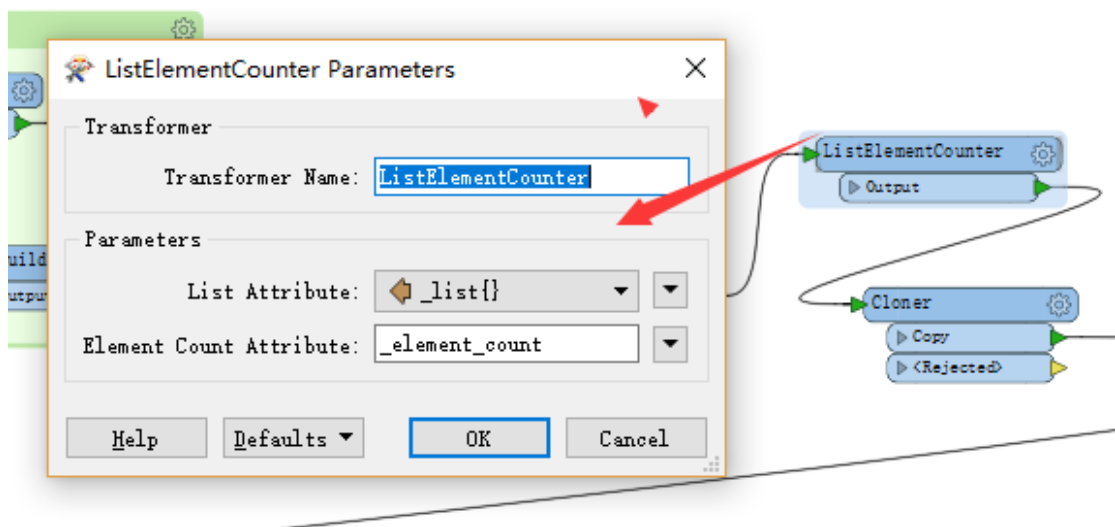


2.2.4 利用 Cloner 进行数据的克隆

因 FuzzyStringComparer 转换器做比较时，是用两个字段对应比较，所以要满足这样一个要求，需要每一条数据克隆总数据量的份数。

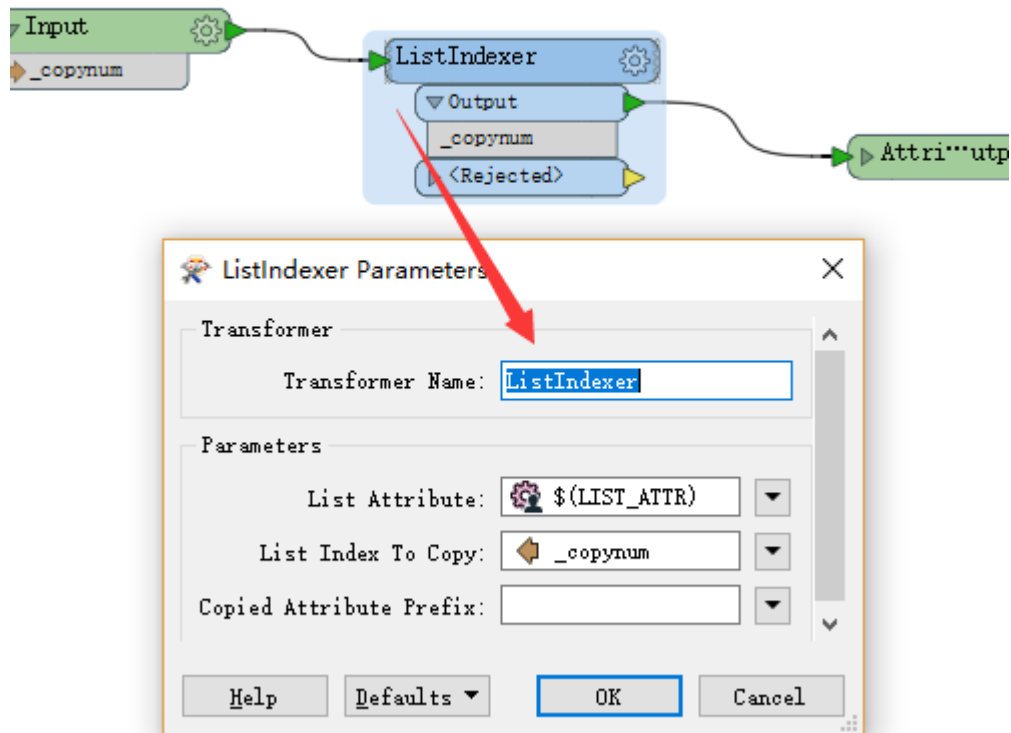
首先需要用 ListElementCounter 创建对象列表，然后用 Cloner 克隆对应的要素数量的份数。

在 ListElementCounter 设置时，创建列表的属性用“_list{ }”，在 Cloner 设置时，克隆的数量，用前面生成的“_element_count”值。



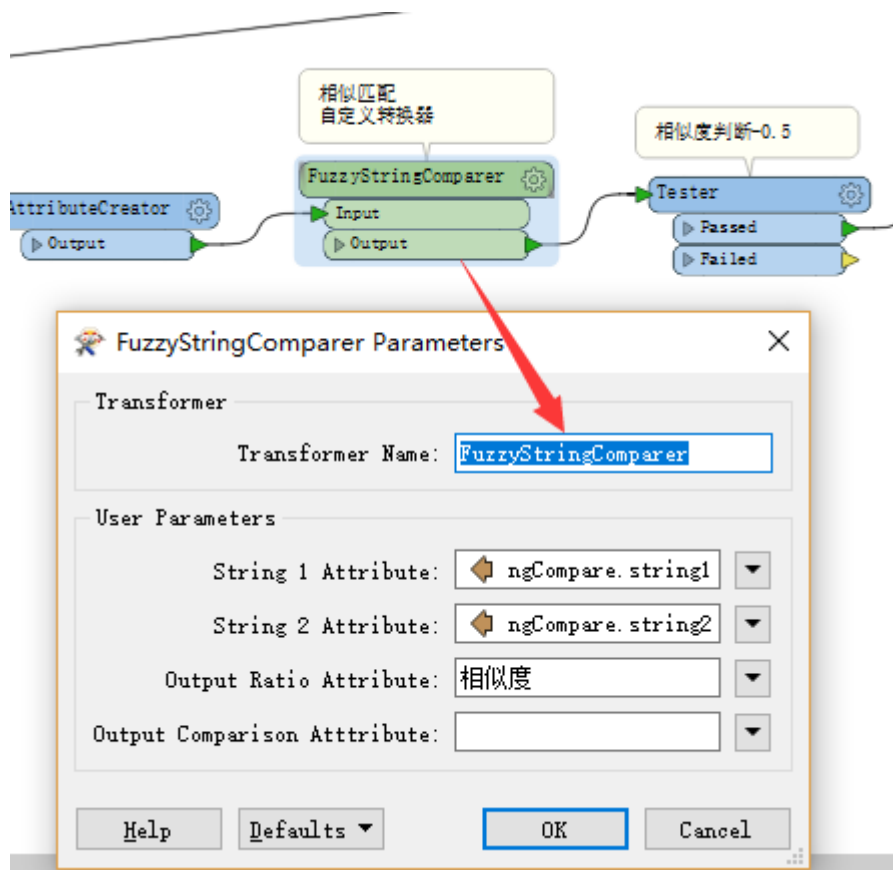
2.2.5 利用 ListIndexer 暴露属性

利用 ListIndexer 依次把列表中的元素暴露出来，这里用到了一个自定义转换器，是因为需要用到分组属性。



2.2.6 属性重命名，满足 FuzzyStringComparer 比较要求

FuzzyStringComparer 是一个自定义转换器，搜索即可找到，为满足这个转换器的要求，需要创建两个字段 “FuzzyStringCompare.string1” 和 “FuzzyStringCompare.string2”，此处通过 AttributeCreator 完成，然后进行数据的相似匹配设置即可进行数据的相似匹配，生成一个相似度值。



2.2.7 利用 Tester 进行相似度过滤

数据经过 FuzzyStringComparer 转换器做相似比较后，会生成一系列新的相似度的属性字段，数据值为 0 到 1，此处我们可以自行进行经验判断进行相似度过滤，以方便后续数据整理。

2.2.8 进行数据去重和编号

数据经过 FuzzyStringComparer 处理后，生成的是一个矩阵表，需要对重复数据进行过滤，最终达到输入数据数量和输出数量相等。

最后对相似度一致的数据，进行统一的编号，方便后续数据的检查和处理。

最后结果如下图，相似度大于 0.5 的给一个统一的 BH(编号)。

	BH	▲ 名称	2≡	相似度
193	6	金鼎大厦停车场(出口)		0.8
194	62	新盛大厦B座		0.6153846153846154
195	62	金益大厦		0.5454545454545454
196	62	金鼎大厦		0.5454545454545454
197	68	金益大厦(金融大街)		0.7272727272727273
198	68	金亚光大厦地下停车场(出口)		0.5384615384615384
199	68	金亚光大厦-C座		0.5
200	68	金亚光大厦C座		0.5263157894736842
201	68	金亚光大厦-B座		0.5
202	68	金亚光大厦B座		0.5263157894736842
203	68	金亚光大厦-A座		0.5
204	68	金亚光大厦A座		0.5263157894736842
205	68	金亚光大厦(金融街一号)		1
206	68	金亚光大厦		0.5882352941176471
207	68	金融大街一号		0.5555555555555556
208	68	光与茶(金融街店)		0.5714285714285714

3 经验总结

FME 在读取 POI 数据时，若数据量比较大，达到上千条以上时，处理效率不高(因数据克隆)，期待后续在这方面有所改善。但即便如此，FME 依旧让我们看到了它的强大，万能的 FME！