改良方案

增量地址进入

判断是否有经纬度

1无经纬度

直接进入字符串相似度识别，相似度算法优化，采用余弦相似度，现在好像是最大匹配。匹配示例，数据1

山东省德州市乐陵市山东省乐陵市郑店镇经济开发区钢材大市场

数据2

山东省德州市乐陵市郑店镇经济开发区钢材市场

将两个数据拆分成单个字的集合

｛山东省德州市乐陵市山东省乐陵市郑店镇经济开发区钢材大市场｝

｛山东省德州市乐陵市郑店镇经济开发区钢材市场｝

求出两个集合的并集

｛’山‘’东‘’省‘’德‘’州‘’市‘’乐‘’陵”郑‘’店‘’镇‘’经‘’济‘’开‘’发‘’区‘’钢‘’材‘’大’’场’｝

然后算出每个集合对应的字频的集合，就是一个字出现了几次

｛2,2,2,1,1,4,2,2,1,1,1,1,1,1,1,1,1,1,1,1,1｝

{1,1,1,1,1,3,1,1,1,1,1,1,1,1,1,1,1,1,1,0,1}

算余弦相似值

sim=0.988623 不去除多余省市区影响不大，可能空格标点影响也不大

2有经纬度

进入距离识别找兄弟

1在表2中找到兄弟地址，获得兄弟地址的Group ID

现在的方案

通过Group ID在表三中进行查找

输出与表3相同Group ID的数据

、 改良方案

在表2相同group ID的兄弟地址们中，运算支付识别算法

算出来与兄弟地址们每一个的相似度值，然后进行排序，选取最大的一个数据输出

2 在表2中找不到兄弟地址，获得新的Group ID

最为新地址加入到表2表3中生出新表2新表3

输出 未找到匹配地址，将其作为新地址保留