04/19

大数据大作业

实验内容：健康码红码生成模拟

实验平台：可以用学校大数据实验平台，也可以各小组自行搭建

实验方式：按小组完成。现在下发实验要求和样例数据，各组开始准备，要求在大数据平台（具体哪些不限）上完成，如果实在无法在大数据平台上完成，也可以java直接编程，但是排名靠后。实际数据会在5月份下发。

计时说明：数据准备不算时间，可以认为在拿到感染人文件之前，做尽可能多的准备，最后模拟从拿到感染人文件就开始计时，找出所有密接人后停止计时，这个过程要全程对计算机录像，尽量在屏幕显示多的信息且录制下来，该录像在报告的时候要播放，作为认定结果的重要依据。

实验验收和评分：第14周二前各组在对分易提交结果数据文件和方案说明PPT（报告分享会用，里面包括完成时长，时长录像不需要提交到对分易，但是要带到答辩现场播放）。按正确率/完成时长的值大小排名，越大越靠前，但结果需要在课上各小组抽1人做报告说明，以认可结果，如果说明不清楚，小组可以换抽人补充，但每换1人，排名延后1名。小组按排名折算成绩，全组人相同，占90分。另外，实验后，每人提交一份实验报告，主要说明自己在实验中做的事情和心得，按个人算，占10分。

数据集1：地点

数据集2：手机号

新冠抗疫离不开大数据的支持，在发现感染病人后，可以通过其手机的漫游信息，发现和其行程可能有交集（都在同一个基站注册过）的人员，健康码标注为红色，促使密接人员做核酸筛查。

cdinfo.txt文件为基站下载汇总的人员漫游信息，其中数据如下所示：

55001525849,20220507045223,1,13771874017

55001525849,20220507060705,1,14987849191

55001525849,20220507062150,1,18953076900

55001525849,20220507083519,2,17950088996

......

其中4列信息分别为 cellid(就是基站的编号)，时间（到秒），注册状态（1表示注册入基站，2表示离开基站），手机号码。

另外，会下发一个infected.txt文件，其中数据如下所示：

17718961269

19281886829

14384778402

18515965265

......

一列数据为发现被感染人员的手机号码，这个文件数据不会很多，可能就几个号码。

现在贴在文件中的是一个小样本数据，两个文件都有，供大家调测用。由于5G基站数目庞大，基站切换非常频繁，这种数据量是非常大的，正式数据为20G左右，下发方式为给U盘拷贝和在实验平台包中提供。通用实验平台开放的时间另外通知。

实现要求：从下发数据中找到需要标红码的人员列表，即与感染人员同时间在一个基站的手机列表，用于精确标红码，按号码升序导出到redmark+组号.txt文件中，文件格式和infected.txt相同。



1. 读取两个文件的内容，将其转换为(key, value)格式的RDD，其中key是基站id，value是(手机号,时间戳,状态)的元组。这一步的时间复杂度是O(n)，其中n是cdinfo.txt文件中的记录条数。
2. 对RDD按照key进行分组，得到每个基站的所有记录。再对每个基站的记录按照时间戳进行排序，得到有序的列表。这一步的时间复杂度是O(nlogn)，其中n是cdinfo.txt文件中的记录条数。
3. 读组中。这一步的时间复杂度是O(m)，其中m是infected.txt文件中的记录条数。
4. 对持久化的RDD进行映射操作，对每个基站的记录进行遍历，找出与感染人员有过重叠停留时间的其他手机号码，并将其放入一个列表中。这一步的时间复杂度是O(nm)，其中n是cdinfo.txt文件中的记录条数，m是infected.txt文件中的记录条数。
5. 对新的RDD进行扁平化操作，将列表中的元素展开成单个元素。然后对新的RDD和感染人员手机号码的RDD进行合并操作，并去除重复元素。这一步的时间复杂度是O(n+m)，其中n是cdinfo.txt文件中的记录条数，m是infected.txt文件中的记录条数。
6. 将最终结果保存到一个文件中。这一步的时间复杂度是O(n+m)，其中n是cdinfo.txt文件中的记录条数，m是infected.txt文件中的记录条数。

综上所述，整个代码的时间复杂度是O(nlogn+nm+n+m)，其中n是cdinfo.txt文件中的记录条数，m是infected.txt文件中的记录条数。