05/07

大数据大作业及实验4

实验内容：健康码红码生成模拟

实验平台：大数据实验平台，自选具体用的系统

实验方式：分小组完成（每组4人，1-2班，3-4班内可以混编，最多有3组5人），现在下发实验要求和测试数据，各组开始准备，实现方式不做限制，最后在各小组的平台上能够完成功能就可以

实验验收：到实验那天，统一下发数据后，各组完成在微信群提交结果，按结果正确率及提交次序评分，正确的情况下越快越好（按小组算，全组一致）。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）

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，可能就几个到几十个号码。

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，届时会根据机器情况，下发类似于中等城市一天的模拟数据。下发方式为开放下载源，大家可以用wget获取。实验开始前大家可以事先将自己小组准备的东西放在平台上，平台原则上一直开放，具体另外通知。

基本要求：从下发数据中找到需要标红码的手机列表，按号码升序导出到redmark+组号.txt文件中（其中组号为A或B（A：1-2班，B：3-4班）xx（两位数字组号，必须2位，如1组，就是01），文件格式和infected.txt相同，上传到微信大班群，以微信中提交文件的时间为准。如果你小组用大数据平台完成实在有困难，可以在平台上直接编程完成也可以，或者哪怕手工找出一部分，只要在微信提交了，就有成绩。

提升要求：各组必须完成基本要求，本次大作业才可以得分，但如果有感兴趣的组可以准备提升要求，就是要更精确防控，只找与感染人员同时间在一个基站的人员列表，精确标红码，输出文件名superredmark+组号.txt，也是上传到微信群，按正确率和上传时间排名，完成提升要求的组可以获得奖励分。

