04/30

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（快速实训中的大数据综合实训平台），也可以各小组自行搭建。不能用其他云平台。

实验方式：按小组完成。现在下发实验要求、样例数据和一个网盘大样例数据，各组开始准备，要求必须在大数据平台（具体用什么不限）上完成，实际数据在小组准备好之后单独下发（第十三周开始）。

计时说明：第十三周开始，做好准备的组长加老师微信，老师给百度网盘链接，就算开始计时，小组完成，组长将结果文件在微信里面发给老师，就停止计时。

其他要求：小组注意文件保密，不要给别的组。小组从拿到文件开始处理，要对关键步骤截屏，比如下载、下载完成、导入、导入完成、处理开始、处理中的关键节点、处理完成、导出结果等，如多台机器共同工作，各台机器也要有一些截图，截图要带时间，放到汇报胶片中。

实验评分：第13周日晚12点前各组在对分易提交方案说明PPT（报告分享会用，里面包括完成时长（组长和我微信确认）和方案及执行过程说明（贴过程截图））。按正确率（（找出正确条数/正确条数）\*（找出正确条数/找出总条数））/完成时长的值大小排名，越大越靠前，但结果需要在14周周一和周二晚上的实验课上各小组抽1人做报告说明，以认可确实是所选取的大数据平台完成，如果说明不清楚（大家也可以提问，但以老师的判断为准），小组可以再抽人补充，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（相当于实验八报告），主要说明自己在实验中做的事情和心得，按个人算，占10分。

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，压缩后接近4G，百度网盘链接提供，目前先提供一个网盘样例数据，和实际数据大小类似：

链接：https://pan.baidu.com/s/1CvWeZbGid3QWd\_bchjDNQw?pwd=8ik3

提取码：8ik3

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（比如一班1组为 11组，四班5组为45组）.txt文件中，文件格式和infected.txt相同。

