1.利用用户画像，给用户打标签，利用相同标签找到目标人群

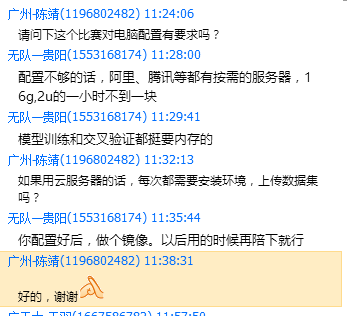
实例：美的豆浆机通过Youmi DSP进行了Look-alike人群扩展投放

有米广告取得美的家电第一方消费者数据，涵盖浏览、购买行为等ID信息。通过导入Youmi DMP进行全库记录匹配，找到个体的在线历史大数据。经由人群分析模型，有米洞察到美的用户的个性倾向特征，通过标签算法挖掘，将数据库中拥有高相似画像的人群列为一类精准用户。根据标签模型，得出这些用户具有较多且重合的“健康”“时尚”“亲子”“女性”“中高收入”“一二线城市”等细分人群画像。

分析：利用用户画像给用户打上各类标签。根据种子人群分析大部分种子用户具有的标签特征 例如：家庭女性、30-40岁、已婚，未生小孩，健康。那么对于一个标签为：上班族，30-40岁、已婚，未生小孩，健康 女性就是其目标人群。

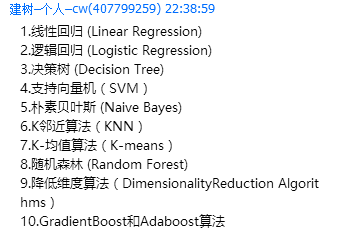
2.利用分类算法来实现的：种子人群为正样本，候选对象为负样本，训练分类模型，然后用模型对所有候选对象进行筛选。

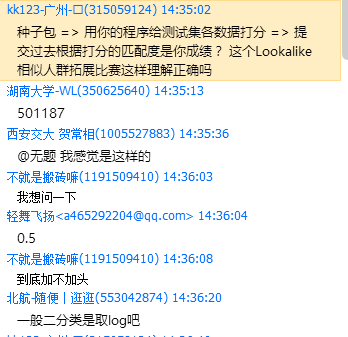
显然候选样本并发所有的样本都是负样本，所有这是一个典型的PU learning问题



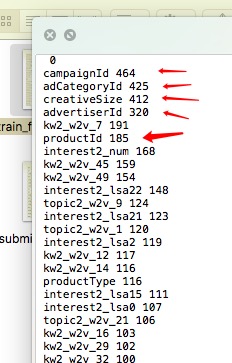
一个解决内存的问题，流式读取？

算法list（可以以后用）：

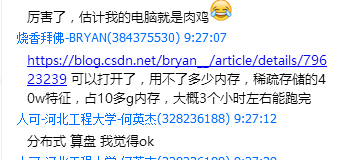


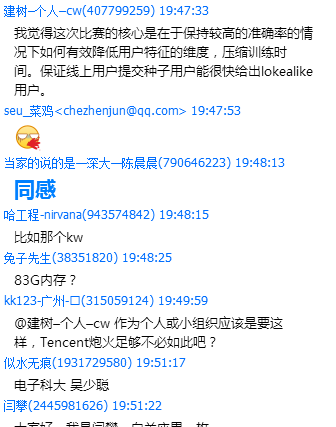


数据量大先做特征筛选

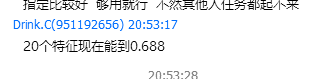


相关性计算，余弦距离？

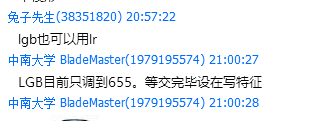




进程池： pool [https://morvanzhou.github.io/tutorials/python-basic/multiprocessing/5-pool/](qq://txfile/)

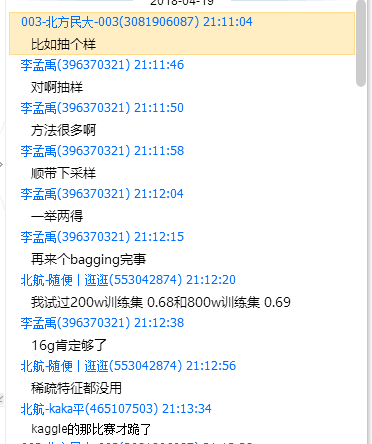




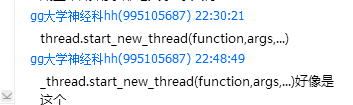




数据大可以降维：



Python多线程·：



内存不够：流式，特征降维（PCA好像就是，w2v？word2vec？好像也能），进程池。

独热编码怕是不行，太吃内存了

还有特征可能需要转格式，手册里有个github，群里疑似也有办法，但我没看到

以上群记录看到了20号……后面的再看吧，看不动了……