Thinking1 什么是近似最近邻查找，常用的方法有哪些

答： "ANN，Approximate Nearest Neighbor，近似最近邻检索，在牺牲度接受度范围的情况下提高检索效率。近似最近邻检索是线性复杂度，处理大规模数据可以使用。

其中LSH，Locality-Sensitive Hashing，局部敏感哈希是ANN的一种。

其中我们使用minhash降维来解决数据特征维度太高。再使用LSH降维来解决用户候选集太多。

Minhash：找到feature vector里第一个为1的行号，得到signature vector，多次随机置换行，组成signature matrix，做到了降维。使用signature matrix的jaccard来近似原始matrix的jaccard。（实际使用可以通过打擂法）。

LSH：将一个文档vector切分成多个band，其中再同一个band中，文档的段相同的话，就被分到同一个桶，相同段对应的文档相似度就会很高。

如果你要查到一个文档相似的文档，你可以把这个文档同样安装上面的步骤，找到这个文档桶里的其他文档就是相似的文档。"

Thinking2 为什么两个集合的minhash值相同的概率等于这两个集合的Jaccard相似度

答： "minhash的定义是找到特征向量中第一个为一的行号。

两个向量MinHash值相同的概率的算法是：值都为1的行/值不相等的行。

Jaccard相似度：向量元素的交集/向量元素的并集。

上面两个的定义所表示的数学公式相同。所以他们相等。"

Thinking3 SimHash在计算文档相似度的作用是怎样的？

答： "SimHash计算过程：

设置simhash的位数。

初始化simhash，位数都为0。

提取文本中的特征。

计算各个word的hash code。

Hashword中位为1的，则simhash乘以他的权重，否则就为0。

将他们相加计算最后得到的simhash，如果大于1，则设为1，否则为0。

Simhash相似度：

假设我们设hamming距离小于等于3的文档为相似文档。

将候选集文档切分为k+1=4份。（抽屉原理，四个中只要有一个相等，那么他的hamming距离就不会大于3

那么一个文档同样和切分为k+1=4份，与候选集做对比，一旦发现有相同的段，拿出该段对应的候选集，将所有的候选集拿出后再做精细对比。"

Thinking4 为什么YouTube采用期望观看时间作为评估指标

答： 因为采用期望观看时间是因为一个人看视频的时间越长，商业化的收益越高，因为中间可以插入的广告越多。

Thinking5 为什么YouTube在排序阶段没有采用经典的LR（逻辑回归）当作输出层，而是采用了Weighted Logistic Regression？

答： 如果没有权重，预测的结果是0,1。对观看时长赋予权重，这样可以得到观看时长的期望。我们想要预测的不是用户是否点击视频，而是他观看的时长。