**Thinking1:什么是近似最近邻查找，常用的方法有哪些?**

答: 近似最近邻查找,就是根据数据的相似性,在牺牲可接受范围内的精度的情况下提高检

索效率.

常用的方法有LSH,MinHash,SimHash等等…

**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: 同一行的两个元素的情况有三种:

X.两者都为1 Y.一个1一个0 Z.两者都为0

易知Jaccard相似度为|X|/(|X|+|Y|).另一方面,若排列是等概率的,则第一个出现X

中元素出现在Y中元素之前的概率也为|X|/(|X|+|Y|),而只有这种情况下两集合的

minhash值想同.

**Thinking3:****SimHash在计算文档相似度的作用是怎样的？**

答: SimHash更适用与长文本中的检索,通过simhash计算两个文档的汉明距离相似度,进行

搜索文档的快速去重.

**Simhash计算过程:**

1.设置SimHash的位数，比如32位，需要综合考虑存储成本以及数据集的大小

2.初始化SimHash，将各位初始化为0

3.提取文本中的特征，比如采用2-Shingles

4.使用传统的hash函数计算各个word的hashcode

5.对各word的hashcode的每一位，如果该位为1，则simhash相应位的值加它的权重

（通常是出现的频率）；否则减它的权重

6.计算最后得到的32位的SimHash，如果该位大于1，则设为1；否则设为0

**如何通过文档的SimHash计算文档之间的相似度**

1.基于传统的IR方法，将文章转换为一组加权的特征值构成的向量

2.通过SimHash算法得到每篇文档的指纹（fingerprint）

3.计算不同文档之间的Haming距离

**Thinking4:为什么YouTube采用期望观看时间作为评估指标?**

答: 因为普通的CTR指标对于对于视频搜索具有一定的欺骗性,很多时候如果只是点击,就

可当作评估指标,会对最终想要的结果出现一定的偏差.

**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: 首先,由于YouTube的推荐场景是以观看时长为优化目标而设定的.

普通的逻辑回归只能够预测一个用户是否观看了视频,也就是CTR的预测,而增加了观看时长作为权重之后, Weighted Logistic Regression变成了预测期望观看时长.

**Action1: 使用MinHashLSHForest对微博新闻句子进行检索 weibo.txt  
针对某句话进行Query，查找Top-3相似的句子**

答:



**Action2: 请设计一个基于DNN模型的推荐系统**

**阐述两阶段的架构（召回、排序）**

**以及每个阶段的DNN模型设计：**

**1) DNN输入层（如何进行特征选择）**

**2) DNN隐藏层结构**

**3) DNN输出层**

答: 召回阶段:

1. DNN输入层:选取用户的基本信息,用户行为信息(浏览记录,偏好等等),结合时间维度的偏好特征,作为基础特征,然后对特征进行embedding处理,基本采用word2vec的方式,对embedding的向量进行数据标准化处理
2. 选取合适的激活函数,接入DNN输入层的数据作为DNN隐藏层,3层塔型结构
3. 对DNN的输出结果,作为粗排序的阶段的样本

排序阶段:( DeepFM)

1. 选择要推荐商品的的信息和用户与商品之间的交互信息作为特征,加入到DNN输入

层,embedding选择与召回阶段相似的方式,并对数据进行标准化处理

1. 使用DeepFM模型,对推荐结果进行预测,加入召回阶段的结果,然后选择Top-N推荐