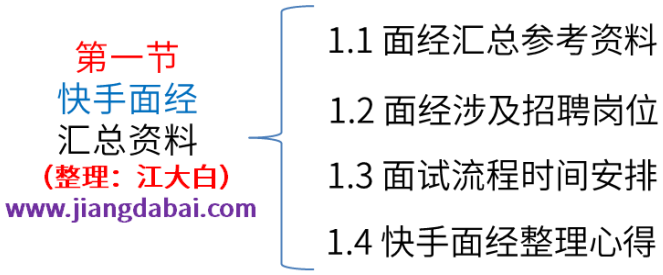
14|快手算法岗武功秘籍

1 快手面经汇总资料



1.1 面经汇总参考资料

① 参考资料：

（1）牛客网：快手面经-72 篇，网页链接

（2）知乎面经：点击进入查看

（3）面试圈：点击进入查看

② 面经框架及参考答案：

（1）面经框架及参考答案：点击进入查看

（2）大厂目录及整理心得：点击进入查看

1.2 面经涉及招聘岗位

（1） 实习岗位类

【快手杭州商业化算法实习生】、【算法测试实习生】、【用户增长算法实习生】

（2） 全职岗位类

【计算机视觉工程师】、【机器学习算法工程师】、【广告算法工程师】、【openday 社区科学部算法工程师】、【推荐算法工程师】、【图像增强算法工程师】、【y-tech 部门工程师】、【数据挖掘算法工程师】、【视觉识别算法工程师】、【商务化业务部广告算法工程师】、【语

音算法工程师】、【快手游戏 AI 算法工程师】、【NLP 算法工程师】

1.3 面试流程时间安排

快手面试流程-整理：江大白			
	面试类型	面试流程	备注（侧重点）
第一面	技术面	自我介绍+项目/实习经验 +技术问答+算法编程	/
第二面	技术面	自我介绍+项目/实习经验 +技术问答+算法编程	有的考官会深挖项目 细节及创新点
第三面	技术Leader面	自我介绍+项目经验+公司发展	/
第四面	HR面	基础人力问题	/

PS：以上流程为大白总结归纳所得，以供参考。

1.4 快手面经面试心得汇总

★ 对简历上自己实习或者做过的项目的细节，要知道清楚，会针对项目问得很深

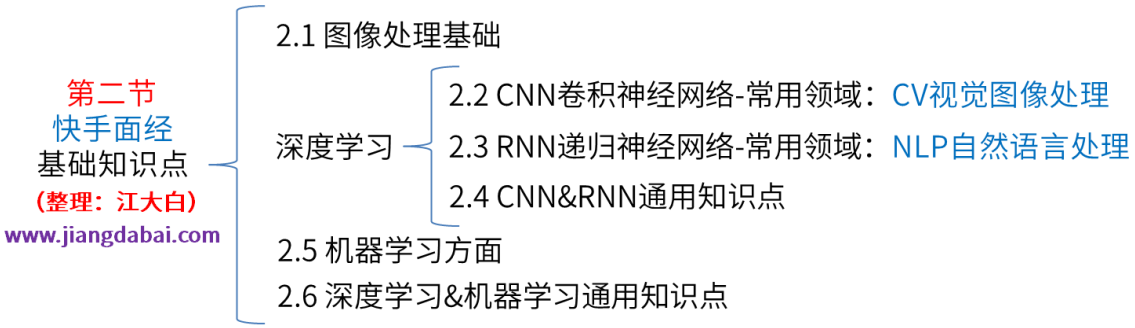
★ 二面和面试官交流了很多，深刻认识到数学在算法方面的重要性，“数学能力决定上限，代码能力决定下限”。

★ 感觉快手都没咋问基础知识，基本就是项目经历+算法，感觉比头条简单一些。

★ 还是得多刷算法题，常见题目必须得秒写，二叉树，链表，二分查找，这是目前这么多面试出现频率最高的题目，可能是因为实习生的要求不高，问的都比较基础，估计校招的时候就会问一些 DP 相关的题目了。

★ 有人说的 HR 面的问题，基本都是一模一样，先介绍部门然后问我什么城市选择、职业规划、三个词形容自己等。

2 快手面经涉及基础知识点



2.1 图像处理基础

- 手写直方图拉伸？
- 手写均值滤波？

2.2 深度学习：CNN 卷积神经网络方面

2.2.1 讲解相关原理

2.2.1.1 卷积方面

- 什么是 dropout？让你实现一个 dropout 你怎么实现？
- dropout 原理讲一下？作用以及训练测试时的不同？

2.2.1.2 池化方面

- 池化层的作用？池化层反向传播的时候是怎么计算的？

2.2.1.3 网络结构方面

- 介绍常用的网络结构，如何改进？

2.2.1.4 其他方面

- BN 层解释一下？有啥用为啥好用？
- BN 的原理、作用以及训练测试时的不同？

- BN 为什么能加快收敛速度?BN 解决了什么问题？是怎么计算的？基于什么计算的？训练和测试的时候有什么不同？测试时候的均值和方差是怎么来的?
- 为什么输入网络之前数据要做归一化?
- 梯度爆炸的原因及解决方法
- BN 的 gama labada 意义

2.2.2 数学计算

- 卷积时间复杂度?

2.2.3 激活函数类

- 介绍 softmax，对 softmax 的理解
- 证明 softmax 的特征规划问题，在撕代码的界面推

2.3 深度学习：RNN 递归神经网络方面

2.3.1 讲解相关原理

- LSTM 各类门结构?
- LSTM 的门是什么，激活函数是什么?
- RNN 的反向传播 + bp 推导

2.4 深度学习：CNN&RNN 通用的问题

2.4.1 基础知识点

- 训练集测试集和验证集怎么划分?
- 训练数据太少怎么办?
- 模型 attention 怎么做，结果效果?

2.4.2 模型评价

- AUC 啥意思？横纵坐标什么？物理含义什么数据分布改变时，AUC 有什么影响（没啥影响，我居然说我不太清楚？）
- AUC 的定义和计算方法、auc 实际上衡量的是什么能力，roc 曲线是否会出现先增后减的情况?

2.5 传统机器学习方面

2.5.1 讲解相关原理

2.5.1.1 数据准备

- 介绍一下自己在数据清洗这一块有什么心得方法?

2.5.1.2 特征工程

无

2.5.1.3 有监督学习-分类和回归方面

① 分类回归树（集成学习）

- gbdt 和 rf 的区别?
- 讲一下 rf 和 lightgbm 区别?

A.基于 bagging：随机森林

- 分析一下随机森林可不可以剪枝?
- 随机森林是什么原理?

B.基于 boosting：Adaboost、GDBT、XGBoost

- 介绍一下 LightGBM 与 Xgboost 的区别？Lightgbm 相对于 xgboost 的优化?
- 介绍一下 GRU、RF、GBDT、XGBoost
- 如果判断树模型 lgb 过拟合，怎么调整树的参数?

- XGBoost 的特征重要性是如何得到的?
- GBDT+LR

② K 近邻（KNN）

- KKT 条件具体是什么?

③ 逻辑回归 LR

- 问了标准化，那些模型需要对数据进行标准化，标准化有什么用，LR 中是否一定需要标准化 ？

④ SVM（支持向量机）

- 通俗的讲一下 SVM?
- 讲 SVM 原理，SVM 损失函数是什么?
- SVM 为什么能够求解对偶问题，求解对偶问题为什么和原问题一样？为什么要求解对偶问题？svm 的公式是什么？如果线性不可分怎么办?
- 问了支持向量机的基础思路，支持向量机的损失函数，了解支持向量机的核函数吗？核函数需要满足什么样的条件？核函数怎么进行构造?
- SVM 解决线性问题还是非线性问题?

⑤ 朴素贝叶斯（Naive Bayes）

- 贝叶斯的原理是什么?
- 朴素贝叶斯公式是怎么得到的，推导一下?
- 贝叶斯公式
- 先验概率和后验如何用到朴树贝叶斯上的?

⑥ 决策树（DT）

- 你了解决策树么？讲一下?
- 决策树是怎么样进行划分的，决策树的损失函数?
- DNN 和树模型，哪个用于输入是类别特征时效果更好，为什么?
- 树模型如何输出特征重要性得分？物理含义是什么?

2.5.1.4 无监督学习-聚类方面

- 介绍一下 kmeans 算法，然后问在分布式的环境下怎么实现？
- 说说 k-means 算法和高斯混合模型的关系？
- kmeans，K 值选择，初始点？
- k-means 的数据应该符合什么条件
- 说说高斯混合模型以及参数估计过程？

2.5.2 手推算法及代码

- 写一下逻辑回归的损失函数，并且推导一下权重更新公式？
- 推导 LR、LR 损失函数是什么，数学含义是什么？
- 写线性回归的损失函数，并推导权重更新公式。
- 手推 XGBoost

2.6 深度学习&机器学习面经通用知识点

2.6.1 损失函数方面

- 神经网络有哪些常用的损失函数？

2.6.2 激活函数方面

- RELU 有什么缺点？
- 推导 softmax 的梯度，和 tanh 的梯度？
- 反向传播梯度推导？（relu、sigmoid）

2.6.3 网络优化梯度下降方面

- 如果不用梯度下降优化，还能怎么优化，我说了一些优化算法，他说牛顿法怎么具体更新参数的。
- 说说最优化方法有哪些，以及具体解释一下 adam 解决了什么问题（从动量以及学

习因子自适应的角度解释了一下）

- 具体解释 adam，二阶矩是什么？为什么要用二阶矩？
- 剃度优化 bfgs 的原理
- 写 sgd 和 momentum 优化的方法？
- lda 调参中的参数是什么，α，β 怎么设置？

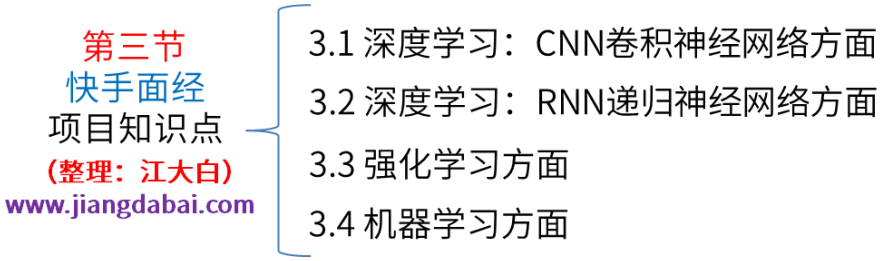
2.6.4 正则化方面

- 正则化有哪些方法？详细介绍下 L1 和 L2？
- 正则化与 bias，variance 的关系？
- 说说正则化技术，为什么 L1 正则化可以保持参数的稀疏性？（我回答了两个角度，一个是画约束图，另一个是 L1 正则化本质上是假设样本的先验分布服从拉普拉斯分布，画出拉普拉斯分布曲线解释），然后又推了一下为什么 L1 正则化本质上是假设样本服从拉普拉斯分布。
- 从两种数学角度解释 L1、L2（一个是画图 一个是数值角度）

2.6.5 过拟合&欠拟合方面

- 过拟合和欠拟合的原因及解决方式？

3 快手面经涉及项目知识点



3.1 深度学习：CNN 卷积神经网络方面

3.1.1 目标检测方面

- RCNN、Fast-RCNN、Faster-RCNN 讲一下？
- 对 anchor 的理解？
- 手写 NMS 代码

3.1.2 图像分类

- 介绍图像分类用到的 Loss
- 一般做分类我们都用交叉熵损失，交叉熵损失是否可以替换为 MSE 损失呢？

3.1.3 人脸识别

- center loss 作用
- 小论文中改进了 center loss，介绍 center loss
- 开放问：为什么 center loss 没有得到广泛的应用？
- 问到在处理上万的类别的分类任务上，centerloss 能否使用，arcface 有什么问题？
- 人脸 arcface 中的 arc 为啥比 cosface 那些好，为啥比 triplet 好？

3.1.4 音频算法

- 音频部分基础知识：

（1）MFCC、FBANK 提取过程、区别。

（2）帧移，为什么选择这个范围

采样率、带宽、 奈奎斯特定理等

- 问几个现实场景中的问题，说几个解决方案或思路。

（1）说话人识别注册语音少怎么解决：数据扩充、自适应等。

（2）应对噪声场景：数据扩充、降噪、GAN 网络、说话人分离、DNN VAD 等。

3.2 深度学习：RNN 递归神经网络方面

3.2.1 自然语言处理 NLP

① Bert

- Bert 的 Position Embedding 是怎么处理的?
- Bert 了解么，position embedding 是什么，哪种效果好?

② Attention

- attention 计算时为什么要除根 dk?

③ Word2vec

- 讲一下 word2vec 原理?
- word2vec 具体的细节，损失函数?

④ 其他

- tfidf 公式写一下
- 评估两个句子的相似度，有什么方法

3.3 强化学习

- 生成式模型与判别式模型的区别?
- 说一说基于值函数和基于策略梯度的 RL 算法的区别，什么时候用值函数 RL 算法合适，什么时候用基于策略梯度的合适?
- 说说 onpolicy 和 offpolicy 的区别，并分别举几个代表性算法?
- 写一写 q 函数的原始定义，并说说 q 函数的意义?
- 你对 GAN 有了解是吧，写写 GAN 的损失函数吧，并解释一下 G 和 D 的训练过程。
- 写一下 RL 中 Q 函数的表达式，并说说 Q 函数的意义。
- 你在项目中用了 DDPG，说说 DDPG 算法吧，（我说的时候可能表述的不够清楚，面试官让我画出几个网络之间的输入输出关系）。面试官接着问，这是 offpolicy 算法吧，

你先解释下 offpolicy，并说一下你在项目中怎么实现的。（我回答的核心是在 π 策略上加了噪声，形成采集样本策略）面试官说加的噪声少了会怎么样，加的大了会怎么样，有什么影响）

- 你对 DQN 了解吗，对 DQN 的改进算法了解多少？说说它对 Q-learning 的改进在哪些方面?
- 项目是关于 DQN 和 DDQN 的，后来问了 DDQN 能从根本上解决 DQN 的高估现象吗?

3.4 机器学习方面

3.4.1 推荐系统

3.4.1.1 讲解原理

- 讲一下 DeepFM 的原理?
- 协同过滤，userCF、itemCF 讲一下?
- 推荐方法用了哪些？冷启动怎么实现
- CTR 损失函数及意义?CTR 评价指标?
- youtubu 召回模型
- deepfm 模型
- 推荐系统的各个环节（召回、精排、重排）都是干啥的，你有什么理解?
- 如果给你一个 DNN 或逻辑回归模型，怎么输出特征的重要性得分?
- 广告算法岗：

（1）了解计算广告的常见收费方式吗？ eCPM 是什么含义?

（2）了解竞价机制吗？解释一下广义第二竞价机制?

（3）为什么要用广义第二竞价机制呢？（优点是什么）

（4）了解哪些常见的广告形式?

- 聊到推荐系统，我说了 FM 和 FFM，面试官问我 FFM 的第一个 F 代表什么，我回

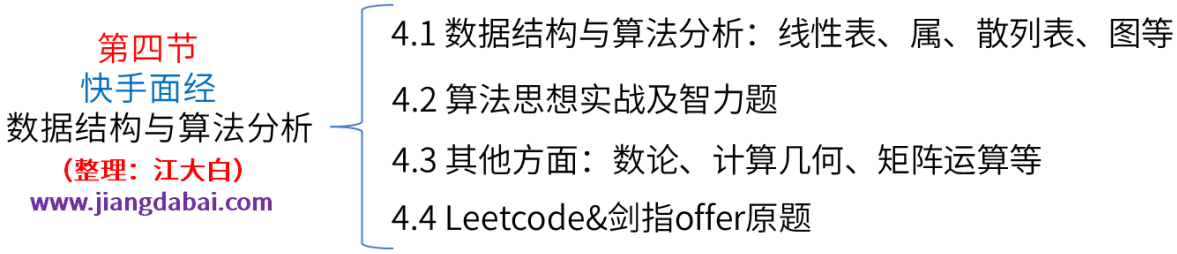
答说 Field。面试官让我解释一下为什么要用 FFM?

- 讲一下 FM、FFM、DeepFM 的流程?

3.4.1.2 手写代码

- fm 的公式推导，怎么写成矩阵形式
- 手推 FM/FFM 公式

4 数据结构与算法分析相关知识点



4.1 数据结构与算法分析

4.1.1 线性表

4.1.1.1 数组

- 给定一个数组 [a,b,b,c,c,c,b,a,b,c]，按照字母序排序 ‘a’,’b’,’c’，变成 [a,a,b,b,b,b,c,c,c,c] ？
- 循环数组找最小值
- 旋转数组的查找
- 合并有序数组（C++）
- 小和问题：在一个数组中，每一个数左边比当前数小的数累加起来，叫做这个数组的小和。求一个数组的小和。

例子：

[1,3,4,2,5]

1 左边比 1 小的数，没有；

3 左边比 3 小的数，1；

4 左边比 4 小的数，1、3；

2 左边比 2 小的数，1；

5 左边比 5 小的数，1、3、4、2；

所以小和为 1+1+3+1+1+3+4+2=16

- AABBCDDEE 给一个数组，除一个数组外均为重复出现，要求找出单独的字母，时间复杂度尽可能低?
- 以时间复杂度 O(n)从长度为 n 的数组中找出同时满足下面两个条件的所有元素：

（1）该元素比放在它前面的所有元素都大；

（2）该元素比放在它后面的所有元素都小

- 给一个长度为 N 的数组 里面有 1-N+1 这些 unique 的数字，其中少了一个 怎么找出少的这个，不允许用额外 space

4.1.1.2 链表

- 链表反转
- 判断一个链表是否有环，有环的话输出环的第一个节点，没有的话输出空。
- 写一个合并 K 个排序链表的代码?
- 合并两个排序链表并删除重复值?
- 做了将 K 个链表排成一个排序好的单链表

4.1.1.3 字符串

- 删除字符串中多余空格，连续空格都变成一个，要求时间空间复杂度尽可能小
- 找出字符串中最长的不含重复字符的子串长度，用双指针轻松解决
- 字符串全排列

4.1.2 树

4.1.2.1 二叉树

- 判断一个二叉树是否为二叉搜索树。优化空间
- 一颗二叉树，找到找到叶子节点的值等于给定值的那条路径?
- 非递归实现二叉树翻转、分割子树求乘积 max?
- 求二叉树从根节点到叶子节点路径等和等于 target 的路径
- 二叉树的友视图
- 翻转二叉树
- 层次构建二叉树，然后先序遍历输出
- 层序遍历二叉树
- 二叉树的蛇形遍历
- 二叉树 Z 形遍历
- 二叉树非递归前序遍历
- 求树中两个节点的最近公共父节点，时间复杂度
- 给一个二叉搜索树，和一个区间，删掉不在区间内的节点?

4.1.2.2 堆

- 最大堆实现
- 堆和栈的区别，应用

4.1.3 排序

- 归并排序(只写了 O(n)空间的)，O(1)空间能做?
- 归并排序的平均复杂度是多少？最坏复杂度是多少?
- 最大的 k 个数
- 求 top(k) return sorted(A)[::-1][:k]?

- 链表快排、为什么数组排序都用快排不用归并？
- 手写快排
- 快排的平均时间复杂度，最坏情况是什么，复杂度多少？

4.1.4 搜索

- 快排 dfs

4.2 算法思想实战及智力题

- 从文件中读取 n 个数，求最大的 K 个数（约为 K 个即可）写入新文件；要求：n,k 都很大，无法直接装入内存，空间复杂度 O(1)，尽量减少 IO 次数；思路：利用概率密度
- 给几种硬币，凑一个数，求最少硬币数
- 最长递增子序列
- dfs 全排列
- 有重复的数组，两个数相加=n，求两个下标，列出所有可能情况
- 连续最长子序列
- 二分查找：有序递增不重复序列，找出第一个缺失的元素，加强版：如果可以重复呢？
- 最大值和最小值的差小于或者等于 num 的子数组数量
- 实现编辑距离，O(n)空间复杂度

4.3 其他方面

4.3.1 数论

- 凸函数是什么，有什么良好的性质？极值是什么？
- 解释一下什么是凸函数？（我回答了 Hession 矩阵半正定就行，）面试官接着问如果函数不可导怎么判断？

- 最小二乘法原理?
- 极大似然函数和极大后验函数是啥?
- 极大似然估计，条矩阵问题：AX=Y，A 列数大于行数，为什么会有无穷个解，有无穷个解的条件，如何得到一个最优解 ？
- 如果有 2 组独立同分布的数据，但模型分布不一样，怎么解决？EM 算法，混合高斯模型?
- K 个独立高斯同分布随机变量的结果是?
- ax=b，求 x 的方法。（求逆矩阵，如果不可逆怎么办。什么情况下可逆。）
- 求 a.b 独立且服从 0-1 均匀分布，求|a-b|的期望
- 两个独立同分布的骰子，求骰子数积的期望?
- 扔硬币，连续扔出 2 次的最大期望值，连续扔出两个正就停止
- 求两个数的汉明距离?
- 求 y = sqrt(x)

4.3.2 概率分析

- 打可乐,机器坏了每次打 0-1 杯,如果＜1/2 再打一次,最多打两次……最后＜2/3 就投诉,问最后投诉的概率?
- 写一个随机函数发生器，随机产生（1,2,3,4）四个数，当采集了无穷多数以后，产生的数概率服从（0.1,0.2,0.3,0.4）分布?
- 54 张牌抽出四张，抽出为 1234 的概率
- 三个人斗地主，出现王炸的概率
- 一个孤岛重男轻女，直到生出男孩为止，男女出生率 1 比 1，初始比例也是 1 比 1，问最终的男女比例?
- 某疾病发病概率 1/1000，患者有 95%的概率检测出患病，健康者有 5%的概率被误诊，问若一个人被检测出患病，实际患病概率是多少?

4.3.3 矩阵运算

- 说说矩阵分析里面特征值和特征向量的意义？
- 矩阵问题：AX=Y，A 列数大于行数，为什么会有无穷个解，有无穷个解的条件，如何得到一个最优解？
- Leetcode 困难难度:一个矩阵，1 能走，0 不能走，可以上下左右走，问最短路径，时间复杂度(nm),空间复杂度(nm)

4.3.4 其他

- K 个独立高斯同分布随机变量的结果是？
- 最大连续子序列和，并返回开始和结束的位置
- 两个均匀分布相加是什么分布

4.4 Leetcode&剑指 offer 原题

- Leetcode 23
- LeetCode 42：接雨水
- Leetcode 103
- Leetcode 113
- Leetcode 199
- Leetcode 206
- Leetcode 300
- Leetcode 322
- Leetcode 542
- Leetcode 原题：两数之和
- Leetcode 原题：用栈实现队列
- Leetcode 原题： 最长回文子串

5 编程高频问题：Python&C/C++方面

第五节
快手面经
编程高频问题
（整理：江大白）
www.jiangdabai.com

5.1 Python方面：网络框架、基础知识、手写代码相关

5.2 C/C++ 方面：基础知识、手写代码相关

5.1 python 方面

5.1.1 网络框架方面

5.1.1.1 Pytorch 相关

- Pytorch 实现把一个 Tensor 中大于 0 的数字都置为 0

5.1.1.2 Tensorflow 相关

- tensorflow 原理，keras 和他的区别?

5.1.2 基础知识

5.1.2.1 线程相关

- 讲一下 python 的多线程和多进程

5.1.2.2 讲解应用

- Python 引入其他库，底层实现是怎么样的?
- Python 的 sort，如果两个次数相同，sort 会改变他们的先后顺序吗?
- 判断字符、取字符、输出二维数据的行列

5.2 C/C++方面

5.2.1 基础知识

5.2.1.1 内存相关

- 堆和栈存储
- 两个栈实现队列

5.2.1.2 讲解原理

- 多态（虚函数）
- 空类的大小
- 静态函数，静态变量
- 类里面的变量+static 有什么作用?
- C：指针占用内存大小、strlen 和 sizeof 的区别、堆栈的区别、double 型与 0 怎么比较
- C++：STL 容器、虚函数、类的一些知识
- 什么是深拷贝浅拷贝?

5.2.1.3 讲解应用

- vector 和 list 底层实现?

6 操作系统高频问题：数据库&线程&常用命令等

第六节
快手面经
操作系统高频问题
（整理：江大白）
www.jiangdabai.com

6.1 数据库方面：基础知识、手写代码相关

6.2 操作系统方面：TCP、线程&进程、常用命令相关

6.1 数据库方面

无

6.2 操作系统方面

6.2.1 常用命令

- 怎么设置环境变量: .bashrc，然后 source 一下?
- 统计当前文件夹下面.jpg 文件的数量：ls -l | grep '.*\.jpg' | wc -l
- 计算当前文件夹的大小：du -sh

6.2.2 其他问题

- 数据挖掘算法岗问到的问题：

（1）spark 执行机制

（2）spark sql 写两个表字段相同 id 和 ip，统计两个表中不一样的 ip 的出现次数排序

7 技术&产品&开放性问题

7.1 技术方面

- 介绍天池比赛（讲训练验证数据划分以及原因、残差特征、lgb 和 xgb 的选用、具体的模型融合方法、AUC、FM 和 LR 的区别、LR 的损失函数推导、交叉熵公式、AUC 高一定预测的转化率准确吗？）
- 介绍在腾讯实习的项目（讲 DeepFM、讲 word2vec 原理、word2vec 中的优化：哈夫曼树 and 负采样 and 频繁下采样、讲协同过滤、AUC、连续值分桶）
- 如何判断一张图片含有什么噪声?
- 场景题：诈骗项目，想选出 topk 的诈骗类（没有标签）

主要类别有电商诈骗，虚拟交易诈骗，婚恋诈骗，主播诈骗等

数据有：举报人 ID，被举报人 ID，私信数据，评论数据，浏览记录

- 样本不均衡的处理方法

权重调整，采样，对样本不均衡不敏感的指标，

面试官补充：对样本不均衡不敏感的模型，比如 SVM，XGB

然后讨论了一波为什么 XGB 对样本不均衡不敏感？

某个项目为什么效果不好，有何反思，未来如果再做你会怎样改进？

- 场景设计题：现在我们要给快手上的视频用半监督方法打标签，target 是印度风的视频，已知印度快手上传的视频中印度风 bgm 的比例远高于其他国家，现在给你视频的 bgm 和上传国家两个特征，设计一个方法给这些视频打上是否是印度风的标签。个人感觉这个有点像半监督召回问题。

- 如何理解数据挖掘/机器学习/数据科学之间的关系？

- 在实际的业务数据中，现在有一个文件不断保存业务过来的流式数据，如何从文件中等可能的取出想要的某个数据，回答了挺久但好像都不是他想要的回答。

- 你如何解决一段文本去找相关图片的任务？

答：通过文本摘要得到关键字，通过搜索引擎去搜索相关图片，然后利用图片描述，与相关文本做相似度计算，取其 TopN

7.2 产品方面

- 风控部门，如何识别那些经过加工方式上传的恶意视频