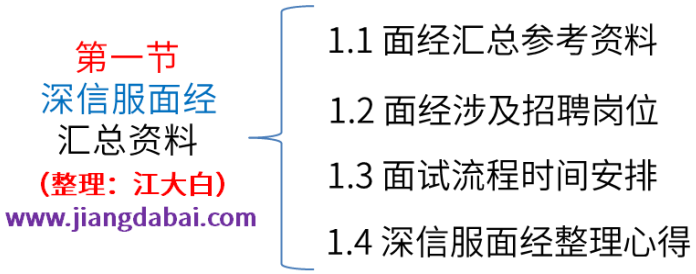
16|深信服算法岗武功秘籍

1 深信服面经汇总资料



1.1 面经汇总参考资料

① 参考资料：

（1）牛客网：深信服面经-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 面经参考答案：

（1）面经答案：点击进入查看

1.2 面经涉及招聘岗位

（1）全职岗位类

【机器学习工程师】、【NLP 算法工程师】、【大牛计划算法工程师】

1.3 面试流程时间安排

深信服面试流程-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	主要是项目细节及创新点 关注解决问题的思路
第二面	技术面	自我介绍+项目/实习经验 +技术问答+算法编程	主要是项目细节及基础知识
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。

其他注意点：

- 第三面技术 Leader 面有的时候会有

1.4 深信服面试心得汇总

★ 比较看重及觉问题的能力以及项目落地开发的能力

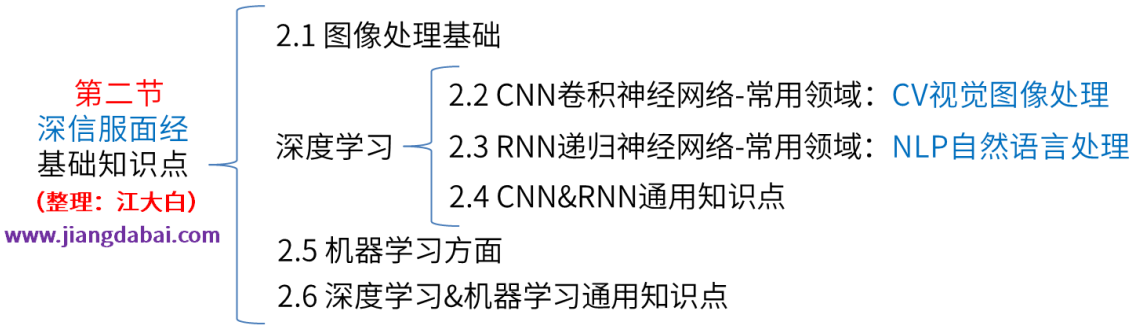
★ 深信服是我面了这么多公司中感觉面试最难的一家公司，真的很考验技术。

★ 很多都是 python 开发及前端、后端开发的，统计的主要还是统计机器学习相关的

★ 感觉机器学习问的多一些，深度学习的不多

★ 不同公司定义算法工程师不一样，深信服的是做什么，他说是做安全和云计算，所以更多的是会做一些安全工作，比如使用算法进行漏洞检测，攻击检测这些

2 深信服面经涉及基础知识点



2.1 图像处理基础

无

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

- 讲一下 CNN 的原理?
- 对于 dropout 的认识

2.2.1.2 网络结构方面

- 谈一谈 Resnet?

2.2.1.3 其他方面

- 深度学习模型的初始化要注意哪些问题？（权重随机初始化，避免网络退化，初始化范围要小，缩小样本空间，输入样本要 BN，防止梯度消失）
- 什么是 BN?
- 是否了解自动调参 auto ml?

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 的门结构有哪些？门的输入是什么，输出是什么，怎么在网络里面使用？

2.4 深度学习：CNN&RNN 通用的问题

2.4.1 基础知识点

- 注意力模型的原理？
- 如何解决样本不平衡问题？
- 有一个分类任务有几千个类，应该怎么去训练模型？
- Attention 机制讲一下？

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

无

2.5.1.2 特征工程

① 特征降维

- 特征如何降维？
- LDA 的实现方法、LDA 中的奇异值分解矩阵实现？

② 特征选择

- 在项目讲述过程中问了几个问题：（1）特征是什么样的？（2）怎么构建特征工程的？（3）为什么选择这个模型？
- 机器学习中有哪些算法需要进行归一化？

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）

- 介绍一下集成学习，以及你选择的方法原因?

A.基于 bagging：随机森林

- 随机森林原理，如何选择最优分裂点，列抽样为什么可以缓解过拟合?

B.基于 boosting：Adaboost、GDBT、XGBoost

- 为什么要模型融合，模型融合的几种方法，模型融合的优点?

（bagging,boosting,stacking,还有我自己项目里面的方法）

- 白样本多，黑样本少，选择哪个模型更适合一些，为什么？如何评价效果（抽样，训练多个分类器，可以采用 bagging，如果 LR 和 SVM 里面选择 SVM,LR 对样本过于敏感，SVM 只处理支持向量）
- xgboost 相对于 GBDT 的优点，为什么会快些、xgboost 如何做并行?
- 讲一下 xgboost（从提升树开始讲，讲了一通）？为什么要二阶展开？xgboost 采样的时候怎么采样的?

② 逻辑回归 LR

- LR 和 SVM 的区别？（处理点、计算方式、损失函数、自带正则等）
- LR 过拟合是什么样的情形，如果样本有很多重复的特征，对于 LR 训练效果有没有影响?
- LR 使用什么损失函数，为什么不用差平方，用差平方与交叉熵差别在哪?

③ 决策树（DT）

- 信息熵的定义?

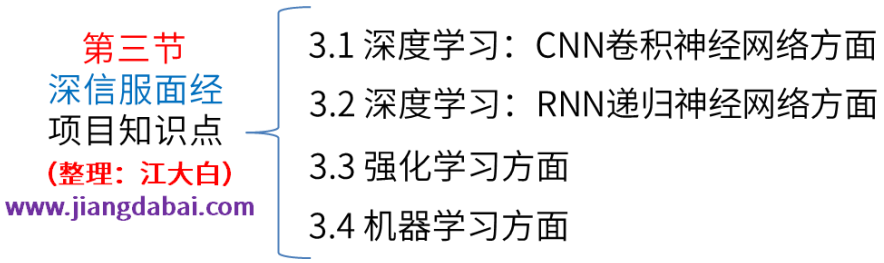
2.5.1.4 无监督学习-聚类方面

- 问了 kmeans 原理，优缺点，如何不自己设置 k 就能知道 k 取多少?

2.6 深度学习&机器学习面经通用知识点

- 训练的模型过拟合了，怎么办?
- 有一个分类任务有几千个类，应该怎么去训练模型？陷入局部最优值怎么办?

3 深信服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

- Focal loss F1 怎么解决样本不平衡的问题?

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

- 你觉得 BERT 有哪些可以改进的地方？我说您是说 BERT 的缺点吗?
- 文本分类中，传统机器学习 tfidf+lr/svm 和 Bert 的区别?

② Word2vec

- word2vec 原理?
- word2vec 和 onehot 的区别?

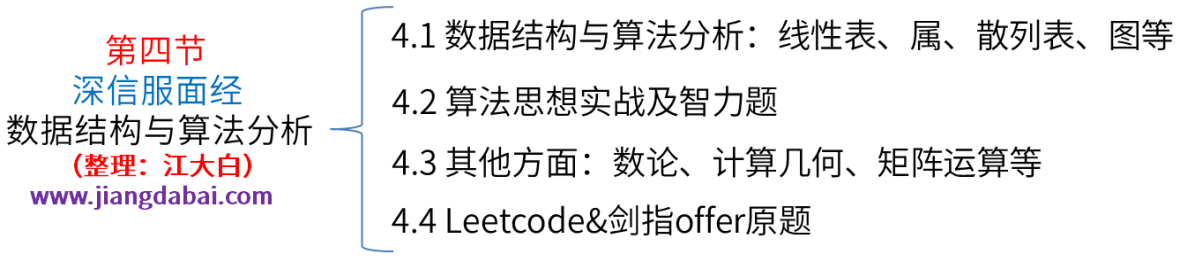
3.3 强化学习

无

3.4 机器学习方面

无

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 arr，tagert，求数组元素的组合之和=target 的所有可能。（不能重复）
- 数组中无重复的和为 target 的所有组合（回溯+去重）
- 删除数组中重复的数字，相对位置不变输出，且保证输出元素满足从小到大顺序的个数最多？
- 给定一个数组，求所有重复的数。用字典，问还有没有其他方法，提示用 hash。

4.1.1.2 字符串

- 有一篇英文文章，找前 10 个出现次数最多的字母？

我回答先遍历一遍论文，简历字母为 key 出现次数为 value 的哈希表！然后从 values 中找到前 10 个最大的元素！面试官又问，怎么找？我说用最小堆。

- 一个特别特别大的字符串，怎么统计给定字符集里面的字符在字符串中首次出现的位置，然后加快效率更好？我说 map-reduce，他说更好的呢，用数据结构实现？我说

hashtable，在对应位置上面存首次出现的位置就行。

- 英文文本，得到前 K 大频率的单词，如果文本很大怎么办？写伪代码

4.1.2 树

- 平衡二叉树与数组获取元素谁比较快？（如果给定下标，数组快，给定值，平衡二叉树快，原因时间复杂度）

4.1.3 排序

- 排序算法介绍，它们的复杂度？
- 常使用的排序算法，说一些你熟悉的并口头实现？（说的使快排，复杂度 O（nlogn））
- 说一说你知道的时间复杂度为 O(nlogn)的算法。以快排为例，具体解释一下为什么它的时间复杂度是 O(nlogn)呢？
- 类似快排的这些算法思想都属于什么算法？（分治算法）请你说一说这一类算法的时间复杂度有什么规律？快排是分成两组 1，是否可以分为更多组呢？为什么？

4.2 算法思想实战及智力题

4.2.1 算法思想实战

- 场景题目：公司组织出游，每个人时间不同，让给出算法求出一个时间段尽可能满足最多人的要求？
- 利用二分法，判断一个数是否在给定的 list 里，并问时间复杂度？
- 红，绿，蓝三种颜色，n 个各自，相邻不能一样，首位不能一样，求填充方法数？

4.2.2 智力题

- 四个人分别花 1,2,5,8 分钟过桥，桥上只能容纳两个，且过河要手电筒！问最短过河方案！
- 想了一下为 8+1+5+1+2，面试官说这是第二优的解法，但不是第一优的。
- 打灯过桥的问题，4 个人过桥的时间分别为 1 分钟，3 分钟，7 分钟，9 分钟(具体

时间忘记了，但思路不变)。只有一个手电筒，每次只能两个人过桥，过桥时间以两个人中最慢的那个人过桥时间为准，问这四个人全部过桥最少需要多少时间?

4.3 其他方面

4.3.1 数论

- 给定 0-100 个数，再给定一个数，怎么判断这个数是不是重复了

4.3.2 计算几何

- 二维平面有 n 个点，求点 p，其到所有点距离之和最小?
- 给定坐标系内的一个矩形和一个点，怎么判断该点是否在矩形内部?

4.3.3 概率分析

- 一个公交站 1 分钟内有车经过的概率是 q，那么 3 分钟内有车经过的概率是多少?
- 如果一个人在公交车站台一分钟内能等到公交车的概率为 p，那么这个人在三分钟里能等到这辆公交车的概率为多少? (从反面去思考 1-(1-p)*(1-p)*(1-p))

4.3.4 矩阵运算

- 一个 100*100 矩阵，里面所有数都是正数，判断从左上角到右下角是否存在一条路径和为奇数的路径?

4.4 Leetcode&剑指 offer 原题

- Leetcode 65

5 编程高频问题：Python&C/C++方面

第五节
深信服面经
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 基础知识

5.1.1.1 内存相关

- python 里面线程与进程、进程如何共享内存?
- python 的内存处理机制有哪些？分别介绍一下。
- Python 了解么？说一下内存管理机制

5.1.1.2 区别比较

- Python 中的进程，线程和协程的区别?
- 深拷贝和浅拷贝的区别
- python 中 is 和==有什么区别?

5.1.1.3 讲解原理

- python 里面的进程、线程、协程特点-说了进程是资源最小单元，线程分配最小单元 线程相互影响?
- 了解 python 装饰器吗？它的作用?
- python 装饰器知道吗？装饰器的原理是什么？本质是？闭包?
- 判断一个值是否在数组里，用 set 快还是 list 快?
- Python 中常用的排序有哪些?
- Python 中的自动化测试用过么?
- 统计各个函数执行的时间（就是想问装饰器，我不会），python 装饰器
- 如果结构体中，声明了一个 char 型变量，一个 int 型变量，那么这个结构体占多少字节?
- python 中的内存管理知道吗，介绍一下?
- python 中 self 的用法?

- python 中的可变对象和不可变对象有哪些，特点、用法?

5.1.1.4 讲解应用

- python 里常用的数据类型有哪些
- python 字典中有 1000 万条数据，如何取所需要 value 值的 100 万条?

5.1.2 手写代码相关

- 介绍下 lambda

5.2 C/C++方面

5.2.1 基础知识

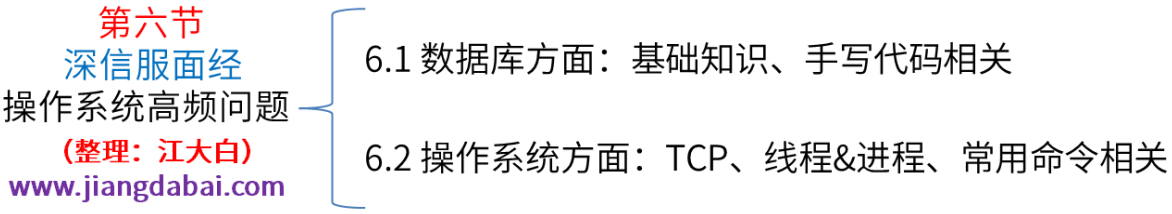
5.2.1.1 区别比较

- size of 和 stl：len 的区别

5.2.1.2 讲解原理

- 结构体：内存对齐

6 操作系统高频问题：数据库&线程&常用命令等



6.1 数据库方面

无

6.2 操作系统方面

6.2.1 TCP 协议相关

- tcp /udp 的区别
- 三次握手讲一下?
- 为什么三次握手，为什么四次挥手?

6.2.2 线程和进程相关

6.2.2.1 区别比较

- 进程，线程的区别?
- 深拷贝、浅拷贝

6.2.2.2 讲解原理

- 计算机网络，介绍知道的网络知识?
- linuxs 线程和内存管理

6.2.3 常用命令

- Linux 命令了解么？Linux 中软链接和硬链接的区别?
- Linux 中文件权限是怎么样的?

7 技术&产品&开放性问题

7.1 技术方面

- 有一个大的文本序列，求出现次数最多的前 k 个单词？复杂度？太多怎么保存?
- 说一下无监督学习?
- 考虑过读取 10M 大小的 json 文件，内存会占用多少？怎么保存的么?
- 垃圾短信多分类任务(如何分开发票，广告，商铺信息等)，有什么思路

- 如果有 n 种类别(比如新闻类，体育类等)的网站，目前收集到一些网站，及其网站中不良信息的位置，那么新来一个网站，如何判断该网站中是否含有不良信息，若含有，不良信息在哪个位置？
- 如果收集到一些网站的语料，如何判断这些网站中是否有不良信息？

他说如果有一些语料，然后还有一些关键词，如何判断这些语料中是否含有这些关键词？

- 大牛计划岗位：怎么设计一个查重系统？
- 场景题，一张多个人的合影，怎么确定人的位置，以及输入工号得到位置，怎么设计网络等等。