面试部分说明

NLP面试过程将主要集中在项目上，这也应当是你简历中的主要内容。而在项目中，问题又可以分为以下几个类型：项目软知识，模型数据问题，模型对比问题，模型优化问题。下面将针对不同的项目和问题类型列出具体的可能问题。

项目软知识

- 问题1：简单阐述一下自己的工作经历和项目经历(自我介绍)。
- 问题2：这个项目预计对产生的公司价值，以及最后产生了怎样的价值(项目价值)。
- 问题3：项目团队有多少人，你在其中主要负责哪个模块。

在线医生

模型数据与评估指标问题

- 问题1：在做命名实体审核或识别的时候，都做了那些数据预处理?

!!! Tip "提示" 基本上所有的模型训练之前都要进行数据预处理，而NLP的数据预处理大体上可以包括：文本处理基本方法（如：分词），文本张量的表示方法（如：word2vec），文本数据分析（如：长度分布），文本数据增强方法（如：回译）。根据自己对项目中模型的理解进行回答。

- 问题2：遇到过一些样本不均衡的问题吗，是如何解决的，效果怎样?

!!! Tip "提示" 类似于样本量少或者样本不均衡的情况，均使用上采样(增加样本数量)或下采样(减少样本数量)解决，一般建议使用上采样，方法就是回译数据增强，一般情况下，回译数据增强能为模型验证准确率带来1%-5%区间范围的提升。

- 问题3：命名实体识别模型的评估指标是什么?

!!! Tip "提示" NLP领域的模型评估指标并不多，准确率，召回率，精度，F1，都要记住!

- 问题4：还知道哪些模型评估指标，比如AUC，代表什么含义?

!!! Tip "提示" 这是一道向传统机器学习拓展的问题，你面试的公司，可能正在使用机器学习解决问题中，因此，面试官很容易问出和当前系统方法有关的问题。其中，若他们正在使用AUC为评估指标，那么你需要解释什么是AUC，以及为什么使用AUC。这里给出AUC的含义：对于任意一对正负例样本,模型将正样本预测为正例的可能性 大于将负例预测为正例的可能性的概率。

- 问题5：在实体审核的时候，正负样本是如何定义的，条数大概是多少?

!!! Tip "提示" 这是一道细节题，考察数据的真实性，实体审核的这个样本来自结构化网站抓取，负样本则是对应的乱序文本，对于条数，无论是否记得清晰，因为是使用BERT进行微调的训练方式，且微调网络并不复杂，一般数据量应该在2万-10万条皆可。

模型对比问题

- 问题1：有进行一些传统机器学习的尝试吗，为什么没有选择LR或者xgboost这样的模型?

!!! Tip "提示" 又是一次与机器学习的对比情况，但凡是这种对比，无论你是否作为对比实验，建议回答为：基于BERT的微调与LR等效果相当，但随着数据量的增大，BERT模型的优势将会更加明显。

- 问题2：有和一些当前表现更好的模型做对比实验吗，比如XLNET，roBERTa等等?

!!! Tip "提示" 这是与其他大型模型的对比实验，这里注意，因为我们解决的是中文问题，除了BERT的chinese和多语言对中文友好，其他大型模型都没有明确的权威的中文文本实现方案，包括号称已经在CLUE上取得中文优秀得分的ALBERT，都没有成熟到工业使用的程度。

- 问题3：BERT模型相比LSTM的优势是什么?

!!! Tip "提示" BERT相比LSTM的优势，可以直接用BERT主要结构Transfomer相比LSTM的优势来回答。

- 问题4：在做NER时，如果只使用BiLSTM是不是也可以产生结果，为什么还要加CRF?

!!! Tip "提示" 这是使用LSTM+CRF基本必问的问题，因为CRF能利用标签序列的信息，更具体的回答展开方式可自查。

模型优化问题

- 问题1：模型训练过程中做过哪些优化?

!!! Tip "提示" 训练过程的优化一般有两种目的，第一是提高训练速度，第二是提升评估指标。如何提升训练速度呢，当然是分布式训练（模型分布或者数据分布），关于分布式训练的实现大家可以参考NLP案例库中的案例。而在训练过程中提升评估指标（比如准确率）往往是最主要的，一般情况，使用基于贝叶斯的超参数调优方法可以在原有基础上获得改进。

- 问题2：模型部署过程中做过哪些优化提升推断延迟?

!!! Tip "提示" 模型部署过程的优化往往只有一个目的：就是提升推断速度，一般使用的方法就是模型量化。关于量化内容，大家可参考NLP案例库：应用于bert模型的动态量化技术。

智能文本分类

模型数据与评估指标问题

- 问题1: 简单说一下模型数据的来源?

!!! Tip "提示" 因为智能文本分类是多任务多模型系统，语料来源十分惹人注意，正如项目中所说，公司性质能够为我们提供充足的语料。

- 问题2：对原始数据做过哪些数据清洗工作吗?

!!! Tip "提示" "数据清洗"在含义范围上是大于"数据预处理"的，当然这种问题你直接回答数据预处理的内容也是可以的。一般NLP中的数据清洗过程是都会包括人工审查，去停用词，以及数据预处理。

- 问题3：模型正负样本的数量都是多少?

!!! Tip "提示" 智能文本分类并不是一个模型，而且每个模型的样本数量也不相同，可以给面试官举一些模型的例子，比如：影视判别模型，正负样本都是8000条。

模型对比问题

- 问题1：为什么用fasttext模型？优势是什么?

!!! Tip "提示" 一般在线深度分类模型都会以fasttext作为初始尝试，因为其训练速度和推断速度能够满足实际应用要求。

- 问题2：fasttext模型的损失函数是什么？这种损失函数有什么好处?

!!! Tip "提示" 交叉熵损失，其优势一般是指与MSE的优势，在机器学习阶段应该学习过。

- 问题3：fasttext模型的优化器选择的是什么，有做过一些对比吗?

!!! Tip "提示" Adam，对比就是描述一下Adam优化器与其他的比如Adagrad，SGD相比的改进点。

模型优化问题

- 问题1：模型训练过程中做过哪些优化?

!!! Tip "提示" 多进程训练，训练的时间加速了，你就有机会尝试更多的参数，也就有机会调出更好的模型。

- 问题2：模型部署过程中做过哪些优化提升推断延迟?

!!! Tip "提示" 模型部署的优化就是为了更快的推断，我们这里使用了多线程预测。

- 问题3：你认为这个项目的最大亮点是什么?

!!! Tip "提示" 多进程训练与多线程预测
