NLP 自然语言

一、简答题(每个 1.5 分)

1. 请详细介绍一下 Word2Vec 的结构。
2. 请介绍一下 HMM 模型中的 Vitebi 算法原理及应用场景。
3. 请详细介绍一下 RNN、LSTM、GRU 结构的差异性。
4. 请解释一下为什么 Attention 注意力机制特征提取能力强的原理。
5. 请详细介绍一下 Transformer 的网络结构。
6. 请列出至少三种常用的分类损失函数，以及它们的应用场景或优点。
7. 请列出至少三种激活函数，以及它们的优缺点。
8. 请解释一下梯度消失、梯度爆炸、模型退化的产生原因，如何判断该问题的出现以及该问题的解决方案。
9. 请解释一下深度学习中模型欠拟合、过拟合产生的原因以及解决方案。
10. 请详细描述一下你所了解的批归一化操作原理，以及不同批归一化方式的优缺点。
11. 请详细介绍一下 Bert 网络结构、和其它语言模型的区别以及为什么采用 position embedding。
12. 请详细说明一下为什么序列建模中更偏向于使用 CRF 而不是 softmax?
13. 什么样的数据集不适合深度学习?
14. 请列出至少三种常用的深度学习优化算法，并介绍它们的优缺点。
15. 数据不平衡会不会影响模型效果，如果影响模型效果，如何解决?
16. 如果已经明确原始数据集中存在大量异常数据，请问接下来的模型训练如何处理?
17. 请描述一下学习率、批次大小、惩罚性系数对模型训练的影响以及如何调整优化这些超参数?
18. 请详细说明一下 RNN 为什么存在长时依赖问题？为什么 LSTM 可以缓解长时依赖问题?
19. LSTM 中的激活函数可以更改吗？为什么?
20. 激活函数在神经网络中是不是必须存在的？为什么?
21. 请描述对 fine-tuning 的理解以及如何进行有效的 fine-tuning。
22. 在 NLP 应用场景中，如果需要进行批归一化处理，一般常用的批归一化的方式是什么？为什么?
23. 文本相似度计算的方式有哪些？如何具体实现?
24. 当使用 jieba 进行分词操作，如果出现分词效果不佳的时候，如何进行处理？以及如何判断分词效果好不好?
25. 逻辑回归的原理及步骤介绍、手写逻辑回归的损失函数；逻辑回归的损失函数中为什么是连乘？相加是否可以?
26. NLP 常用的数据增强方式有哪些?
27. 请详细介绍一下 AlBert 体系算法的优化，以及为什么不建议在 AlBert 中使用 Dropout 操作?
28. 如果想自己实现一个分词模型，你觉得应该如何实现?
29. 请使用 PyTorch 框架实现 multi-head self-attention。
30. 请使用 PyTorch 框架实现 Batch Normalization。

二、编程题

1. 基于 LSTM+CRF、Bert+LSTM+CRF 分别实现命名实体识别(10 分)

    a. 可任选其它类型的数据集；
    b. 最终提交：数据(50-100 条数据即可)、项目代码文件(离线模型训练、在线接口部署代码)、文档(接口描述文档、代码说明文档) —> 最终目的就是一个只是懂一点 python、linux 的开发工程师看到提交的代码&文档可以复现。
2. 任选一种算法实现关系抽取，并详细说明为什么采用这种算法(10 分)
    a. 可任选其它类型的数据集；
    b. 最终提交：数据(50-100 条数据即可 + 标注文件)、项目代码文件(离线模型训练、在线接口部署代码)、文档(接口描述文档、代码说明文档) —> 最终目的就是一个只是懂一点 python、linux 的开发工程师看到提交的代码&文档可以复现。
3. 请在以下领域中任选一个领域，详细描述该领域中可能会用到 NLP 的功能，以及实现该功能的技术方案。(15 分)
    a. 领域可选：互联网、工业制造、军事、农业、航空、安防等领域，也可选自己熟悉的领域。
    b. 最终提交：技术方案文档，包括领域、功能、实现的整体架构、用到的各个算法技术点、结合领域可能存在的问题对算法的可能改进点。
4. 任选一种课堂上没有讲过的 NLP 算法进行结构的分析、改进点的分析、优缺点的分析以及最终的代码复现应用。(20 分)
    a. 要求课堂上没有讲过的任意 NLP 算法均可。
    b. 最终提交：结构图、分析文档、复现的代码以及代码的项目应用(类似 2.1 和 2.2)