2023 NLP 小组考核要求

1 实操题三选一

1.1 对自然语言处理（或者多模态但必须包括自然语言文本模态）的数据(公开数据、自采集)进行包括数据处理、分析以及制定研究问题（例如，句子分类、文本生成）。提供部分公开数据集：

- https://ai.stanford.edu/~amaas/data/sentiment/
- http://qwone.com/~jason/20Newsgroups/
- https://www.clips.uantwerpen.be/conll2003/ner/
- https://catalog.ldc.upenn.edu/LDC2013T19
- https://www.statmt.org/wmt20/translation-task.html

1.2 将源代码转换为抽象语法树（AST）是程序分析的基本步骤，请从下方数据集链接中选择至少一种语言，对其进行可视化（可选），对其 AST 进行不同方面分析（例如，Path 和 Type）。

数据集：

- https://s3.amazonaws.com/code-search-net/CodeSearchNet/v2/{python,java,go,php,javascript,ruby}.zip

参考转换：

- https://docs.python.org/3/library/ast.html
- https://github.com/javaparser/javaparser
- https://github.com/jquery/esprima

1.3 复现和优化一个将自然语言问题转换为 SQL 查询的模型。这个实操题有助于了解自然语言处理和数据库查询之间的关系，以及如何将深度学习技术应用于实际问题。

- 代码地址：https://github.com/rhythmcao/text2sql-lgesql
- 复现代码：要求仔细阅读 README 文件，并按照给定的步骤安装所需的环境、下载数据集和运行代码。在复现的过程中，可能需要解决一些依赖问题或调试代码。要求撰写实验报告，记录在复现过程中的经历、所遇到的问题以及如何解决问题。
- 优化模型：要求研究现有模型的架构和训练方法，并尝试提出优化建议。优化可以从（模型架构、训练策略、数据增强）几个方面进行。

2 理论题

从下方的 Paper List 中选择一篇论文进行阅读总结，并制作 Slides 进行汇报。

Paper List:

- Phased LSTM: Accelerating Recurrent Network Training for Long or Event-based Sequences
- Self-Attention with Relative Position Representations
- The expressive power of pooling in Graph Neural Networks

3 说明

实操题在题目设置范围内鼓励大家进行开放性地回答，提交更具竞争力、特色的回答。理论题也不限定讨论范围，欢迎展示更深刻的汇报。请大家于 2023 年 5 月 31 日前把完成资料提交到邮箱（boyanxu@qq.com）。