**从零开始如何快速进入NLP的研究工作**

**任何问题，争取通过搜索、交流、和查阅文献得到解决**

**不要以书本为解决方案，基于书本的学习方式在不再适合科学研究**

1. 首先具有较强的编码能力
   1. 了解机器翻译，按照[流程](http://www2.statmt.org/moses/?n=moses.baseline)（这是一个超链接）跑出统计机器翻译模型SMT
   2. 按照[流程](https://github.com/pytorch/fairseq/tree/main/examples/translation" \l "training-a-new-model)（这是一个超链接）跑出神经机器翻译模型NMT（需要使用哈工大深圳的GPU集群；高性能云服务中心GPU集群进入试运行阶段，试运行阶段我校师生可免费申请使用，试运行过后将严格按照相关规范运营。请需要使用的师生登录平台，使用统一认证方式登录激活账号。使用过程如有问题可加用户交流qq群936399193。集群网址http://hpc.hitsz.edu.cn/（校外需开启VPN才能访问），帮助文档http://hpc.hitsz.edu.cn/help/index.html）
   3. 将0.1和0.2中的数据集（训练集，开发集和测试集）调换；也就是基于0.1中的数据集跑NMT，基于0.2中的数据集跑SMT

**完成0部分之后，用Latex（Overleaf网站）写个报告，包括且不限于数据的准备，环境搭配，代码运行调试中遇到的问题，评价指标，结果分析（精美的图和表），个人收获等等。**

**做完之后即可联系我。**

**完成了0之后，尽快完成以下内容：**

1. 了解各种文本分类任务，去下载SST2或者SST5句子分类数据
2. 手写实现一个基于ngram的最大熵分类器 (c++ & numpy)

抽取ngram特征，转化成为特征向量，实现模型参数矩阵，分类打分，计算概率，最大熵损失，计算模型参数反向梯度，最大熵优化，测试准确率，调参环节

1. 利用深度学习库（例如pytorch）实现ngram最大熵分类
2. 1-gram分类，使用预训练词向量进行句子分类，了解pooling，理解深度学习
3. 实现CNN，理解CNN和ngram的关系
4. 实现LSTM，理解LSTM和CNN的关系
5. 实现SelfAttention，理解pooling和SelfAttention关系
6. 实现Transformer，了解Transformer的各个组件
7. 理解word2vec以及语言模型，动手实现word2vec
8. 学会使用ELMo
9. 学会使用BERT，了解BERT分词机制
10. 查找相关论文，了解自己的分类模型和SOTA的区别，撰写文档
11. 了解NER与序列标注，下载conll03和GermEval2014数据
12. 实现BiLSTM-CRF模型
13. 实现BERT-(BiLSTM)-CRF模型，学会处理异常情况
14. 了解中文分词，查找相关数据，实现分词模型
15. 查找相关论文，了解对应数据SOTA结果，撰写文档
16. 了解各种序列标注任务