项目计划书

一 任务简介

NLP（自然语言处理），即让计算机去理解人类的自然语言（文本、语音等），进而完成各种各样的任务（NER、文本分类、机器翻译、阅读理解、问答系统、智能对话、搜索推荐系统等等），被誉为人工智能皇冠上的明珠。自然语言处理任务总结可以分为：自然语言生成和自然语言理解。

文本分类是 NLP 的一项基础任务，属于自然语言理解，旨在对于给定文本文档，自动地分配正确的标签。文本分类在许多方面的应用很多，例如：信息检索、自然语言推理、情感分析、问答等。文本分类任务从分类目的上可以划分为三类：二元分类、多类别分类以及多标签分类。

多标签分类，不同于多类别分类，多标签分类的总标签集合大，而且每个文本都包含多种标签，即将多个标签分配给特定文本。由于不同文本分配的标签集不同，给分类任务带来一定程度的困难。另外，当总的标签集合数目特别大的时候，这种情况可以算作为一种新的多标签分类任务，即极端多标签文本分类（Extreme multi-label text classification (XMTC)）。多标签文本分类的重点在于标签之间的共现性和层次性。

本项目旨在提高编程能力：了解如何在程序中划分数据集；如何建立网络模型结构；如何选择并使用评估方法；如何进行结果分析评估。

二 任务方法综述

本项目从两方面进行：简单文本分类模型和多标签文本分类模型。参考网上开源的代码，使用网上开源的数据集。为使任务方法具有比较性，相同任务采用同种数据集。所有任务的代码都用 python 和 pytorch 编写。

2.1 预训练语言模型

词向量是目前 NLP 领域最常用的文本向量嵌入方法，所以在此需要了解掌握最经典的方法：word2vec 和 glove。

至于其他经典的方法以了解为主：比如 EMLO、GPT、BERT、GPT2.0、

ALBERT、XLNET、GPT3.0。网上或其公司有开源服务，了解如何调用 API；了解如何针对文本分类任务进行微调。

2.2 文本分类模型

希望可以从多方面了解文本分类的实施流程。所以选用多种网络结构来了解。

CNN：TextCNN、CharCNN；

RNN：RNN、LSTM、GRU、Bi-LSTM；

Bi-LSTM+Attention；

RCNN；

Transformer；

Fasttext。

2.3 多标签文本分类模型

希望可以从多方面了解多标签文本分类的实施流程。所以选用多种网络结构来了解。

CNN：CNN、XML-CNN、层次性-CNN；

RNN：GRU；

C-RNN；

SGM。

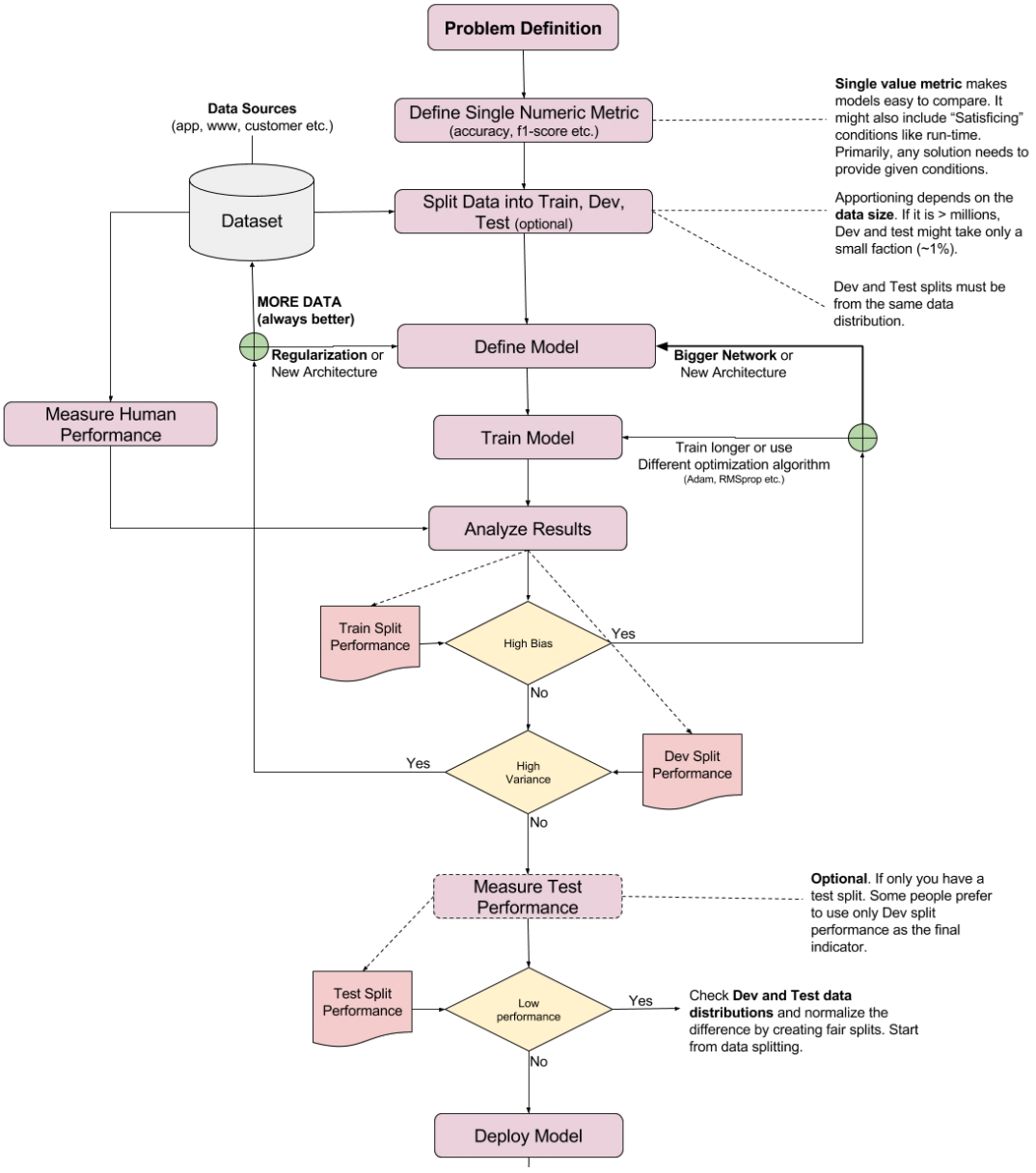
2.4 其他

由于网络模型中常用一些 tricks，所以在此项目任务加入一些简单高效的 trick，来提升网络性能，例如正则化、数据预处理方面、更高效的优化方法。

三 项目流程综述

时间：四个月（8 月-12 月）。历时四个月，其中文本分类任务 2 个月，多标签文本分类任务 2 个月。在每月中旬书写项目进度完成情况。

任务流程（见下图）：



参考资料

论文资料：上述所有模型的论文资料在个人电脑中。

网上课程

文本分类：

博客：文本分类实战 https://www.cnblogs.com/jiangxinyang/p/10241243.html

Github 文本分类：https://github.com/CoreJT/TextClassificationSystem

Github 中文文本分类：https://github.com/649453932/Chinese-Text-Classificatio

n-Pytorch

多标签文本分类

CNN: https://github.com/zhangfazhan/Multi_Label_TextCNN

https://github.com/moxiu2012/PJ_NLP

SGM：https://github.com/lancopku/SGM

待续

在 paperwithcode、Github 上寻找相关资料。