哈尔滨工业大学（深圳）

大一年度项目中期检查报告

项目名称： 通知内容管理 App 的设计与实现

项目负责人： 王靳 **学号：** 220111012

联系电话： 15816870583 **电子邮箱：** 220111012@stu.hit.edu.cn

学　　院： 计算机科学与技术学院

指导教师： 吴宇琳 **职称：** 助理教授

联系电话： 18126282493 **电子邮箱：** wuyulin@hit.edu.cn

学　　院： 计算机科学与技术学院

填表日期：　2023　年　6　月　30　日

一、项目基本信息（包括项目负责人、按顺序）

姓名	性别	所在学院	学号	联系电话	本人签字
王靳	男	计算机科学与技术学院	220111012	15816870583	王靳
吴语诗	女	计算机科学与技术学院	220110928	15967167116	吴语诗
蔡德林	男	理学院	220810316	15919094899	蔡德林
邹悦	女	理学院	220810424	18820366233	邹悦

立项背景

随着信息化发展，通知更多借助网络渠道。学校目前使用的飞书 App 仍存在重要通知被淹没、通知对象针对性不强、通知本身信息冗杂等问题。学长学姐开发的 App 不支持自动生成关于截止日期的提醒。大多数同学需要在群消息中反复寻找、查看同一条通知，时间利用效率低。项目计划设计并实现一个通知内容管理 App，实现学校通知精准分类、要点捕捉与简化、重要通知收藏与推荐、日程安排表个性化生成五项功能，希望服务于学院通知发布工作。

项目研究内容及实施方案

（一）研究内容

本项目通过对 “文本挖掘”的研究，利用相关算法将学院以大段文本形式呈现、信息糅合一体的通知抽象成一个个简单标签，然后根据使用者的身份定位，将简化后的通知信息标签与使用者一一对应。此外还需要研究如何编写 App 前端，以及如何汇聚多个 App（例如飞书、微信、QQ）中所有通知群的信息到开发的 App 中。

（二）实施方案

（1）相关知识学习：python 的基础用法，利用 pytorch 框架进行模型的训练与调试，学习 scapy 爬虫的使用方法，学习前端开发和项目部署。

（2）前端开发：做出一个好看的 UI，并让后端的功能与前端 UI 中的按钮对应。

（3）后端开发：为相关文本挖掘算法调整合适的参数，并对模型加以训练，做出一个通过调用简单指令实现将通知简化与分类、能够通过模糊搜索查看完整通知等功能的后端软件。

（4）前后端对接与部署：将整个客户端打造成 docker 镜像，能直接一键部署在本地。

二、项目研究中期报告

（一）项目实施的进展情况及取得的成果

在知识学习方面，小组学习了 Python 基础知识和一些高级用法（例如正则表达式等），对于接下来的工作非常有帮助。

在前端方面，小组已经取得了一些进展。通过对 Qt / C++的学习并根据相关教程进行简单的实践，小组成功编写的了一个前端，并留下了用 lambda 表达式完成的接口，如图 1-1、图 1-2、图 1-3 所示。经过目前很多应用（如 Jupyter / ALIST 等）的启发，项目前端将会应用一些 Web 框架（例如 Flask）重新设计。

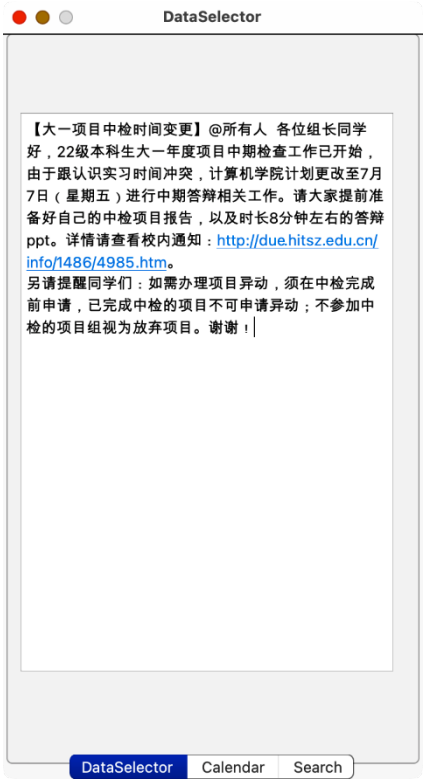

图 1-1 数据采集界面

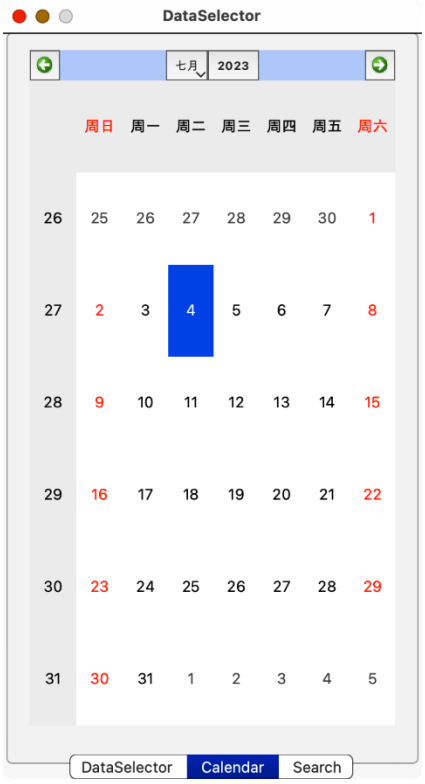

图 1-2 日期输出界面

图 1-3 通知搜索界面

在后端方面，小组学习了迁移学习相关技术，如何利用 pytorch 搭建神经网络，API 文档编程等。此外还学习了机器学习的传统算法（如 KNN、Kmeans、决策树等），深度学习的经典算法（如 CNN、RNN、LSTM），NLP 的经典算法及模型（如 Transformer、BERT、T5 等）。

小组了解到在 NLP 应用中，Transformer（如图 1-1）、BERT 和 T5 已经成为了重要的模型。Transformer 是一种基于自注意力机制的神经网络（如图 2 所示），旨在解决序列到序列学习问题，优点在于能够处理长序列数据和能够并行计算，因此在训练和推理速度方面具有优势，更加高效和灵活（RNN 和 LSTM 由于无法有效处理长序列数据且在训练和推理速度方面较慢，所以已经逐渐被淘汰）。BERT 是一种基于 Transformer 的预训练语言模型，可用于文本分类、命名实体识别等任务。T5 则是一种基于 Transformer 的 seq2seq 语言模型，可用于文本摘要等任务。这些模型在 NLP 领域具有广泛的应用前景。小组将继续努力学习和实践，为开发出更加优秀的 NLP 应用而努力。目前，小组能够使用 hugging_face 提供的预训练模型，并跟着 hugging_face 官方的教程，完成了 Amazon 商品评论摘要的项目。

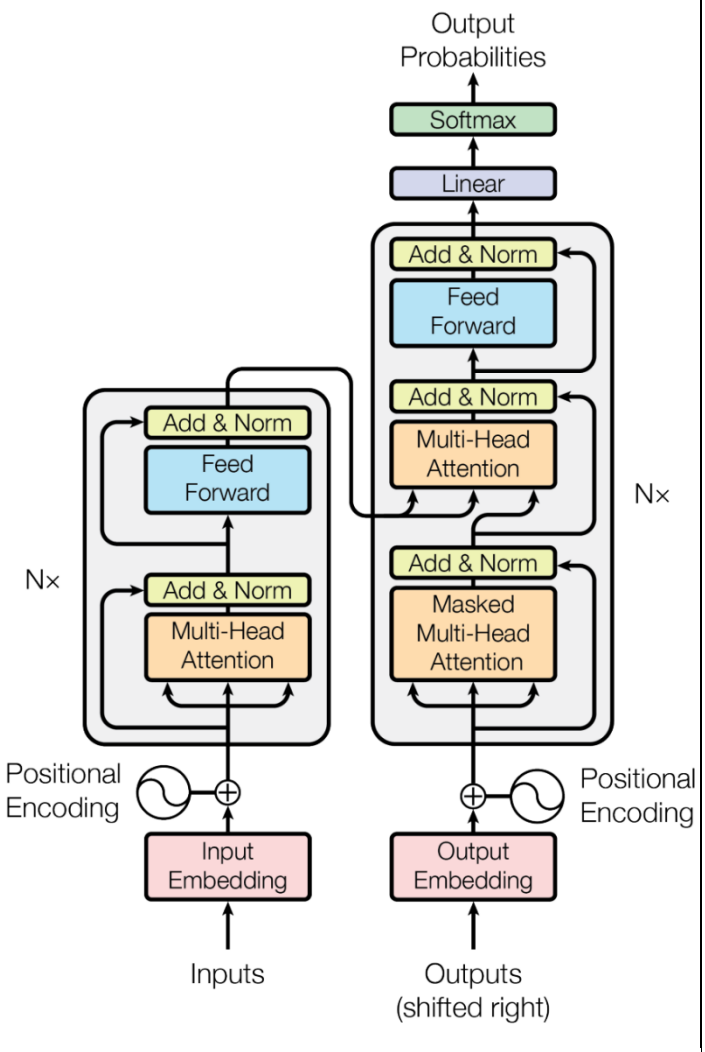

图 2 Transform 模型结构图

在项目部署方面，小组已经熟悉了 Linux（Arch）环境，并能够使用 docker-compose 等命令行工具完成项目的部署。还能够编写一些简单的 Shell 脚本。

（二）遇到的困难及下一步工作计划

（1）遇到的困难

不可否认，当前项目进度确实与原定计划有所滞后。对此，深感抱歉并进行了反思和检讨。在此基础上，已经采取了一系列措施来加快项目进度。

在前端方面，项目目前基于 Qt/C++编写的前端过于粗糙，需要进行改进。因此，决定后续使用基于 flask 框架的 web 前端。这个决策是基于小组成员使用 jupyter, alist 等 web 前端的启发，认为这些前端无论是审美还是效果都很好。

在后端方面，小组成员学习了相关的算法，例如一些传统的机器学习的算法（决策树 / kmeans / XGBoost 等）却发现它们更多的适用于结构化数据的分类或回归问题。同时，小组自身水平存在限制。例如数学水平不足以匹配反向传播、梯度下降、transformer 的学习，导致进展缓慢。由于项目是有关通知内容的分类，不同于很多现成的新闻分类数据集，因此需要自制数据集。制作过程比较繁琐，即便将计科飞书群通知和哈工大官方通知作为数据集，数量也仍然有限（或许只有千条左右），并不能保证模型训练出来的效果如何。此外训练模型的算力不够、互联网上资料良莠不齐等问题也给带来了很多挑战。

在项目部署方面，由于后端尚未成型，暂未遇到任何困难。

（2）下一步工作计划

针对上述问题，小组制定了相应的解决方案如下：

在后端方面，首先，小组成员将尽快提升自己的数学水平，确保能够正确理解、掌握并运用项目的相关算法。其次，实现本项目时需要通过一些技术手段来制作数据集。最后是进行模型训练。在 Hugging Face 上有许多现成的 BERT 模型可供使用，而做一个文本分类项目使用 BERT 模型完全能够满足需求，于是项目计划采用迁移学习等技术来实现模型的训练。这样可以极大地缩短训练时间，提高训练出来的模型的准确率。为实现文本摘要的目标，小组计划直接使用文本摘要现成的训练好的模型，例如 mT5_multilingual_XLSum。在算力不够的情况下，将在本地机上执行小样本的训练，并在调试完毕以后，在算力服务器上执行大样本的训练。

具体深度学习的开发大致分为以下几步：

- 利用 Python 的 scapy 爬虫技术从飞书和哈工大官网上爬取一些通知，并使用 Doccano 标记工具或 ChatGPT 提供的 API 来实现数据的自动化标记。
- 将使用 dynamic padding 和 dynamic truncation 技术将输入的通知转化为固定长度，并利用 word2vec 模型将文章转化为词向量（一般词向量有 768 维）。

（以上两步可以使用 hugging_face 提供的封装好的包——tokenize 来实现）

- 在 BERT-Chinese-wmm（分类）预训练模型和 PEGASUS-T5（文本摘要）预训练模型的基础上进行迁移学习与微调。
- 使用 pytorch 封装好的 metrics 包进行模型评估。

以上是文本分类和文本摘要的大致步骤。此外项目还计划使用正则表达式等技术匹配截止日期并投递到日历相应的日期中。

（3）进度安排

具体进度安排如表 1-1 所示。

表 1-1 进度安排

时间	工作安排
2023 年 9 月 1 日之前	完成后端，文本分类，文本摘要，日期输出
2023 年 9 月 15 日之前	完成前端开发，基于 flask 框架的 web 前端
2023 年 10 月 31 日之前	完成前后端对接与部署

（三）结题预期目标

实现通知内容管理 App，支持通知信息分类、信息精简、重点关注、模糊搜索、生成日程安排五项功能。

（四）经费使用情况

购买了《HuggingFace 自然语言处理详解》（李福林等著），花费 56.6 元，剩余 1943.4 元。

三、指导教师意见

团队工作认真，已完成部分前端工作和后端相关工作，后期需加快后端开发进度。

签　　名：吴亭琳 2023 年 7 月 5 日

四、项目专家组意见

1．是否达到中期目标（在□内打√）：

□达到中期目标　　□基本达到中期目标　　□未达到中期目标

2．成绩评定（在□内打√）：

□合格

□不合格，项目实施
- □改进后可继续执行
- □提出警告、观察后再定继续执行或中止
- □中止实施

3. 其他意见和建议：

组长签名：　　年　　月　　日