通知内容管理App的设计与实现

指导老师：吴宇琳

汇 报 人：王靳

组　　员：王靳、吴语诗、邹悦、蔡德林

目 录

CONTENTS

01 项目进展情况

02 问题与解决方案

03 下一步计划

04 小结

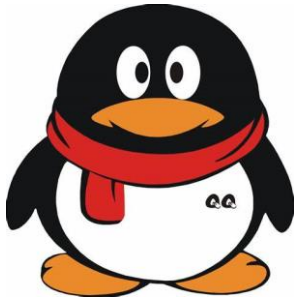
Part 01

项目进展情况

1.1 项目概述

（1）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。



1.1 项目概述

（2）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。

1.2 项目进展情况及成果

（1）相关知识学习方面：

学习了Python基础知识和一些高级用法（例如正则表达式，装饰器，多线程等），对于接下来的工作非常有帮助。

例如：正则表达式可以用于日期的匹配，可以帮助我们完成将日期输出至ddl日历中。装饰器普遍用于Flask框架的应用中。

1.2 项目进展情况及成果

（2）前端方面：

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。

受到一些应用（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。

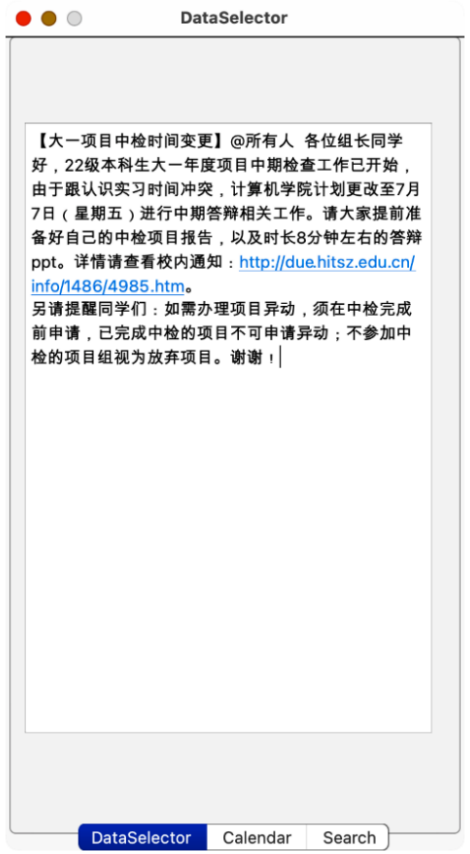


图 1-1 数据采集界面

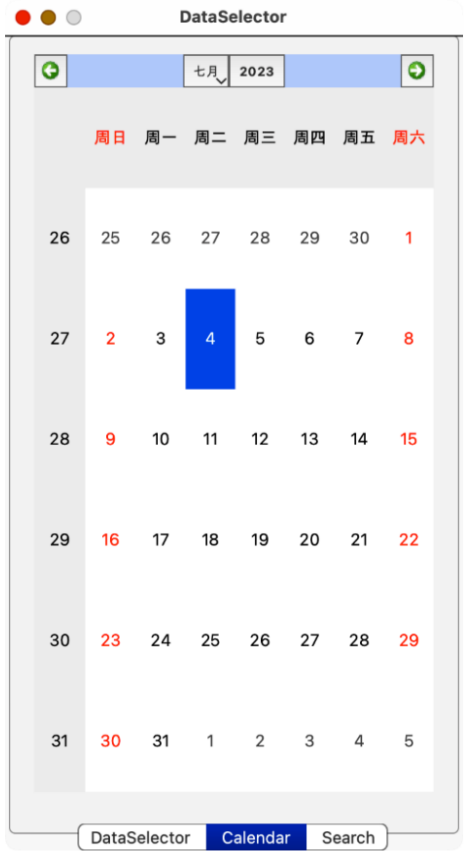


图 1-2 日期输出界面

图 1-3 通知搜索界面

1.2 项目进展情况及成果

（3）后端方面：

- 迁移学习相关技术
- 利用pytorch搭建神经网络
- 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）**

1.2 项目进展情况及成果

Transformer

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）。

参考论文：https://dl.acm.org/doi/10.5555/3295222.3295349
Attention is all you need

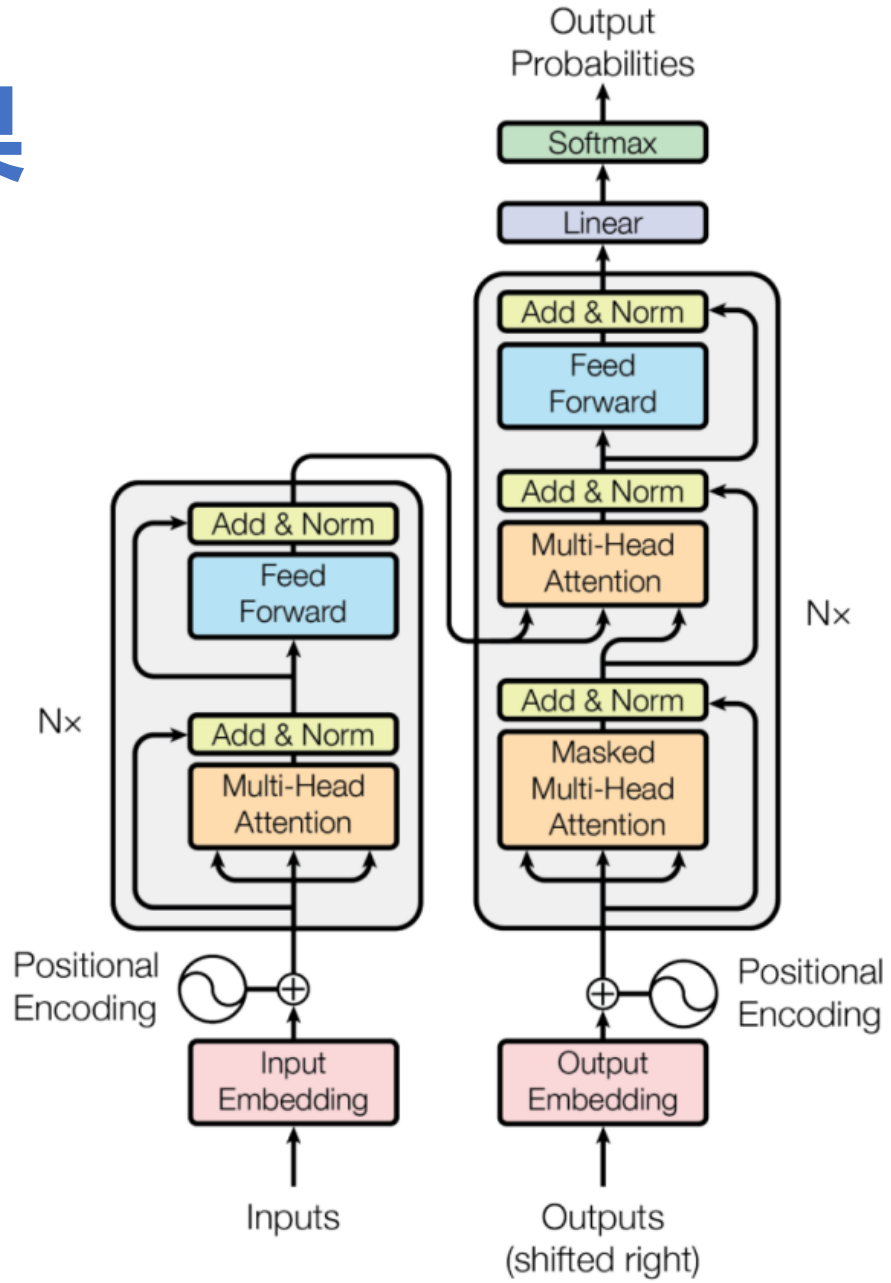


图 2 Transform 模型结构图

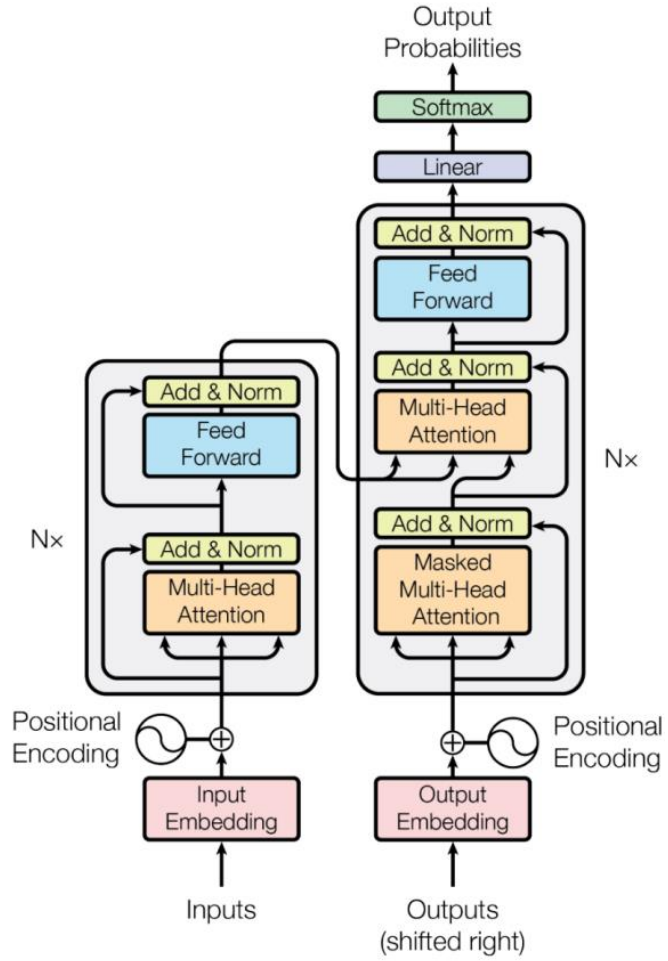
1.2 项目进展情况及成果

文本摘要使用：T5模型

参考论文：
https://dl.acm.org/doi/abs/10.5555/3455716.3455856
Exploring the limits of transfer learning with a unified text-to-text transformer

中文预训练模型：https://arxiv.org/abs/1912.08777
PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization

Github：https://github.com/ZhuiyiTechnology/t5-pegasus



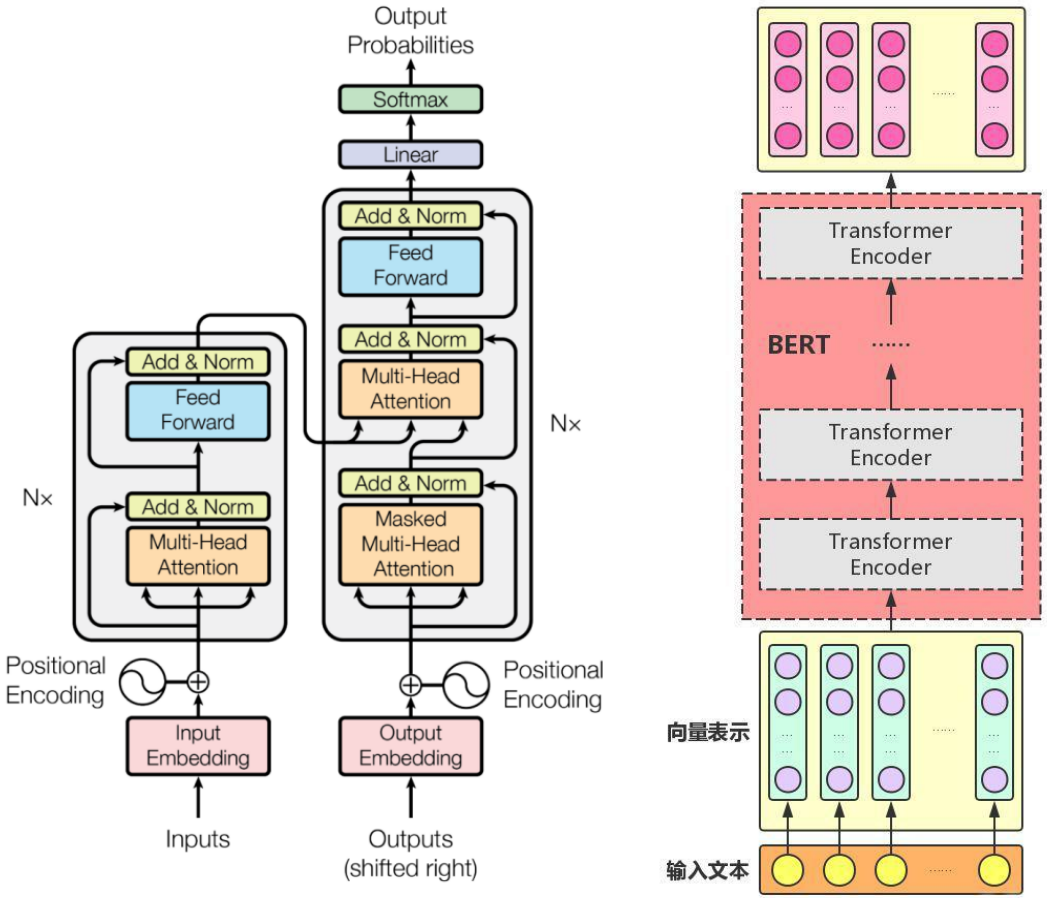
T5模型实际上是一个完整版的Transformer，实现了Seq2Seq。Encoder和Decode都用到了

1.2 项目进展情况及成果

文本分类使用：BERT模型（如右图）

参考论文：https://arxiv.org/abs/1810.04805
BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

中文预训练模型：
https://arxiv.org/abs/1906.08101
Pre-Training with Whole Word Masking for Chinese BERT
（由哈工大和科大讯飞联合发布）



1.2 项目进展情况及成果

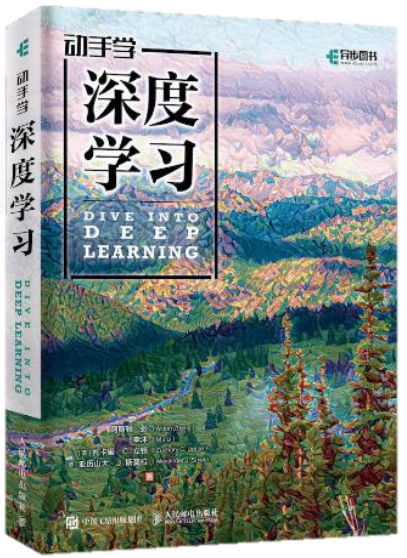
使用Hugging_Face提供的预训练模型，并跟着Hugging_Face官方的教程，完成了Amazon商品评论摘要的项目。 https://huggingface.co/learn/nlp-course

参考书籍：

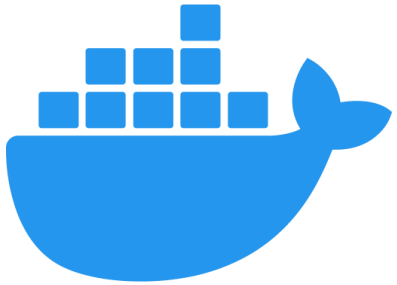
李沐《动手学深度学习》

https://zh.d2l.ai/chapter_preface/index.html 书籍开源地址

李福林《HuggingFace自然语言处理详解》

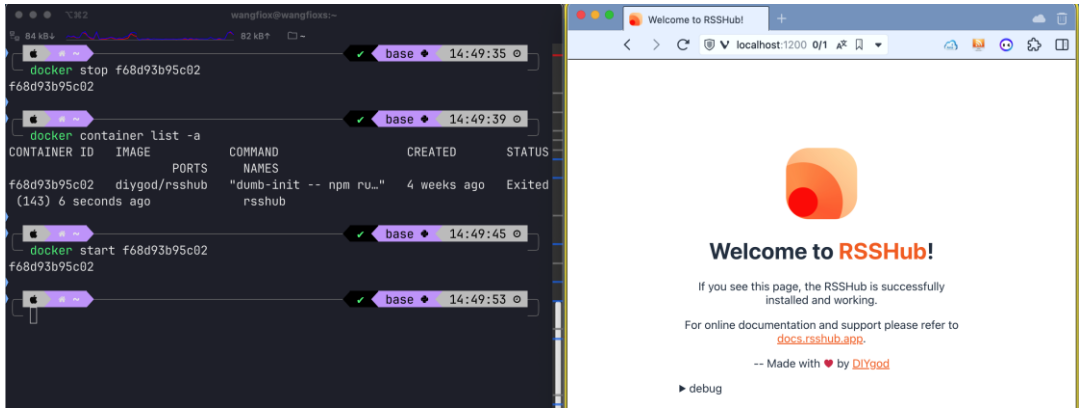


1.2 项目进展情况及成果

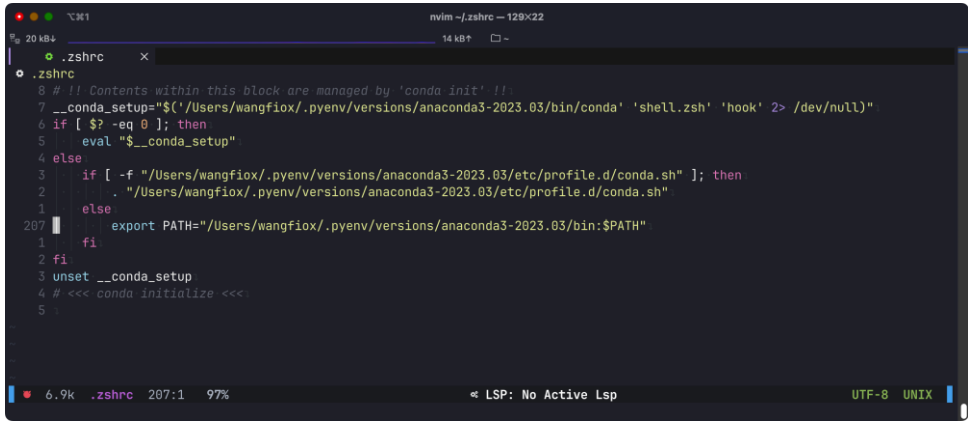


（4）前后端对接和部署方面：

- 熟悉了Linux（Arch）环境。
- 能够使用docker-compose等命令行工具完成项目的部署。
- 能够编写一些简单的Shell脚本。



Docker的简单使用



简单的shell脚本的编写

Part 02

问题与解决方案

2.1 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。

2.2 后端方面

- **学习方面走了弯路。**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。

- **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。

- **数据集制作难题。**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。

Part 03

下一步计划

3.1 后端方面

- **尽快提升自己的数学水平**
- **通过一些技术手段来制作数据集**
- **进行模型训练**

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 PEGASUS T5。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。

3.2 具体深度学习的开发

- 数据搜集与预处理：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。
- 特征工程：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）。
- 迁移学习：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。
- 模型评估：使用pytorch封装好的metrics包进行模型评估。

使用Hugging_Face提供的封装好的包——Tokenizer和word2vec来实现

3.3 前端方面

前端开发：学习如何使用Flask框架。

在现代Web开发中，框架是一个非常重要的概念。Flask是一个使用Python编写的轻量级Web应用框架，它被广泛应用于开发Web应用程序和API。

Flask的优点之一是其灵活性。它不强制使用特定的工具或库，而是允许开发人员使用自己喜欢的工具。Flask还提供了大量的扩展，可以轻松地添加各种功能，例如身份验证、数据库集成和表单验证。

3.4 进度安排

2023年9月1日前

完成后端开发

- 文本分类
- 文本摘要
- 日期输出

2023年9月15日前

完成前端开发

基于flask框架的web前端

2023年10月31日前

完成前后端对接与部署

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。

Part 04

小结

4.1 项目进展情况总结和展望

- 项目进度滞后可能会影响整个计划的实施，需要时常对照计划安排及时采取措施加快进度。
- 仔细分析项目进度滞后的原因，制定相应的解决方案，确保项目及时完成。
- 积极与组员、学长、老师沟通和寻求帮助，共同推进项目进程。

本项目的目标是为学院通知发布工作提供服务，实现以下五项功能：
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。

我们将致力于让学校通知更加高效、便捷地传达给广大师生，提高沟通效率和信息传递的准确性。

谢　　谢

T H A N K　Y O U

哈爾濱工業大學
HARBIN INSTITUTE OF TECHNOLOGY