博学谷
www.boxuegu.com
黑马程序员线上品牌

基于BERT+P-Tuning文本分类介绍

一样的教育，不一样的品质

目录

Contents

01

项目背景

项目背景

文本是信息传播的重要途径和载体，将文本数据正确归类，从而更好地组织、利用这些信息，具有重要的研究意义。文本分类致力于解决上述问题，是自然语言处理（Natural Language Processing，NLP）领域的经典任务之一，被广泛应用于舆情监测、情感分析等场景中。

目前实现文本分类的方法很多，如经典的应用于文本的卷积神经网络（Text-CNN）、循环神经网络（Text-RNN)、基于BERT等预训练模型的fine-tuning等，但是这些方法多为建立在具有大量的标注数据下的有监督学习。在很多实际场景中，由于领域特殊性和标注成本高，导致标注训练数据缺乏，模型无法有效地学习参数，从而易出现过拟合现象。因此，如何通过小样本数据训练得到一个性能较好的分类模型是目前的研究热点。

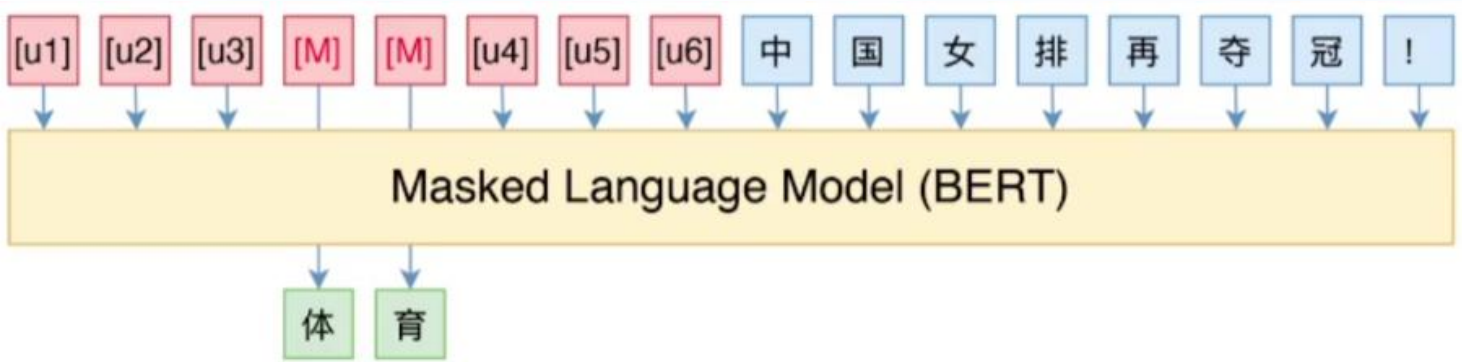
本章我们将以"电商平台用户评论"为背景，基于BERT+P-Tuning（软模版）方法实现评论文本的准确分类，这样做的目的在于提升用户体验。通过深入了解用户对不同商品或服务的评价，平台能够快速回应用户需求，改进产品和服务。自动分类也为个性化推荐奠定基础，帮助用户更轻松地找到符合其偏好的商品。同时，这项技术降低了运营成本，替代了繁重的人工处理工作。通过评论分析，电商平台还能迅速获取市场反馈，为商家提供有针对性的数据，助力制定精准的运营策略。

02

P-Tuning回顾

定义

P-Tuning（Pattern-Tuning）是一种连续空间可学习模板，P-Tuning的目的解决PET的缺点，使用可学习的向量作为伪模板，不再手动构建模板。



以新闻分类任务为例：原始文本：中国女排再夺冠！P-Tuning可学习模板：[u1] [u2] …[MASK]…[un]，Label：体育/财经/时政/军事

P-Tuning的实现过程

将模版（用特殊字符代替自然语言，特殊字符可以自由学习）与原始文本拼在一起输入预训练模型，预训练模型会对模板中的mask做预测，得到一个label。

P-Tuning的特点

优点

- 可学习模板参数，全局优化学习到更好的模板表示
- 缓解人工模板带来的不稳定性

缺点

- 超多分类任务场景：预测难度大
- 蕴含任务（给定两句话，让模型判断两句话的逻辑关系）等不适合基于模板方式解决

03

环境准备

环境准备

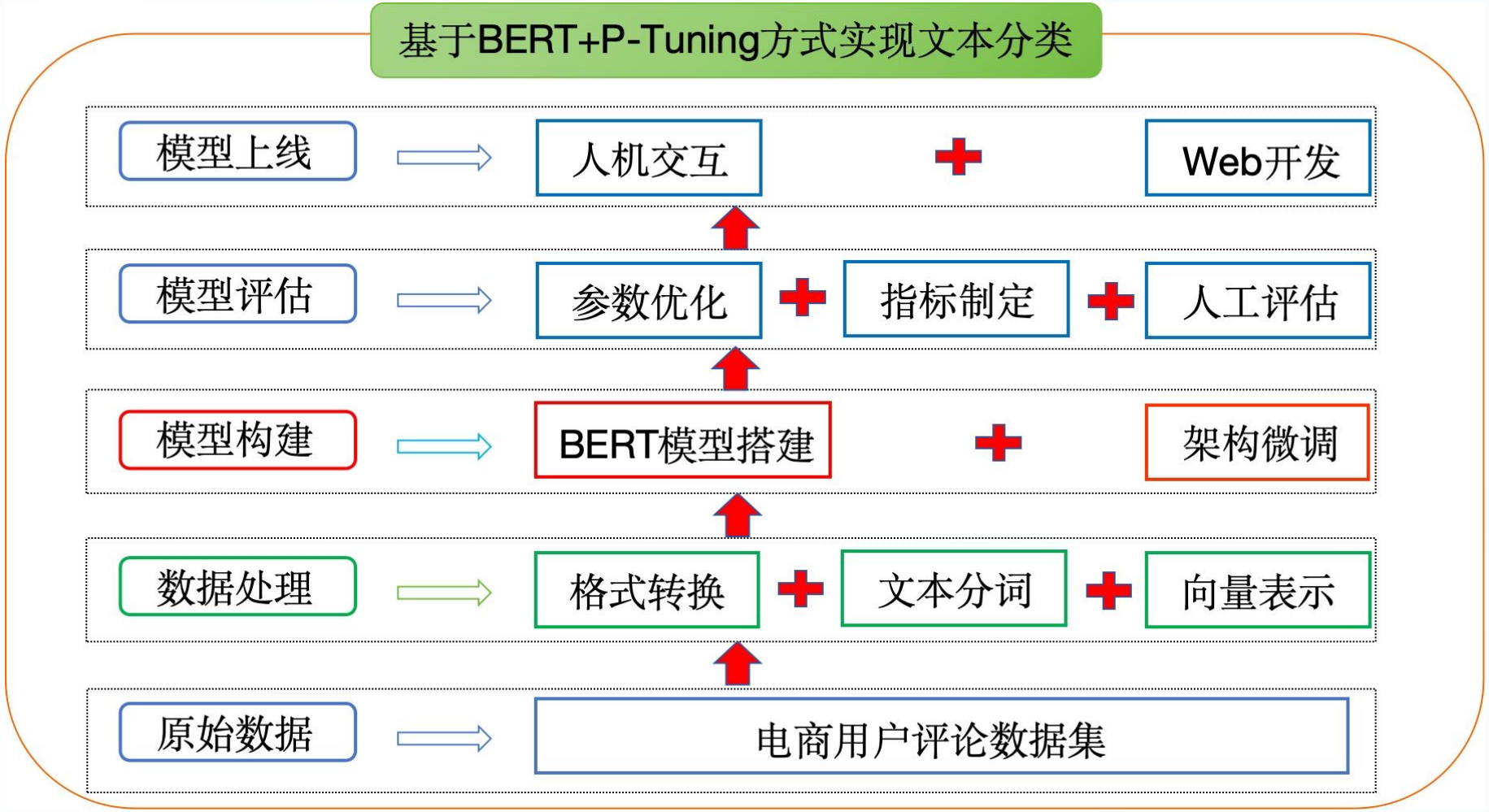
本项目基于 pytorch + transformers 实现，运行前请安装相关依赖包：

- torch
- transformers==4.22.1
- datasets==2.4.0
- evaluate==0.2.2
- matplotlib==3.6.0
- rich==12.5.1
- scikit-learn==1.1.2
- requests==2.28.1

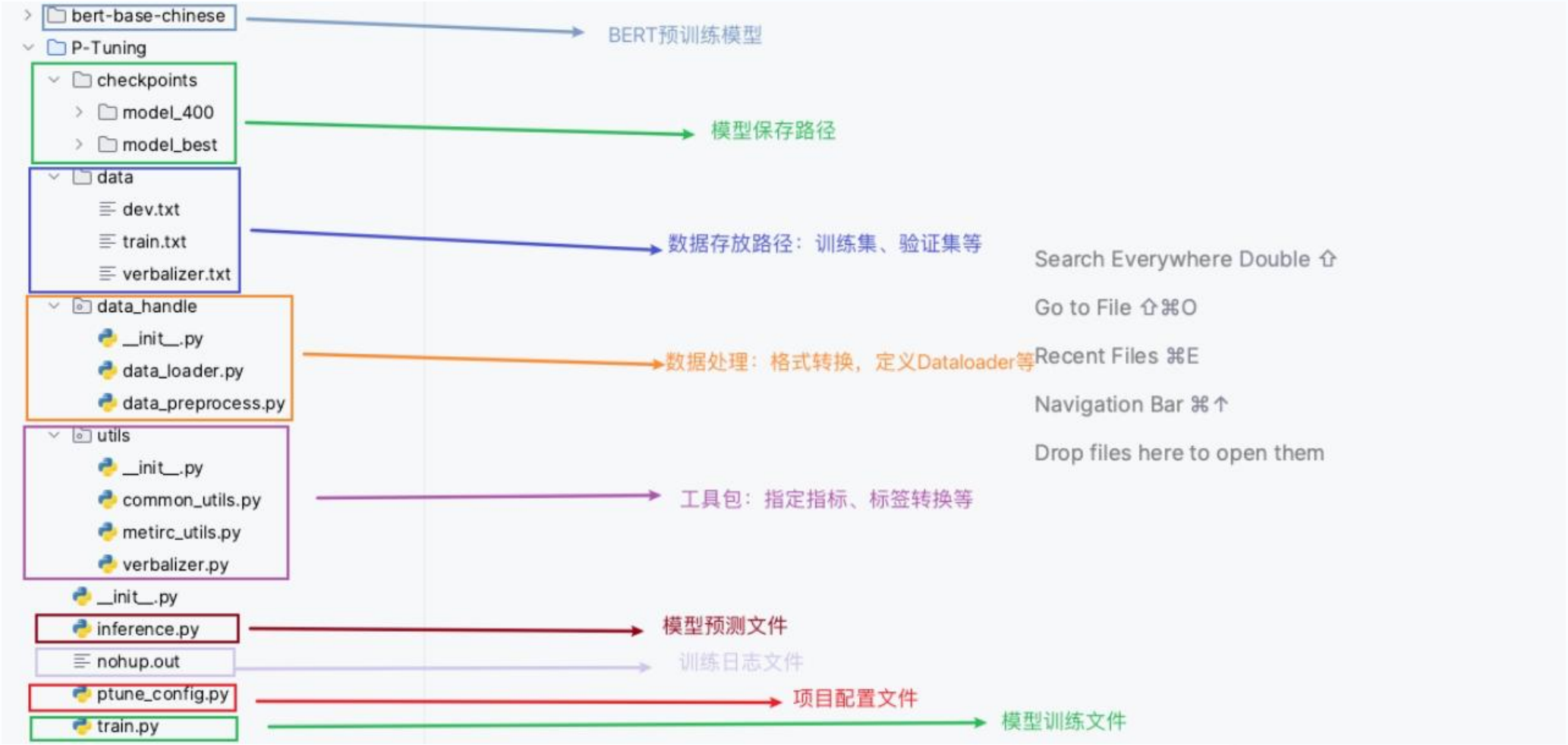
04

项目架构

项目架构流程图


基于BERT+P-Tuning方式实现文本分类
模型上线
人机交互
Web开发
模型评估
参数优化
指标制定
人工评估
模型构建
BERT模型搭建
架构微调
数据处理
格式转换
文本分词
向量表示
原始数据
电商用户评论数据集

项目整体代码介绍


bert-base-chinese
BERT预训练模型
P-Tuning
checkpoints
model_400
model_best
模型保存路径
data
dev.txt
train.txt
verbalizer.txt
数据存放路径：训练集、验证集等
data_handle
__init__.py
data_loader.py
data_preprocess.py
数据处理：格式转换，定义Dataloader等
utils
__init__.py
common_utils.py
metirc_utils.py
verbalizer.py
工具包：指定指标、标签转换等
__init__.py
inference.py
模型预测文件
nohup.out
训练日志文件
ptune_config.py
项目配置文件
train.py
模型训练文件
Search Everywhere Double ⇧
Go to File ⇧⌘O
Recent Files ⌘E
Navigation Bar ⌘↑
Drop files here to open them

博学谷
www.boxuegu.com
黑马程序员线上品牌

Thanks!

扫码关注博学谷微信公众号