博学谷
www.boxuegu.com
黑马程序员线上品牌

BERT+PET文本分类模型搭建

一样的教育，不一样的品质

目录

Contents

01

实现模型工具类函数

基本介绍

01

目的

模型在训练、验证、预测时需要的函数

02

代码路径

/Users/**/PycharmProjects/llm/prompt_tasks/PET/utils

03

脚本

utils文件夹共包含3个py脚本：verbalizer.py、metirc_utils.py以及common_utils.py

verbalizer.py

目的：定义一个Verbalizer类，用于将一个Label对应到其子Label的映射。

部分代码展示：
具体参考代码
文件

```
# -*- coding:utf-8 -*-
import os
from typing import Union, List
from pet_config import *
pc = ProjectConfig()
```

common_utils.py

目的：定义损失函数、将mask_position位置的token logits转换为token的id。

脚本里面包含两个函数：mlm_loss()以及convert_logits_to_ids()

部分代码展示：
具体参考代码
文件

```
# coding:utf-8
# 导入必备工具包
import torch
from rich import print
```

metirc_utils.py

目的：定义（多）分类问题下的指标评估（acc, precision, recall, f1）。

定义ClassEvaluator类

部分代码展示：
具体参考代码
文件

```
from typing import List


import numpy as np
import pandas as pd
from sklearn.metrics import accuracy_score, precision_score,
f1_score
from sklearn.metrics import recall_score, confusion_matrix
```

02

实现模型训练函数，验证函数

简介

01

目的

实现模型的训练和验证

02

代码路径

/Users/**/PycharmProjects/llm/prompt_tasks/PET/train.py

03

脚本

脚本里面包含两个函数：model2train()和evaluate_model()

代码实现

定义model2train()函数

定义evaluate_model函数

部分代码展示：
具体参考代码文件

```
import os
import time
from transformers import AutoModelForMaskedLM, AutoTokenizer, get_scheduler
from pet_config import *
import sys
sys.path.append('/Users/ligang/PycharmProjects/llm/prompt_tasks/PET/data_handle')
sys.path.append('/Users/ligang/PycharmProjects/llm/prompt_tasks/PET/utils')
from utils.metirc_utils import ClassEvaluator
from utils.common_utils import *
from data_handle.data_loader import *
from utils.verbalizer import Verbalizer
from pet_config import *
pc = ProjectConfig()
```

代码实现

模型训练结果部分展示：

```
.....
global step 40, epoch: 4, loss: 0.62105, speed: 1.27 step/s
global step 50, epoch: 6, loss: 0.50076, speed: 1.23 step/s
global step 60, epoch: 7, loss: 0.41744, speed: 1.23 step/s
...
global step 390, epoch: 48, loss: 0.06674, speed: 1.20 step/s
global step 400, epoch: 49, loss: 0.06507, speed: 1.21 step/s
Evaluation precision: 0.78000, recall: 0.76000, F1: 0.75000
```

结论：BERT+PET模型在训练集上的表现是精确率=78%

注意：本项目中只用了60条样本，在接近600条样本上精确率就已经达到了78%，如果想让指标更高，可以扩增样本。

03

实现模型预测函数

代码介绍

目的

加载训练好的模型并测试效果

代码路径

/Users/**/PycharmProjects/llm/prompt_tasks/PET/inference.py

代码实现

定义inference()函数

部分代码展示：
具体参考代码
文件

```
import time
from typing import List

import torch
from rich import print
from transformers import AutoTokenizer, AutoModelForMaskedLM
import sys
sys.path.append('/Users/**/PycharmProjects/llm/prompt_tasks/PET/data_handle')
sys.path.append('/Users/**/PycharmProjects/llm/prompt_tasks/PET/utils')
from utils.verbalizer import Verbalizer
from data_handle.template import HardTemplate
from data_handle.data_preprocess import convert_example
from utils.common_utils import convert_logits_to_ids
```

代码实现

结果展示

{
'天台很好看，躺在躺椅上很悠闲，因为活动所以我觉得性价比还不错，适合一家出行，特别是去迪士尼也蛮近的，下次有机会肯定还会再来的，值得推荐'：'酒店'，
'环境，设施，很棒，周边配套设施齐全，前台小姐姐超级漂亮！酒店很赞，早餐不错，服务态度很好，前台美眉很漂亮。性价比超高的一家酒店。强烈推荐'：'酒店'，
'物流超快，隔天就到了，还没用，屯着出游的时候用的，听方便的，占地小'：'平板'，
'福行市来到无早集市，因为是喜欢的面包店，所以跑来集市看看。第一眼就看到了，之前在微店买了小刘，这次买了老刘，还有一直喜欢的巧克力磅蛋糕。好奇老板为啥不做柠檬磅蛋糕了，微店一直都是买不到的状态。因为不爱碱水硬欧之类的，所以期待老板多来点其他小点，饼干一直也是大爱，那天好像也没看到'：'水果'，
'服务很用心，房型也很舒服，小朋友很喜欢，下次去嘉定还会再选择。床铺柔软舒适，晚上休息很安逸，隔音效果不错赞，下次还会来'：'酒店'
}

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