

# 预训练模型学习情况周报10

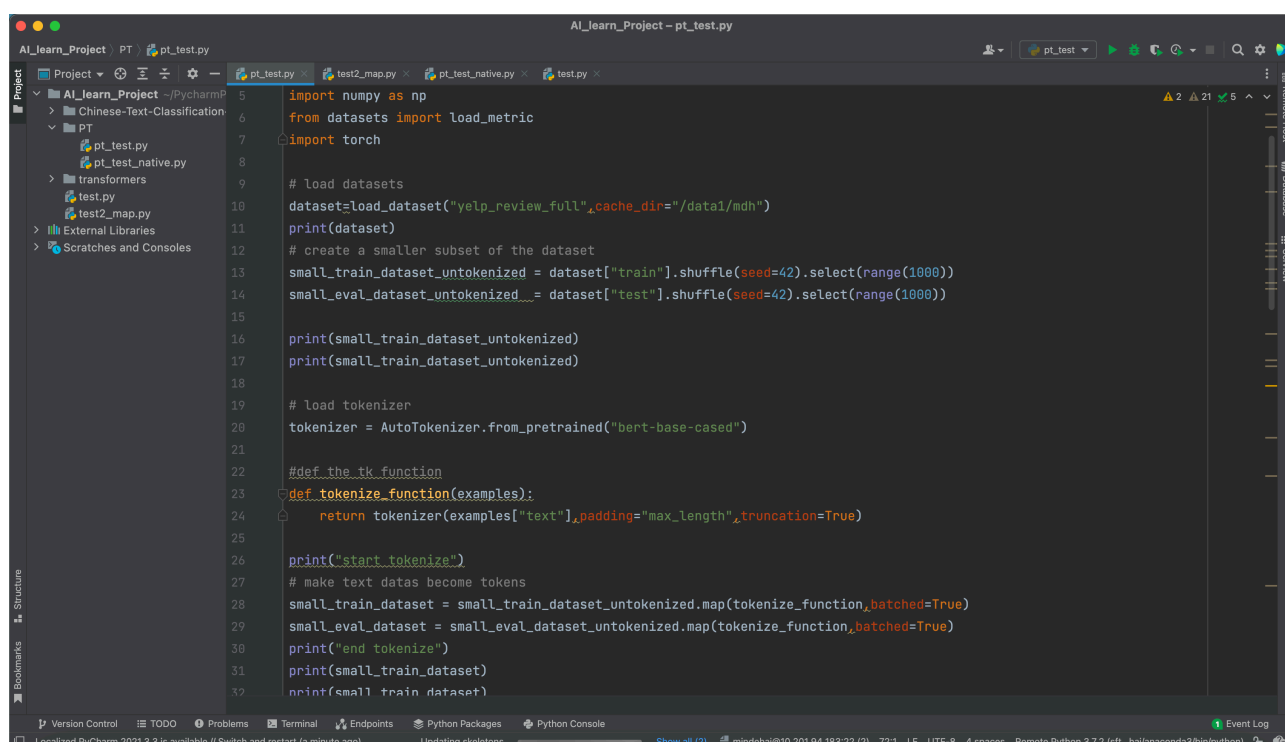
汇报人：闵德海

## 本周主要学习内容：

本周重点学习了使用HuggingFace 中的AutoModel、AutoTokenizer、datasets 去用Task-based 数据集Fine-tune 已有的预训练模型。

并整理了该小节的学习笔记在个人博客上：[预训练模型的使用-实战笔记](#)

同时阅读了huggingface平台上Fine-tune a pretrained model 、Auto Classes，并学习了使用HF库中的list\_metrics去做Evaluate predictions。



```
AI_Learn_Project - pt_test.py
5 import numpy as np
6 from datasets import load_metric
7 import torch
8
9 # load datasets
10 dataset=load_dataset("yelp_review_full",cache_dir="/data1/mdh")
11 print(dataset)
12 # create a smaller subset of the dataset
13 small_train_dataset_untokenized = dataset["train"].shuffle(seed=42).select(range(1000))
14 small_eval_dataset_untokenized_ = dataset["test"].shuffle(seed=42).select(range(1000))
15
16 print(small_train_dataset_untokenized)
17 print(small_train_dataset_untokenized)
18
19 # load tokenizer
20 tokenizer = AutoTokenizer.from_pretrained("bert-base-cased")
21
22 #def the tk function
23 def tokenize_function(examples):
24     return tokenizer(examples["text"],padding="max_length",truncation=True)
25
26 print("start tokenize")
27 # make text datas become tokens
28 small_train_dataset = small_train_dataset_untokenized.map(tokenize_function,batched=True)
29 small_eval_dataset = small_eval_dataset_untokenized.map(tokenize_function,batched=True)
30 print("end tokenize")
31 print(small_train_dataset)
32 print(small_train_dataset)
```

目前进度如下，红色对勾为当前完成的任务，圆圈代表正在进行中的任务。

