自然语言处理大作业

环境配置 ——OS:Window11

注意！！！项目目录不能有中文路径！！！

Sentence-bert

python=3.9.18 torch=2.0

首先使用conda创建虚拟环境py39pt20并激活,然后下载主目录下的requirements中的包

```
conda create -n py39pt20 python==3.9
conda activate py39pt20
pip install -r requirements.txt -i https://pypi.tuna.tsinghua.edu.cn/simple
```

SimCSE

python=3.9.18 torch=2.0

```
conda install pytorch==2.0.1 torchvision==0.15.2 torchaudio==2.0.2 pytorch-cuda=11.8 -c pytorch -c nvidia
pip install jsonlines -i https://pypi.tuna.tsinghua.edu.cn/simple
pip install loguru -i https://pypi.tuna.tsinghua.edu.cn/simple
pip install scipy -i https://pypi.tuna.tsinghua.edu.cn/simple
pip install tqdm -i https://pypi.tuna.tsinghua.edu.cn/simple
pip install transformers -i https://pypi.tuna.tsinghua.edu.cn/simple
```

STT&TTS

python=3.9.18 torch=2.0

```
pip install pyaudio -i https://pypi.tuna.tsinghua.edu.cn/simple
pip install baidu-aip -i https://pypi.tuna.tsinghua.edu.cn/simple
pip install chardet -i https://pypi.tuna.tsinghua.edu.cn/simple
pip install pygame -i https://pypi.tuna.tsinghua.edu.cn/simple
#注意使用时关闭代理
```

(应该是所有的包了，如有缺少，自行conda/pip install即可)

项目说明

- exp：实验结果记录文件，保存了实验过程截图，同时录制了STT&TTS的demo视频，以及配置环境运行项目的视频
- data: 处理后的数据
  - jieba_data_exp_10_17 :10月17日进行最后一次修改的数据，为数据清洗后的jieba分词的结果，各个csv文件的含义见命名。
  - data_final_exp_10_29:10月29日进行最后一次修改的数据，为最终实验中使用的数据。
    - label_exp_train_10_29:清洗后的训练集

        - label_exp_dev_10_29:清洗后的评估集
        - label_exp_test_10_29:清洗后的测试集
        - dev_data_F1:用于得到F1得分的数据，其中test_final.csv为本次任务预测的结果
        - result:最终提交结果
    - origin_data：原始数据
    - process:数据预处理的相关python脚本文件
        - csv_to_json:csv文件转为json文件
        - clean：数据清洗函数
        - process_json：清洗json文件保存为csv文件
        - process_xlsx:清洗xlsx文件保存为csv文件
        - （-）dev&test_label_generate:生成dev和test负样例
        - （-）train_label_generate:生成train的负样例
- origin_data：原始数据
- data_pre_process_10_17: 数据预处理，对比了一下是否去除停用词的效果
- pytorch：学习pytorch的框架，做了一些练习，记录了一下
- lab01-04: 借鉴的部分开源项目代码
- STT&TTS：语音模块
    - STT：语音转文字
    - TTS：文字转语音
    - predict：运行训练好的simcse模型，得到预测的结果并返回TTS模块
- test01:SimCSE模块
    - chinese_roberta_wwm_ext_pytorch: 存放预训练的chinese_roberta_wwm_ext_pytorch模型
    - clean:数据清洗模块
    - config.yaml:超参数配置文件，可以配置文件路径
    - data_load:数据读取以及装入到dataset
    - eval:评估函数，数据来自data_process构建的验证集
    - main:训练主函数
    - model:定义SimCSE模型
    - predict:进行test数据集的预测
    - run.sh:脚本文件
- test02 : sentence-bert模块
    - s-bert-finetuning:微调s-bert模型
    - s-bert：运行s-bert模型进行评估
    - result：将预测结果写入test.csv文件
- test03 : word2vec模块（最后舍弃了）
- demo：语音展示、环境配置展示、项目运行展示

demo命令

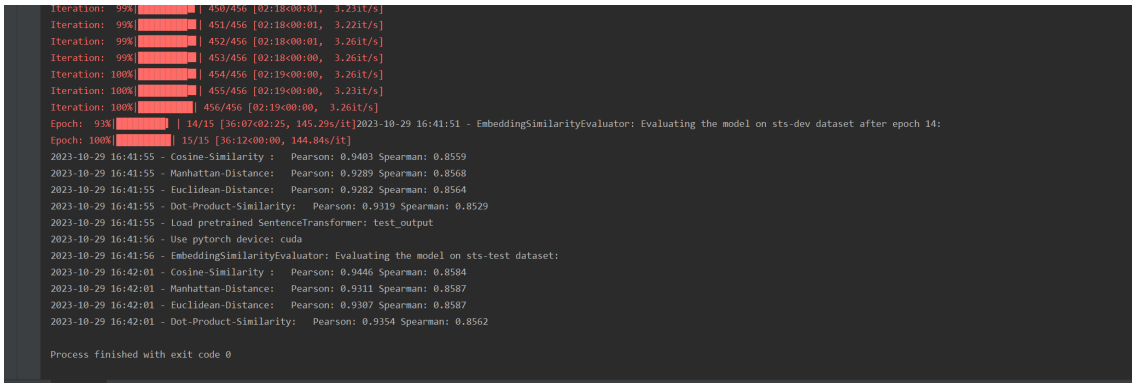
- STT&TTS

```
conda activate py39pt20
cd STT&TTS
python TTS.py
```

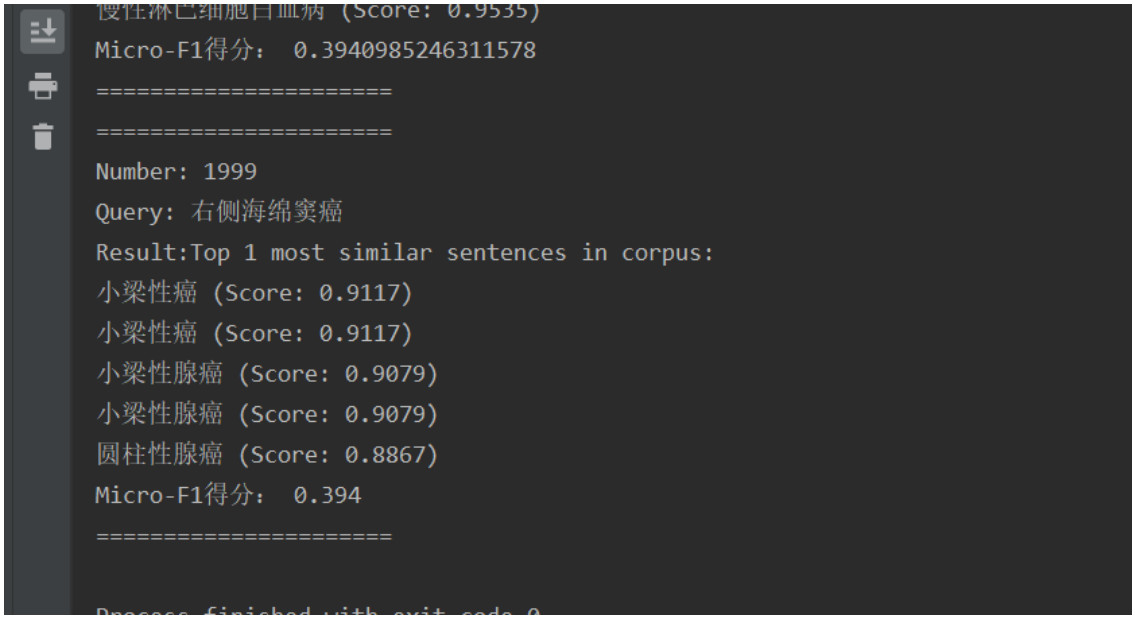
  运行后，您可以等待提示回车录制您想要标准化的文本输入，稍等片刻，会得到标准化匹配后的语音回复。

- Sentence-bert

```
conda activate py39pt20
cd test02
python s-bert-finetuning.py #开始微调
python s-bert.py #加载微调模型dev数据集上的预测并给出F1得分
python result.py #生成test.csv
cd ..
cd data
cd process
python csv_to_json.py#记得修改路径
```



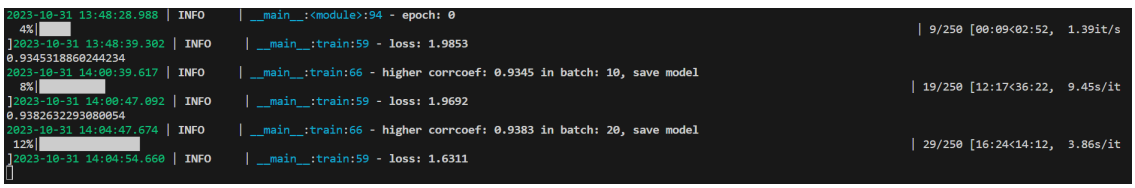
图示为训练过程



图示为评估过程

- SimCSE

```
conda activate py39pt20
cd test01
bash run.sh
```



图示为训练过程，评估过程没来得及做完

- word2vec（这个并没有用到，是一开始想用的，后来发现无监督的效果不好，舍弃了，不过自己写了一个shell脚本，学了一下shell脚本的编写，学习了一下yaml文件使用，日志文件在output目录下，感兴趣可以运行）

```
cd test03
cd model_exp_word2vec_10_20
bash run.sh
```

参考开源仓库：

https://github.com/princeton-nlp/SimCSE.git

https://github.com/vdogmcgee/SimCSE-Chinese-Pytorch

https://github.com/DataArk/CHIP2021-Task3-Top3.git

https://github.com/Lisennlp/two_sentences_classifier.git