



博学谷

www.boxuegu.com

黑马程序员线上品牌

基于GPT2 搭建医疗问诊机器人

一样的教育，不一样的品质





目录

Contents

1. 项目简介
2. 数据介绍
3. 数据处理
4. 模型搭建
5. 模型训练和验证
6. 模型预测（人机交互）

01

项目简介

项目背景

目标

聊天机器人是一种基于自然语言处理技术的智能对话系统，能够模拟人类的自然语言交流，与用户进行对话和互动。聊天机器人能够理解用户的问题或指令，并给出相应的回答或建议。其目标是提供友好、智能、自然的对话体验。

应用

当前，聊天机器人在多个领域得到广泛应用。首先，它们常用于在线客服系统，能够快速、准确地回答用户的常见问题，解决疑问。其次，聊天机器人可以作为个人助手，提供个性化的推荐、建议和日程安排等服务，提升用户体验。此外，聊天机器人还被应用于社交娱乐、语言学习、旅游指南等领域，为用户提供有趣、便捷的对话体验。

项目背景

常见的相关聊天机器人产品



微软小冰

微软公司开发。它具备自然语言处理、情感分析和对话生成等功能，能够与用户进行智能对话，提供情感支持和娱乐等服务。



阿里云小蜜

阿里云公司推出，提供了丰富的智能对话服务。它具备自然语言处理和对话管理能力，支持多领域的应用场景，如在线客服、智能助手和虚拟导购等。



百度智能云小度

百度智能云开发，提供了多领域的智能对话能力。小度机器人可应用于家庭助理、智能音箱和移动应用等场景，通过语音和文本交互与用户进行智能对话，提供信息查询、音乐播放和日程安排等功能。

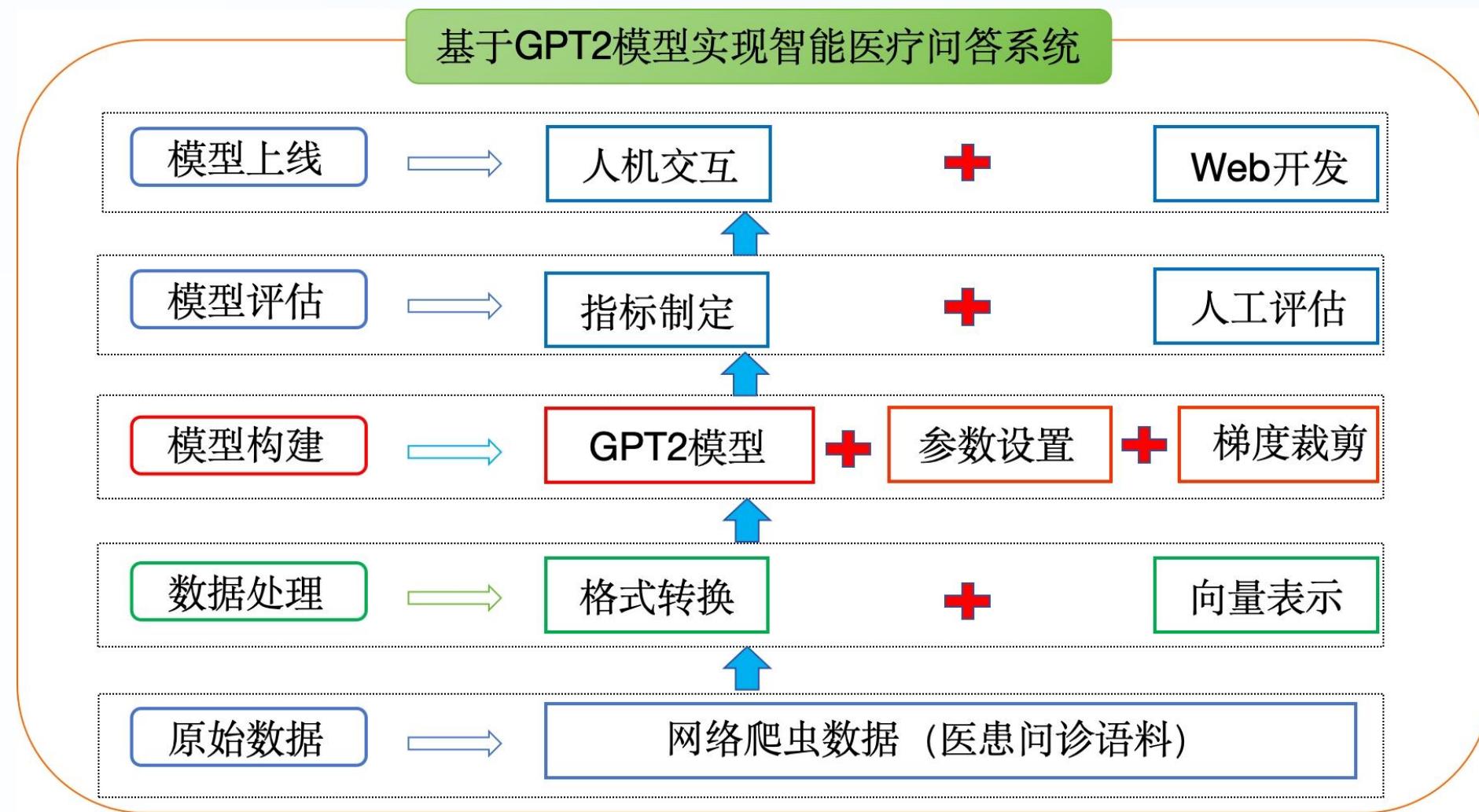
本项目基于医疗领域数据构建了智能医疗问答系统，目的是为用户提供准确、高效、优质的医疗问答服务。

环境准备

- python>=3. 8
- transformers>=4. 2. 0
- pytorch>=1. 7. 0



项目整体结构



02

数据介绍

数据介绍

数据存放位 置

数据存放位置：
/Users/**/PycharmPr
ojects/11m/Gpt2_Cha
tbot/data

data文件夹里面包含2
个medical_train.txt
medical_valid.txt

txt文档

数据介绍

数据示例：

帕金森叠加综合征的辅助治疗有些什么？

综合治疗；康复训练；生活护理指导；低频重复经颅磁刺激治疗

卵巢癌肉瘤的影像学检查有些什么？

超声漏诊；声像图；MR检查；肿物超声；术前超声；CT检查

数据格式：每段医患对话之间用\n隔开，一共约3w对聊天记录

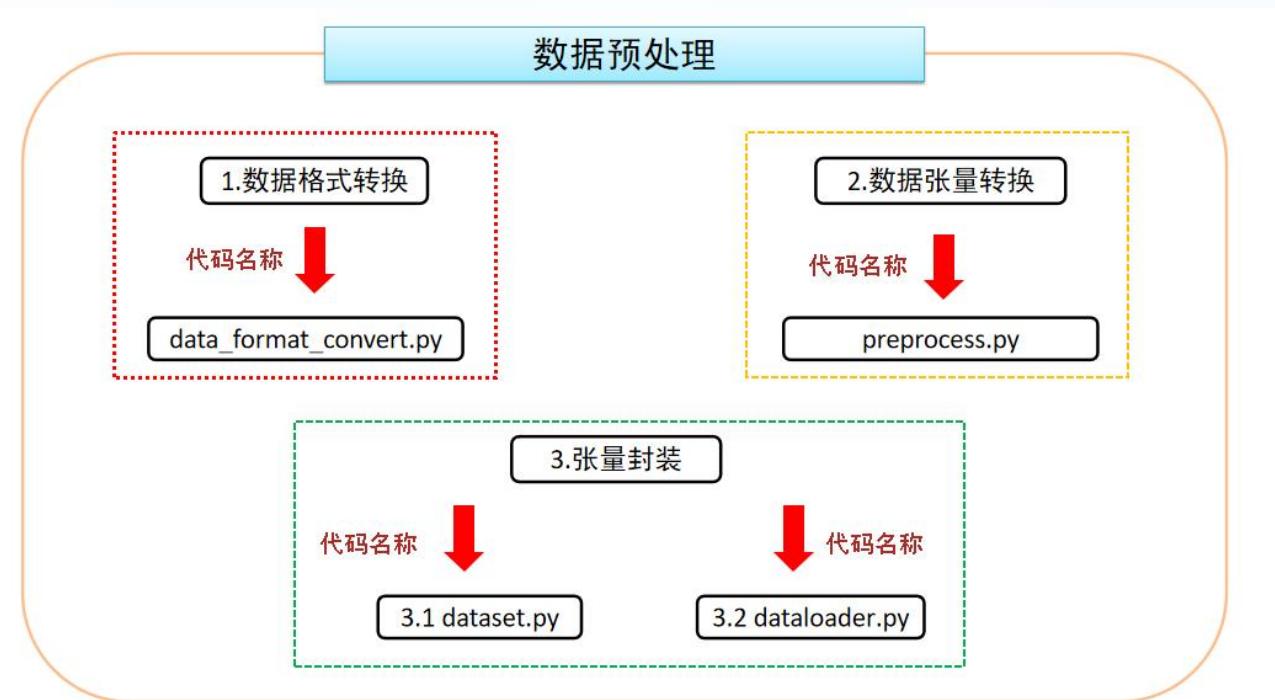
03

数据处理

数据处理简介

目的：将中文文本数据处理成模型能够识别的张量（数字）形式。

基本流程如下：



注意：我们这里省略的了数据格式的转换，目的，将非txt的文档转换为txt的形式，并符合上述数据格式要求

代码路径

数据处理脚本: /Users/user/PycharmProjects/11m/Gpt2_Chatbot/data_preprocess/preprocess.py

构建dataset脚本: /Users/user/PycharmProjects/11m/Gpt2_Chatbot/data_preprocess/dataset.py

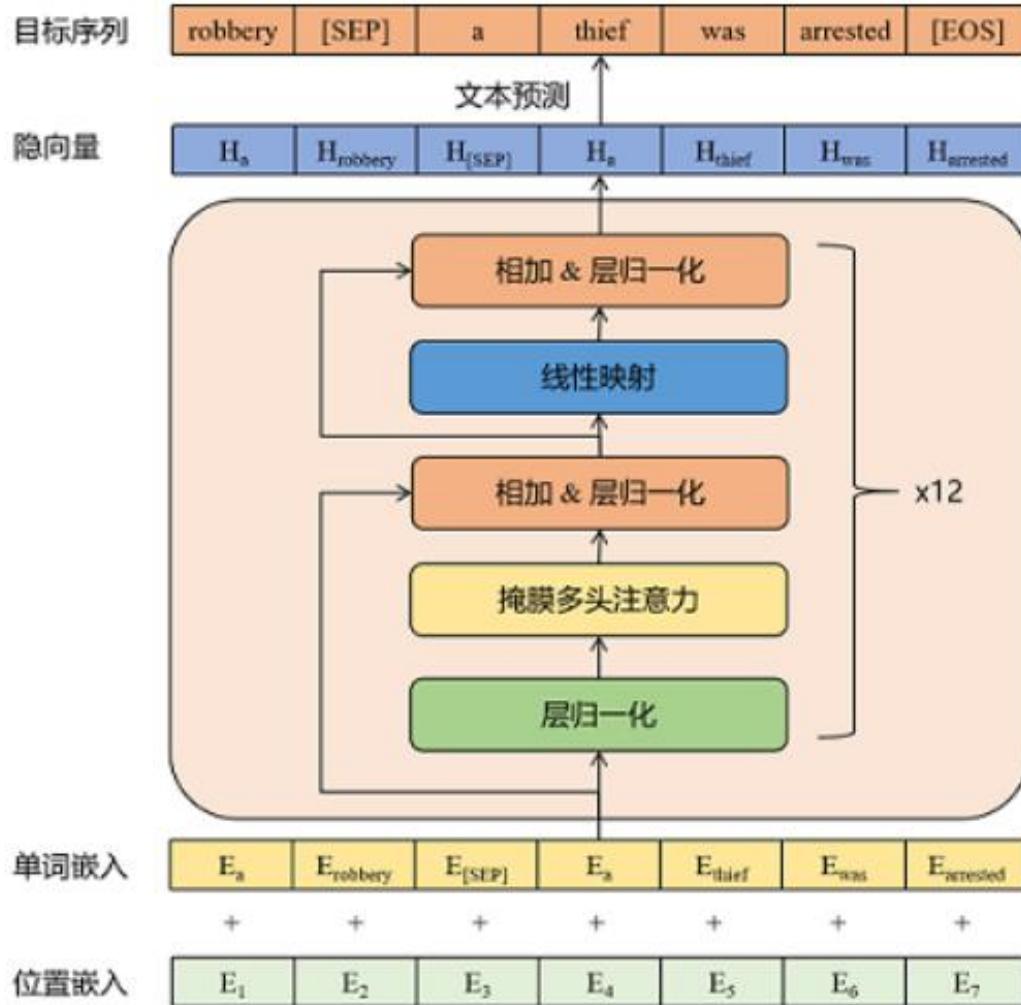
构建dataloader脚本: /Users/user/PycharmProjects/11m/Gpt2_Chatbot/data_preprocess/dataloader.py

注意: 具体代码实现参考附属资料

04

模型搭建

模型架构介绍



模型架构解析:

- **输入层:** 词嵌入层: WordEmbedding +位置嵌入层: PositionEmbedding
- **中间层:** Transformer的Decoder模块---12层
- **输出层:** LayerNorm层+线性全连接层

模型主要参数简介(详见模型的config.json文件):

- n_embd: 768
- n_head: 12
- n_layer: 12
- n_positions: 1024
- vocab_size: 21128/13317 (可以选择和 config.json对齐)

GPT2模型准备

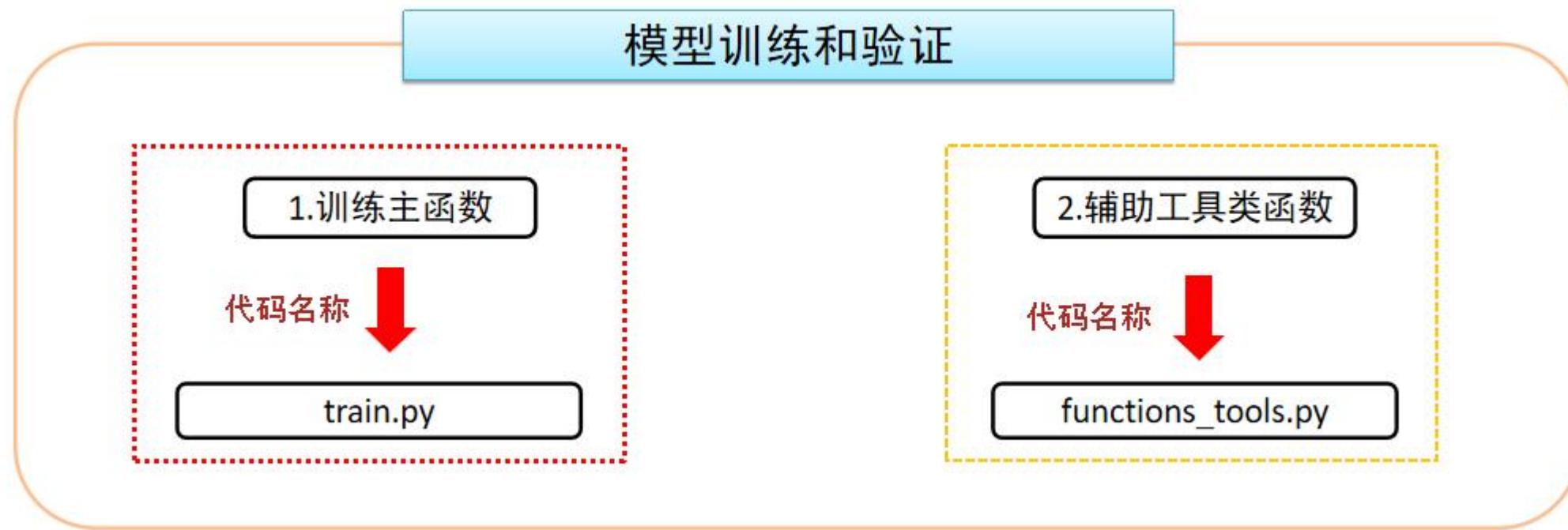
本次项目使用GPT2的预训练模型，因此不需要额外搭建Model类，下面代码是如何直接加载使用GPT2预训练模型。

```
from transformers import GPT2LMHeadModel, GPT2Config
# 创建模型
if params.pretrained_model:
    # 加载预训练模型
    model =
        GPT2LMHeadModel.from_pretrained(params.pretrained_model)
else:
    # 初始化模型
    model_config =
        GPT2Config.from_json_file(params.config_json)
    model = GPT2LMHeadModel(config=model_config)
```

05

模型训练和验证

主要代码



代码路径

训练脚本: /home/user/ProjectStudy/Gpt2_Chatbot/data_preprocess/train.py

辅助工具类脚本: /home/user/ProjectStudy/Gpt2_Chatbot/data_preprocess/functions_tools.py

注意: 具体代码实现参考附属资料

06

模型预测（人机交互）

模型预测

运行 interact.py，使用训练好的模型，进行人机交互，输入Ctrl+Z结束对话之后，聊天记录将保存到sample目录下的 sample.txt 文件中。

脚本路径：/Users/user/PycharmProjects/11m/Gpt2_Chatbot/interact.py

基于flask框架，实现web界面的交互。

脚本路径：/Users/user/PycharmProjects/11m/Gpt2_Chatbot/flask_predict.py

注意：具体代码实现参考附属资料



黑马程序员线上品牌

Thanks !



扫码关注博学谷微信公众号

