博学谷
www.boxuegu.com
黑马程序员线上品牌

新媒体行业文本评论分类与信息抽取系统

一样的教育，不一样的品质

博学谷
www.boxuegu.com
黑马程序员线上品牌

目录

Contents

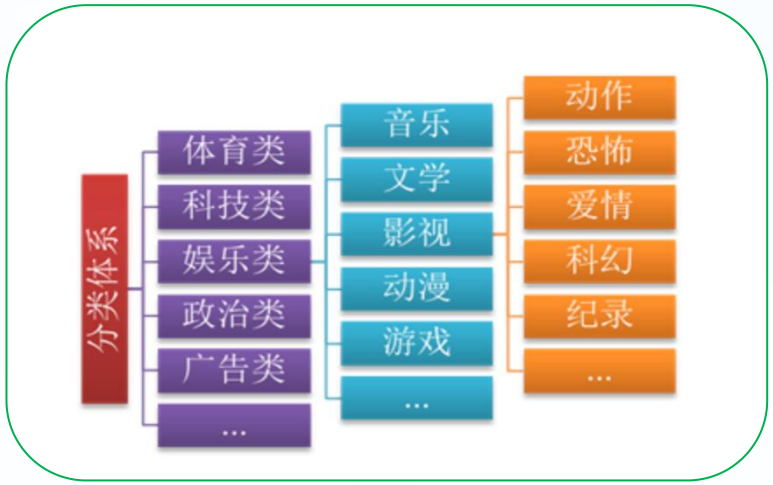
博学谷
www.boxuegu.com
黑马程序员线上品牌

01

项目简介

项目背景

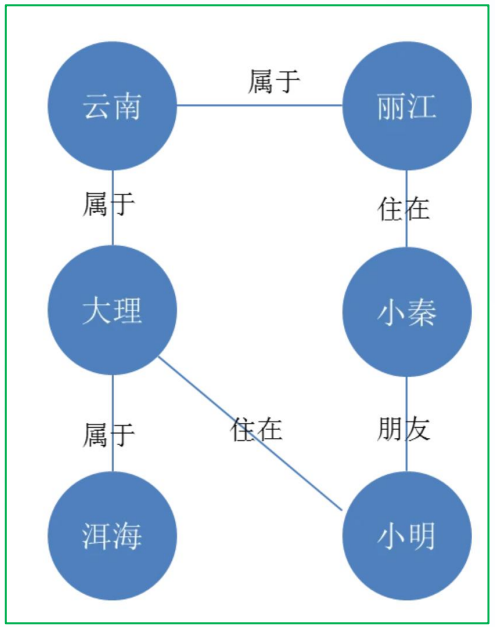
文本分类概念：



- 文本分类是指将一段或多段文本按照其内容或主题特征划分到不同的类别或标签中的过程。
- 在实际工作中文本分类应用非常广泛，比如：新闻分类、简历分类、邮件分类、办公文档分类、区域分类等诸多方面，还能够实现文本过滤，从大量文本中快速识别和过滤出符合特殊要求的信息。

项目背景

信息抽取概念：



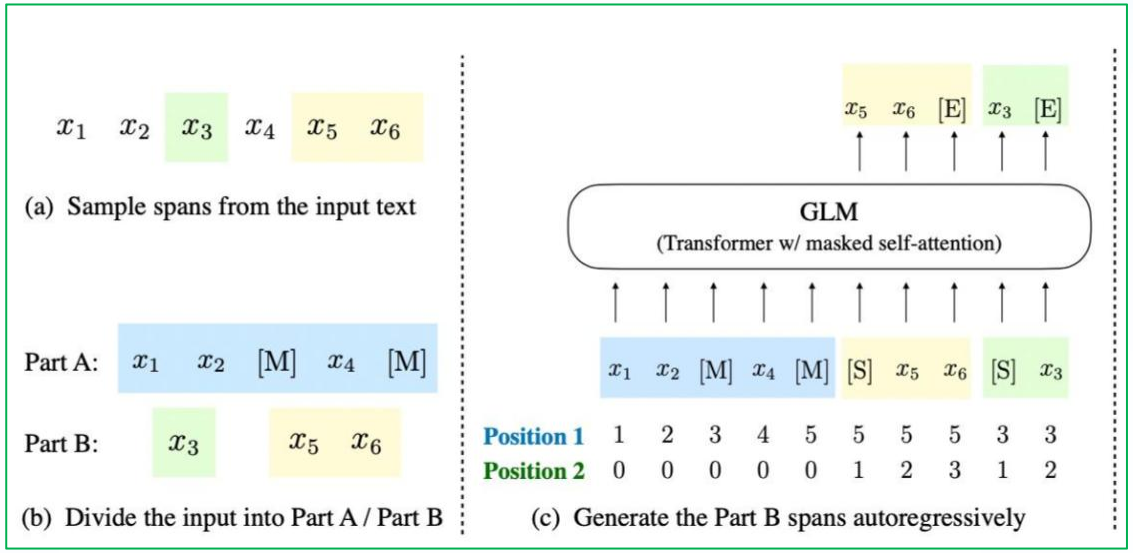
- 信息抽取，是从无结构或半结构化的自然文本中识别出实体、关系、事件等事实描述，以结构化的形式存储和利用的技术。
- 以“小明和小秦是很好的朋友，他们都属于云南人，小明住在大理，小秦住在丽江。”为例，可以得到如：<小明，朋友，小秦>和<小秦，住在，丽江>和<小明，住在，大理>等三元组信息。

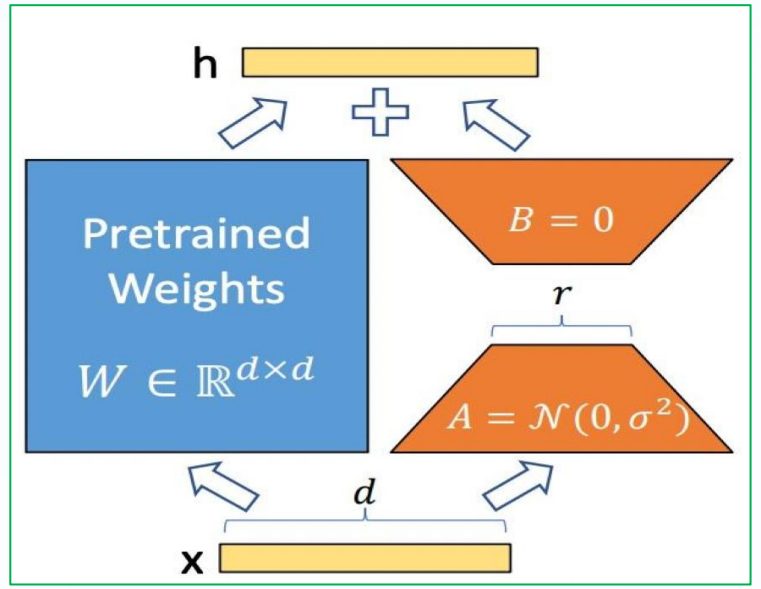
项目背景

随着互联网技术的快速发展，新媒体行业已经成为信息传播的主要平台之一。在这个信息爆炸的时代，人们通过社交媒体、新闻客户端、博客等多种形式获取信息。然而，随着信息量的不断增加，如何高效地管理和利用这些信息成为了亟待解决的问题。

本项目基于部分“新媒体行业”数据为背景，通过文本评论的分类和信息抽取，帮助新媒体行业从海量的信息中快速准确地获取有用的信息，并进行合理的分类和管理。这不仅有助于新媒体平台提升用户体验，还能够为信息生产者提供更精准的数据分析和决策支持。

技术选型





基于ChatGLM-6B模型+LoRA微调方法，实现文本分类及信息抽取的联合任务的开发

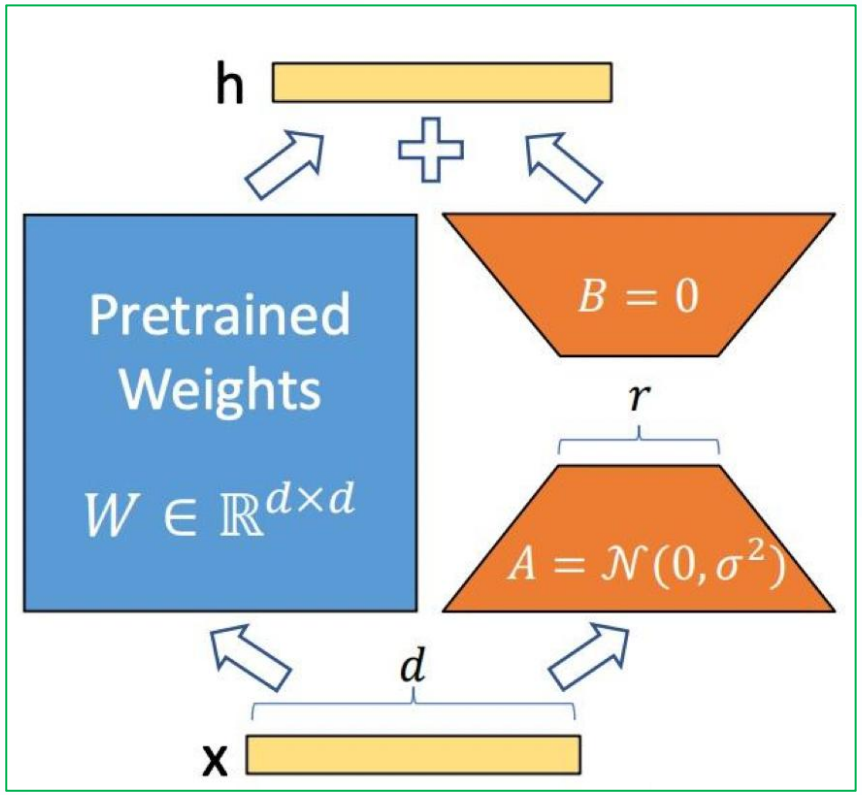
ChatGLM-6B回顾

ChatGLM-6B 是清华大学提出的一个开源、支持中英双语的对话语言模型，基于 General Language Model (GLM) 架构，具有 62 亿参数。该模型使用了和 ChatGPT 相似的技术，经过约 1T 标识符的中英双语训练(中英文比例为 1:1)，辅以监督微调、反馈自助、人类反馈强化学习等技术的加持，62 亿参数的 ChatGLM-6B 已经能生成相当符合人类偏好的回答（目前中文支持较好）。

相比原始Decoder模块，模型结构有如下改动点：

embedding 层梯度缩减	为了提升训练稳定性，减小了 embedding 层的梯度。梯度缩减的效果相当于把 embedding 层的梯度缩小了 10 倍，减小了梯度的范数.
layer normalization	采用了基于 Deep Norm 的 post layer norm.
激活函数	替换ReLU激活函数采用了 GeGLU 激活函数.
位置编码	去除了绝对位置编码，采用了旋转位置编码 RoPE.

LoRA 原理回顾



LoRA技术冻结预训练模型的权重，并在每个Transformer块中注入可训练层（称为秩分解矩阵），即在模型的Linear层的旁边增加一个“旁支”A和B。其中，A将数据从d维降到r维，这个r是LoRA的秩，是一个重要的超参数；B将数据从r维升到d维，B部分的参数初始为0。模型训练结束后，需要将A+B部分的参数与原大模型的参数合并在一起使用

环境准备

本次环境依赖于趋动云https://platform.virtaicloud.com/算力

- 操作系统：CentOS 7
- CPUs：8 core(s)，内存：48G
- GPUs：1卡， A800， 80GB GPUs
- Python：3.9
- Pytorh：1.11.0
- Cuda：11.3.1
- 价格：13.58元/小时

环境准备

1. 创建一个虚拟环境，您可以把 `llm_env` 修改为任意你想要新建的环境名称：

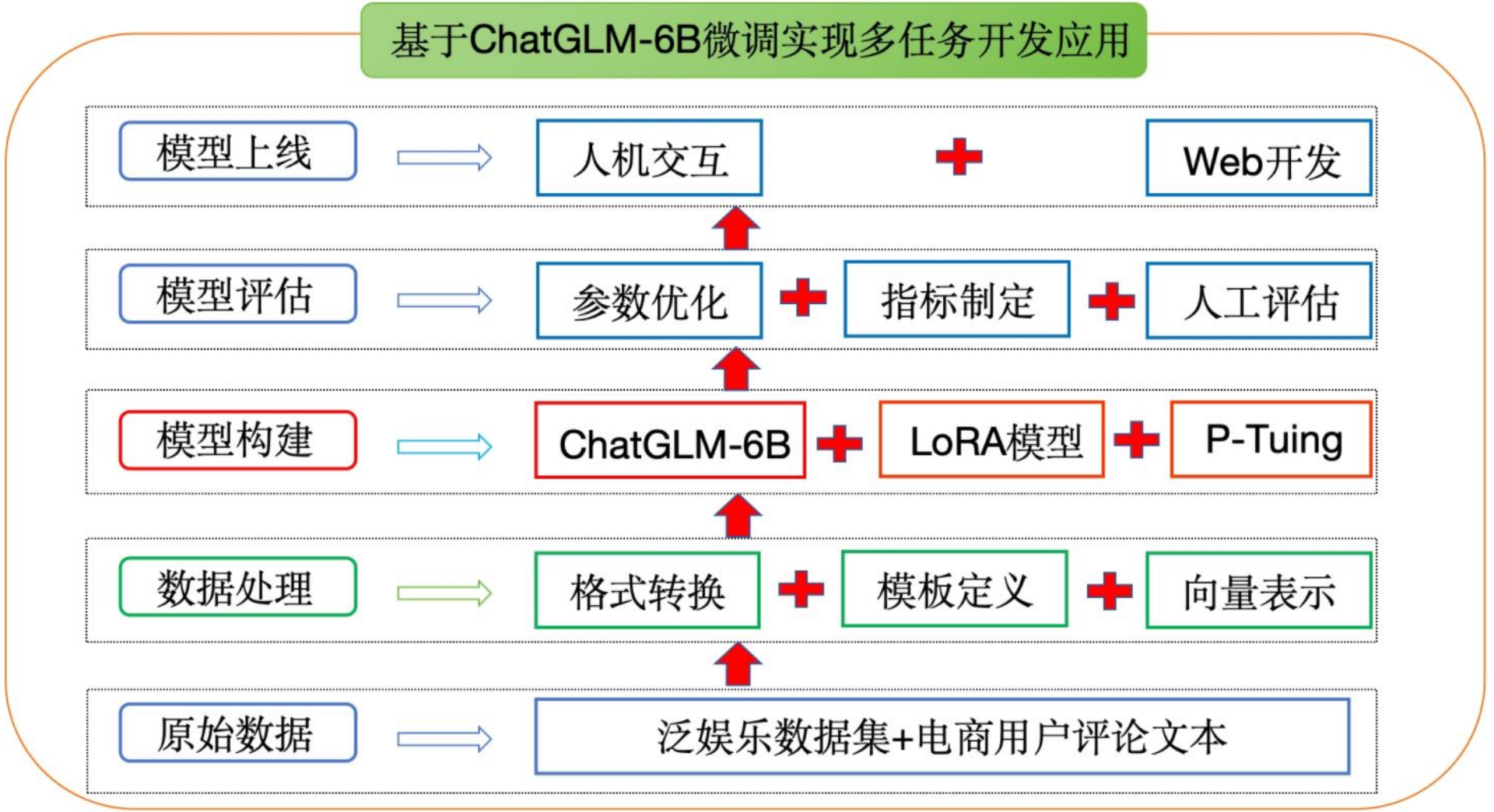
```
conda create -n llm_env python=3.9
```

2. 激活新建虚拟环境并安装响应的依赖包：

```
conda activate llm_env
pip install -r requirements.txt（requirements.txt文件内容如下所示）
```

- protobuf>=3.19.5,<3.20.1
- transformers>=4.27.1
- icetk
- cpm_kernels
- streamlit==1.17.0
- matplotlib
- datasets==2.10.1
- accelerate==0.17.1
- packaging>=20.0
- psutil
- pyyaml
- peft

项目整体结构


基于ChatGLM-6B微调实现多任务开发应用
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
ChatGLM-6B
LoRA模型
P-Tuing
数据处理
格式转换
模板定义
向量表示
原始数据
泛娱乐数据集+电商用户评论文本

博学谷
www.boxuegu.com
黑马程序员线上品牌

02

数据介绍

数据介绍

数据存放位置

数据存放位置：
/Users/**/PycharmProjects/llm/ptune_chatglm/data

data文件夹里面包含
mixed_dev_dataset.jsonl
mixed_train_dataset.jsonl

jsonl文档

数据介绍

数据示例(信息抽取)：

{"context": "Instruction: 你现在是一个很厉害的阅读理解器，严格按照人类指令进行回答。\nInput: 帮我提取出下面句子中所有的SPO，并输出为json，不要做多余的回复：\n\n张国荣主演电影大片《霸王别姬》荣获法国戛纳国际电影节最高奖项金棕榈大奖，成为首部获此殊荣的中国影片。\nAnswer: ", "target": "```json\n[{\"predicate\": \"主演\", \"object_type\": \"人物\", \"subject_type\": \"影视作品\", \"object\": \"张国荣\", \"subject\": \"霸王别姬\"}]\n```"}

数据格式：字典样式；context内容代表：原始输入文本（prompt）；target指向：目标文本

数据介绍

数据示例(文本评论分类)：

{"context": "Instruction: 你现在是一个很厉害的阅读理解器，严格按照人类指令进行回答。\nInput: 下面句子中的主语是什么类别，输出成列表形式。\n\n朋友说京东水果新鲜，我就买了，真的很失望，不新鲜，下次不会再买了\nAnswer: ",
"target": "[\"水果\"]"}

数据格式：字典样式；context内容代表：原始输入文本（prompt）；target指向：目标文本

数据介绍

训练数据集：mixed_train_dataset.jsonl

共计包含：902条样本

验证数据集：mixed_dev_dataset.jsonl

共计包含：122条样本

博学谷
www.boxuegu.com
黑马程序员线上品牌

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。

数据处理脚本：/Users/user/PycharmProjects/llm/ptune_chatglm/data_handle/data_preprocess.py

构建dataloader脚本：/Users/user/PycharmProjects/llm/ptune_chatglm/data_handle/data_loader.py

具体代码详情看附件资料代码

博学谷
www.boxuegu.com
黑马程序员线上品牌

04

模型搭建

ChatGLM-6B模型准备

本次项目使用ChatGLM-6B的预训练模型，因此不需要额外搭建Model类.

```
tokenizer = AutoTokenizer.from_pretrained(pc.pre_model, trust_remote_code=True)

config = AutoConfig.from_pretrained(pc.pre_model, trust_remote_code=True)

if pc.use_ptuning:
    config.pre_seq_len = pc.pre_seq_len
    config.prefix_projection = pc.prefix_projection
model = AutoModel.from_pretrained(pc.pre_model,
                                  config=config,
                                  trust_remote_code=True)
```

具体代码详情看附件资料代码

博学谷
www.boxuegu.com
黑马程序员线上品牌

05

模型训练和验证

代码路径

训练脚本：/Users/user/PycharmProjects/llm/ptune_chatglm/train.py

辅助工具类脚本：/Users/user/PycharmProjects/llm/ptune_chatglm/utils/common_utils.py

具体代码详情看附件资料代码

博学谷
www.boxuegu.com
黑马程序员线上品牌

06

模型预测（人机交互）

模型预测

预测脚本：/Users/user/PycharmProjects/llm/ptune_chatglm/inference.py

具体代码详情看附件资料代码

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