武彪

男 | 年龄：23岁 | 13028007779 | 13834377890@163.com
求职意向：算法工程师 | 期望薪资：22-25K | 期望城市：北京

个人优势

熟悉 LLM 算法原理与应用，掌握 RAG（检索增强生成）、Agent 构建、Function Call 开发等关键技术，具备基于 LangChain 框架的开发实战经验。
精通多类型文档的向量化处理流程，熟练使用 Milvus 向量数据库进行向量数据建模与检索，具备 Zilliz 云端部署与调优经验。
掌握主流大语言模型的微调方法，包括 P-Tuning、LoRA、QLoRA 等，熟悉 RLHF、Prompt Engineering 等下游任务对齐策略。
具备知识图谱构建与应用经验，熟悉实体识别、关系抽取等核心技术，掌握 Neo4j 图数据库的使用与查询优化。
熟悉 Agent 与 Workflow 的搭建流程，能够快速基于现有框架解决实际业务问题，实现项目快速迭代与落地。
熟悉 Transformer 架构及 Attention 机制的原理与实现，掌握 CNN、RNN、LSTM、GRU 等深度学习模型的应用场景与实现方式。
熟悉常见机器学习算法，包括决策树、随机森林、GBDT、XGBoost 等，具备从特征工程到模型评估的全流程建模能力。
掌握模型压缩技术，如量化、剪枝、蒸馏等，能够在保证性能的前提下实现轻量化部署。
精通 Python 编程语言，熟练使用 NumPy、Pandas、Matplotlib 等库进行数据清洗、分析与可视化处理。
熟悉 BERT、GPT、ChatGLM 等主流预训练模型的结构与原理，具备文本分类、意图理解、对话问答、相似度匹配、机器翻译等 NLP 任务的实战经验。
熟练使用 PyTorch、Hugging Face Transformers、FastText、Ollama 等工具进行模型训练与部署，掌握 Coze、RAGFlow、Dify 等 RAG 工具链，具备从文档预处理、向量化到检索增强生成的全流程开发经验。
擅长基于 FastAPI 构建高性能异步服务接口，熟悉 Flask 框架及 Docker 容器化部署流程，具备微服务架构设计与部署经验。
熟练使用 Linux 操作系统，掌握常用命令与脚本编写，具备良好的环境配置与调试能力。
英语 CET-4，具备良好的英文读写能力，可阅读官方英文文档与论文，具备一定的英文技术交流能力。

项目经历

基于Graph+RAG的医疗知识问答系统　算法工程师　2024.02-2025.02

内容:

项目背景:
针对医疗问答中知识分散、表述多样的问题,通过信息抽取构建医学知识图谱,结合RAG实现结构化知识与自由文本的混合检索与生成,融合NER,关系抽取,语义匹配与生成优化等NLP技术,有效提升问答系统在医疗领域的专业性与语言理解能力
项目流程:
一、知识图谱
1.数据获取数据部门提供的数据,经过标注后共获得样本30w条左右
2.信息抽取
(1)Pipline方式
实体识别:使用BiLSTM+CRF模型,准确率95.23%
关系抽取:使用BiLSTM+Attention模型,准确率89.54%,
(2)Joint方式使用CasRel模型,基于参数共享方式直接从文本提取出SPO三元组
关系抽取准确率92.62%。
实现了5种实体类型,8w条边
3.知识融合
基于TE-IDE相似度计算实现实体消歧

基于规则实现实体统一、基于同义词映射进行关系对齐

4.知识图谱搭建

基于Neo4j图数据库实现SPO三元组数据的存储

5.问答系统搭建

分别构建自然语言理解(NLU)、对话管理(DM)、自然语言生成(NLG)模块,使用Flask框架进行部署上线。

二、RAG智能问答系统

(1)数据预处理:

总数据:共处理2.6万份医疗相关文档,涵盖临床指南、药品说明书、疾病诊疗方案等相关知识

处理策略:文档整理、数据清洗、Paddle-0CR提取信息

(2)知识入库:

基于LangChain加载器对文档进行加载,然后进行文档分块切割;

使用bge-large-zh模型对文档进行词向量转化,并存入Milvus向量数据库支持3个collectios:基础医疗,临床决策,医疗管理

(3)问答检索:

实现用户意图识别:基于BERT模型实现用户意图识别,支持4种类型,F1-score=92.3%

混合向量检索:针对用户提问进行BM25(粗排)+BGE-reranker(精排)的混合向量检索

检索增强:构建Prompt,送入ChatGLM3-6B大模型获得答案

(4)系统评估:

基于RagAS工具实现RAG系统的评估:案相关性0.89

人工评估1000条测试集,Acc=90.6%

(5)服务部署:

使用VLLM框架进行模型推理加速

使用Flask+Docker框架进行部署上线

项目优化:

检索(R)优化:query改写,文档进行多级索引

增强(A)优化:Graph+RAG,把query和从neo4j、Milvus中检索出的结果合并,增强prompt

生成(G)优化:KVCache

大模型推理加速优化:使用vLLM框架

业绩:

该系统提升了医疗服务人员的工作效率40%,减少了医疗资源浪费,显著提高了诊断与治疗决策的精准度

用户售后问题分类　算法工程师　2023.10-2024.02

内容:

项目背景:

为了优化服务流程，通过用户输入的问题描述进行售后问题分类，例如商品质量问题、物流问题、退换货问题等

项目流程:

1. 获取数据，进行数据分析和处理

2. 进行模型搭建：使用 随机森林和Fasttext 搭建 baseline 模型进行初步判断

3. 基于Bert构建分类模型，进行多分类，acc指标93.6%。

4. 对Bert模型进行量化压缩，模型大小减少245MB，模型预测速度大幅提升，acc指标92.4%

5.选用 textCNN 模型进行知识蒸馏,提升模型预测效果

项目职责:

1.对数据进行清洗，以及重构以符合fasttext的输入标准。

2.项目baseline解决方案的实现，选用随机森林和fasttext对项目评估基线。

3.负责模型的设计及优化，开始选用BERT系列，但是发现模型过于庞大，模型预测速度不理想，期间对模型尝试了剪枝、量化、知识蒸馏等处理，最终在发现使用textCNN时，模型达到了速度和精度的双平衡

项目成果：

1. 最终老师模型 F1-Score = 0.953

2. 学生模型大小缩小近 20 倍，推理速度提升 10 倍，F1-Score = 0.927

业绩：

减少了人工成本，显著提升了用户满意度和售后服务效率。

教育经历

山西农业大学 本科 计算机科学与技术 2021-2025