2024.03 - 至今 **地图数据作业软件使用情况数据分析Agent**

算法工程师

项目背景

传统地图数据作业软件使用情况，基于数据埋点采集数据后上传BI系统。产品经理定期查看数据，基于图表数据进行分析决策，进而采取行动的链路较长。本项目旨在构建从BI系统数据采集到输出决策建议并发送邮件的Agent，提升数据分析的智能化及信息决策触达效率。

工作内容

本项目基于Lanhchain框架进行智能体开发，主要包含以下模块：

1.数据采集模块

- 通过API接口获取每日软件使用情况相关数据。

2.数据分析模块

- 使用大模型ERNIE-4.0 对采集到的数据进行分析。
- 基于数据分析产生数据预警、行动建议等。
- 结合数据及决策建议自动生成图文并茂的数据分析报告。

3.邮件发送模块

- 分析数据预警及数据分析报告需要触达的人员，并自动发送邮件。

项目成果

- 数据预警有效率86%，决策采纳率55%。每周节省数据分析2/人天。

2024.01 - 2024.05 **地图数据生产支持问答系统三期**

算法工程师

项目背景

在地图数据作业员的培训过程中，涉及大量图片和视频资料。为了提高培训效果和问每解决效率，在二期问答系统的基础上进行**多模态**升级。不仅能够通过文字检索返回与查询相关的图片和视频内容，还支持用户通过截图进行内容检索。这种增强型的问答系统将显著提升对话的质量和互动性，帮助作业员更快速、准确地获取所需信息，从而加速学习过程并提高工作效率。

工作内容

1.图片、视频数据处理及入库

- 使用大模型为图片、视频数据生成对应的文字描述信息，并存储在MongDB数据库中。

2.模型微调

- 基于作业台账的截图及问答、描述等构建图文数据集。
- 对Fuyu-8B模型进行微调，地图数据场景下的文本-图像匹配的准确率从45%提升至76%。

3.截图搜索功能实现

- 集成PaddleOCR技术，用于从截图中提取文本信息；
- 结合提取到的文本信息和微调后的模型进行图文检索，准确率85%。

4.集成上线

- 将图片、视频搜索及图文检索功能集成到问答系统中。
- 使用Flask+Docker框架进行部署上线。

项目成果

作业效率提升31.5%，准确率提升21.2%

2023.03 - 2023.12 **地图数据生产支持问答系统二期**

算法工程师

项目背景

基于知识图谱的地图数据生产问答系统仅可支持地图数据作业手册相关问答，文心一言发布之后，使用**RAG**技术实现了综合大模型基础能力、地图数据作业手册知识图谱、交通通识、智慧城市、规章制度等一体的问答系统。该系统实现对作业员的智能对话式支持，提升了作业效率、准确率及用户满意度。

工作内容

1.数据预处理

- 原始PDF处理：使用PaddleOCR将原始PDF文档（502M）转换为文本数据。
- 数据清洗：对txt、doc、ppt等格式的数据（1.2T）进行清洗，去除特殊字符及冗余信息并进行实体消歧。

2.创建多级索引

- 根据换行符切分chunks。
- 使用embedding v1模型进行文本向量化。
- 构 Milvus向量数据库存储索引，支持3个collections（交通通识、智慧城市、规章制度），共145万条。

3.问题改写与意图理解

- 对用户问题进行改写和拆解，为准确进行意图识别做准备。
- 使用Lora微调后的Bert 模型（2.7w条数据）进行用户意图识别，支持5种类型，F1-score=90%。
- 使用Bert模型从用户输入中提取实体列表，用于知识图谱检索。

3.RAG架构实现

- 检索模块（R）：文本检索（BM25）+语义检索（Embedding）+ReRank（Ernie_matching）。
- 增强模块（A）：将检索到的信息进行结构化处理，整合成统一的格式后构 Prompt。
- 生成模型（G）：将上述步骤构 的Prompt，输入ERNIE-4.0模型，生成文本回答。

4.内容安全合规检查

- 基于内容审核平台实现输入、输出文本内容合规检查。

5.历史索引和内容缓存

- 通过构 结构化索引，可以快速查询和定位已处理或检索过的内容，提高查询效率。
- 保存已处理或检索的结果到缓存中，当遇到相同或相似的查询时，直接从缓存中提取结果，减少响应时间。

5.模型评估

- 基于RagAS工具实现RAG系统的评估；
- answer relevancy=0.85+人工评估1000条测试集acc=90%。

6.Web展示

- 新作业员培养周期从1.5周缩短到1周；作业效率提升72%；准确率提升15%。

2021.09 - 2023.02 **地图数据生产支持问答系统一期**

职位

项目背景

地图数据作业手册知识繁多，传统依赖人工查阅文档及组长培训，新作业员培养周期长，作业效率低且容易出错。为此开发了基于地图数据作业手册**知识图谱**的智能问答系统。该系统通过构建知识图谱，实现对作业员的智能化支持，显著提高了数据作业的准确性和效率。

工作内容

1.数据获取

- 全要素地图数据作业手册，经过分词、标注后共获得样本20w条；实体类型46种，关系7种。

2.信息抽取

- 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；
- 实体识别准确率 95.1% ，关系抽取准确率 92.6%；综合F1-score=90.5%。

3.知识融合

- 基于TF-IDF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（节点：2273640个；关系：2528515条）。

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+Docker框架进行部署上线。

工作成效

新作业员培养周期从2周缩短到1.5周；作业效率提升67%；准确率提升30%。

项目一：餐饮软件故障智能问答系统（RAG 系统）　　2024.02-2024.12

项目背景

基于 RAG 的餐饮软件故障智能问答系统是一个为相关运维人员提供专业的故障解答的系统，提

升了软件故障的解决效率，提高了其准确性。

解决主要痛点：问题定位困难、故障排查时间长、严重依赖开发人员、知识和经验不足。

项目职责

使用餐饮软件数据集对大模型进行微调，构建向量数据库，实现 RAG 应用及模块优化，提升软件故障系统的智能问答能力。

(1)数据预处理：

- 总数据：2.6 万个文档，公司餐饮软件使用手册、运维手册、项目文档等相关知识
- 处理策略：文档整理、数据清洗、Paddle-OCR 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-reranker-large 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 collections：功能故障（错）、性能故障（慢）、可靠性故障（崩）

(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：针对提问进行 BM25+ bge-reranker-large 的混合向量检索，进行重排
- 检索增强：最后构建 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；答案相关性 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

项目成果：该系统不仅准确回答了软件故障知识，还提供了丰富的案例，工作效率提升了 33%。

项目二：基于餐饮领域知识图谱的个性化问答系统　　2023.06-2024.02

项目背景

该问答系统为用户提供个性化推荐的餐饮服务，满足用户的个性化需求，获取详细的菜品、餐厅信息以及用户评价，旨在提升线上餐饮服务的智能化和个性化水平。

工作内容

1.数据获取

- 数据部门提供的数据，经过标注后共获得样本 306798 条还会给

2.信息抽取

- 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%。
- 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；
  - 实体识别准确率 91.25% ，关系抽取准确率 92.62%。

- 实现了 8 种实体类型，10w 条边

3.知识融合

- 基于 TF-IDF 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。

工作成效：平台上用餐好评率上升 23.18%

故障处理流程知识库问答系统 NLP算法工程师 2024.03-2024.11

内容：

项目背景:集团要求构建"集中化监控+属地化支撑"运维体系,以省为单位统一管理故障全流程。广东省EMOS平台提供了技术支持,为实现高效工单派发奠定了基础。引入基于RAG技术的本地知识库问答机器人,可进一步优化工单派发与闭环处理流程,提高用户满意度与运维质量。

技术栈:Chatglm3-6b、Milvus、BGE-M3、Langchain、Gradio、Docker

业绩：

项目职责：

一、知识库构建：

1、基于相关部门提供的原始数据，精心筛选并整理了 100M故障工单文档，涵盖了整个自动化全流程业务线，确保数据的全面性和代表性

2、利用 OCR 和 NLTK 技术，全面加载并预处理文档，包括文本清洗、内容分块和向量表示等，最后将数据导入 Milvus 数据库，为后续模型处理奠定坚实基础。

二、数据检索：

1、基于用户的Query，通过意图识别，识别Milvus中的collections类别

2、结合 BM25 和 BGE生成混合向量索引，实现对文档进行高效检索。通过重新排序（rerank）技术，进一步提升了检索结果的相关性与准确性，为后续的数据分析和模型训练提供支持

三、LLM模型微调：

构建QA Pair数据集，使用LLMA-Factory训练框架对ChatGlm3-6b+P-Tuning作为基线模型微调，

后续迭代调优使用ChatGlm3-6b+P-TuningV2+Lora微调，综合使用了权重衰减、Warm up 、

梯度裁剪、梯度检查等手段优化模型训练

四、模型评估与部署：

基于RagAS工具实现RAG系统的评估；使用Gradio进行模型部署

信用卡知识图谱问答系统 NLP算法工程师 2023.04-2023.12

内容：

项目背景:随着银行信用卡业务的复杂性增加,客户在处理申请、审批、激活、使用、还款等多项服务时常面临信息宜杂与处理不及时的问题。尤其在面临不同信用卡政策,额度调整和优惠活动时,客户难以快速找到准确答案。为此,我们设计并构建了一个智能问答系统,旨在提升银行信用卡业务领域的客户服务效率。

技术栈:Bilstm+crf、Viterbi、Joint、Casrel、Neo4j、flask、streamlit

业绩：

项目职责:

一、负责关系抽取模块的研发

1.使用pipeline方法实现关系抽取,基于BiLSTM+Attention模型进行训练和评估,模型最终ACC为80%

2.使用joint联合抽取方法实现,基于Casrel模型解决了pipeline方法的SEO\EPO问题,模型最终ACC为94%

二、负责KBQA的NLU模块开发。通过将LR与GBDT两种算法的预测结果进行集成融合提升闲聊意图模型性能。整合ner和Classifier模型,构成NLU模块。

三、负责模型的优化:采用不同策略进行优化,badcase处理,超参数选择,正则化

项目成效:

1.项目中所使用到的模型整体ACC达到了93%以上,体现了模型问答相关任务中高效性能与稳定表现

2、使用预测结果融合、模型优化策略,有效改善了模型的鲁棒性和泛化能力

- 熟悉LLM算法、RAG开发、Agent、Function Call开发，熟悉Langchain框架；
- 掌握LLM等多种微调方法，如P-Tuning、Lora 等，掌握 RLHF、Prompt 等下游任务对齐策略；
- 熟悉NLP基本算法：知识图谱构建、意图理解、文本分类、相似性匹配、对话问答、机器翻译等；
- 掌握知识图谱相关技术，包括实体、关系抽取、Neo4j数据库等；
- 熟悉Bert、GPT、ChatGLM、ERNIE 等预训练模型的结构和原理；
- 掌握Transformer的原理，Attention 机制的原理及实现；
- 掌握深度学习算法，如CNN、RNN、LSTM、GRU等；
- 掌握常见的机器学习算法，如决策树、集成学习（随机森林、GBDT）等；
- 掌握模型压缩技术，如模型量化、模型蒸馏等；

- 熟练使用Pytorch、PaddlePaddle 深度学习框架；
- 熟练使用PaddleNLP、PaddleOCR；
- 掌握Python语言；熟练使用Numpy、Pandas、Matplotlib 进行数据清洗、筛选等操作；
- 熟 Flask、Docker、Gradio等服务部署模块；
- 熟练使用Linux 操作系统，熟 Linux常用命令；
- 熟练使用SQL各种常用函数，实现各种复杂业务；
- 掌握数据爬取技术；