高祥云 | NLP 算法工程师

基本信息

联系电话：18377302284 |邮箱：18377302284@163.com |学历：本科 |专业：软件工程

职业技能

- **RAG**：掌握端到端架构设计、多阶段检索链路优化（数据处理、混合检索、重排序）、评估迭代。
- **大模型**：掌握大模型微调技术（SFT, LoRA），熟悉Qwen、DeepSeek等主流模型的原理与应用。
- **NLP基础**：掌握文本分类/抽取、Attention、Transformer技术原理，BERT、GPT等模型应用。
- **数据工程与处理**：熟悉 Hadoop 生态（HDFS、MapReduce、Hive、Spark），可独立完成数据抽取、建模、ETL与性能调优。
- **基础工程与工具**：Python、Java、SQL、Flask、Docker、Linux、Git

工作经验

雷特科技 | NLP 算法工程师 | 2020.08 - 2025.06

灵犀智能问答助手 | 2024.01 - 2025.06

项目背景

- 为解决公司内部知识孤岛与检索效率低下的核心痛点，构建基于RAG架构的智能问答平台。平台整合了SOP、产品规格、会议纪要等多源异构文档， 通过自然语言查询，赋能运营及客服团队实现精准、高效的内部知识获取。

技术栈

- LangChain, Qwen2.5-7B-Instruct, bge-m3, Milvus

主要职能

- 系统构建与基线确立 (整体合格率: 0.43)
  - 异构文档（PDF/Word/扫描件/复杂表格）的智能解析与结构化处理。
  - “固定分块+向量检索”快速构建MVP，并基于200个真实场景下2000个问答对建立RAGAS整体合格率0.43的基线，为后续优化提供数据依据。
- 多阶段检索链路优化 (整体合格率: 0.83)
  - 分块与索引：设计内容自适应分块策略（文档结构、语义、Q&A），替代固定分块，上下文相关性提升34%。
  - 查询与召回：引入LLM查询改写与元数据过滤的混合检索,解决术语鸿沟和多条件复杂查询的召回难题。
  - 融合与精排：RRF融合，Reranker模型对召回结果进行精排，Top-3文档准确率提升近30%。
- 反馈闭环与模型微调 (整体合格率: 0.86)
  - 通过用户点赞/点踩功能，建立Bad Case分析机制，驱动模型持续迭代。
  - Embedding微调：基于Bad Case挖掘业务场景下的难负例，微调bge-m3模型，精准识别“公司黑话”等领域术语，提升检索召回的精准度。
  - LLM微调：利用高质量三元组对Qwen2.5进行指令微调，使生成内容更符合业务口吻。

项目成果

- 基于200次AB测试（旧流程vs新系统），客服工单平均处理时长5.2min→2.8min，效率提升46.2%。
- 目标部门（运营、客服）的日活用户渗透率达到70%，并在季度内部工具满意度调研中获得4.5分（满分5分）的高分评价，成为部门日常工作的核心辅助工具。

实时舆情分析系统 | 2022.06 - 2023.12

项目背景

- 基于主流电商及社交平台的多渠道爬虫数据（共12个来源），构建情感分析模型，实现企业舆情的实时监控与反馈机制。通过识别用户评论中的正面、负面及中性评价， 辅助产品设计、营销策略及售后服务优化，助力管理层做出数据驱动的精准决策。

技术栈

- Jieba, TF-IDF, FastText, BERT, TextCNN, Flask, Docker

主要职能

- 基于 FastText 构建 baseline 模型，负面召回率 88.35%
- 使用 BERT + 数据增强（同义词替换、句式重构）提升召回率至 93.28%
- 通过知识蒸馏（BERT → TextCNN）与量化技术，将模型响应时间从 700ms 降至 300ms
- 部署模型至生产环境（Flask+Docker+Linux）

项目成果

- 公司负面舆情处置时效从小时级别缩短至分级别
- 上线一年后.通过舆情驱动的产品改进,目标SKU退货率降低18%

智能审计 | 2020.08 - 2022.05

项目背景

- 传统审计依赖人工比对手机号、税号、天眼查平台关联信息（如法人关系、处罚记录），不仅效率低下（单次分析耗时数分钟），且易出错。 为提升审计效率与风险识别能力，构建基于知识图谱的问答式审计系统，通过自动化采集外部数据并融合内部业务数据，实现秒级响应的风险关系识别

技术栈

- CasRel, BERT, TextCNN, Flask, Neo4j, Docker

主要职能

- **设计知识图谱 Schema**： 根据审计业务需求，定义了四大实体类型（公司、人、新闻、处罚记录）和十种关系类型（如人-人亲属关系、人-公司法人关系、公司-新闻拥有关系、公司-处罚记录处罚关系），建立可扩展的图谱结构。
- **SPO三元组抽取与关系建模**： 基于 CasRel 模型（以 BERT 为基座）对 8 万条公司动态新闻进行三元组抽取，每条文本长度约 200~400 字，初始 Recall 达 86.7%。 引入数据增强（同义词替换、句式重构），模型泛化能力显著提升，Recall 提高至 89.25%。
- **构建知识图谱**： 将抽取结果与公司内部数据库中的客户、合同、员工信息融合，构建包含 22 万个实体 和 98 万条关系 的知识图谱，使用 Neo4j 实现可视化展示与高效查询。
- **用户意图识别与语义槽填充**： 使用 BERT 构建多任务分类器，支持 7 类审计意图识别（准确率 98.78%）及关键槽位提取（如公司名、人名、时间范围），填充准确率达 97.62%。

- **模型性能优化与部署上线**： 对 BERT 模型进行知识蒸馏（TextCNN 作为学生模型）和量化处理，将整体推理延迟从 800ms 缩短至 250ms，采用 Flask + Docker 架构部署于 Linux 服务器

项目成果

- 业务审计效率提升 60%,每日审计案例处理量从 20 件提升至 33 件,显著降低人工操作成本。

中软国际科技有限公司 | 大数据开发 | 2018.07 - 2020.06

康泰精算保险 | 2019.05 - 2020.06

项目背景

- 保险精算项目需要计算海量明细保单数据,以便生成财务报表。项目使用SparkSQL来计算,时效大大提高,增强保险公司的商业信誉。项目将多部门的业务数据库同步到hive数据集市, 使用SparkSQL加载源数据表,计算保单的保费、现金价值、准备金等明细,提供给财务部门收费或支出,最后对保单汇总计算（业务发展类指标,成本费用类指标等）,并向业务人员做数据展示。

技术栈

- Hadoop2.7.5、Hive2.1.0、Sqoop1.4.7、SparkSQL2.4.5、Azkaban

主要职能

- 使用Sqoop同步各异构数据源,到Hive数仓的ODS层。
- 使用SparkSQL加载保单信息,计算保单的保费,使用复杂的UDAF函数等技术。
- 使用Spark的视图,缓存表技术,结合迭代算法计算保单的现金价值、生存金、准备金等明细费用。
- 使用Shuffle调优,缓存持久化等措施对程序进行不断优化提速。

金泰商超新零售 | 2018.07 - 2019.04

项目背景

- 该项目基于一家大型连锁超市研发的大数据分析平台.项目主要围绕销售、履单、会员、商品和客服等零售环节中涉及的数据、信息等。 通过大数据分析可以提高履单效率、减少运营成本、更有效地满足客户服务要求,实现库存优化和增加营收的目标,并针对数据分析结果,提出具有中观指导意义的解决方案

技术栈

- CDH 6.2.1：Zookeeper、Hadoop、Hive、Hue、Sqoop

主要职能

- 参与商超新零售项目的环境搭建
- 负责完成整个源系统的数据抽取工作
- 完成销售模块、用户模块、商品模块、促销模块数据建模工作
- 满足公司日常运营的80%的数据需求和报表需求

教育背景

桂林航天工业学院 | 计算机科学 | 本科 | 2014.09 - 2018.06

自我评价

- 技术驱动型工程师，擅长从底层逻辑出发解决问题，对 AI 新技术有强烈好奇心与学习能力。

- 有韧性与执行力，曾完成 800 公里长途骑行，面对压力能保持冷静并持续突破。
- 关注 AI 前沿发展，热爱分享与交流。