b42c2edf4643c96f1Xd609W-EFtVwY-2UfqWWOGrlvbTNhFm2A~~

联系人

电话：
13262854888

电子邮件：
lantureman@163.com

爱好

游泳，篮球，唱歌，电影，旅游

个人信息

- 个人性格：开朗有活力，性格随和，易相处，乐观上进
- 技术扎实：多年 AI 算法经验，专注 NLP 方向，对 cv 方向略有涉及。
- 团队协作：善于跨部门沟通，具备项目管理能力。
- 持续学习：关注前沿技术（如大模型、以及大模型运用，如 chatgpt，llama，qwen 等 RAG，Agent 等运用和开发

教育背景

姓名：蓝贤文　　性别：男　　出生日期：1989 年 11 月 05

2008/9-2012/6　江西师范大学　　本科|材料化学

技能

- **自然语言处理（NLP）**：实体识别、文本分类、情感分析、文本纠错、问答系统、关系抽取。
- **机器学习/深度学习**：熟悉 CRF、LSTM、GRU、BERT、Transformer，logtis，deepFM，xgboost，随机森林，线性回归，k-means 等聚类分类。
- **编程与工具**：Python、pytorch、NumPy、Pandas、Scikit-learn、Matplotlib，基本 linux c++。
- **搜索与推荐**：语义召回、排序模型、特征工程、粗排精排，机器学习。
- **CV 方向：**目标检测，目标识别等
- **大模型：langchain，llamaindex， RAG，Agent 等**

工作经验

2013/5-2013/11	小牛普惠金融	金融专员
2014/3-2018/2	上海房金所金融信息服务有限公司	产品经理
2018/4-2021/5	中国指数研究院	nlp 算法工程师
2021/8-2022/3	上海竞动科技有限公司	nlp 算法工程师
2022/6-2023/11	上海恒格科技股份有限公司	nlp 算法工程师

工作内容及描述：

- 评估项目需求，通过需求分析，结合 AI 算法模型提出解决方案
- 收集数据，对数据进行分析，清洗，预处理
- 根据需求分析建立模型，各模块开发以及衔接，测试部署等

项目经验（上海恒格科技股份有限公司）

百科搜索文本转写

项目背景

为提升搜索引擎召回率，设计文本转写系统，通过同义词扩展将用户查询改写为多组语义相似的关键词组合，解决用户表达差异导致的搜索漏召回问题。

技术方案及实现：

1. 词向量模型训练与优化

基于数据库百科、新闻等语料（100 万原始数据）训练 Word2Vec 模型，优化负采样策略，使 Top-3 同义词准确率达 85%

设计"领域词表过滤"机制，结合业务关键词库约束同义词搜索空间，专有名词准确率提升 32%

2. 实现多粒度改写策略

分词→词性过滤（保留名词/动词/形容词等）→同义词扩展→TF-IDF 权重排序

3. 文本转写验证

基于历史搜索记录，抽取 4000 条数据，关键词分析改写，通过人工审核改写准确率 83%，主要错误来源于专有名词泛化错误较多，通过规则提高专有名称泛化能力

通过文本转写，相比单独 ES 召回，TOP5，TOP10，TOP20，平均提升 12%

百科搜索实体提升策略以及意图识别

项目背景

为提升企业级搜索引擎的召回精准度，设计并实现基于实体和意图的结构化召回服务，覆盖日期、人名、保险产品名、城市等 10+类实体，优化 Elasticsearch 查询策略，支持多维度搜索需求。

技术方案及实现：

1. 实体识别与归一化

构建多类别实体识别模型（日期、人名、城市等），采用 Bilstm-CRF 模型架构，F1 值达 88.5%。

日期归一化：将“下周节目”“国庆节”等相对时间转换为标准格式 yyyyMMdd，支持绝对时间和节假日逻辑。

城市归一化：识别用户输入中的地点片段（如“南京栖霞区”）并标准化为“南京市”，准确率 92%。

采用实体对搜索关注板块以及关键词策略进行权重提升，增强搜索能力

2. 意图识别与搜索策略设计

基于知识库文档数据预训练所得模型基础上，将搜索问题通过 bert 向量化

基于规则引擎和分类模型，构建标准意图库，通过预训练模型向量化，与用户搜索问题 cosine 计算相似度，识别用户搜索意图，

对用户意图识别分类，准确率 89%。

设计实体权重动态调整策略，如保险产品名匹配时优先召回条款文档（权重提升 30%）。

Elasticsearch 集成与优化

将实体与意图拼接为 ES 布尔查询条件，支持多字段组合召回（如(城市:上海 AND 意图:天气)。

搜索引擎召回率提升 25%，核心实体（如保险产品名）准确率提升至 90%

通过文本转写，相比单独 ES 召回，TOP5，TOP10，TOP20，平均提升 7%

百科搜索语义召回粗排精排

项目背景

为解决传统关键词搜索的语义鸿沟问题，设计基于向量检索的语义召回引擎，提升搜索引擎对用户意图的理解能力。

技术方案及实现：

1. 在 bert 中文预训练模型基础上，基于百科领域知识库文档数据进行再次预训练，得到相关领域预训练模型
2. 在领域内所得预训练模型，使用 Bert 双塔语义召回模型开发，对 query-doc 召回
3. 基于 pytorch 构建 Query-Document 双塔模型，将用户 Query 与文档内容编码构建语义向量。
4. 使用 5 万条标注数据训练模型，优化对比损失函数（Contrastive Loss），单独语义召回 TOP5、TOP10 召回准确率分别为 60%，66%
5. 设计 Milvus 分布式存储架构，对语义理解层面对海量文档数据进行向量搜索，采用 cosine 相似度检索，支持实时语义编码。。
6. 根据用户搜索时间与文档库时间需求的差异性，通过高斯时间衰减算法，动态调整搜索结果权重进行重新计算得分

项目背景

为解决基于向量检索的语义召回引擎，提升搜索引擎对用户意图的理解能力，对召回文档进行精排

技术方案及实现

采用 Cross-Encoder 为精排模型框架，问题-文档对的相关性计算

裁剪 BK-PTM 网络，减少网络层数，提升计算速度，得到精排模型

在实体识别，文本转写，关键词，意图识别，语义召回的基础上，进行精排后，最终效果提升，TOP5、TOP10 召回准确率分别至 72%，77%，TOP20 至 89%

百科搜索重排

项目背景

为提升企业级搜索引擎的个性化排序能力，设计基于用户特征（部门、岗位、分公司）和文档特征（数据源、模块）的二次排序模型，优化搜索结果与用户偏好的匹配度。

技术方案及实现

逻辑回归模型开发与优化

构建二分类逻辑回归模型预测文档点击率（CTR），通过 Sigmoid 函数映射概率，损失函数采用对数损失，模型 AUC 达 0.664 。

实现特征组合优化，整合用户属性（政治面貌、分公司）与文档属性（所属模块、数据源），特征维度扩展至 50+。

引入正则化（L2）防止过拟合，模型训练迭代效率提升 30%。

规则引擎与特征工程

设计基于规则的推理网络，动态调整特征权重（如分公司匹配权重提升 20%）。

开发特征归一化模块，支持用户输入特征（如“政治面貌”）与文档特征的标准化映射。

项目背景

为提升企业级搜索引擎的点击率预测精度，在逻辑回归上进行优化重排策略，采用高阶交叉特征建模

技术方案及实现

基于 AutoInt 网络结构，通过多头自注意力机制捕捉用户-文档特征交互，模型 AUC 提升至 0.77（较基线逻辑回归模型提升 12%）。

实现稀疏特征 Embedding，对用户政治面貌、部门等离散特征进行 One-Hot 编码，生成低维稠密向量，特征维度压缩 50%。
特征工程与数据增强
设计动态负采样策略，结合随机负样本与回填负样本，缓解数据稀疏性问题，训练集样本量扩展至 210 万条。
构建多维度特征组合（如"分公司+模块"），支持实时特征拼接，特征覆盖率提升至 95%。

CTR 预测模型 AUC 提升至 0.77，搜索结果的用户点击率提升 22%

香港法律诉讼文件分类

项目背景

香港英文不同类型保险法律诉讼函，诉讼回复，结案理赔等文件归类，总体 50 多个类别

技术方案及实现

结合业务规则以及数据类型大小，采用 simhash 无监督文本分类、关键词枚举、英文法律相关 bert 预训练模型等方案进行分类比较
Bert 法律英文法律预训练模型结合保险及法律诉讼文件业务知识加入预训练模型，选择标注类别为标准库文本，通过语义相似度计算分类，类别平均准确率 60%，不够预期
Simhash 结合 ngram 分词，构建关键词权重表，采用汉明距离计算文本相似度，类别平均准确率达到 75%
策略分析 simhash 每个类别得分，设定阈值，对得分高且符合阈值的 100%分类的，直接划分类别，部分分类鉴别不清，结合业务规则关键词枚举实现，结合以上方法，多个类别实现 100%分类，少部分低于 70%，无法鉴定，模糊不清则返回到其他，由业务自行判断
通过结果和应用场景对比，采用 simhash 算法和关键词枚举结合实现业务

大模型实现邮件信息提取

项目背景

公司内部邮件通知信息提取，内部员工沟通软件增加日程信息提醒功能

技术方案及实现

基于本地部署大模型 chatglm 进行调用，调试 promt 提取关键信息，实现邮件信息提取，为员工提供日程提醒功能

项目经验（上海竞动科技有限公司）

金融舆情监控

项目背景

捕捉行业热点与市场情绪，量化板块表现，辅助宏观策略制定，涉及金融舆情监控，股市波动预警，企业公告结构化，行业趋势分析等应用场景
技术方案及实现
基于 bert 模型结构，结合 Grid Tagging Scheme 的细粒度观点，实现三元组提取模型
数据处理：基于舆情新闻文本数据，进行清洗和分析，数据支持六类标签（A/O/POS/NEU/NEG/N），直接建模三元组关系。
基于 BERT 利用 Transformer 捕捉长距离依赖关系，注意力机制增强词对关联性，提升 aspect-opinion 关系建模。

采用端到端标注，通过统一网格标注同时提取 aspect term、opinion term 及情感极性，避免管道方法误差传播

指标： F1-score（OTE 任务）达 87.58%

问询函违规案例法律知识回复和法律知识结构化提取

应用场景

对来自证监会、交易所发布的证券法规和及相关国家单位对上市公司发布涵盖信息披露、市场监管、公司治理等多个方面的问询函，上市企业就问询函立即启动内部应对机制，需参考问询函所涉及法律法规，建立合规内控体系，确保及时、准确回应监管问询，降低风险

项目技术及方案

结合业务方向，方案构建分别为法律法规结构化知识构建，法律法规条文对比，通过命名实体识别进行知识结构化查询和语义及其他文本相似度方法对法律条文进行拆分比对

基于 bert+crf 构建实体识别模型，结合规则提取法律发条实体和关系，通过 simcse 模型构建有监督语义相似度

模型应用于实现数据库法律法规结构化提取，数据归一化处理，机器人问答回复，针对问询函进行法律法规提取，为上市公司就问询函相关问题进行提供相关法律法规数

项目经验（中国指数研究院）

房企满意度情感分析

- **目标背景**：收集物业以及楼盘满意度调查，实现自动化分析业主以及顾客反馈情感倾向。
- **技术方案**：
  - 基于 bert 的 Aspect-Based 情感分析，联合建模方面术语与极性词。
  - 采用 Mean-Max Pooling 特征融合，Acc 达 93%。
- **成果**：节省 70%人工审核成本，提升客服效率。

文本纠错系统

- **目标背景**：修复用户输入错误和同义词替换，提升问答体验。
- **技术方案**：
  - Bi-GRU（检测）+ BERT（纠错）联合模型，引入 SoftMask 技术。
  - 动态学习率与损失加权优化。
- **成果**：纠正输入错误和同义词替换，搜索准确率提升 9%