



CCF BDCI

2024 CCF 大数据与计算智能大赛^{12th}

决赛答辩评审会

《基于运营商文本数据的知识库检索》

队伍名称: kelvincjr

模板说明：本模板章节及内容仅供参考。

请结合实际情况按需修订、排序和阐述。

目录



CCF BDCI

01 赛题解读

02 解决方案

03 效果展示

04 创新点总结

05 应用探索和规划

06 团队介绍

一、赛题解读

赛题任务和评测标准

- 本赛题要求选手使用运营商相关的文档构建知识库，根据用户问题检索知识库并返回答案所在的文本块。

特别说明：本赛题意在优化 RAG（除 LLM 答案生成及之后部分）流程，选手应重点关注文本解析、文本切片、检索过滤排序等环节，意在消除大模型的使用对大家的不公平性。如在复现环节发现团队使用违反赛题要求的手段生成结果，将取消团队参赛成绩。

- 本赛题使用的运营商相关文档包括三类：
 1. 新闻稿件150篇，部分文档经过改写或扩写，因此内容可能与现实情况不符，可能不具有现实意义，仅允许在本次比赛中使用。
 2. 公开报告25篇，包括年度及半年度报告摘要、季度报告、股东大会会议资料等。
 3. 通信行业权威研究报告及白皮书15篇（来源：中国信息通信研究院）。

■ 问题评分

$$\text{score}_i = [\text{weight}_i^{\text{kw}} \times \text{score}_i^{\text{kw}} + (1 - \text{weight}_i^{\text{kw}}) \times \text{score}_i^{\text{es}}] \times p_i$$

$\text{score}_i^{\text{kw}}$: 关键词评分。

$\text{score}_i^{\text{es}}$: 向量相似度评分。

$\text{weight}_i^{\text{kw}}$: 关键词评分占总评分的权重。通常为0.1至0.9之间的权重，表格查数问题的关键词得分权重为1。

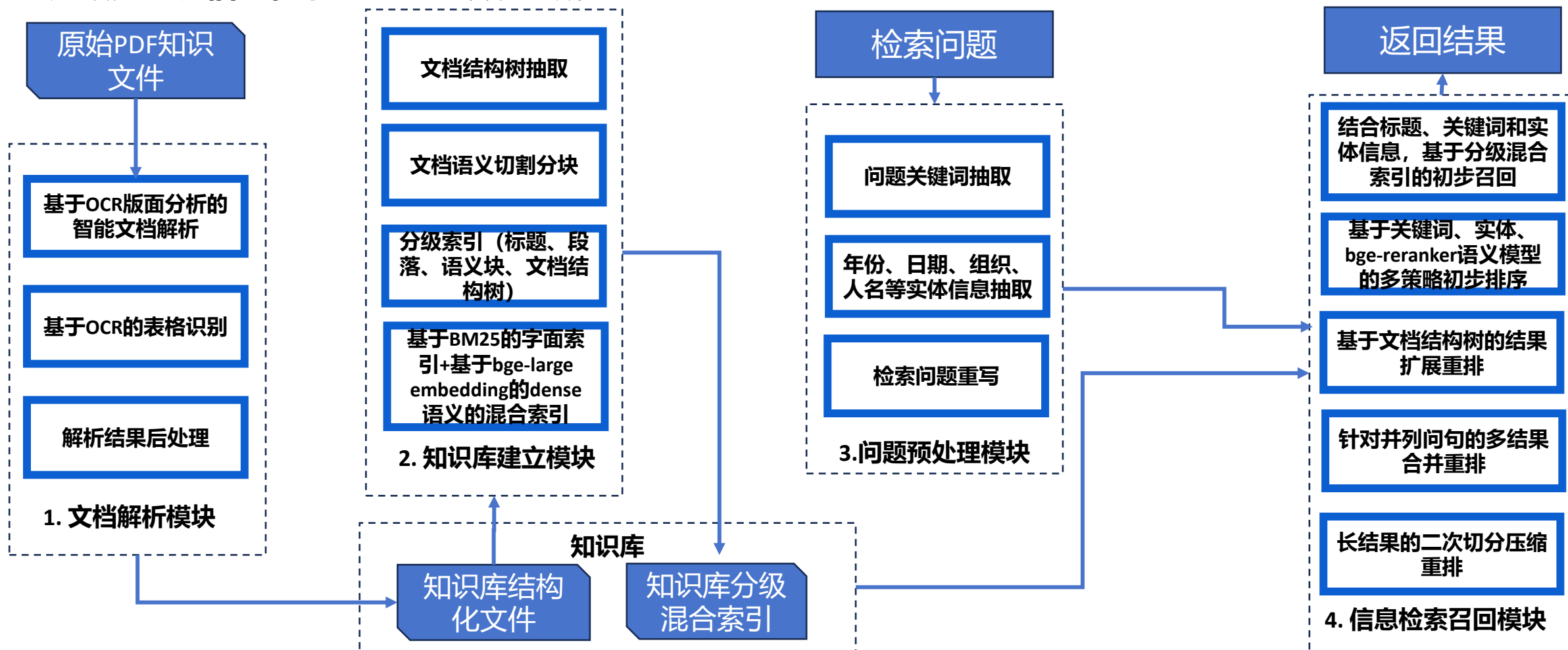
p_i : 长度惩罚项。

解决思路

- 首先赛题主要聚焦在RAG的检索部分，从评测标准来看，需要命中关键词句，而且召回长度有惩罚限制，在不使用大模型的前提下，对检索的要求很高，需要准确定位到答案对应的文本片段和句子。
- 我们基于智能文档解析和自然语言处理技术搭建了一套完整的文档信息检索和问答系统，通过**可视化精细化的文档解析**、**多维度多层次的知识索引**、**结合关键词和实体的问题改写**和**多策略的信息召回排序**四大模块，保证最终返回结果的完整性和准确性。

二、解决方案 – 总体解决方案概览

基于智能文档解析技术和自然语言处理技术构建了一个完整的文档知识检索系统，由文档解析、知识库建立、问题预处理、信息检索召回四个模块组成。



二、解决方案 – 文档解析模块

文档解析模块主要包括基于OCR版面分析的智能文档解析、基于OCR的表格识别和解析结果后处理等流程。

基于OCR版面分析的智能文档解析

中国联通网络通信股份有限公司 2022 年年度报告

第二节 公司基本情况

1 公司简介

股票种类	股票上市交易所	股票简称	股票代码
A股	上海证券交易所	中国联通	600050

联系人和联系方式	董事会秘书	证券事务代表
姓名	李玉焯	雷晓旭
办公地址	北京市西城区金融大街21号	北京市西城区金融大街21号
电话	010-66259179	010-66259179
电子信箱	dongmi@chinaunicom.cn	ir@chinaunicom.cn

2 报告期公司主要业务简介

2022 年是党的二十大召开之年，也是中国联通全面实施战略规划起跑之年。中国联通坚持以习近平新时代中国特色社会主义思想为指导，认真学习宣传贯彻党的二十大精神，坚决贯彻落实党中央、国务院决策部署，坚持以“数字信息基础设施运营服务国家队、网络强国数字中国智慧社会建设主力军、数字技术融合创新排头兵”的企业定位，在“强基固本、守正创新、融合开放”的战略指引下，大力发展“大联接、大数据、大应用、大安全”五大主责主业，扎实推进“1+9+3”战略规划体系先行见效，实现战略规划全面实施第一年的良好开局。

“大联接”方面，中国联通抢抓“双千兆”“物超人”发展机遇，坚持量质质效协同发展。战略定位、政策设计、资源配置同向发力，实现基于规模的价值经营、基于质量的合规发展、基于结构的融合发展、基于效能的有效发展，围绕全量用户融合化发展全力推进端网业协同的价值经营，聚力做大联接规模和价值。截至 2022 年 12 月，用户规模再创新高，“大联接”用户累计达到 8.6 亿户，宽带用户跨越 1 亿户历史关口，5G 套餐用户累计达到 2.1 亿户，推出“格物”设备管理平台，深入智慧城市和工业互联网两大领域，为客户提供便捷专业的设备管理服务，物联网率先实现“物超人”，物联网终端连接累计达 3.86 亿户。

“大计算”方面，联通云全面升级至 7.0 版本，核心技术攻关取得突破，计算、存储、网络核心性能指标增强，云原生容器平台获得运营商唯一金融场景卓越认证，自研操作系统 CUIlinux 及自研数据库成功上线，深化“5+4+31+X”云资源布局，基于微节点进一步下沉，逐步实现一市一池算力布局，由行业云、私有云、信创云组成的全新联通云产品体系，全面满足客户全场景上云需求，支撑数字政府、智慧城市、智慧城市建设，助力科技强国，积极参与央企数字化转型。

“大数据”方面，持续提升数据智能融合创新能力，聚合数据治理、数据安全、数据可视化服务能力，完善产品体系，发布“数治”政务大数据平台，11 项能力入围信通院数据治理产业图谱。在数字政府、数字金融、智慧文旅、数据安全等领域，实现省市级标杆项目规模复制，联通大数据业务保持高速增长，市场的持续领跑行业，区块链专利储备齐全。

“大应用”方面，公司致力于打造更丰富、更好用的创新产品，深化面向个人和家庭市场的产

中国联通网络通信股份有限公司 2022 年年度报告

第二节 公司基本情况

1 公司简介

股票种类	股票上市交易所	股票简称	股票代码
A股	上海证券交易所	中国联通	600050

联系人和联系方式	董事会秘书	证券事务代表
姓名	李玉焯	雷晓旭
办公地址	北京市西城区金融大街21号	北京市西城区金融大街21号
电话	010-66259179	010-66259179
电子信箱	dongmi@chinaunicom.cn	ir@chinaunicom.cn

2 报告期公司主要业务简介

2022 年是党的二十大召开之年，也是中国联通全面实施战略规划起跑之年。中国联通坚持以习近平新时代中国特色社会主义思想为指导，认真学习宣传贯彻党的二十大精神，坚决贯彻落实党中央、国务院决策部署，坚持以“数字信息基础设施运营服务国家队、网络强国数字中国智慧社会建设主力军、数字技术融合创新排头兵”的企业定位，在“强基固本、守正创新、融合开放”的战略指引下，大力发展“大联接、大数据、大应用、大安全”五大主责主业，扎实推进“1+9+3”战略规划体系先行见效，实现战略规划全面实施第一年的良好开局。

“大联接”方面，中国联通抢抓“双千兆”“物超人”发展机遇，坚持量质质效协同发展。战略定位、政策设计、资源配置同向发力，实现基于规模的价值经营、基于质量的合规发展、基于结构的融合发展、基于效能的有效发展，围绕全量用户融合化发展全力推进端网业协同的价值经营，聚力做大联接规模和价值。截至 2022 年 12 月，用户规模再创新高，“大联接”用户累计达到 8.6 亿户，宽带用户跨越 1 亿户历史关口，5G 套餐用户累计达到 2.1 亿户，推出“格物”设备管理平台，深入智慧城市和工业互联网两大领域，为客户提供便捷专业的设备管理服务，物联网率先实现“物超人”，物联网终端连接累计达 3.86 亿户。

“大计算”方面，联通云全面升级至 7.0 版本，核心技术攻关取得突破，计算、存储、网络核心性能指标增强，云原生容器平台获得运营商唯一金融场景卓越认证，自研操作系统 CUIlinux 及自研数据库成功上线，深化“5+4+31+X”云资源布局，基于微节点进一步下沉，逐步实现一市一池算力布局，由行业云、私有云、信创云组成的全新联通云产品体系，全面满足客户全场景上云需求，支撑数字政府、智慧城市、智慧城市建设，助力科技强国，积极参与央企数字化转型。

“大数据”方面，持续提升数据智能融合创新能力，聚合数据治理、数据安全、数据可视化服务能力，完善产品体系，发布“数治”政务大数据平台，11 项能力入围信通院数据治理产业图谱。在数字政府、数字金融、智慧文旅、数据安全等领域，实现省市级标杆项目规模复制，联通大数据业务保持高速增长，市场的持续领跑行业，区块链专利储备齐全。

“大应用”方面，公司致力于打造更丰富、更好用的创新产品，深化面向个人和家庭市场的产

基于LayoutLMv3模型

基于OCR的表格识别

联系人和联系方式		董事会秘书	证券事务代表
姓名		李玉焯	雷晓旭
办公地址		北京市西城区金融大街21号	北京市西城区金融大街21号
电话		010-66259179	010-66259179
电子信箱		dongmi@chinaunicom.cn	ir@chinaunicom.cn

```
<html>
<body>
  <table>
    <tr>
      <td rowspan=1 colspan=1>联系人和联系方式</td>
      <td rowspan=1 colspan=1>董事会秘书</td>
    </tr>
    <tr>
      <td rowspan=1 colspan=1>姓名</td>
      <td rowspan=1 colspan=1>李玉焯</td>
    </tr>
    <tr>
      <td rowspan=1 colspan=1>办公地址</td>
      <td rowspan=1 colspan=1>北京市西城区金融大街21号</td>
    </tr>
    <tr>
      <td rowspan=1 colspan=1>电话</td>
      <td rowspan=1 colspan=1>010-66259179</td>
    </tr>
    <tr>
      <td rowspan=1 colspan=1>电子信箱</td>
      <td rowspan=1 colspan=1>dongmi@chinaunicom.cn ir@chinaunicom.cn</td>
    </tr>
  </table>
</body>
</html>
```

联系人和联系方式 董事会秘书 证券事务代表

姓名	李玉焯	雷晓旭
办公地址	北京市西城区金融大街21号	北京市西城区金融大街21号
电话	010-66259179	010-66259179
电子信箱	dongmi@chinaunicom.cn	ir@chinaunicom.cn

基于SLANet-plus模型

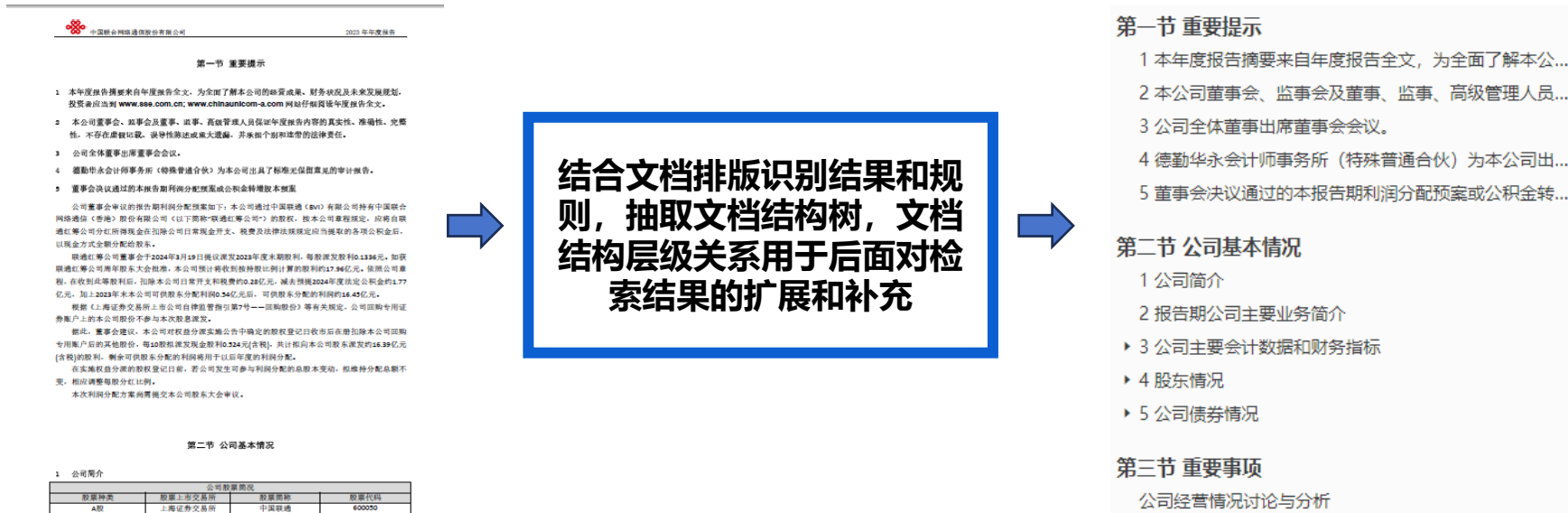
解析结果后处理

主要包括段落分页合并，表格分页合并，OCR识别内容丢失补全等处理流程。

二、解决方案 – 知识库建立模块

知识库建立模块主要包括文档结构树提取、文档语义切割分块和分级混合索引建立等子流程。

➤ 文档结构树提取



➤ 文档语义切割分块

- 运营商的文档有很多长段落，如果用段落作为单一索引单位，召回的信息会太粗，所以需要进一步对段落信息进行切分。为了使切分后的块具有完整独立的语义，我们使用了阿里的nlp_bert_document-segmentation_chinese-base模型。

坚持贯彻新发展理念，践行 5G 高质量发展。5G 中频基站新增 22 万站，4G 共享基站超过 200 万站。与友商在新疆启动全球首个 5G 异网漫游试商用，进一步提升边远地区网络资源使用效益；努力推动传输线路、管道光缆等基础设施的共建共享共维取得积极成效。以实际行动响应“双碳”政策，2023 年节省运营开支约人民币 390 亿元，减少碳排放约 1,150 万吨，累计节省资本开支约人民币 3,400 亿元。

二、解决方案 – 知识库建立模块

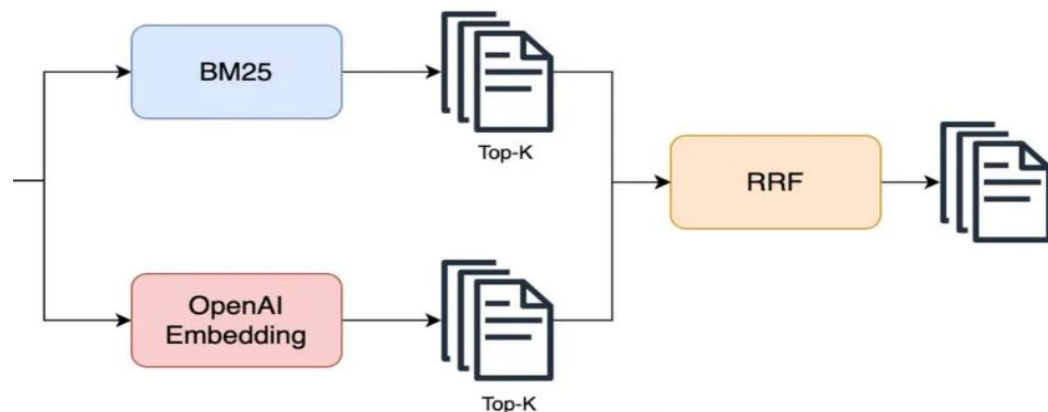
➤ 分级混合索引建立

我们的方案采用多级分层混合索引的方式，分别对标题、段落、语义块、文档结构树建立不同类型的索引，以实现高效精准检索。

- ✓ 对于标题，目前主要是针对年报类专题，按年份、报告类型（年报、第一/三季度报、上半年报）等实体字段进行索引；

中国联合网络通信股份有限公司2023年年度报告摘要

- ✓ 对于段落和语义块，主要采用基于BM25的字面索引+基于bge-large embedding的dense语义索引组合的混合索引方式；



- ✓ 对于文档层次结构，直接采用文档结构树的树状结构来进行索引，能快速检索到文档元素之间的层次关系。

二、解决方案 – 问题预处理模块

问题预处理模块主要包括问题关键词抽取、实体抽取（年份、日期、组织、人名等）和检索问题重写等几个流程。

➤ 问题关键词提取

准确的关键词识别对于基于BM25的字面信息检索效果有很大影响，方案中我们采用hanlp的关键词短语提取功能。

2022年联通在“大联接”和“大数据”业务上取得了什么成果？

30/800

TopK:

提取

关键词短语

编号	关键词短语	权重
1	什么成果	95.29%
2	“大联接”	79.11%
3	“大数据”	63.52%
4	2022年	41.20%
5	联通	26.78%

➤ 多实体信息提取： 采用正则结合分词词性的方法来识别问题中的年份、日期、组织、人名等实体信息。

➤ 检索问题重写： 采用关键词、实体拼接原问题的方式来组成最终的检索问题。

二、解决方案 – 信息检索召回模块

作为整个系统的核心，信息检索召回模块采用多种召回和重排策略结合的方式来保证最终返回结果的完整性和准确性。

改写后的检索问题：月户均数据流量 手机用户 根据联通年度报告，2019年和2020年手机用户月户均数据流量分别为多少？

topk召回

基于标题和实体的文档召回

基于分级混合索引的片段召回

```
ques_id 21, question 根据联通年度报告，2019年和2020年手机用户月户均数据流量分别为多少？,
match_titles: ['中国联合网络通信股份有限公司2020年年度报告摘要',
'中国联合网络通信股份有限公司2019年年度报告摘要'], match_files: ['AY06', 'AY11'],

*** rank: 1, Score: 16.688133335113523, File: [AY11], Title: [
中国联合网络通信股份有限公司2019年年度报告摘要], Text: [公司持续落实“提速降费”政策，手机上网流量
单价同比大幅下降，手机上网总流量和手机用户月户均数据流量持续显著增长。2019年，手机上网总流量增长46%
，手机用户月户均数据流量达到8GB；手机上网收入达到人民币1,028亿元。]

*** rank: 2, Score: 8.897784012556075, File: [AY06], Title: [
中国联合网络通信股份有限公司2020年年度报告摘要], Text: [公司价值经营卓有成效，移动业务发展动力持
续释放，移动主营业务收入及用户ARPU已连续三个季度双双环比上升。2020年，移动主营业务收入达到1,567亿
元，同比止负转正；移动用户ARPU达到42.1元，同比提升4.1%；手机上网总流量增长16.5%，手机用户月户均数
据流量达到约9.7GB。营销费用进一步下降，提质增效步伐稳健，盈利能力持续改善。]
```

多策略排序

基于关键词、实体、bge-reranker语义模型的多策略排序

基于文档结构树的结果扩展重排

针对并列问句的多结果合并重排

长结果的二次切分压缩重排

```
=====ques_id: 21, rerank result=====
*** rank: 1, Rerank_Score: 3.890625, Fuzz_Score: 20, Kp_Score: 1.0, Answer_kps:
['月户均数据流量', '手机用户'], Title_Score: 0.6666666666666666, Title_matched_info:
['2019年', '年年度'], Token_Score: 4, Overlap_tokens: {'手机用户', '月', '数据', '户均',
'流量'}, Recall_Score: (1, 16.688133335113523), File: [AY11], Title: [
中国联合网络通信股份有限公司2019年年度报告摘要], Text: [公司持续落实“提速降费”政策，手机上网流量
单价同比大幅下降，手机上网总流量和手机用户月户均数据流量持续显著增长。2019年，手机上网总流量增长46%
，手机用户月户均数据流量达到8GB；手机上网收入达到人民币1,028亿元。]

content_to_nodes: 公司持续落实“提速降费”政策，手机上网流量单价同比大幅下降，手机上网总流量和手机
用户月户均数据流量持续显著增长。2019年，手机上网总流量增长46%，手机用户月户均数据流量达到8GB；手机
上网收入达到人民币1,028亿元。
*** rank: 2, Rerank_Score: 1.94921875, Fuzz_Score: 17, Kp_Score: 1.0, Answer_kps:
['月户均数据流量', '手机用户'], Title_Score: 0.6666666666666666, Title_matched_info:
['2020年', '年年度'], Token_Score: 4, Overlap_tokens: {'手机用户', '月', '数据', '户均',
'流量'}, Recall_Score: (2, 8.897784012556075), File: [AY06], Title: [
中国联合网络通信股份有限公司2020年年度报告摘要], Text: [公司价值经营卓有成效，移动业务发展动力持
续释放，移动主营业务收入及用户ARPU已连续三个季度双双环比上升。2020年，移动主营业务收入达到1,567亿
元，同比止负转正；移动用户ARPU达到42.1元，同比提升4.1%；手机上网总流量增长16.5%，手机用户月户均数
据流量达到约9.7GB。营销费用进一步下降，提质增效步伐稳健，盈利能力持续改善。]
```

```
=====ques_id: 21, combined rerank result=====
match combined rerank case
*** Find Combined rank: 2 and 1, Cb_kp_score: 1.0, Cb_answer_kps: ['月户均数据流量',
'手机用户'], Cur_kp_score: 1.0, Cur_answer_kps: ['月户均数据流量', '手机用户'],
Cb_title_score: 1.0, Cur_title_score: 0.6666666666666666, Title_matched_info: ['2020年',
'年年度'], Text: [公司持续落实“提速降费”政策，手机上网流量单价同比大幅下降，手机上网总流量和手机
用户月户均数据流量持续显著增长。2019年，手机上网总流量增长46%，手机用户月户均数据流量达到8GB；手机
上网收入达到人民币1,028亿元。
公司价值经营卓有成效，移动业务发展动力持续释放，移动主营业务收入及用户ARPU已连续三个季度双双环比上
升。2020年，移动主营业务收入达到1,567亿元，同比止负转正；移动用户ARPU达到42.1元，同比提升4.1%；手
机上网总流量增长16.5%，手机用户月户均数据流量达到约9.7GB。营销费用进一步下降，提质增效步伐稳健，盈
利能力持续改善。]

*** Combined rank: 3 and 1, Cb_kp_score: 1.0, Cb_answer_kps: ['月户均数据流量',
'手机用户'], Cur_kp_score: 0.0, Cur_answer_kps: [], Cb_title_score: 0.6666666666666666,
Cur_title_score: 0.6666666666666666, Title_matched_info: ['2019年', '年年度'], Text: [公司
持续落实“提速降费”政策，手机上网流量单价同比大幅下降，手机上网总流量和手机用户月户均数据流量持续显
著增长。2019年，手机上网总流量增长46%，手机用户月户均数据流量达到8GB；手机上网收入达到人民币1,028
亿元。]
```

多结果合并重排



二、解决方案 – 信息检索召回模块

改写后的检索问题：我国区块链发展 挑战 我国区块链发展面临的主要挑战有哪些？

topk召回

基于标题和实体的文档召回

基于分级混合索引
的片段召回

```
=====ques_id: 32, recall result=====
query: 我国区块链发展 挑战 我国区块链发展面临的主要挑战有哪些？
*** rank: 1, Score: 12.90481327176094, File: [AZ07], Title: [区块链白皮书(2023年)], Text: [
六、我国区块链发展面临的挑战与展望]

*** rank: 2, Score: 12.709654307365417, File: [AZ07], Title: [区块链白皮书(2023年)], Text:
[六、我国区块链发展面临的挑战与展望.....34]
```

多策略排序

基于关键词、实体、bge-reranker
语义模型的多策略排序

基于文档结构树的结果扩展重排

针对并列问句的多结果合并重排

长结果的二次切分压缩重排

```
=====ques_id: 32, rerank result=====
*** rank: 1, Rerank_Score: 5.40625, Fuzz_Score: 8, Kp_Score: 0.0, Answer_kps: [],
Title_Score: 0, Title_matched_info: [], Token_Score: 2, Overlap_tokens: {'区块链',
'我国'}, Recall_Score: (21, 0.21194795966148375), File: [AZ07], Title: [
区块链白皮书(2023年)], Text: [区块链技术自主研发能力存在短板，难以支撑应用深度创新。我国在共识算法、数据存储等方面自主创新能力尚需进一步强化，身份安全、授权管理和生物识别等配套技术有待提升，以解决未来数字资产、数字身份等应用创新过程中提出的新问题。此外，我国仍存在开源社区话语权较弱、核心专利较少、基础设施与海外差距较大等困境，难以支撑未来数字资产交易规模化增长和自主化应用。]

content_to_nodes: 区块链技术自主研发能力存在短板，难以支撑应用深度创新。我国在共识算法、数据存储等方面自主创新能力尚需进一步强化，身份安全、授权管理和生物识别等配套技术有待提升，以解决未来数字资产、数字身份等应用创新过程中提出的新问题。此外，我国仍存在开源社区话语权较弱、核心专利较少、基础设施与海外差距较大等困境，难以支撑未来数字资产交易规模化增长和自主化应用。
*** rank: 2, Rerank_Score: 5.30859375, Fuzz_Score: 52, Kp_Score: 1.0, Answer_kps:
['我国区块链发展', '挑战'], Title_Score: 0, Title_matched_info: [], Token_Score: 4,
Overlap_tokens: {'挑战', '区块链', '我国', '发展'}, Recall_Score: (1, 12.90481327176094),
File: [AZ07], Title: [区块链白皮书(2023年)], Text: [六、我国区块链发展面临的挑战与展望]

match rank2 rank3 case2
content_to_nodes: 六、我国区块链发展面临的挑战与展望
before_all_child_content
```

六、我国区块链发展面临的挑战与展望

- (一) 主要挑战
- (二) 发展展望

基于文档结构树对
结果进行扩展，返
回该级标题下子标
题及相关片段内容

```
ques_id: 32, rank1_score: 5.30859375, rank1_result: 六、我国区块链发展面临的挑战与展望
ques_id: 32, extend_result1: 六、我国区块链发展面临的挑战与展望
回望过去，我国区块链产业发展已近十载，从初期以公有链为主导构建全球最为活跃的加密资产交易生态，到将重心转向联盟链推动区块链技术与实体经济深度融合，再到数据要素、Web3.0等新理念推动数字信任价值潜力释放，区块链一直在质疑中前进，在挑战中发展

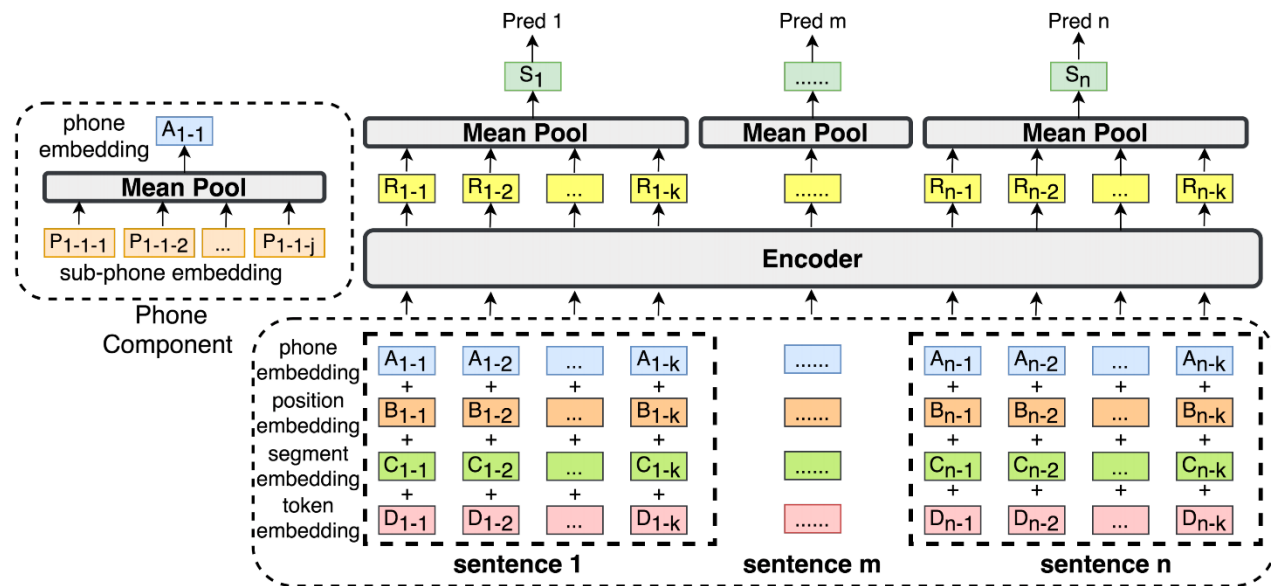
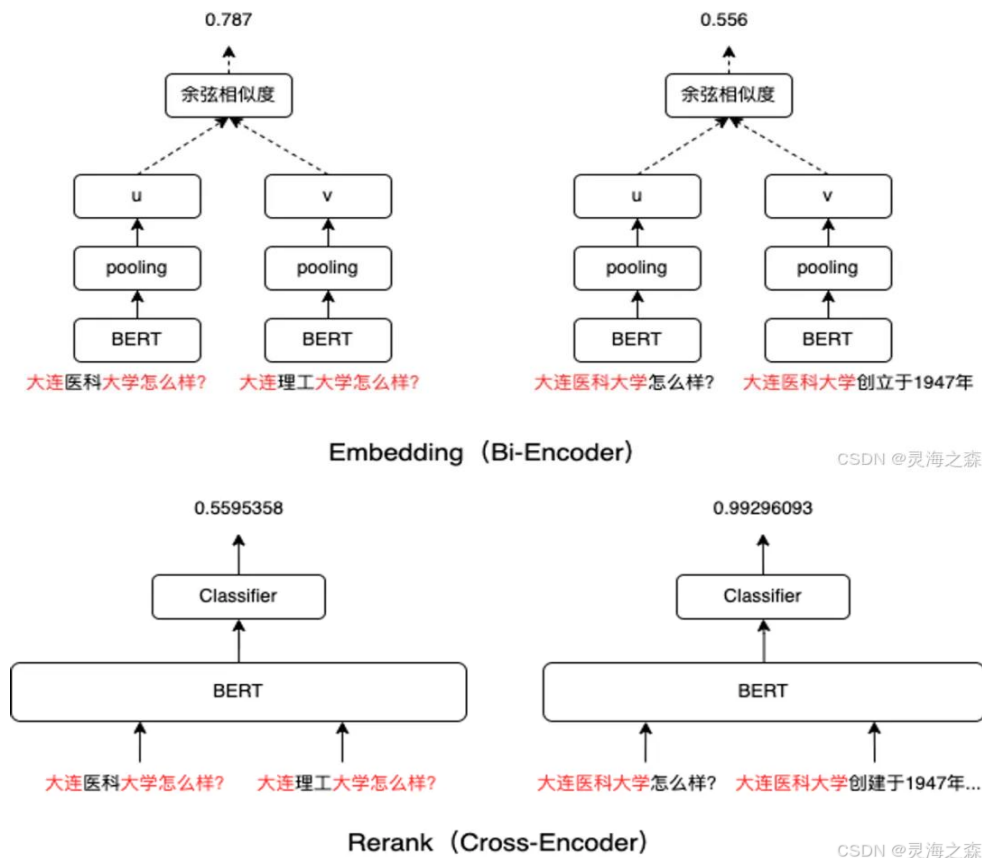
(一) 主要挑战
区块链技术自主研发能力存在短板，难以支撑应用深度创新
联盟链缺乏有效治理机制，产业规模化增长受限
区块链基础设施建设缺乏统筹，影响规模化应用

(二) 发展展望
在构建以数据为关键要素的数字经济的大背景下，我国区块链技术应用和产业发展取得了积极进展，技术研发、应用创新和产业发展持续突破，其服务多方业务高效协作、建立可信价值网络的关键作用进一步获得业界认可
支持大规模网络节点和高频交易的开放联盟链加快演进步伐
服务更大范围的价值流通的仍是区块链应用创新的重要方向
面向Web3.0和数据要素市场，区块链将迎来全新发展机遇
```

二、解决方案 – 主要使用到的模型

主要使用到的模型包括：

- ✓ bge-large embedding模型用于语义召回
- ✓ bge-reranker-large模型用于精排
- ✓ nlp bert document-segmentation chinese-base模型用于文档语义切分



- nlp_bert_document-segmentation_chinese-base模型基于阿里在论文《Sequence Model with Self-Adaptive Sliding Window for Efficient Spoken Document Segmentation》中提出了SeqModel
- SeqModel利用BERT对多个句子同时编码，建模了更长的上下文之间依赖关系之后再计算句向量，最后预测每个句子后边是否进行文本分割

三、算法效果

A榜和B榜成绩均排名前列，复现后以第二名的成绩入围赛题决赛，答辩后最终获得赛题第一进入总决赛。

A榜B榜

我的成绩

到目前为止，您的最好成绩为76.00150300分，第3名，在本阶段中，您已超越112支队伍。

排名	排名变化	队伍名称	有效提交次数	最高分提交时间	最高得分
	-	default13205898	1	2024-11-15 15:32	76.99024000
	-	yichi	1	2024-11-15 21:40	76.89067500
	↓ 1	default13172956	1	2024-11-15 23:36	76.00150300
4	↑ 1	default13197902	1	2024-11-15 23:14	74.78379200
5	-	default7681215	1	2024-11-15 23:59	74.78173800

B榜测评成绩排行榜

入围决赛名单

经过客观评测及代码审核，最终入围决赛队伍公布如下：

序号	队伍	队伍id
1	Babata AI (yichi)	336793
2	kelvincjr (default13172956)	336620
3	default13197902	347829
4	晓悟 (default7692247)	346318
5	default13296736	341579

复现后入围决赛排名（第二）

获奖名单

奖项	获奖团队
一等奖	kelvincjr (ID:336620) 拟入围总决赛
二等奖	南湖类脑RAG (ID:347829) Babata AI (ID:336793)
三等奖	晓悟 (ID:346318) HappyRAG (ID:341579)

赛题决赛后最终排名第一，入围总决赛

四、创新点总结

作为RAG的最佳实践，通过清晰完善的AI系统架构，构建出由文档解析、知识库建立、问题预处理、信息检索召回四个模块组成的一个完整的智能文档知识检索系统。除了完善的体系性外，该系统还有以下四个创新点：

可视化精细化的文档解析

- **基于OCR的版面分析：**通过OCR版面分析，识别出文档的标题、段落、表格、图片等元素，最大化还原文档内容结构。
- **准确的表格识别：**采用多种可视化表格识别算法，准确还原多种复杂的表格结构和内容。

结合多信息的问题重写

- **结合关键词和多种实体信息的问题重写：**通过提取关键词和多种实体信息，结合原问题进行重写，最大化突出需要召回的关键信息。

多维度多层次的知识索引

- **文档结构树：**创新性地通过模型+规则识别出文档的层次目录结构树，并基于结构树中文档元素的关系建立索引。
- **基于语义的文档分块：**对比传统的定长分块和段落分块，语义分块更加精细化，使得召回的文档片段具有完整的文章结构边界和具体语义。
- **分级+混合索引：**基于标题、段落、语义块、文档结构树建立分级索引，并采用关键词BM25和BGE dense embedding的混合索引结构。

多策略的信息召回排序

- **基于多策略综合打分的召回排序：**基于分层混合索引和BGE Reranker模型，结合标题、关键词、实体和语义综合打分，进行召回和初步排序。
- **有针对性的精确排序策略：**针对不同场景的精确排序策略，包括基于文档结构树的结果扩展重排策略、针对并列问句的多结果合并重排策略、针对长结果的二次切分压缩重排策略，保证结果的完整性和精准性。

五、应用探索和规划

- 通过使用RAG检索增强生成技术将私域数据作为大模型的外接知识库，可以很好地解决大模型知识幻觉的问题，这也是当前大模型应用落地的主要方式之一。
- 但使用大模型生成最终的结果，始终具有不确定性，而采用**智能信息检索技术**精准返回知识库中的**回答问题需要的知识片段**，**辅助机器人和人工**回答问题在实际中更具有应用价值。

潜在的应用场景：

- **智能客服和座席辅助场景：**主流的电商、银行和运营商，主要都是通过智能客服机器人和人工座席共同为客户提供服务，背后都是基于统一的知识库系统，知识检索是其中的关键。但传统的基于关键词的全文检索方式无法满足座席基于客户的当前问题精确检索定位到回答问题需要的知识片段。而**基于自然语言问答方式的智能知识检索可以有效解决这一问题，提升检出知识的完整性和准确度。**
- **企业内部的文档检索场景：**在企业内部会有很多不同类别的内部政策、规章和流程文档，如何使得员工能快速检索到相关的信息，极大影响到企业内部的运营效率。通过利用这些企业内部文档构建内部知识库，并**建立基于自然语言问答方式的企业知识搜索引擎**，能大幅提升员工的查找准确性和效率，实现企业内部的降本增效。

六、团队介绍

蔡嘉荣 (队长) - kelvincjr

- 中山大学计算机硕士，研究方向数据挖掘、机器学习。
- 目前任职某银行信用卡中心，数据科学专家，负责数据科学团队的管理。
- 多年的软件开发、架构经验，近几年专注于数据科学、机器学习、NLP和大模型等领域的研究和开发。
- 近年积极参加各种数据挖掘和算法类比赛，多次获得名次和奖项。



CCF BDCI

2024 CCF 大数据与计算智能大赛^{12th}

感谢观看!