软件课设中期报告

目录：

一、项目名称

法智先锋——智能法律咨询服务机器人

二、学生团队

易俊哲，龙文振，陈梓彧，郭珺碧

华中科技大学 电子信息与通信学院 电信2202班

三、指导老师

谭运猛

四、项目背景

随着中国特色社会主义进入新时代，全面依法治国已成为国家治理体系和治理能力现代化的重要标志。提升公民的法治意识和法律素养，对于维护社会稳定、促进社会和谐具有重要意义。然而，尽管我国在法治建设方面取得了长足进步，但由于法律知识的专业性和复杂性，普通民众在面对具体法律问题时，往往感到困惑和无助，无法迅速获取到准确有效的法律咨询信息。此外，传统的法律咨询服务方式（如律师事务所咨询、电话咨询等）存在成本高、效率低等问题，难以满足大众日益增长的法律需求。

为此，我们计划开发一款名为**“法智先锋”**的AI法律咨询服务机器人。该机器人专注于提供即时、高效的在线法律咨询服务，利用先进的自然语言处理技术和检索增强生成能力，能够理解用户的法律问题并快速给出专业解答。相比于传统法律咨询，我们需要研发的产品可以做到在低咨询成本、快回复速度的前提下提供高质量的解答。

五、需求分析

- 设计并实现一个Web界面，包括对话区、用户信息显示区。
- 实现基于大规模语言模型的对话生成功能。
- 构建本地知识库，实现与大模型的结合使用。
- 实现基于用户画像的个性化回答生成。
- 设计并实现中国法律领域的专业问答能力。
- 系统应能通过数据库实现信息的存储与交互。
- （选做）设计管理员后台，支持知识库的在线更新和系统性能监控。
- （选做）能够使用自己本地部署的开源大模型进行调用回答用户问题。

六、实施方案论证

本产品的实现逻辑为：加载文件 -> 读取文本 -> 文本分割 -> 文本向量化 -> 问句向量化 -> 在文本向量中匹配出与问句向量最相似的 `top k` 个 -> 匹配出的文本作为上下文和问题一起添加到prompt中 -> 提交给LLM生成回答

技术模型选择

1. 大规模语言模型
   选择：Qwen大模型
   理由： Qwen大模型具备强大的语言理解和生成能力，能够处理复杂的自然语言文本。本地部署能够提高生成回答的速度和质量，满足智能客服场景的需求。
   优势：
   - 相比于OpenAI免费。

- 强大的语言理解和生成能力。
- 本地部署提高系统稳定性和响应速度。

2. 知识库构建
   选择：**Milvus**
   理由： Milvus是高性能的向量搜索引擎，支持多种距离度量方法，能够高效处理大规模向量数据。
   优势：
   - 高效的向量检索能力。
   - 支持多种索引类型。
   - 易于扩展和维护。
3. 嵌入式模型
   选择：**text2vec**
   理由： text2vec是一个文本向量表征工具，把文本转化为向量矩阵，实现了Word2Vec、RankBM25、Sentence-BERT、CoSENT等文本表征、文本相似度计算模型，在Github上有4.5k个stars，是一个非常流行的开源工具。
   该工具实现了句子的Word2Vec向量表示。
   优势：
   - 提供高质量的向量表示。
   - 支持中文文本。
   - 便于后续处理。

检索增强生成

1. 结合知识库信息
   系统在生成回答时结合知识库中的信息，提高回答的准确性和详细程度。
2. **LLM**
   结合知识库的信息后，作为增强的Prompt，能让生成更专业的回答。

综上所述，通过选择Qwen大模型、Milvus和Text2vec，构建的智能客服系统在技术上是完全可行的。

目前已找到的数据源

- 北京大学开放研究数据平台上的法律数据
- 中国检查网：起诉书等
- 中国裁判文书网：裁决书、裁定书、决定书等
- 司法部国家司法考试中心：行政法规库、法考真题等

- 国家法律法规数据库：官方法律法规数据库
- 中国法律智能技术评测（CAIL）历年赛题数据
- 中国法研杯司法人工智能挑战赛（LAIC）历年赛题数据
- 百度知道法律问答数据集：约 3.6w 条法律问答数据，包括用户提问、网友回答、最佳回答
- 法律知识问答数据集：约 2.3w 条法律问答数据
- 中国司法考试试题数据集：约 2.6w 条中国司法考试数据集
- LaWGPT 数据集 @pengxiao-song：包含法律领域专有词表、结构化罪名数据、高质量问答数据等
- 法律罪名预测与机器问答 @liuhuanyong：包括罪名知识图谱、20w 法务问答数据等
- 法律条文知识抽取 @liuhuanyong：包括法律裁判文书和犯罪案例
- 中国法律手册 @LawRefBook：收集各类法律法规、部门规章案例等
- 刑法最新罪名一览表：记录2021年最新刑法罪名

七、系统架构

前端展示层

拟按照以下内容设计前端

- **对话区：**
    - 法律术语高亮：自动识别并高亮显示法律术语，点击术语可查看简短定义的解释。
    - 案例参考链接：对于某些类型的问题，AI可以提供相关案例的链接，用户点击后可在新窗口或对话框中查看案例详情。
    - 反馈机制：每个回答下方都有一个反馈按钮，用户可以选择`有帮助`或`无帮助`，并可添加具体意见，帮助系统自主迭代更新。
- **用户信息显示区：**
    - 基本信息展示：显示用户的用户名、头像等基本信息。
    - 法律咨询记录：提供一个单独的页面或面板，用户可以查看自己的咨询历史，包括问题描述、回答、相关案例链接等。

用户画像层

1. **用户信息：**
    - 标签化用户信息。
    - 提供贴心服务。
2. 在本产品中可能会用到的几个维度：
    - 用户的年龄，不同年龄段的用户对法律的需求通常不同。
    - 用户的教育程度，鉴于大多老年人受教育程度低，了解法律的程度也不同。
    - 用户的收入水平，用户处于何种社会阶层，通常会影响他们的法律需求。
    - 用户的兴趣爱好，如果用户喜欢购物，则可能需要电商相关的法律咨询。

数据处理层

1. **嵌入式模型：**
    - 使用Text2vec模型处理知识库文本。
    - 向量存储到Milvus。
2. **Milvus向量数据库：**
    - 快速检索相似问题。
    - 数据管理与监控。
3. **Qwen大模型：**
    - 本地部署，提高速度。
    - 定期更新模型。
4. **知识库系统：**
    - 存储结构化知识，支持快速查询。

任务分工

团队四人通力协作，同力完成所有工作！

进度安排

- **阶段一：** 产品方案论证，团队成员搜集资料，讨论方案，确定产品最终方案。
- **阶段二：** 搭建系统后端，团队成员共同完成后端模型的搭建。
- **阶段三：** 完善前端功能，美化UI设计，增强用户体验。

- **阶段四：** 排查错误验收，进行安全测试，实现并发功能。