项目名称		密级
文脉医典——医学文献智能识别与检索系统		仅供收件方查阅
项目编号	版本	文档编号
		Project ID_INIT_002

文脉医典——医学文献智能识别与检索系统

项目立项报告

拟制	翟苌坤 饶艺 谭雪瑶 昌宇	日期	2024-12-23
评审人	王文鑫	日期	2024-12-23
批准	王文鑫	日期	2024-12-23

武汉市软酷网络科技有限公司

Revision Record

修订记录

Date 日期	Revision Version 修订版本	CR ID /Defect ID CR/ Defect 号	Sec No. 修改章节	Change Description 修改描述	Author 作者

Catalog

目 录

1 项目提出

项目名称：文脉医典——医学文献智能识别与检索系统

项目简介：

医药行业医药代表的日常工作中，不仅要向客户提供产品注册证等资质文件，也同时需要提供更多专业性文献资料。这些医学文献不仅数量巨大，渠道众多，而且文件中含有大量的医学图片，普通的检索功能难以满足医药代表和医生的日常工作需求。为了医疗企业的合规性要求，更及时地服务于内外部，进一步推进企业的数智化转型，打造全新的医学文献智能识别检索系统迫在眉睫。医学文献作为医疗企业的核心资料，工作人员需要向医药代表快速且精准提供相关医学信息与文献已经成为日常工作，借助人工智能技术打造医学文献识别检索系统势在必行。当前医疗行业中医学文献检索的痛点：（1）大量的医学文献依靠工作人员人工阅读记录并筛选，效率低下，且造成巨大的人力消耗与浪费。（2）对于工作人员有极高的专业判断水平要求。（3）常规系统的检索功能无法识别到医学影像图片中的文字，导致大量的重要医学文献与信息无法及时提供项目为解决当前医学文献检索的痛点，借助图像文字识别技术（OCR）打造"医学文献智能识别检索系统"，实现对大量有医学影像图片的医学文献进行智能识别与管理，精准快速地获取医药代表与医生所需的文献资料。

项目目标：

本项目的主要目标是利用先进的 OCR（光学字符识别）技术，设计并实现一个医学文献智能识别与检索系统。该系统旨在解决当前医学文献检索过程中存在的效率低下、人力消耗大以及医学影像图片中文字无法有效识别等问题。具体目标包括：**实现医学文献的智能识别**：通过 OCR 技术，系统能够自动识别并提取医学文献中的文字信息，特别是医学影像图片中的关键诊断结果和病理描述等。**构建高效的检索机制**：基于识别出的文字信息，系统应提供多种检索方式，如关键词搜索、作者搜索、发表日期搜索等，以便用户能够快速找到所需的医学文献。**提升用户体验**：系统设计应简洁易用，提供友好的用户界面和交互体验，确保医药代表和医生能够轻松上手并高效使用。**确保数据安全与合规性**：系统应严格遵守医疗行业的合规性要求，确保医学文献数据的存储、处理和访问都符合相关法规和标准。通过实现这些目标，本项目旨在提高医学文献的检索效率，降低人力消耗，推动医疗行业的数智化转型，并为医药代表和医生提供更加便捷、高效的医学文献检索服务。

系统边界：

1. **功能边界**：

- 检索功能：本产品的核心功能，用以实现用户通过关键字检索 PDF 文件，并返回根据文献关键字，文献名称，文献内容优先级排序的查询结果。包括全文检索、关键词高亮、结果排序等。
- 用户功能：用户中心提供用户注册和登录服务，并允许用户查询和修改自己的基本信息。
- 后台功能：收集医学文献数据集，可能包括 PDF 文件、图像文件等。对数据进行预处理，如格式转换、清洗、去重等，以确保数据质量。将文本并通过大模型转化为数据库内容存储数据库，并供后台数据驾驶舱进行数据分析，提取数据关键字以供用户检索。
- 数据驾驶舱：开发后台数据分析和展示界面，包括数据可视化、统计报表、用户行为分析等，并管理文本、用户的数据以及状态信息。

2. **技术边界**：

前后端开发采用 flask，数据库使用 mysql，集成 OCR 技术，将图像中的文字转换为可检索的文本，并实现高效的搜索算法。

- **数据采集范围**：主要包括 PDF 文献、图像文件和纯文本文件，所有数据需符合版权及隐私保护要求。**图片转文本**：使用 tesseract 库进行图片转文本处理**预处理技术**：涵盖格式转换、数据清洗、去重、分词与标注、图像处理等技术，确保数据的结构化和高质量。**存储技术**：使用关系型或非关系型数据库进行数据存储，支持高效检索与可扩展性。**检索技术**：采用 LDA 进行文本关键字提取。

工作量估计：

模块	子模块	工作量估计（人天）	说明
用户中心	用户注册与登录	2	用户中心提供用户注册和登录服务，注册功能可能包括手机注册和邮箱注册，需要发送验证码和验证邮件来完成注册流程登录功能则支持多种方式，包括账号密码登录、手机登录、邮箱登录以及第三方登录（如微信、QQ、微博）
	基本信息查询与修	2	用户中心允许用户查询

	改		和修改自己的基本信息。
	用户信息、权限管理	2	用户中心负责管理用户的个人信息，还涉及账号的创建、管理、注销以及状态控制等功能。
数据库设计		5	设计数据库模型，包括用户表、文献表、检索记录表等。需要考虑数据的规范化、索引优化以提高检索效率，以及数据的安全性和备份策略。
数据集收集调整		5	收集医学文献数据集，可能包括PDF文件、图像文件等。需要对数据进行预处理，如格式转换、清洗、去重等，以确保数据质量。
核心检索模块		36	开发核心检索功能，包括全文检索、关键词高亮、结果排序等。需要集成OCR技术，将图像中的文字转换为可检索的文本，并实现高效的搜索算法。
后台数据驾驶舱		36	开发后台数据分析和展示界面，包括数据可视化、统计报表、用户行为分析等。需要选择合适的数据可视化工具和库。
总工作量（人天）：	88		
模块	子模块	工作量估计（人天）	说明

表 1 工作量估算

备注：“人天”即 1 个人工作 8 小时的量就是 1 人天

2 开发团队组成和计划时间

项目计划：2024 年 12 月 23 日 - 2025 年 1 月 11 日 （计 1 月）

项目总监： 1 人 姓名：王文鑫

项目经理： 1 人 姓名： 王文鑫

项目成员： 4 人

人员来源：

昌宇 华中师范大学计算机学院 2021 级本科生

饶艺 华中师范大学计算机学院 2021 级本科生

翟茬坤 华中师范大学计算机学院 2021 级本科生

谭雪瑶 华中师范大学计算机学院 2021 级本科生

3 项目预计支出

设备,场地占用费:

无

本地人员工资(管理费)：

(average salary+ management fee)* number of staff * months= 10000*4*1=40000 元

(平均工资+管理费)*人员数目*月份= 10000*4*1=40000 元

外协人员工资：

无

加班费：

无

交通费：

无

住宿费：

无

其它费用（如业务交往,招待,办公等）：

无

总计：4 万元

说明： 无

4 风险评估和规避

技术风险：OCR 技术识别精度问题：风险描述：OCR 技术在识别医学影像图片中的文字时，可能受到图片质量、字体大小、排版等因素的影响，导致识别精度不高，进而影响检索效果。解决方法：采用先进的 OCR 算法，如深度学习模型，进行训练和优化，提高识别精度。同时，对医学影像图片进行预处理，如去噪、增强对比度等，以改善识别效果。**数据格式多样性**：风险描述：医学文献可能来自不同的出版社、数据库或学术平台，数据格式和编码方式可能不同，这可能导致数据导入和处理时出现问题。解决方法：建立统一的数据格式标准，对接收到的医学文献进行格式转换和编码统一处理，确保数据能够顺利导入系统。

管理风险：

1.团队协作问题：风险描述：项目团队成员可能来自不同的专业背景，协作时可能出现沟通障碍或理解偏差。解决方法：加强团队建设，提高团队成员之间的沟通和协作能力。定期组织项目会议，分享进展和经验，及时解决问题。

其它风险：

数据安全风险：风险描述：医学文献中可能包含敏感信息，如患者隐私、药物配方等，若系统安全措施不到位，可能导致数据泄露。解决方法：加强系统安全防护，采用加密存储、访问控制等技术手段，确保数据安全。同时，建立数据备份和恢复机制，以应对数据丢失或损坏的情况。**合规性问题：**风险描述：医疗行业的合规性要求严格，若系统不符合相关法规和标准，可能面临法律风险。解决方法：在项目设计和实施过程中，严格遵守医疗行业的合规性要求，确保系统的合法性、合规性和安全性。同时，加强与行业监管机构的沟通和合作，及时了解法规变化和政策导向。