

《个人知识库助手系统》

用 户 需 求 说 明 书

v1.0

第三小组

|  |
| --- |
|  |

**版 本 历 史**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 版本/状态 | 作者 | 参与者 | 日期 | 备注 |
| 1.00 | 万雨晴 | 刘御玺、汪子涵 | 2025-08-25 | 创建 |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

**目 录**

[第一部分 引言 4](#_Toc2105874386)

[一、说明 4](#_Toc1868694499)

[二、定义 4](#_Toc1738970877)

[第二部分 综述 5](#_Toc1063998642)

[一、项目背景 5](#_Toc413415867)

[二、建设目标 6](#_Toc610448315)

[三、建设原则 6](#_Toc216971224)

[四、用户业务需求说明 8](#_Toc691148066)

[1、整体业务需求示意图 8](#_Toc538791828)

[2、需求详细说明 8](#_Toc1423935600)

[2.1 导入文档流程描述 8](#_Toc150787258)

[2.2文档检索流程描述 10](#_Toc1054651529)

[.2.1 输入关键词检索 10](#_Toc2096854889)

[.2.2 用标签检索 10](#_Toc1637275467)

[用 10](#_Toc1033589580)

[2..3 利用知识库内容智能回答用户问题流程描述 11](#_Toc403791447)

[用户自然语言输入内容，大模型将内容转化成向量，向量在faiss数据库中找到相关内容再找到对应的markdown文件返回给用户，用户可以选择合适的部分或者全部文件作为上下文给ai生成回答再返回给用户。 11](#_Toc1247390223)

[第三部分 需求分析 12](#_Toc1601165253)

[一、用例分析 12](#_Toc252075733)

[1、 用户用例 12](#_Toc834541101)

[1.1 用户登录用例描述 12](#_Toc884952330)

[二、界面风格 14](#_Toc1573349950)

[第四部分 验收标准 16](#_Toc1221252640)

[一、功能范围定义 16](#_Toc1048594374)

[二、性能指标定义 17](#_Toc1482883854)

[第五部分 环境和部署要求 18](#_Toc7131711)

[一、网络部署图 19](#_Toc1977376269)

[二、应用部署图 19](#_Toc1692114586)

[四、运行环境说明 21](#_Toc1539172856)

[1、服务器 21](#_Toc2137720829)

[2、客户机器环境 21](#_Toc503861674)

第一部分 引言

### 一、说明

|  |
| --- |
| 编写本说明书的目的是为了准确阐述项目具体业务需求和需求边界，本说明书的作者是【个人知识库助手】项目组，本说明书的确认者是【项目经理】负责人，本说明书的读者是项目所有直接干系人。本说明书是指导项目实施的重要指导性文件，也是用户最后进行验收（终验）的依据，说明书中内容一旦确认双方将以此为基础开展工作。如果需要变更说明书内容，必须走变更流程，变更必须得到甲乙双方书面确认，最后变更内容将作为本文的一部分，在项目实施过程中得以体现。 |

### 二、定义

|  |
| --- |
| **云端同步：**通过云服务器，实现数据在不同设备间的自动更新与备份，保证多终端访问数据的一致性，支持实时更新，提升使用流畅度。  **文本向量(Embedding)：**将文本转化为一组高纬度的数字向量，用来表示文本的语义特征，相似的文本会在向量空间中更接近，通过计算向量之间的距离来度量文本语义相似度，从而实现语义搜索和知识库检索。  **向量数据库**：一种专门用来存储和检索向量的知识库，支持大规模向量的快速相似度检索，适合与Embedding结合，实现语义级别搜索。  **检索增强生成(RAG)：**一种结合信息检索和大模型生成的技术，能够提高回答的准确性和可信度，减少大模型幻觉的风险。  **MAAS：**通过API的形式向用户提供大模型的能力，用户无需本地训练和维度模型，降低硬件成本与运维成本，用户只需调用接口即可使用模型能力。  **幻觉：**AI“幻觉”指模型生成看似真实可信，但实则为错误或凭空捏造的信息。这并非模型有意为之，而是源于其训练数据及工作方式的固有缺陷。 |

第二部分 综述

### 一、项目背景

|  |
| --- |
| 人工智能教育与科研行业正处于快速发展阶段。随着人工智能、大模型等技术的广泛应用，高效、科研机构以及企业对于知识的获取、整理和利用的需求急剧增加。然而，目前大多数的用户依然依赖于传统的笔记软件和文档管理工具，这些工具普遍存在一些问题，例如：检索依赖关键词，缺乏语义级别的理解；手动管理知识内容，耗费大量时间；难以支撑多端同步和多人协作；缺乏智能问答和个性化的推荐内容。  用户每天接触的知识量巨大，传统方式无法支撑高效的整理与应用。随着大模型对语言文字处理能力的不断增强，加速了智能学习、科研辅助的需求，行业急需低成本可落地的解决方案。从个人、团队到高校，都希望能通过知识管理工具提升教育质量和科研水平。同时，人工智能时代，缺乏智能化知识管理工具的项目逐渐被淘汰，开发一款大模型+向量数据库的个人知识库助手具有迫切性和必要性。  我们致力于打造一款智能的个人知识库助手，实现语义检索和智能问答，突破传统工具的限制，用户能够快速获取关键信息，提升学习效率与科研成果产出。我们希望个人知识库助手能成为人工智能教育与科研行业的基础软件，降低知识管理成本，推动行业从“被动存储”向主动智能转型，具有广阔的市场前景和应用价值。 |

### 二、建设目标

|  |
| --- |
| 我们致力于打造一款智能化的个人知识库助手。该系统支持PDF、MD、Docx等多种市面常见文档的导入，简化信息的聚合过程。在技术上，我们采用RAG（检索增强生成）技术架构，结合Embedding方法与开源向量数据库Faiss来存储和处理数据，这不仅为大型模型提供了精准的外部知识库，有效减少了AI幻觉，还实现了对知识内容的高效模糊搜索，使用户能更快定位到原文相关片段。为优化管理，系统支持灵活的标签分类与标签搜索功能，让知识体系井然有序。同时，我们高度重视数据安全与便捷性，​提供个人数据的云同步与云备份服务，守护您的数字资产。最后，通过支持丰富个性化设置的交互界面，我们致力于满足每位用户的独特需求，最终构建一个真正高效、安全且易用的个人知识管理中心。 |

### 三、建设原则

**（一）实用有用**

|  |
| --- |
| 聚焦知识采集、存储、检索和智能问答等核心功能，确保用户高效管理学习资料、科研文献和工作文档，解决实际“找不到、用不快”的问题。 |

**（二）灵活先进**

|  |
| --- |
| 系统要有一定的灵活性，要能适应用户需求的变化，大模型未来的发展，用户规模的扩大等。在技术上系统要有一定的成熟性和前瞻性，必须考虑跨平台和负载均衡等。既利用前沿AI能力提升检索与问答效果，又通过模块化架构支持功能灵活组合，满足不同用户需求。 |

**（三）界面友好**

|  |
| --- |
| 充分考虑用户实际需求，前端设计以“简洁易用”为原则，提供对话式问答界面、文档上传入口和标签管理功能，使用户无需复杂培训即可快速上手。 |

**（四）兼容扩展**

|  |
| --- |
| 要求系统要有整体规划，能适未来用户的文档管理编辑的更多需求和大模型技术的发展，要能和现有外部系统有限集成，要能提供对后续个人知识库助手平台的支持。系统支持多种数据来源（文档、笔记等）接入，预留多人协作、私有化部署等扩展能力，保证软件可持续发展和适配多样化场景。 |

**（五）安全可靠**

|  |
| --- |
| 系统必须保证网路、硬件、软件和服务体系安全，必须保证系统数据安全，必须提供运行环境自动检测和故障恢复功能。采用API Key管理、权限控制和数据加密措施，保证用户个人知识库在云端同步、多人访问和多人协作时的数据安全与隐私保护。 |

### 四、用户业务需求说明

#### 1、整体业务需求示意图

|  |
| --- |
|  |

#### 2、需求详细说明

##### 2.1导入文档流程描述

1）文件的存储和映射关系。其中原文件和markdown文件保存在MySQL数据库，markdown文件生成的向量保存在faiss数据库，同时记录markdown文件对应的向量的id为vecter\_id在markdown表中。

|  |
| --- |
|  |

1. 数据库关系图：为了实现标签功能，多人协作功能，分别设置了user表，document表，markdown表，tag表。为了实现依靠tag检索设置了关系表documents\_tages。

|  |
| --- |
|  |

##### 2.2文档检索流程描述

###### .2.1 输入关键词检索

户输入检索关键词，利用关键词依靠hnsw方法在faiss数据库检索到对应的向量，和其对应的vecter\_id，再通过vecter\_id在markdown表中检索到对应的markdown呈现给用户。

###### .2.2 用标签检索

用户通过选择一个或多个tag可以依靠documents\_tags表检索到符合要求的相关文档。

##### 2..3 利用知识库内容智能回答用户问题流程描述

##### 用户自然语言输入内容，可以选择使用哪个大模型目前以及支持的大模型有智普ai、deep seek、通意千问，大模型将内容转化成向量，向量在faiss数据库中找到相关内容再找到对应的markdown文件返回给用户，用户可以选择合适的部分或者全部文件作为上下文给ai生成回答再返回给用户。

### 3 市面上主流知识库功能对比及未来展望

#### 3.1 市面上主流知识库功能

当前市面上的智能知识库助手，其核心目标是让信息和知识变得****更容易获取、更高效利用****。它们已经超越了简单的文件存储，更像是一个能理解你、能对话的“****第二大脑****”。

根据它们侧重的技术和应用场景，主要的功能可以归纳为以下几个方面。



#### **3.2我们的知识库完成的功能**

对标市面上主流知识库功能我们的知识库目前已尽做到了，支持多格式文档转化、智能分类与标签化、自然语言问答、多轮对话与上下文理解、内容自动摘要功能。比起市面上主流的智能知识库我们的知识库还有很大不足，主要在知识协作与流转方面以及安全与管控。目前我们的时间和技术水平不支持多人同时协作修改，对于数据安全方面做到了基本的用户密码哈希加密，和保证用户存储内容不会被其他用户访问但是企业级别的数据安全对我们来说还是过于艰巨。目前我们的主要还是偏向文档的智能管理和检索，以及ai对话，但是对文档编辑功能还有待开发，未来我们将继续投入时间开发智能知识库的其他扩展功能，使它可以向市面上的其他知识库助手看齐。

#### 3.3未来展望

对于后续的开发需求，我们准备先加入文档编辑功能，在文档的增删之后实现文档的编辑，在完成单人文档编辑后，我们希望完成多人文档编辑。多人文档编辑最先可以设置临界区，规定同一时间只支持单人编辑，后续可以仿照github支持多人编辑的同时需要用户手动合并冲突部分，实现文件的版本管理。再后续，希望可以和市面上主流文档管理助手例如飞书看齐，支持多人同时在线编辑，且不需要手动合并内容。与此同时多人协作也需要权限管理相关需求开发的跟进，以保证多人协作之后的文档安全问题。

在大模型方面，我们希望未来我们的智能知识库助手可以支持更多的大模型例如豆包、kimi等等给用户更多的选择。

第三部分 需求分析

### 一、用例分析

#### 1、 用户用例

|  |
| --- |
|  |

##### 1.1.1 用户登录用例描述

|  |  |
| --- | --- |
| ID | A001 |
| 用例名称 | 用户登录 |
| 父用例ID | - |
| 主要执行者 | 主用户 |
| 前置条件 | - |
| 事件流 | 1. 打开登录页面 2. 输入用户名及密码 3. 输入验证码 4. 点击登录按钮 |
| 可选事件流 | - |
| 异常事件流 | - |
| 后置条件 | - |

##### 1.1.2 用户注册用例描述

|  |  |
| --- | --- |
| ID | A001 |
| 用例名称 | 用户注册 |
| 父用例ID | - |
| 主要执行者 | 主用户 |
| 前置条件 | - |
| 事件流 | 1. 打开注册页面 2. 输入用户名及密码还有邮箱 3. 点击注册按钮 |
| 可选事件流 | - |
| 异常事件流 | - |
| 后置条件 | - |

##### 1.1.3 用户注销用例描述

|  |  |
| --- | --- |
| ID | A001 |
| 用例名称 | 用户注销 |
| 父用例ID | - |
| 主要执行者 | 主用户 |
| 前置条件 | - |
| 事件流 | 1. 打开注销页面 2. 输入用户名及密码 3. 点击注销按钮 |
| 可选事件流 | - |
| 异常事件流 | - |
| 后置条件 | - |

##### 1.2.1 导入文件用例说明

|  |  |
| --- | --- |
| ID | A002 |
| 用例名称 | 导入文件 |
| 父用例ID | - |
| 主要执行者 | 主用户 |
| 前置条件 | - |
| 事件流 | 1. 选择导入的知识库或创建 2. 选择本地文件导入知识库 |
| 可选事件流 | - |
| 异常事件流 | - |
| 后置条件 | - |

##### 1.2.1 删除文件用例说明

|  |  |
| --- | --- |
| ID | A002 |
| 用例名称 | 删除文件 |
| 父用例ID | - |
| 主要执行者 | 主用户 |
| 前置条件 | - |
| 事件流 | 1. 选择删除文件所在的知识库 2. 选择要删除的文件 |
| 可选事件流 | - |
| 异常事件流 | - |
| 后置条件 | - |

**1.3 增删知识库用例说明**

|  |  |
| --- | --- |
| ID | A003 |
| 用例名称 | 增删知识库 |
| 父用例ID | - |
| 主要执行者 | 主用户 |
| 前置条件 | - |
| 事件流 | 1. 进入目标知识库 2. 选择目标的分段内容 3. 增加、删除该内容 4. 更新知识库 |
| 可选事件流 | - |
| 异常事件流 | - |
| 后置条件 | - |

**1.4 输入检索内容**

|  |  |
| --- | --- |
| ID | A004 |
| 用例名称 | 输入检索内容 |
| 父用例ID | - |
| 主要执行者 | 主用户 |
| 前置条件 | - |
| 事件流 | 1. 用户输入想要检索的内容（ 2. 大模型转化为文本向量 |
| 可选事件流 | - |
| 异常事件流 | - |
| 后置条件 | - |

**1.5 知识库检索内容用例说明**

**用户可以选择检索内容所用的大模型，目前有智普AI、deepseek、通意千问可以作为选择**

|  |  |
| --- | --- |
| ID | A005 |
| 用例名称 | 知识库检索内容 |
| 父用例ID | A004 |
| 主要执行者 | 主用户 |
| 前置条件 | - |
| 事件流 | 1. 选择大模型 2. 大模型计算文本向量相似度 3. 根据topk原则检索知识库 4. 返回k个分段内容 |
| 可选事件流 | - |
| 异常事件流 | - |
| 后置条件 | - |

**1.6 选择上下文AI回答用例说明**

**用户可以选择生成回答的大模型，目前有智普AI、deepseek、通意千问可以作为选择**

|  |  |
| --- | --- |
| ID | A006 |
| 用例名称 | 选择上下文AI回答 |
| 父用例ID | A005 |
| 主要执行者 | 主用户 |
| 前置条件 | - |
| 事件流 | 1. 用户选择大模型 2. 用户选择分段内容作为上下文 3. AI大模型生成回答 |
| 可选事件流 | - |
| 异常事件流 | - |
| 后置条件 | - |

### 二、界面风格

WEB登录界面：

|  |
| --- |
|  |

主页面：

|  |
| --- |
|  |

知识库创建与导入知识功能页面：

|  |
| --- |
|  |

第四部分 验收标准

### 一、功能范围定义

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| # | 产品 | 模块 | 组件 | 规格/型号 | 角色 | 接入 |
| 1 | 个人知识库 | 导入单元 | 导入文件模块 | 支持 txt/pdf/docx/md 等常见格式 | 数据提供方 | 文件上传接口 |
| 2 | 文件预处理模块 | 分块/去噪/分词/清洗规则 | 数据清清洗者 | 数据预处理 |
| 3 | 知识库单元 | 文本向量模块 | Embedding 模型（如 text-embedding-3-large，dim=3072） | 向量生成器 | LLM API |
| 4 | 存储向量模块 | 向量数据库 | 存储管理器 | 向量存储引擎 |
| 5 | 向量检索模块 | 相似度检索 | 检索服务检索API |  |
| 7 | 问答模块 | 用户提问模块 | 支持自然语言输入，文本/语音接口 | 用户入口 | Web前端 |
| 8 | 大模型上下文模块 | RAG（检索增强生成），支持长文本上下文拼接 | 上下文管理者 | 向量检索 |
| 9 | 大模型回答模块 | ChatGPT / DeepSeek / GLM 等通用大模型 | 答复生成器 | LLM API |

### 二、性能指标定义

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| # | 产品 | 模块 | 组件 | 规格/型号 | 性能级别 |
| 1 | 个人知识库 | 导入单元 | 导入文件模块 | 支持txt/docx/pdf/md等常见格式 | 基础 |
| 2 | 文件预处理模块 | 分块/去噪/分词/清洗规则 | 标准 |
| 3 | 知识库单元 | 文本向量模块 | Embedding 模型 | 高级 |
| 4 | 存储向量模块 | 向量数据库FAISS | 高级 |
| 5 | 向量检索模块 | 相似度检索 | 高级 |
| 6 | 索引管理模块 | 支持动态更新/重建索引 | 标准 |
| 7 | 问答模块 | 用户提问模块 | 知识自然语言输入，文本接口 | 基础 |
| 8 | 上下文管理模块 | RAG检索，支持上下文拼接 | 高级 |
| 9 | 回答生成模块 | 大模型回答模块 | 高级 |

第五部分 环境和部署要求

### 一、网络部署图

|  |
| --- |
|  |

### 二、应用部署图

|  |
| --- |
|  |

### 四、运行环境说明

#### 1、服务器

|  |
| --- |
| 1）操作系统​  Windows Server 2016 或更高版本  ​2）内存​  最低 8 GB RAM，建议 16 GB 或以上  ​3）硬盘空间​  系统盘：至少 100 GB 可用空间  数据存储：建议 500 GB 以上 SSD  ​4）运行时环境​  Python: 3.10 或更高版本 |

#### 2、客户机器环境

|  |
| --- |
| 1）硬件​  CPU：Intel i5 9代或同等性能及以上  内存：最低 4 GB RAM，建议 8 GB  ​2）操作系统​  Windows 10 或 Windows 11  ​3）浏览器​  Chrome 90+ / Firefox 88+ / Edge 90+ |