知识引擎原子能力

操作指南

腾讯云

操作指南

最近更新时间：2025-02-28 14:34:32

概述

RAG（Retrieval-Augmented Generation）是一种结合文档检索和生成模型的技术，用于回答复杂的问题。它结合了基于检索的方法和基于生成的方法的优点，在处理长文本和提供详细回答方面表现出色。

工作原理

RAG 通过以下几个步骤实现文档检索和问答：

- **文档加载和解析**

  首先，系统从本地或网络源加载文档。文档可以是任何格式，如 PDF、DOCX、TXT 等。
  文档内容被解析并转换为纯文本格式。
- **文本拆分**

  解析后的文本被分割成较小的文本块（chunks）。这种分割有助于系统更精细地处理和检索文档内容。
- **向量化**

  每个文本块被转换为嵌入向量。嵌入向量是文本内容的向量表示，用于计算文本之间的相似度。
- **向量数据库**

  嵌入向量存储在向量数据库（VectorStore）中。向量数据库用于高效地存储和检索嵌入向量。
- **查询处理**

  用户输入查询。系统将查询转换为嵌入向量。
- **相似度计算**

  系统计算查询向量与文档内容向量之间的相似度。通过相似度计算，系统可以找到与查询最相关的文本块。
- **提示模板生成**

  系统将相关的文本块与查询合并，生成一个提示模板。这个模板将用于生成最终的回答。
- **生成回答**

  提示模板被输入到大型语言模型（如混元大模型）中，语言模型根据提示模板生成最终的回答。

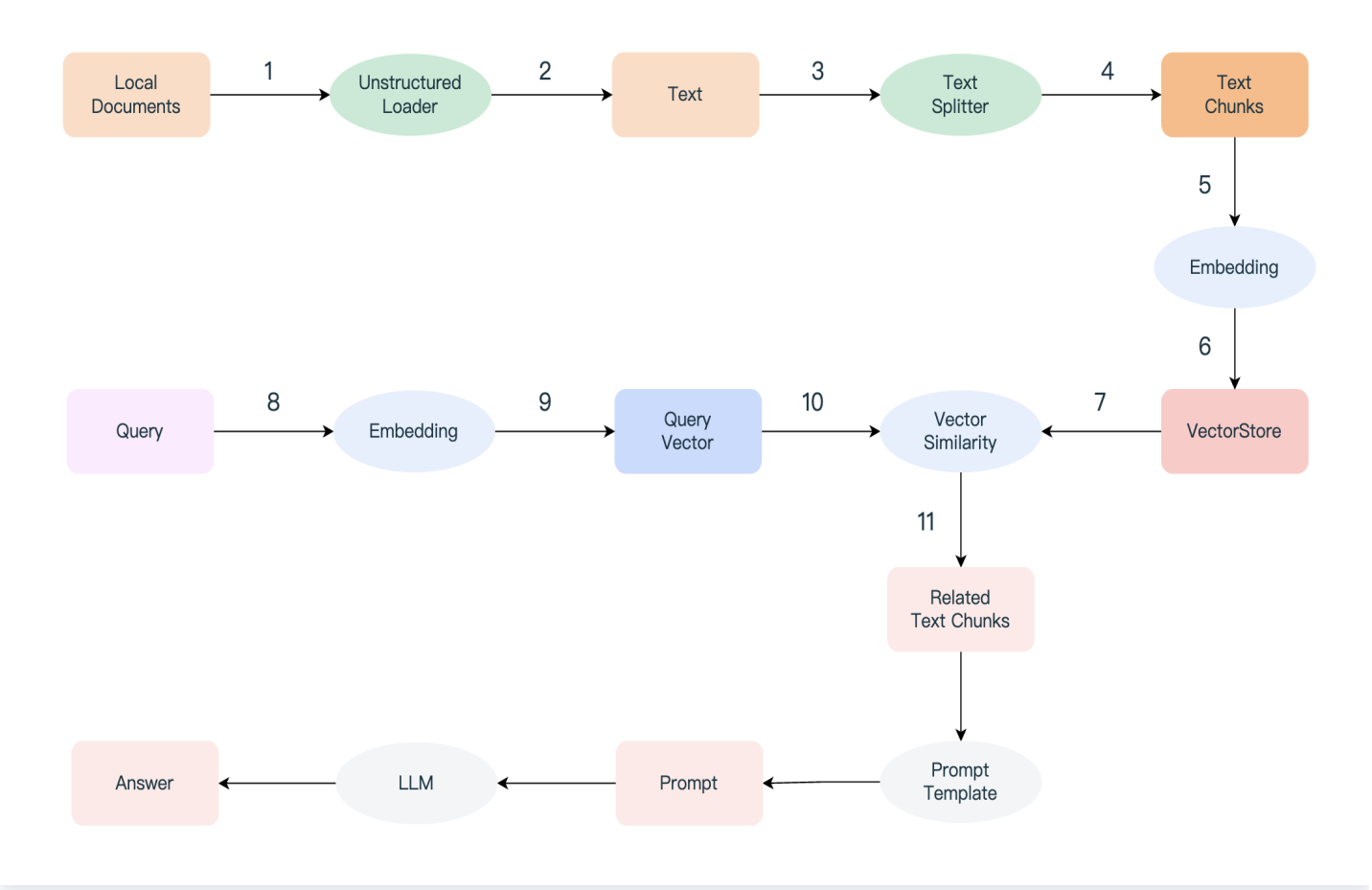
流程图

- **文档解析功能（步骤1-2）**

  涵盖从本地文档的加载到文本内容的提取。这一阶段的目的是将非结构化的文档内容转换为纯文本形式，便于后续处理。
- **文档拆分功能（步骤1-4）**

  包括文档解析功能，并进一步将提取的文本内容分割成较小的文本块。这种分割能够更精细地处理和组织文档内容，提升处理效率和检索效果。
- **RAG 综合能力套件（步骤1-11）**

从文档加载、内容提取、文本分割到嵌入向量的生成和存储，涵盖整个流程。通过结合检索和生成模型的优势，实现高效的文档内容检索和精确的问答生成。



如何开通功能

知识引擎原子能力为大模型知识引擎子产品，需开通大模型知识引擎体验权限后进行使用。大模型知识引擎的开通使用需要先通过 腾讯云企业实名认证 或者 腾讯云个人实名认证 。通过实名认证后，首次在 大模型知识引擎产品介绍页 单击**产品体验**，即可开通知识引擎原子能力使用权限。详情请查看 快速入门 。

文档解析功能流程说明

1. **提交文档**

   首先，您需要将需要解析的文档提交到系统中(CreateReconstructDocumentFlow)。支持的文档格式包括 PDF、DOCX、TXT 等，具体文档格式请参考接口文档说明。
2. **查询解析任务**

   提交文档后，您可以通过查询接口(GetReconstructDocumentResult)检查文档解析任务的状态。系统将会解析文档内容并生成相应的结构化数据。

文档拆分功能流程说明

1. **提交文档**

   首先，您需要将需要拆分的文档提交到系统中(CreateSplitDocumentFlow)。支持的文档格式包括 PDF、DOCX、TXT 等，具体文档格式请参考接口文档说明。
2. **查询拆分任务**

   提交文档后，您可以通过查询接口检查文档拆分任务的状态(GetSplitDocumentResult)。系统将会解析并拆分文档内容并生成相应的结构化数据。

RAG 综合能力套件功能流程说明

1. **创建知识库**

   首先，您需要创建一个知识库(CreateKnowledgeBase)，用于存储问答对或文档。
2. **上传问答对或文档**

   接下来，将问答对(CreateQA)或文档(UploadDoc)上传到知识库中。系统会对这些内容进行处理和索引。
3. **查询文档状态**

   上传完成后，您可以通过查询接口(DescribeDoc)检查文档处理的状态。系统将解析和索引文档内容。
4. **进行检索**

   一旦文档处理成功，您即可通过检索接口(SearchKnowledge)进行内容检索，获取相关的问答或文档内容。