MaxKB

第三节 知识库管理

飞致云 培训认证中心
2024 年 12 月

FIT2CLOUD 飞致云

目录

CONTENTS

01

RAG 技术相关概念

LLM 存在的问题

FIT2CLOUD 飞致云

幻觉问题

LLM 文本生成的底层原理是基于概率的 token by token 的形式，因此会不可避免地产生“一本正经的胡说八道”的情况。比如：你说，”博物馆下周一开门吗？“，很有可能给你回复：”开门”，到时候就是白跑一趟了。

时效性问题

LLM的规模越大，大模型训练的成本越高，周期也就越长。那么具有时效性的数据也就无法参与训练，所以也就无法直接回答时效性相关的问题，例如“帮我推荐几部热映的电影？”

数据安全问题

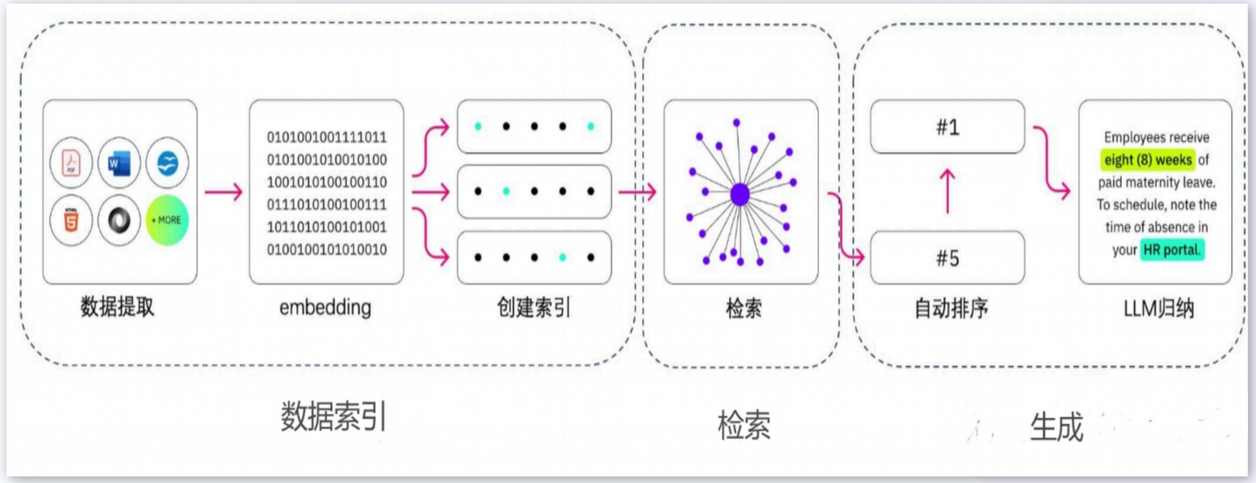
通用的LLM没有企业内部数据和用户数据，那么企业想要在保证安全的前提下使用LLM，最好的方式就是把数据全部放在本地，企业数据的业务计算全部在本地完成。而在线的大模型仅仅完成一个归纳的功能。

RAG 是什么？

FIT2CLOUD 飞致云

RAG是大模型的“外挂”

RAG是 “Retrieval-Augmented Generation” 的缩写，中文可以翻译为“检索增强生成”。这是一种结合了检索（Retrieval）和生成（Generation）的自然语言处理技术，用于提高语言模型在特定任务上的性能和准确性。在加上一个数据向量和索引的工作，我们对RAG就可以总概方式地理解为“ 索引、检索和生成 ”。



- 检索（Retrieval）：在这个阶段，模型会从预先构建的大规模数据集中检索出与当前任务最相关的信息。这些数据集可以是文档、网页、知识库等。
- 生成（Generation）：在检索到相关信息后，模型会使用这些信息来生成答案或完成特定的语言任务。这个阶段通常涉及到序列生成技术，如基于Transformer的模型。

RAG 适用场景

检索增强生成（Retrieval Augmented Generation，RAG）技术

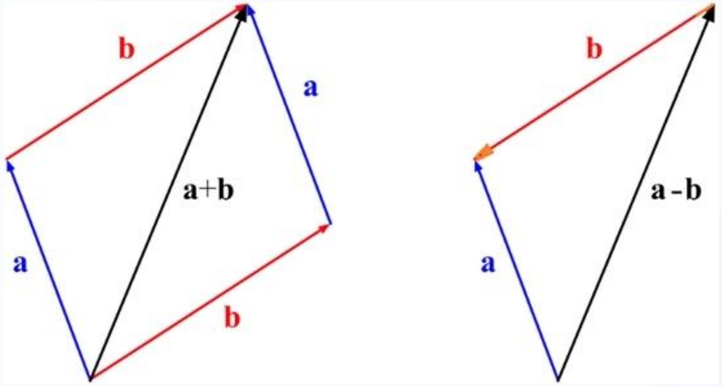
考虑因素	RAG 适用场景
数据类型	非结构化的文本（例如，新闻文章、社交媒体帖子），关注点在于文档内容而非文档间关系
需求场景	需要从大量文档中提取特定信息，如FAQ系统、智能指导问答系统等
查询复杂性	单一或多文档内的信息抽取和相似度检索
知识连贯性	信息通常是独立的，不依赖其他文档的上下文内容
运行性能	对大规模数据的快速检索
更新频率	文档更新频繁，需要频繁地重新索引文档
应用场景示例	在线客服智能回复、系统使用智能助手、快速文档检索等

基本概念-向量

FIT2CLOUD 飞致云

向量是什么？

在数学中，向量（也称为欧几里得向量、几何向量），指具有大小（magnitude）和方向的量。它可以形象化地表示为带箭头的线段。箭头所指：代表向量的方向；线段长度：代表向量的大小。通过向量嵌入模型（Embedding）将文本创建一个多维的向量数组，将所有维度的数据装在一起。



在三维空间中，一个向量可以表示为（x,y,z），其中z是向量在z轴上的分量。

$$\begin{bmatrix}1\\0\\0\end{bmatrix},\begin{bmatrix}0\\1\\0\end{bmatrix},\begin{bmatrix}0\\0\\1\end{bmatrix}\rightarrow\begin{vmatrix}1&0&0\\0&1&0\\0&0&1\end{vmatrix}=1$$

比如，在空间直角坐标系中，这个空间可以由长宽高均为1的正方体构成，这个正方体的大小为1。这个正方体就是空间直角坐标系（3维空间）中的【元素】，大小为1。

基本概念-向量化

FIT2CLOUD 飞致云

为什么要向量化（Embedding）？

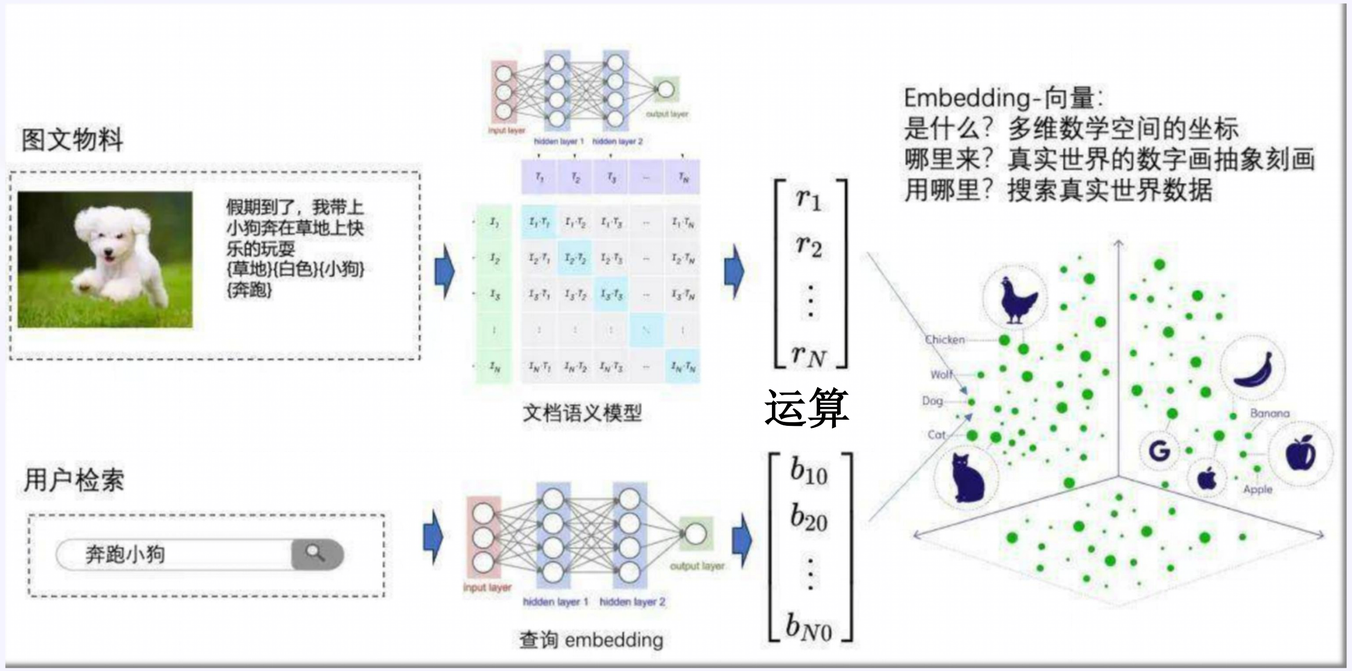
★ 向量跟向量之间，是可以运算的

我们无法对两张图片、两段语音进行运算，但是，可以对这代表这段数据的向量进行运算、处理、进而衍生出更多可能性。

从人脸识别到语音识别，从GPT到AI搜索，当前，几乎所有AI运算，都通过向量运算来实现。

无论是上一拨以计算机视觉为首的CNN卷积神经网络、还是这一拨以大模型为首的Transformer算法，其计算的本质都是对向量进行处理和变化。

尤其是大模型，其Transformer架构本身的Encoder-Decoder（编码-解码）模块设计简直是为向量数据“量身定做”的。

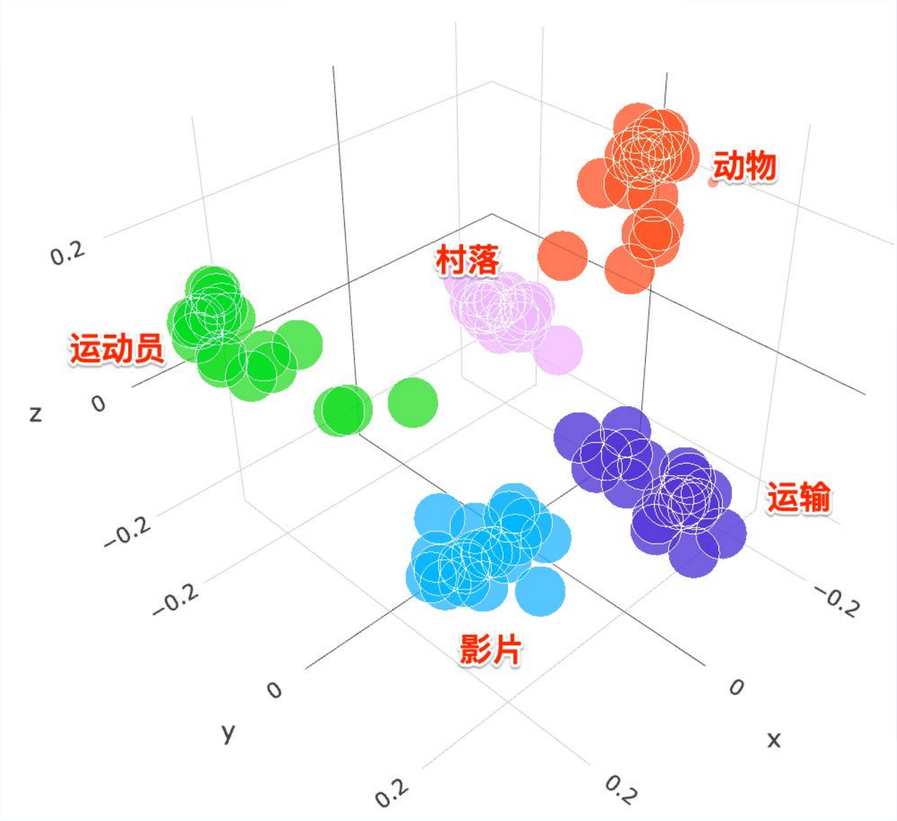


业内有句并不夸张的话：向量以一己之力撑起当代整个AI学科

基本概念-向量相似度

FIT2CLOUD 飞致云

向量相似度计算的使用场景



- 文本分类场景：在文本分类任务中，我们可以使用向量嵌入技术将文本转换为向量，然后使用余弦相似度等算法进行分类。例如，我们可以将新闻文章嵌入到向量空间中，然后根据其与不同类别的中心向量的相似度来确定其类别。
- 信息检索场景：在信息检索系统中，向量嵌入和相似度计算也有广泛的应用。我们可以将用户的查询和文档库中的文档都嵌入到向量空间中，然后通过计算查询向量与文档向量之间的相似度来排序和检索最相关的文档。
- 推荐系统场景：推荐系统可以利用向量嵌入技术将用户和物品嵌入到同一个向量空间中，然后通过计算用户向量与物品向量之间的相似度来生成推荐。例如，电影推荐系统可以将用户的观影历史和电影的特征嵌入到向量空间中，然后推荐与用户向量最相似的电影。

基本概念-文本数据向量化

FIT2CLOUD 飞致云

数据Embedding种类-文本向量化

- 在自然语言处理（NLP）中，Tokenization 是一个关键步骤，它将输入的文本内容（如句子、段落或整个文档）拆分成更小的片段或元素，这些片段通常被称为词元（tokens）。

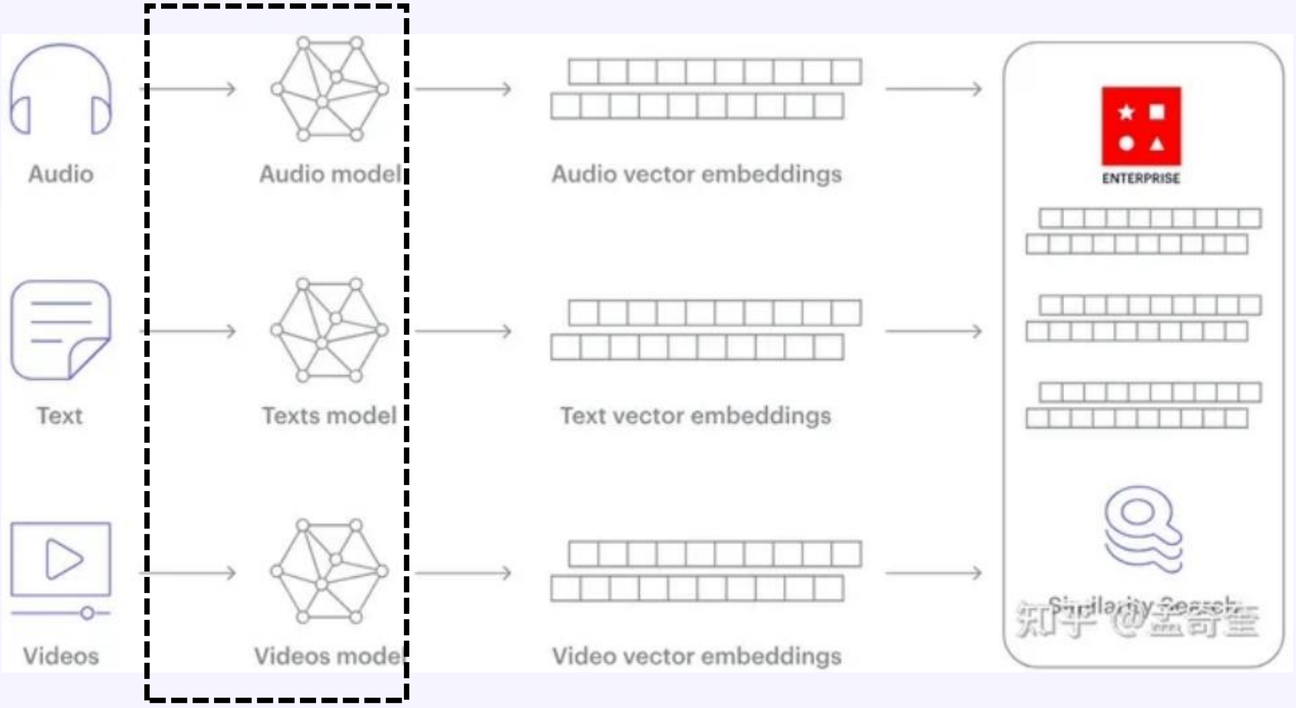
- 词元Tokens：在 NLP 中，tokens 是许多任务的基础，通常是单词，比如大模型回答的一个单词。而大模型的token生成都是依赖前一个token推理下一个token，这个也就是推理计算。

向量化核心-向量模型

FIT2CLOUD 飞致云

embedding models 是一类机器学习模型，
它们的核心功能是将高维、离散的输入数据映射到低维、连续的向量空间中

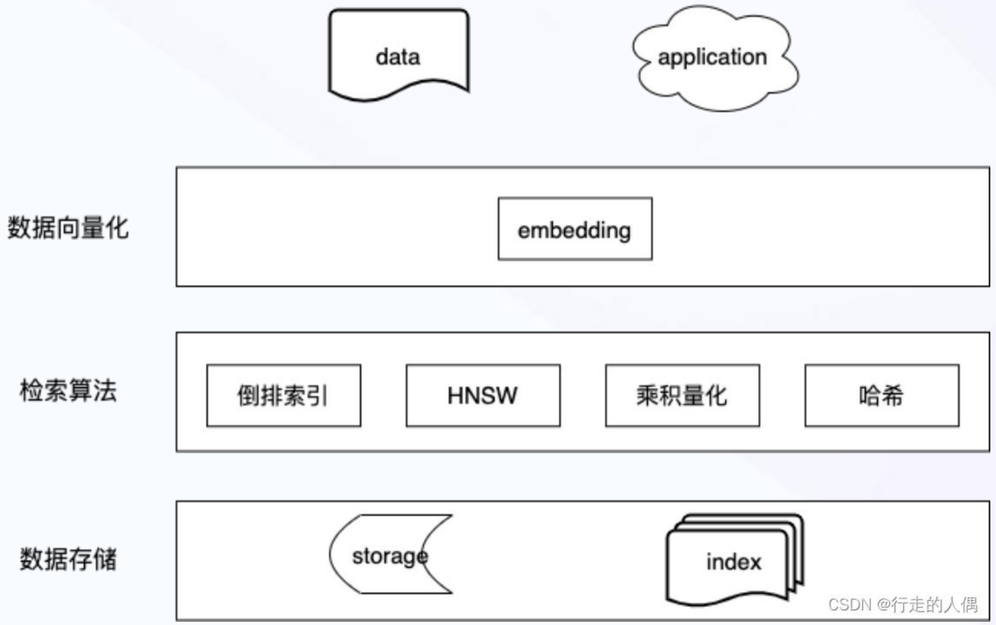
- 如Word2Vec、GloVe和FastText等，能够将词语映射到向量空间中，使得语义相近的词在向量空间中距离较近，从而用于文本分类、情感分析、机器翻译等任务。
- 如在推荐系统中，Embedding模型用于捕获用户兴趣和物品属性的相似性，实现个性化推荐和协同过滤。
- 在计算机视觉领域，图像特征Embedding用于提取图像的紧凑表示，用于图像检索、相似性比较、分类等任务。



MaxKB 默认采用的是text2vec-base-chinese 向量模型

向量化核心-向量数据库

向量数据库（Vector Database），也叫矢量数据库，主要用来存储和检索向量数据。



向量库和传统库使用场景对比：人脸识别举例

- 假设你拍摄了一张100万像素的彩色人脸图片，在传统数据库中，它理论上由100万个像素点组成，每个像素点又需要由R、G、B、A这4组数据表示，所以仅仅是一张图片，就需要有4x100万=400万个数据。
- 现在，数据库里有1万张人脸的照片，请进行比较，最符合的是哪一张照片？由于传统数据库通常只能判断YES/NO，也就是1＝1，1≠3。想象一下，如果在传统数据库中直接进行“暴力”检索，400万x1万这种上百亿级别的计算量会直接让系统崩溃。
- 更困难的是，明明只是同一个人在不同光线下的照片，人类一眼就看得出来，电脑却会因为光照、视拍照角等微小差异而返回“匹配失败”。
- 然而，如果你把这些照片变成向量，神奇的事情发生了。还记得吧？向量跟向量之间，是可以直接运算的。欧式距离、余弦、内积、海明距离……通过计算两个向量之间的距离（相似度），就可以直接找到跟它最接近的一个到多个不等的结果。
- 这可是实打实“降维打击”，不仅计算难度指数级下降，而且还可以开发出向量检索、向量聚类、甚至是将数据库中的高维向量转换成低维向量这类向量压缩与降维的“黑科技”。

文本数据向量化的处理过程

FIT2CLOUD 飞致云

数据索引一般是一个离线的过程，主要是将私域数据向量化后构建索引并存入数据库的过程。
主要包括：**数据提取、文本分割、向量化（embedding）及创建索引**等环节。

数据提取

FIT2CLOUD 飞致云

包括多格式数据（eg：PDF、word、markdown以及数据库和API等）加载、不同数据源获取等，根据数据自身情况，将数据处理为同一个范式；

- Doc类文档：直接解析其实就能得到文本到底是什么元素，比如标题、表格、段落等等。这部分直接将文本段及其对应的属性存储下来，用于后续切分的依据。
- PDF类文档：可以使用多个开源模型进行协同分析来处理图片、表格、标题、段落等内容，形成一个文字版的文档。
- PPT类文档：将PPT转换成PDF形式，然后用上述处理PDF的方式来进行解析。

对源数据进行去重、过滤、压缩和格式化等处理

提提取数据中关键信息，包括文件名、时间、章节title、图片等信息。

文本分割

FIT2CLOUD 飞致云

由于文本可能较长，或者仅有部分内容相关的情况下，需要对文本进行分块切分

1. embedding模型的Tokens限制情况；
2. 语义完整性对整体的检索效果的影响；

1. **固定大小的分块方式：**根据embedding模型的token长度限制，将文本分割为固定长度（例如256/512个tokens），这种切分方式会损失很多语义信息，一般通过在头尾增加一定冗余量来缓解。
2. **基于意图的分块方式：**

- 句分割：以”句”的粒度进行切分，保留一个句子的完整语义。常见切分符包括：句号、感叹号、问号、换行符等；
- 递归分割：通过分而治之的思想，用递归切分到最小单元的一种方式；
- 特殊分割：用于特殊场景。

向量化（embedding）

这是将文本、图像、音频和视频等转化为向量矩阵的过程，也就是变成计算机可以理解的格式。常见的向量模型有：

- ChatGPT-Embedding
- ERNIE-Embedding V1（百度千帆）
- M3E（huggingface）
- BGE（huggingface）

创建索引

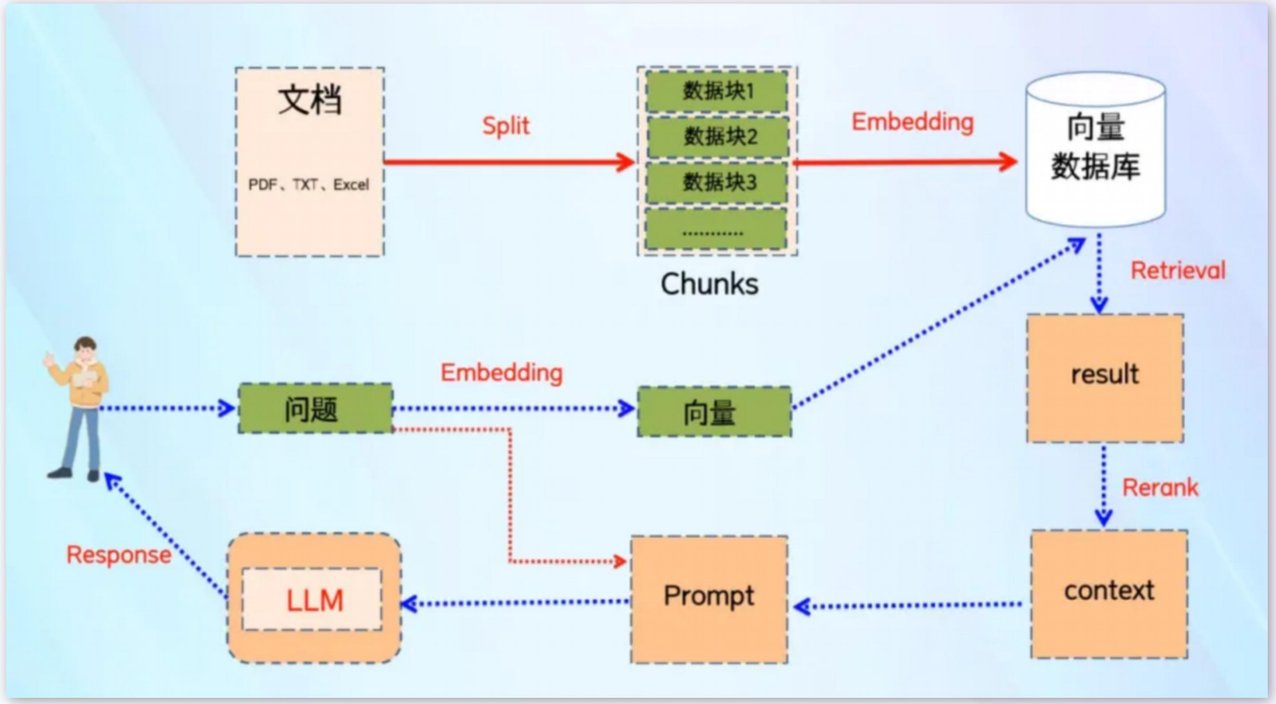
- 思路：数据向量化后构建索引，并写入数据库的过程可以概述为数据入库过程。
- 常用的工具：FAISS、Chromadb、ES、milvus,pgvector等;

注：一般可以根据业务场景、硬件、性能需求等多因素综合考虑，选择合适的数据库。

02

MaxKB 向量模型与知识库检索

文档数据优化的实现原理



- 创建索引：将输入的文档切割成不同的数据块，进行向量化处理后，存储到向量数据库，并创建索引。
- 向量检索：将用户的提问信息向量化，再到向量数据库进行搜索，根据向量相似度的算法，寻找相关性最强的文档片段。

选择合适的向量模型

FIT2CLOUD 飞致云

MaxKB 如何选择合适的向量模型

考虑因素	向量模型的能力要求
语义理解能力	需要能够理解句子或段落级别的语义，而不仅仅是词汇级别的相似度。
运行效率	针对大规模语料的检索需要考虑计算效率和相似度检索时间。
上下文依赖性	选择模型时需要考虑上下文对语义匹配的重要性。
领域适配性	有些模型对特定任务或领域（如法律、医学）需要采用专业领域模型（微调或者现有的）以提供最佳性能。

对接讯飞星火向量模型

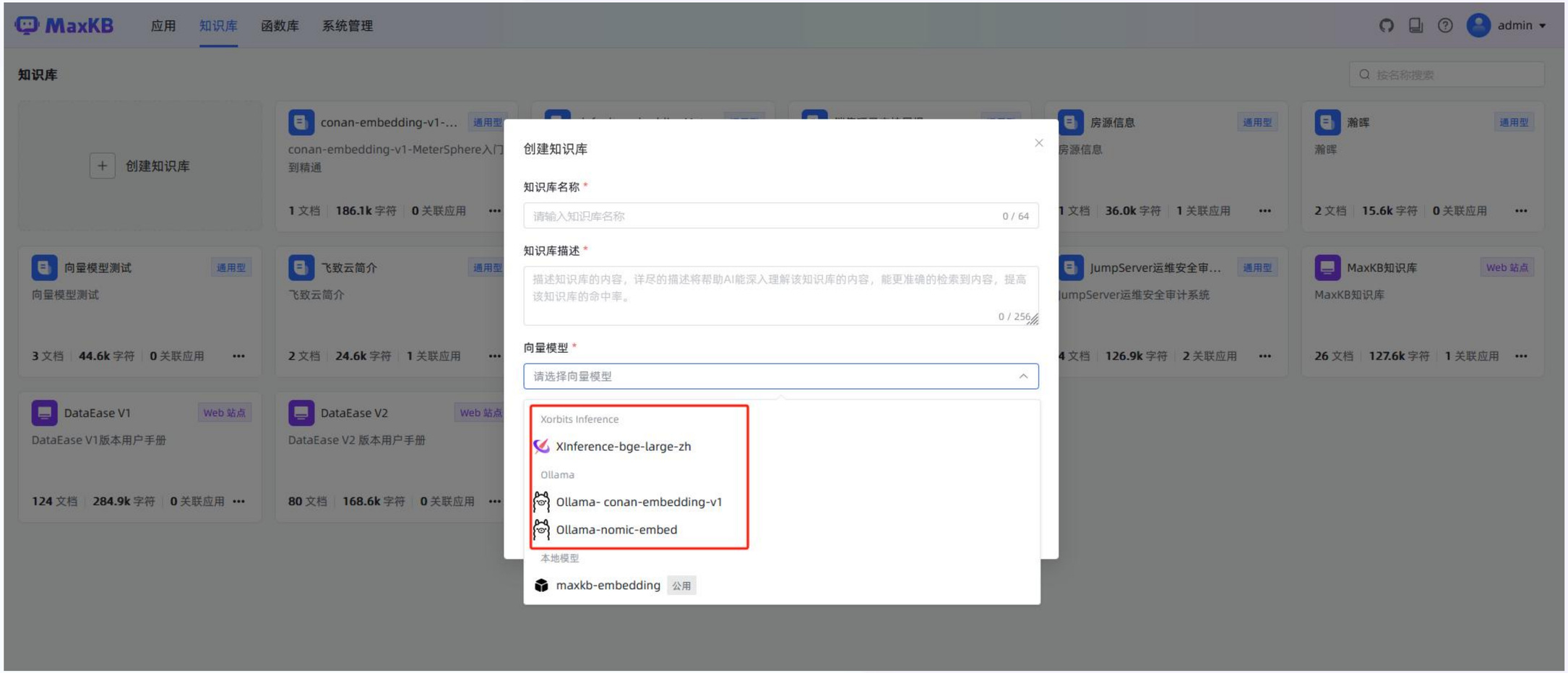
FIT2CLOUD 飞致云

MaxKB 对接讯飞星火向量模型

对接 Ollama 本地向量模型

FIT2CLOUD 飞致云

MaxKB 对接 Ollama 向量模型



知识库检索模式的应用场景

FIT2CLOUD 飞致云

MaxKB 知识库的向量检索、全文检索、混合检索的应用场景

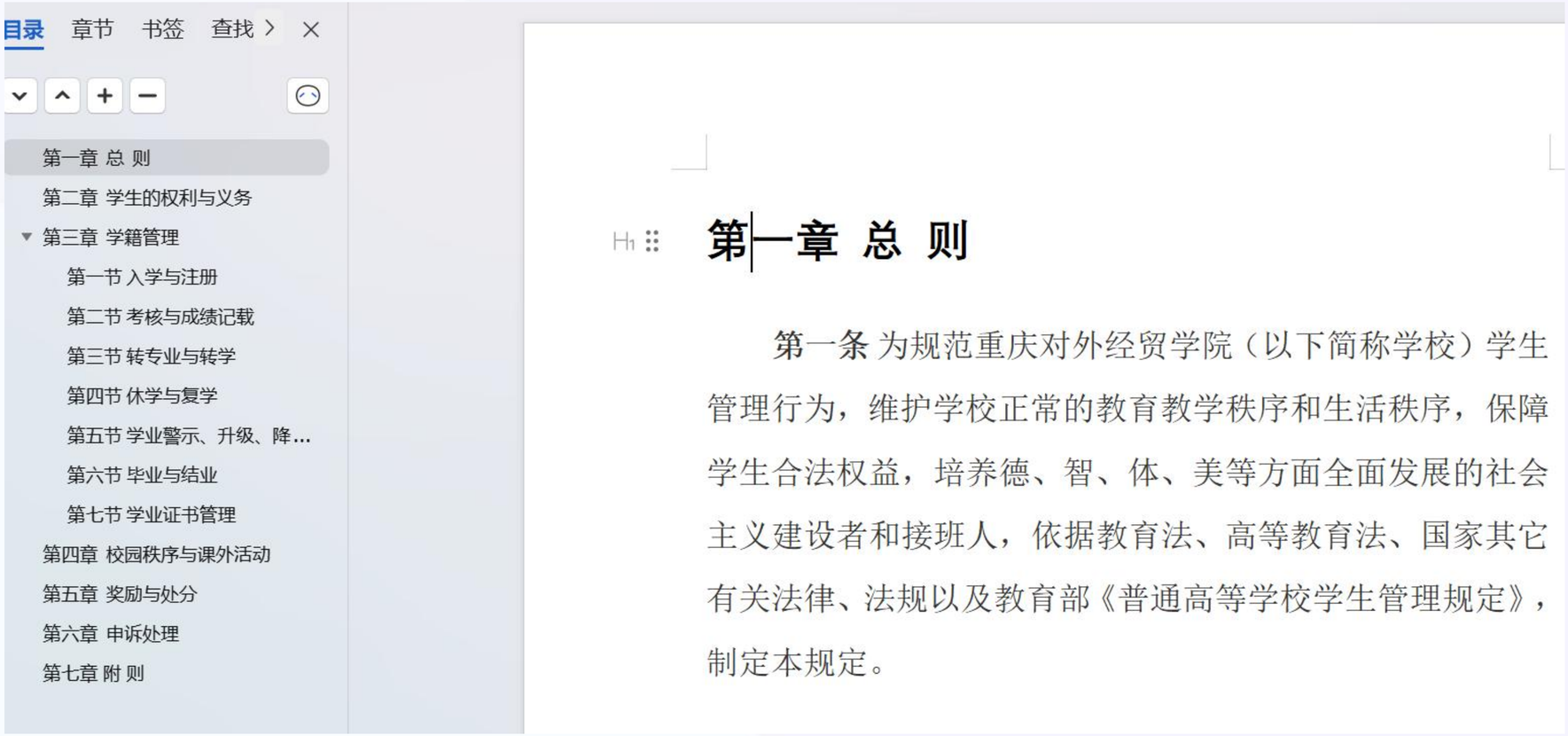
03

MaxKB 文档分段技巧

分段技巧-Word文档优化

FIT2CLOUD 飞致云

举例：高校的内部规章制度



分段技巧-QA问题对

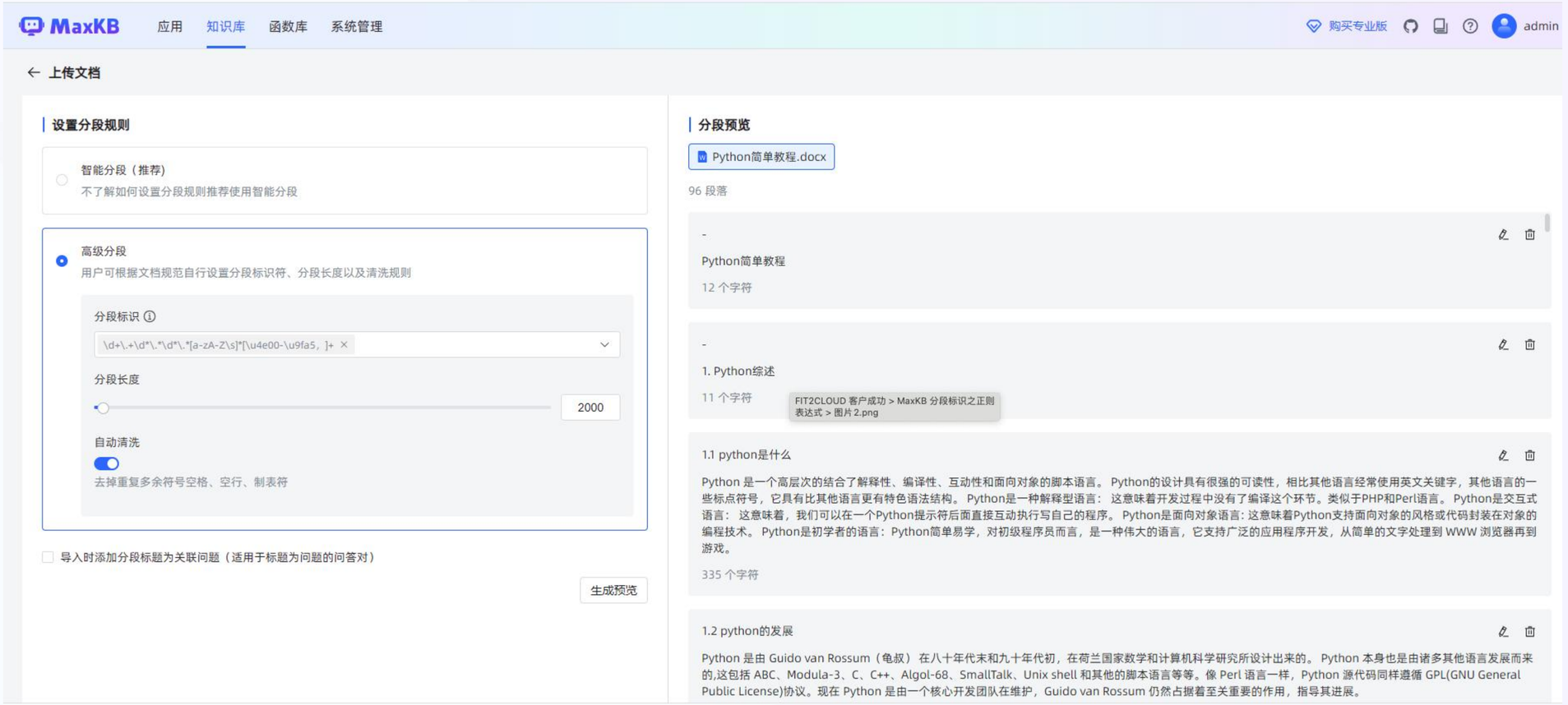
FIT2CLOUD 飞致云

举例：高校的内部规章制度

分段技巧-使用正则表达式

FIT2CLOUD 飞致云

MaxKB 使用正则表达式分段



访问地址 https://kb.fit2cloud.com/?p=bfe242a4-9a77-459c-ac75-4ef078e170c9

练习作业

① 实践操作 MaxKB 知识库的文档分段技巧。

作业要求：

1. 使用本节课的 MaxKB 升级文档，设置分段规则为“高级分段”，输入正则表达式“[一二三四五六七八九十]*[、][\u4e00-\u9fa5a-zA-Z]+”，调整分段长度1000，再导入到 MaxKB 知识库，并提交文档分段的截图。

2. 创建一个简单应用，并关联大语言模型和 MaxKB 知识库，发布应用，输入提问信息“安装 MaxKB 企业版本”，等待应用输出回复内容，提交应用截图。

THANK YOU

www.fit2cloud.com

400-052-0755

扫码申请专业版试用

北京 · 上海 · 深圳 · 广州 · 南京 · 杭州
苏州 · 武汉 · 成都 · 西安 · 长沙 · 济南
郑州 · 厦门 · 合肥 · 青岛 · 重庆 · 天津

技术交流群