一、API基本信息

|  |  |
| --- | --- |
| 属性 | 值 |
| 接口名称 | retrieve\_chunks |
| 接口路径 | <http://127.0.0.1:8000/retrieve/> |
| HTTP方法 | POST |
| API文档标题 | RAGFlow检索API |
| API版本 | 1.0.0 |
| 文档描述 | 基于混合搜索策略（关键词+向量）的知识库检索服务，支持分页和相似度筛选 |

二、接口功能说明

**接口摘要**：检索知识库相关文档片段

**接口描述**：该接口使用混合搜索策略（同时启用关键词匹配和向量相似度分析）从指定数据集中检索与用户查询问题最相关的文档片段。检索结果支持分页返回，并允许通过相似度阈值和权重参数进行精细控制。当前接口固定使用keyword=True参数，这意味着检索过程将始终包含关键词匹配阶段。API返回的文档片段将包含原始文档ID和内容，以便后续处理或展示。

三、请求参数说明

本接口使用JSON格式的请求体，结构如下：

**class** QueryRequest(BaseModel):  
 question: str *# 用户查询问题*  
 page: int = 1 *# 页码*  
 page\_size: int = 10 *# 每页返回数量*  
 similarity\_threshold: float = 0.5 *# 相似度阈值，默认为0.5*  
 vector\_similarity\_weight: float = 0.4 *# 向量相似度权重，默认为0.4*  
 top\_k: int = 512 *# 从前512个候选中筛选，默认为512*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 字段 | 类型 | 必填 | 默认值 | 描述 | 示例值 |
| question | str | 是 | - | 用户输入的查询文本 | “如何配置RAGFlow参数？” |
| page | int | 否 | 1 | 请求的页码，默认返回第一页 | 2 |
| page\_size | int | 否 | 10 | 每页最多返回的文档片段数量 | 15 |
| similarity\_threshold | float | 否 | 0.5 | 相似度阈值，用于过滤低相似度结果（0~1） | 0.65 |
| vector\_similarity\_weight | float | 否 | 0.4 | 向量相似度权重（0~1），剩余权重为BM25权重 | 0.6 |
| top\_k | int | 否 | 512 | 从候选结果中选取的前N个最高相似度片段 | 5 |

**注意事项**： 1. similarity\_threshold和vector\_similarity\_weight的有效范围均为0~1。 2. vector\_similarity\_weight表示向量相似度的权重，剩余权重（1 - vector\_similarity\_weight）自动分配给BM25相似度。 3. top\_k值应大于等于page\_size，以确保足够的候选结果供分页使用。 4. dataset\_id在代码中固定为dc74c94c2bcb11f0bb44129440cf6ebc，因此无需在请求中提供。

四、响应格式说明

成功调用API后，返回的JSON响应格式如下：

{  
 "results": [  
 {  
 "document\_id": "string",  
 "content": "string"  
 }  
 ]  
}

|  |  |  |
| --- | --- | --- |
| 字段 | 类型 | 描述 |
| results | List[dict] | 包含检索到的文档片段的列表 |
| document\_id | str | 文档的唯一标识符 |
| content | str | 检索到的文档片段内容 |

五、示例请求与响应

Headers请求头添加：

key：Authorization

value：Bearer ragflow-FjZTU2MmEyMmJkMTExZjBiMWQ5MTI5ND

**示例请求**：

{  
 "question": "如何设置RAGFlow的相似度阈值？",  
 "page": 1,  
 "page\_size": 10,  
 "similarity\_threshold": 0.5,  
 "vector\_similarity\_weight": 0.4,  
 "top\_k": 200  
}

**示例响应**：

{  
 "results": [  
 {  
 "document\_id": "507f191e810c19729de860ea",  
 "content": "相似度阈值设置在RAGFlow中可以通过retrieval\_params配置，建议根据数据复杂度调整"  
 },  
 {  
 "document\_id": "507f191e810c19729de860eb",  
 "content": "在混合检索中，相似度阈值与向量权重共同决定最终结果的质量"  
 }  
 ]  
}

六、错误处理

当API调用失败时，将返回HTTP 500错误（内部服务器错误），响应内容包含错误原因的简要描述：

{  
 "detail": "查询失败: 无效的相似度权重值"  
}

**常见错误场景**： 1. 参数非法（如相似度阈值超出0~1范围，或权重之和不等于1）。 2. 服务内部错误（如连接到RAGFlow服务器失败）。 3. 无匹配结果（当相似度阈值过高时可能返回空列表）。