Node Guide : PDF To Vector

Overview

The **PDF To Vector** node allows you to process and convert PDF documents into vector embeddings. These embeddings are stored in a vector database and later used for AI-based document retrieval, search, or question answering using RAG (Retrieval-Augmented Generation) techniques.

What This Node Does

- Takes PDF files from a selected document source
- Chunks and embeds the content using a selected embedding model
- Saves the output as vector embeddings into a vector database collection

---

Inputs

This node uses configuration values such as collection name, documents, chunk size, overlap, and embedding model to perform the vectorization process.

Outputs

This node does not return any output in terms of variables. The operation happens silently and results in vectorized data being saved into the specified collection.

Configuration

1. Embedding Model
Select the embedding model from a dropdown of available global configurations.
These are predefined in the Global Configurations Page.

Note: This field is mandatory.

2. Collection Name
Specify the name of the collection where the vector embeddings will be stored.
You can type it directly or select a variable.

3. Documents
Choose a DMS configuration (defined globally) that points to the source PDFs to be processed.

4. Chunk Size
Set the number of characters each chunk should contain.
This controls how documents are broken down for vector generation.

5. Chunk Overlap
Define how many characters should overlap between consecutive chunks.
Useful for preserving context across chunks.

6. Indexing Type
Choose the indexing behavior:

- **Upsert** – Adds new or updates existing content.
- **Re-Index** – Clears and reprocesses the content from scratch.

7. Metadata Fields
Multi-select the metadata fields to be stored with each vector chunk.
Options include:

- Tags
- File ID
- File Name
- File Size
- File Extension

When to Use

Use this node when you want to:

- Prepare PDF documents for AI-powered semantic search or retrieval
- Convert document archives into a searchable knowledge base

- Enable intelligent document Q&A in a RAG-based application

Example Flow

Scenario: You want to index a batch of uploaded legal PDFs for retrieval.
Steps:

1. User uploads documents to DMS
2. Use the **PDF To Vector** node to chunk and embed documents
3. Save the vectors to a named collection
4. Use a Prompt LLM node later to query the vector store

Summary of the Flow

- Connects to your document repository
- Splits and embeds PDFs using the selected model
- Stores the output for downstream AI-based operations