

AI Course | How to Optimize your Files for RAG: Data Structuring



Botpress

24.1K subscribers

Subscribe

53



Share

Ask

Download



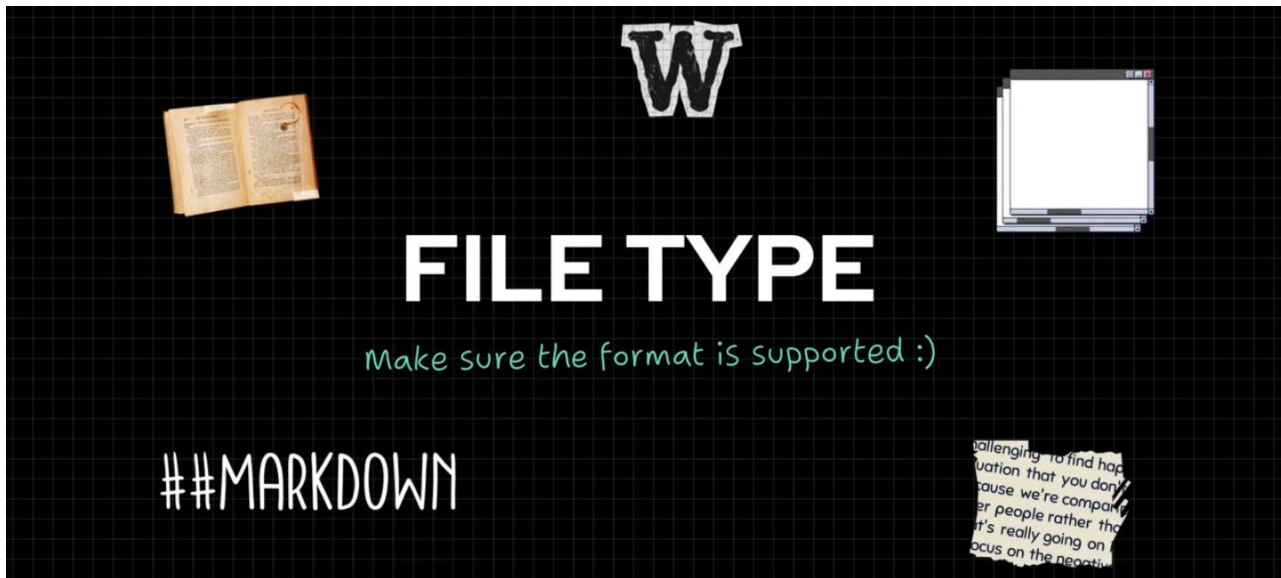
4,996 views Dec 6, 2024

This video explores the foundational steps for preparing data for Retrieval-Augmented Generation (RAG), focusing on document formatting and structure. Proper file types and a logical organization of content are critical for ensuring that an AI agent can accurately process and retrieve information.

Supported file formats include PDFs, Word documents, HTML, Markdown, and plain text. File types that cannot be easily parsed, such as image-based documents with complex formatting, are not suitable for use as a knowledge base, as they limit an agent's ability to generate reliable responses. Botpress automatically converts uploaded files into Markdown format, enabling consistent handling by the AI agent. For optimal results, raw Markdown files or Rich Text knowledge base types are recommended.

Document structure plays a significant role in enhancing an agent's comprehension. Using clear and logical sections, with titles, headings, and subheadings, helps an LLM categorize and retrieve relevant information more effectively. A well-structured document also ensures that vectorization processes group related content correctly, reducing inconsistencies in retrieval performance.

Organizing and standardizing files improves data quality, making it easier for AI agents to process and retrieve accurate information.



Avoid using file types that can't be easily parsed




image-based documents with complex formatting

✓ raw-markdown.md

MARKDOWN

✓ richtext.rtf



Breaking your files into a clear and logical structure

*Headings are
really important*

can greatly enhance your
agent's ability to understand and
retrieve information



🔌 Demos

Enterprise

🔍 young-fish

🤖 Assistant

🌐 Explore Hub

🔗 Share

📄 Publish

Knowledge Bases

New Knowledge Base

Knowledge Base Demo

Knowledge Base Demo

📄 Website

📄 Document

📄 Table

🔍 Web Search

📄 Rich Text

📄 Upload from API

New

📄 KBDemo.pdf

0.01MB

Created 1 minute ago

Variables

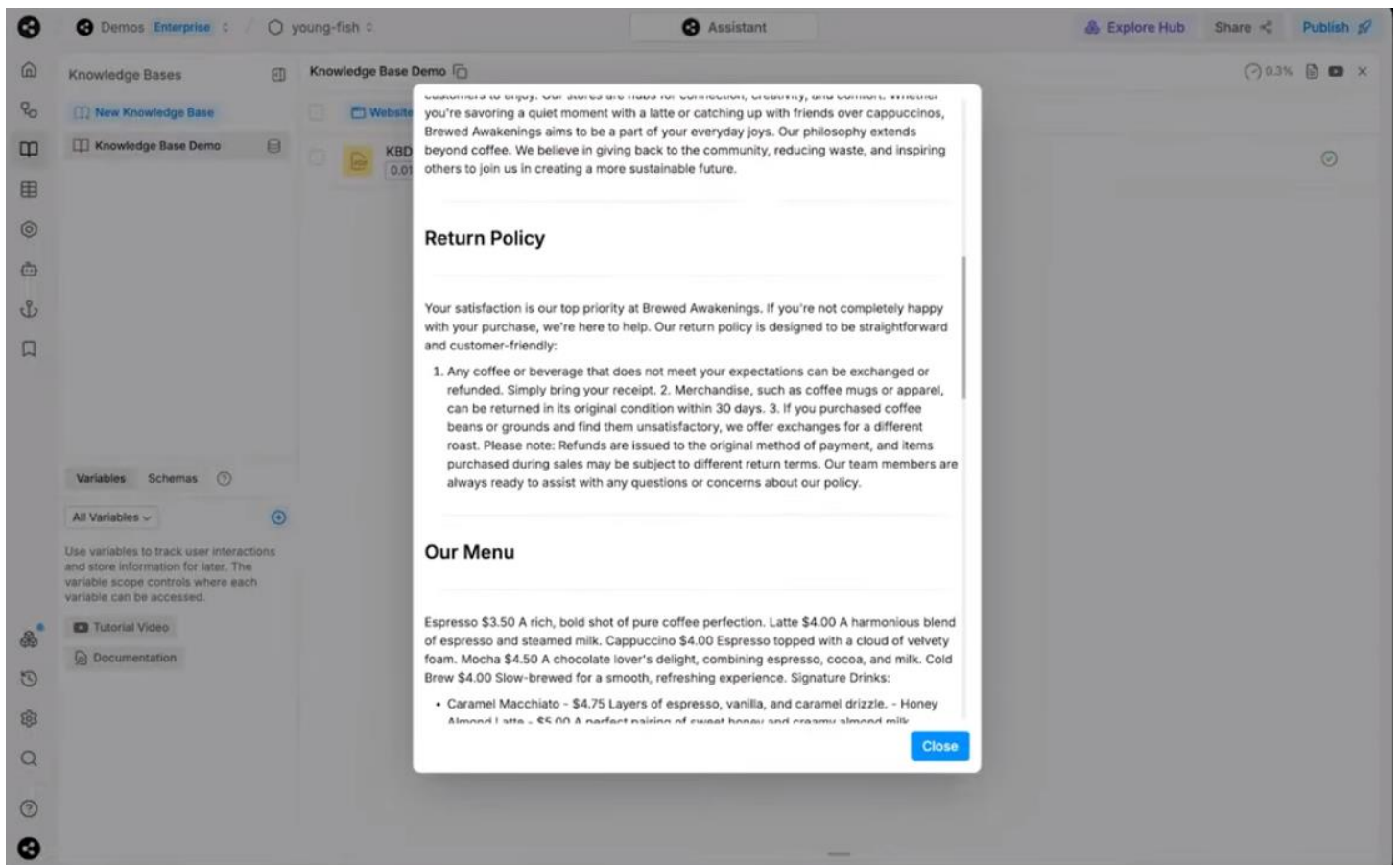
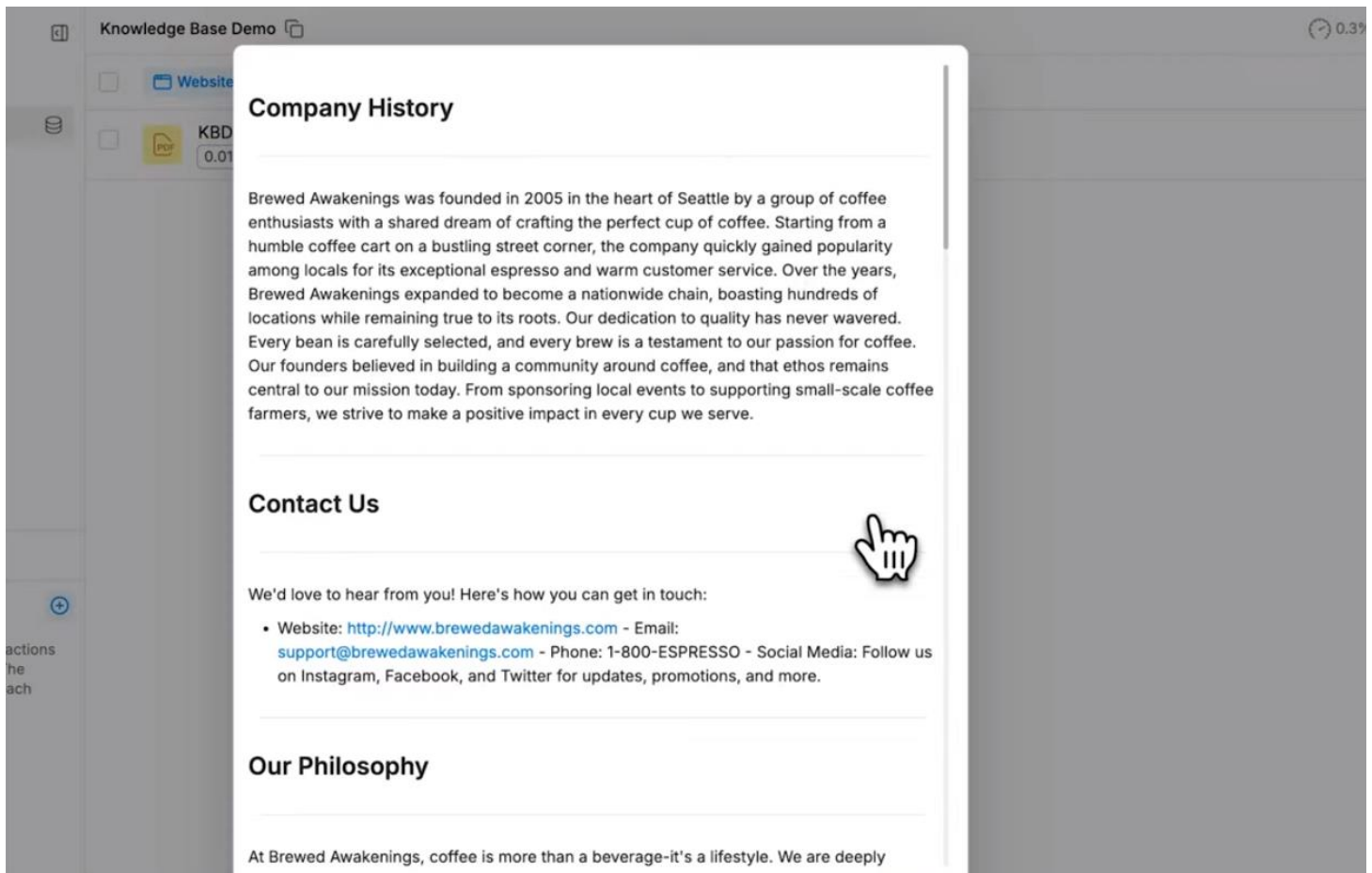
Schemas

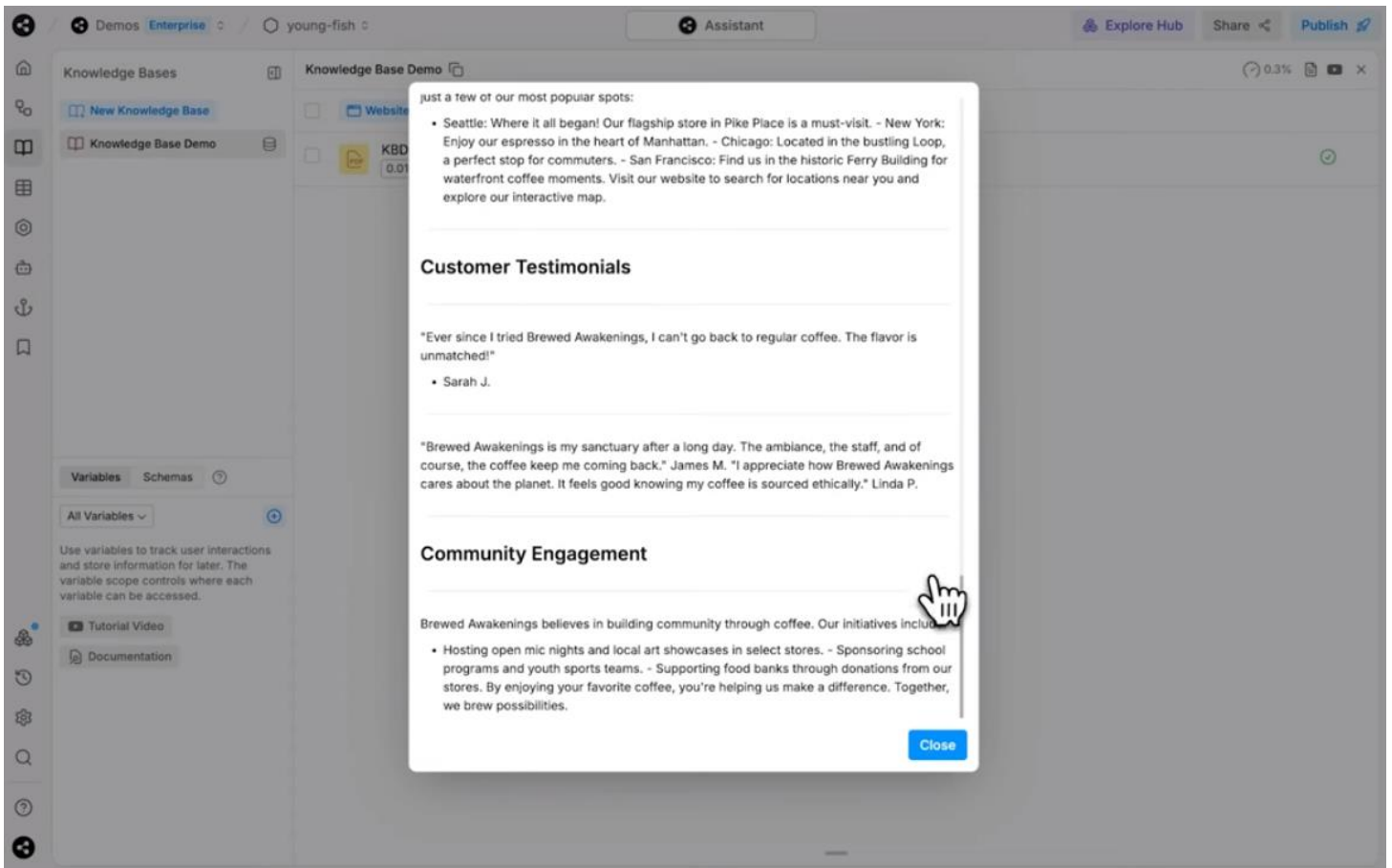
All Variables

Use variables to track user interactions and store information for later. The variable scope controls where each variable can be accessed.

Tutorial Video

Documentation





Properly formatting and organizing documents with clear headings and supported file types ensures your AI agents will parse and retrieve accurate information.