RAG System Enhancement: PDF Support Documentation

Overview

This document demonstrates the PDF support enhancement added to our Retrieval Augmented Generation (RAG) system.

The system can now process both text files and PDF documents, extracting content for embedding and retrieval.

Key Features

1. Universal Document Loader: Handles both .txt and .pdf files seamlessly

2. Page-aware extraction: PDF content includes page numbers for reference

3. Directory scanning: Can process entire directories with mixed file types

4. Error handling: Gracefully handles corrupted or unreadable PDFs

Implementation Details

The PDF support uses the pypdf library for robust PDF text extraction.

Text is extracted page by page and formatted with page markers.

The UniversalLoader class provides a unified interface for all document types.

Distance Metrics

Currently, the system uses cosine similarity for vector comparison.

Future enhancements could include Euclidean distance, Manhattan distance, or dot product similarity.

Each metric has different properties suitable for various use cases.