Docling Java

Docling Java is the official Java client and tooling for Docling — a suite that simplifies document processing and parsing across diverse formats (with advanced PDF understanding) and integrates seamlessly with GenAI frameworks.

What is Docling?

Docling extracts structured information from documents. It understands page layout, reading order, tables, code, formulas, and images, and exports results into convenient formats like Markdown, HTML, and JSON.

Key features

- Parsing of multiple document formats including PDF, DOCX, PPTX, XLSX, HTML, WAV, MP3, VTT, and images (PNG, TIFF, JPEG, ...)
- Advanced PDF understanding: page layout, reading order, table structure, code, formulas, image classification, and more
- Unified, expressive `DoclingDocument` representation
- Multiple export formats and options, including Markdown, HTML, DocTags, and lossless JSON
- Local execution options for sensitive or air-gapped environments
- Integrations with popular GenAI frameworks (e.g., LangChain4j)
- OCR support for scanned PDFs and images
- Visual Language Models support (e.g., GraniteDocling)
- Audio support via Automatic Speech Recognition (ASR) models

Project layout and artifacts

This repository provides a set of artifacts you can mix and match depending on your needs:

- `docling-core`: Java API for working with the data types used by Docling for document representation (see Docling Core).
- `docling-serve-api`: Java API for interacting with a Docling Serve backend (framework-agnostic).
- `docling-serve-client`: Reference HTTP client built with Java `HttpClient` and Jackson for connecting to a Docling Serve endpoint.
- `docling-testing`: Utilities for testing Docling integrations in your codebase.
- `docling-testcontainers`: A Testcontainers module for running Docling Serve in containers.

Links

- Source repository: https://github.com/docling-project/docling-java
- Docling (core project): https://github.com/docling-project
- Supported formats: https://docling-project.github.io/docling/usage/supported_formats/
- DoclingDocument concept: https://docling-project.github.io/docling/concepts/docling_document/

- Community Discord: https://docling.ai/discord

License and contributions

The codebase is released under the MIT License. Contributions are welcome — please see `CONTRIBUTING.md` for guidelines and `CODE_OF_CONDUCT.md` for community standards.