DocWire SDK 2025.08.05 Released – Local AI Embeddings, SentencePiece, Cosine Similarity

Язык оригинала: en

# Оригинал

The 2025.08.05 release brings a major milestone to DocWire SDK: fully local, offline AI-powered text embeddings. With the integration of the multilingual-e5-small model, DocWire now supports multilingual vectorization for advanced NLP tasks—completely offline.  
It also modernizes dependencies by switching from OpenNMT-Tokenizer to Google’s SentencePiece, and includes numerous build and CI improvements for better MSVC and Valgrind support.  
Full release notes:  
https://github.com/docwire/docwire/releases/tag/2025.08.05  
Highlights  
1 · Local AI Embeddings  
Introduces  
local\_ai::embed  
for generating multilingual embeddings using  
multilingual-e5-small  
Powers advanced use cases like semantic search, retrieval-augmented generation (RAG), and document clustering  
CLI-ready via  
--local-ai-embed  
2 · Cosine Similarity Utility  
Built-in cosine similarity function for comparing document/query vectors  
3 · Tokenizer API (SentencePiece)  
Public  
local\_ai::tokenizer  
based on Google’s SentencePiece  
Supports encoding text into token IDs with  
T5Tokenizer  
and  
XLMRobertaTokenizer  
Improvements  
Unified model runner (  
local\_ai::model\_runner  
) now supports both encoder-only and sequence-to-sequence models  
Advanced pooling and L2 normalization for E5-compatible output  
New simplified constructor in  
model\_chain\_element  
with default model  
CLI extended with support for embedding workflows  
Refactors  
Replaced OpenNMT-Tokenizer with a modern SentencePiece integration for improved maintainability and quality  
Fixes  
MSVC: AddressSanitizer (ASan) issues resolved using specific macro definitions  
CI:  
Increased Valgrind timeouts  
Skipped heavy tests under Callgrind  
Abseil leak suppressions added for cleaner reports  
Documentation & Tests  
New end-to-end embedding example (README): document + queries + cosine similarity  
Unit tests for  
local\_ai::tokenizer  
Embedding example is compiled and tested in CI  
Get Started  
GitHub:  
https://github.com/docwire/docwire  
Release:  
https://github.com/docwire/docwire/releases/tag/2025.08.05  
Sourceforge:  
https://sourceforge.net/projects/docwire/files/2025.08.05/  
This update cements DocWire as a serious offline-ready NLP SDK for C++ developers building hybrid pipelines.  
— The DocWire Team

# Перевод на русский

Выпуск 2025.08.05 приносит важную веху для Docwire SDK: полностью местные, автономные текстовые тексты с AI. С интеграцией многоязычной модели 2-й-E5, Docwire теперь поддерживает многоязычную векторизацию для передовых задач NLP-само в автономном режиме.  
Он также модернизирует зависимости, переходя от OpenNMT-Tokenizer на предложение Google, и включает в себя многочисленные улучшения сборки и CI для лучшей поддержки MSVC и Valgrind.  
Полные заметки:  
https://github.com/docwire/docwire/releases/tag/2025.08.05  
Основные моменты  
1 · местные встраивания ИИ  
Вводит  
local\_ai :: ulced  
Для генерации многоязычных встроений с использованием  
Многоязычный-E5-Small  
Powers расширенное использование, такие как семантический поиск, поколение поиска (RAG) и кластеризация документов  
Cli-ready через  
-Локаль-Ай-Эмбби  
2 · Косинусная утилита сходства  
Встроенная функция сходства косинуса для сравнения векторов документов/запросов  
3 · Tokenizer API (предложение)  
Публичный  
local\_ai :: tokenizer  
на основе предложения Google  
Поддерживает кодирование текста в токеновые идентификаторы с  
T5tokenizer  
и  
Xlmrobertatatokenizer  
Улучшения  
Unified Model Runner (  
local\_ai :: model\_runner  
) теперь поддерживает как модели, только для энкодера, так и последовательность в последовательность  
Усовершенствованное объединение и нормализацию L2 для E5-совместимого вывода  
Новый упрощенный конструктор в  
model\_chain\_element  
с моделью по умолчанию  
CLI продлен при поддержке встраивания рабочих процессов  
Рефактории  
Заменил OpenNMT-TOKENICER на современную интеграцию предложения для улучшения обслуживания и качества  
Исправляет  
MSVC: AdderSanitizer (ASAN) Проблемы, решаемые с использованием конкретных определений макросов  
CI:  
Увеличение тайм -аутов Valgrind  
Пропущенные тяжелые тесты под Callgrind  
Подавление утечки. Добавлено для более чистых отчетов  
Документация и тесты  
Новый пример встраивания сквозного встроения (readme): документ + запросы + сходство косинуса  
Модульные тесты для  
local\_ai :: tokenizer  
Пример встраивания скомпилируется и тестируется в CI  
Начните  
GitHub:  
https://github.com/docwire/docwire  
Выпускать:  
https://github.com/docwire/docwire/releases/tag/2025.08.05  
Sourceforge:  
https://sourceforge.net/projects/docwire/files/2025.08.05/  
Это обновление закрепляет DOCWIRE как серьезный автономный NLP SDK для разработчиков C ++, создающих гибридные трубопроводы.  
- Команда Docwire