KARTA PRACY UCZESTNIKA MODUŁ 3 ROZDZIAŁ 2

Tworzenie embeddingów lokalnie

Od zaakceptowanych fragmentów do zapisanego zbioru wektorów

Czas orientacyjny	Pracujesz z	Rezultat
45 min	wynikiem M2, Pythonem i lokalnym Ollama	chunks_with_embeddings.jsonl i dwa pliki opisu przebiegu

Rozszerzysz zaakceptowane fragmenty o embeddingi. Zachowasz identyfikatory i metadane, aby każdy wektor nadal prowadził do właściwego źródła.

Cel i rezultat rozdziału

Utworzysz embeddingi modelem bge-m3:latest lub przeanalizujesz rzeczywisty checkpoint, zachowując jawne pochodzenie wyniku.

Po zakończeniu masz

- chunks_with_embeddings.jsonl z poprawnymi rekordami;
- embedding_manifest.json z konfiguracją i licznościami;
- embedding_errors.jsonl oraz zapis własnego przebiegu.

Zanim zaczniesz

Najpierw wykonaj notebook 00. Otwórz wskazany starter, wybierz MODE i uruchom Przygotowanie. Pola UZUPEŁNIJ wypełnij samodzielnie.

`notebooki/M12_M3_R2_embeddingi_start.ipynb`

Materiał	Do czego służy
dane/knowledge_base_ready.jsonl	Zaakceptowane rekordy M2; bez embeddingów.
dane/POCHODZENIE.md	Pochodzenie dołączonego wejścia referencyjnego.
konfiguracja.json i checkpointy/R2/	Model, ustawienia oraz wariant bez wywołania Ollama.

Gdzie wrócić w prezentacji

W R2 korzystaj ze slajdów „Przygotowanie Ollama i funkcja create_embedding()”, „Minimalny schemat kodu” oraz „Zapisywanie embeddingów”.

Praca krok po kroku

Krok 1 Sprawdź wejście — Wczytaj knowledge_base_ready.jsonl do listy chunks i policz rekordy.

- Sprawdź unikatowe chunk_id, niepusty text i metadata z document_id oraz source.
- Zachowaj version, status, qa_status i lokalizację źródła; nie zmieniaj identyfikatorów.

Krok 2 Przygotuj model — Wybierz MODE: ollama albo checkpoint. Przeczytaj create_embedding() oraz EmbeddingProvider.embed w kod/m3_tools.py.

- Model wynika z konfiguracja.json; używana stała to EMBEDDING_MODEL.
- Parametr model_name trafia do model w wywołaniu klienta Ollama. Checkpoint nie tworzy nowych wektorów.

Krok 3 Sprawdź jedną próbę — Utwórz embedding tekstu jednego rekordu przed przetwarzaniem zbioru.

- Sprawdź wymiar 1024, skończone liczby i niezerową normę.
- Zapisz rzeczywisty tag i digest; w trybie checkpoint zachowaj konfigurację oryginalnego pomiaru.

Krok 4 Przetwórz rekordy — Dodawaj embedding do kopii każdego poprawnego rekordu w embedded_chunks.

- Błędy danych zapisuj z chunk_id i przyczyną; nie zastępuj błędów wektorami zerowymi.
- Przy niedostępnej usłudze przerwij przebieg. Nie pomijaj milcząco całej bazy.

Krok 5 Zapisz trzy pliki — Zapisz rekordy z embeddingami, manifest i rejestr problemów.

- Manifest ma wskazywać wejście, model, backend, wymiar, datę i liczby rekordów wejściowych, poprawnych, pominiętych i błędnych.
- Pusty embedding_errors.jsonl jest poprawny, jeśli nie było problemów. Liczności muszą się zgadzać.

Krok 6 Wczytaj wynik ponownie — Odczytaj zapisany JSONL i porównaj go z wejściem.

- Sprawdź zachowanie chunk_id, tekstu i całych metadanych dla udanych rekordów.
- Zanotuj ścieżkę próby R2. W R3–R5 korzystaj z tego samego, wybranego zbioru.

Zapis próby: komórka Przygotowanie automatycznie tworzy nowy folder wyniki/R2/<proba_id>/. Zachowaj wyświetloną ścieżkę.

Punkty kontrolne pracy

Kiedy	Co sprawdzasz	Co zrobić przy problemie
Pierwszy wektor	Wymiar 1024, liczby skończone, norma dodatnia.	Sprawdź model i treść wejścia.
Cały zbiór	Każdy rekord wejściowy ma jawny wynik przetwarzania.	Porównaj liczności oraz rejestr problemów.
Ponowny odczyt	Wektory nadal prowadzą do tych samych źródeł.	Porównaj identyfikatory i metadane z M2.

Nazewnictwo i zapis wyników

Notebook zapisuje chunks_with_embeddings.jsonl, embedding_manifest.json i embedding_errors.jsonl w wyniki/R2/<proba_id>/. Nie używaj tu nazwy chunks.jsonl.

Identyfikator <proba_id> nadaje notebook. Pliki są chronione przed nadpisaniem; nowy pomiar rozpocznij od Przygotowania. Zapisz notebook przez Ctrl+S i wskaż próbę do oceny.

Lista kontroli przed zakończeniem

- □ Sprawdziłem pochodzenie danych z M2.
- □ Użyłem jednej konfiguracji modelu.
- □ Sprawdziłem wymiary i wartości wektorów.
- □ Zachowałem tekst, chunk_id i metadata.
- □ Zapisałem manifest i jawny rejestr problemów.
- □ Ponownie odczytałem wynik z dysku.

Gdy pojawi się problem

Objaw	Co sprawdzić
Brak połączenia lub modelu	Sprawdź Ollama i wynik ollama list; pobierz bge-m3:latest.
Pusty tekst lub inny wymiar	Zapisz przyczynę, sprawdź wejście i model; nie dopisuj zer.
Przebieg trwa długo	Najpierw sprawdź małą próbkę; czas całej bazy zależy od sprzętu.

Nowy model lub zmieniona konfiguracja wymagają nowego, spójnego przebiegu. Zachowaj poprzednie wyniki. Zapisz także uzupełniony notebook.