KARTA PRACY UCZESTNIKA MODUŁ 2 ROZDZIAŁ 5

Metadane i identyfikacja źródeł

Powiązanie fragmentów z wersją i miejscem w źródle

Czas	Pracujesz z	Rezultat
około 25 min	notebookiem, dokumentami i wynikami wcześniejszych etapów	fragments_with_metadata.json

Połączysz fragmenty z inwentarzem dokumentów. Sprawdzisz, czy z każdego rekordu można wrócić do oryginalnego tekstu, właściwej wersji i miejsca.

Cel i rezultat rozdziału

Zapiszesz jednolite rekordy z metadanymi bez wymyślania nieznanych wartości.

Po zakończeniu masz

- rekordy z polami vector_index, chunk_id, text i metadata;
- strony PDF, sekcje, wersje i lokalizacje źródeł;
- fragments_with_metadata.json oraz kontrolę trzech rekordów.

Zanim zaczniesz

Pracuj na własnej kopii katalogu Student. Zachowaj razem dane, kod, notebooki, karty_pracy i wyniki. Ścieżki są liczone od Student. Ustaw ten sam RUN_ID w każdym rozdziale.

Otwórz notebook

`notebooki/M12_M2_R5_metadane_start.ipynb`

Materiał wejściowy	Do czego służy
wyniki/<RUN_ID>/R4	chunks.json z fragmentami i położeniem.
wyniki/<RUN_ID>/R1	Inwentarz z metadanymi dokumentów.

Ćwiczenie bez LLM: nie potrzebujesz Ollama ani modelu embeddingowego. Najpierw wykonaj test środowiska. Prezentacja M12-M2-v4: slajdy 98–116.

Praca krok po kroku

Krok 1 Połącz dane — Uruchom komórkę Metadane źródła i fragmentu.

- Sprawdź, czy każdy document_id istnieje w inwentarzu.
- Nie łącz metadanych wyłącznie według podobnego tytułu.

Krok 2 Sprawdź schemat — Obejrzyj pola główne i metadata.

- chunk_id jest polem głównym; tytuł i document_id są w metadata.
- vector_index to kolejność od 0, jeszcze bez embeddingów.

Krok 3 Zweryfikuj lokalizator — Odszukaj co najmniej trzy fragmenty w źródłach.

- Dla PDF sprawdź page_number, page_start i page_end.
- Dla TXT, MD i HTML strony są None; korzystaj z section.

Krok 4 Sprawdź wersję i status — Porównaj dane z właściwym dokumentem.

- status dotyczy dokumentu; ocenę fragmentu qa_status dodasz w R6.
- access_level opisuje dostęp, ale sam nie egzekwuje uprawnień.

Krok 5 Opisz braki — Uzupełnij trzy pola słownika kontrola.

- Nie dopisuj dat ani strony 18 znanej tylko z przykładu na slajdzie.
- section_title z formatu pośredniego ma tu nazwę section.

Krok 6 Zapisz i odczytaj JSON — Uruchom Zapis i ponowny odczyt.

- Porównaj listę odczytaną z pliku z listą w pamięci.
- Sprawdź null dla braków i zapisz notebook.

Ważne: nie poprawiaj tylko eksportu JSON bez korekty danych wejściowych. Inaczej ponowne uruchomienie odtworzy ten sam błąd.

Punkty kontrolne pracy

Po etapie	Masz sprawdzić	Gdy wynik nie pasuje
Po połączeniu	Każdy fragment ma właściwy document_id.	Porównaj klucze z inwentarzem.
Po kontroli	Trzy lokalizatory prowadzą do oryginału.	Popraw źródłowe metadane, potem odtwórz wynik.
Po zapisie	Plik JSON zachowuje schemat i polskie znaki.	Odczytaj plik ponownie.

Nazewnictwo i zapis wyników

Zapisz w wyniki/<RUN_ID>/R5/. Pliki: fragments_with_metadata.json, kontrola_metadanych.json.

Zapisz notebook przez Ctrl+S. Zapis plików istniejącej próby jest blokowany: nową próbę nazwij innym RUN_ID lub świadomie ustaw NADPISZ=True. Sposób oddania pracy ustala prowadzący.

Lista kontroli przed zakończeniem

□ Otworzyłem właściwy notebook i zachowałem oryginały źródeł.
□ Uzupełniłem wszystkie wymagane oceny własnymi słowami.
□ Sprawdziłem kryteria z tabeli punktów kontrolnych.
□ Potrafię wskazać źródło i podstawę moich decyzji.
□ Otworzyłem pliki wynikowe i sprawdziłem ich zgodność z próbą.
□ Zapisałem notebook i znam RUN_ID przekazywanej pracy.

Gdy pojawi się problem

Objaw	Najpierw sprawdź
Brak document_id	Wróć do inwentarza i wyniku R4.
Strony są tekstem	Użyj liczb całkowitych lub None.
TXT ma numer strony	Usuń wymyśloną wartość; zachowaj sekcję.
Powtarza się chunk_id	Sprawdź wersję podziału i źródła; odtwórz R4.

Zasada końcowa: zachowaj dane wejściowe i własne uzasadnienie. Błąd wykryty i poprawnie opisany jest wartościowym wynikiem ćwiczenia.