Organizacja i kontrola jakości bazy wiedzy

Audyt fragmentów i wybór zbioru do Modułu 3

Czas	Pracujesz z	Rezultat
około 35 min	notebookiem, dokumentami i wynikami wcześniejszych etapów	knowledge_base_all.jsonl

Połączysz kontrolę techniczną z oceną treści. Oddzielisz błędy, ostrzeżenia i własne decyzje, a następnie zapiszesz zbiór zaakceptowanych fragmentów.

Cel i rezultat rozdziału

Przygotujesz raport oraz sprawdzony plik JSONL przeznaczony do dalszych eksperymentów.

Po zakończeniu masz

- pełny zbiór knowledge_base_all.jsonl oraz zbiór ready;
- raport qa_report.json i własne qa_summary.md;
- ocenę pokrycia pięciu pytań z identyfikatorami fragmentów.

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_R6_audyt_start.ipynb`

Materiał wejściowy	Do czego służy
wyniki/<RUN_ID>/R5	Fragmenty z metadanymi do walidacji.
wyniki/<RUN_ID>/R1/pytania.json oraz dane/zrodla	Pytania oraz oryginały do kontroli ręcznej.

Ćwiczenie bez LLM: nie potrzebujesz Ollama ani modelu embeddingowego. Najpierw wykonaj test środowiska. Prezentacja M12-M2-v4: slajdy 119–141.

Praca krok po kroku

Krok 1 Uruchom walidację — Wczytaj R5 i uruchom Automatyczna kontrola.

- Sprawdź pola, typy, puste teksty, identyfikatory i źródła.
- Próba JSONL sprawdza jeden obiekt w każdym fizycznym wierszu.

Krok 2 Rozpatrz zgłoszenia — Przejrzyj błędy i ostrzeżenia walidatora.

- Długość 80–1000 znaków to próg kontroli, nie definicja jakości.
- Błąd blokuje accepted; ostrzeżenie wymaga uzasadnionej decyzji.

Krok 3 Sprawdź treść — Porównaj cały mały korpus z oryginałami.

- Oceń procedury, warunki, jednostki, granice i metadane.
- Dla większej bazy: max(10, 5%) losowo oraz wszystkie ostrzeżenia.

Krok 4 Podejmij decyzje — Uzupełnij decyzje w kolejności chunk_id.

- Wpisz accepted, review lub rejected i własne uzasadnienie.
- Poprawkę danych wykonaj we właściwym etapie; ponów R5 i audyt.

Krok 5 Sprawdź pokrycie — Uzupełnij pięć pozycji listy pokrycie.

- Wpisz pełne, częściowe, brak lub niejednoznaczność i dowód.
- chunk_ids wskazują accepted; przy braku wpisz pustą listę.

Krok 6 Zapisz pakiet do M3 — Uzupełnij podsumowanie i uruchom zapis.

- Sprawdź ready: tylko accepted, vector_index kolejno od 0.
- Przekaż do M3 dopiero po GO i opisaniu ograniczeń pokrycia.

Ważne: GO oznacza gotowość do eksperymentów M3, nie pełne pokrycie pytań ani dobrą jakość wyszukiwania. Ograniczenia zapisz w podsumowaniu.

Punkty kontrolne pracy

Po etapie	Masz sprawdzić	Gdy wynik nie pasuje
Po audycie	Każde zgłoszenie ma własną decyzję.	Nie akceptuj automatycznie ostrzeżeń.
Po pokryciu	Oceny pytań wskazują właściwe accepted.	Odszukaj informację w podanym fragmencie.
Po zapisie	Ready ma ciągłą numerację i tylko accepted.	Odczytaj JSONL i porównaj raport.

Nazewnictwo i zapis wyników

Zapisz w wyniki/<RUN_ID>/R6/. Pliki: knowledge_base_all.jsonl, knowledge_base_ready.jsonl, qa_report.json, qa_summary.md.

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
Accepted jest blokowane	Popraw błąd albo odrzuć rekord z uzasadnieniem.
Duplikat treści	Sprawdź pochodzenie i wersję; nie usuwaj na ślepo.
Pozostało review	Wyjaśnij problem; nie przekazuj zbioru z NO-GO.
Brak odpowiedzi w korpusie	Zapisz brak pokrycia; nie twórz fikcyjnego źródła.

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