KARTA PRACY UCZESTNIKA MODUŁ 3 ROZDZIAŁ 4

Wyszukiwanie najbliższych fragmentów

Ranking top-k z tekstem i pochodzeniem każdego wyniku

Czas orientacyjny	Pracujesz z	Rezultat
35 min	zbiorem embeddingów z R2 i funkcją wyszukiwania	ranking_results.jsonl i interpretacja wyników

Zbudujesz wyszukiwanie w całym zbiorze fragmentów. Sprawdzisz, czy wysoko oceniony wynik rzeczywiście zawiera odpowiedź i czy można wskazać jego źródło.

Cel i rezultat rozdziału

Zwrócisz uporządkowane wyniki top-k z identyfikatorem, tekstem, metadanymi i wynikiem podobieństwa.

Po zakończeniu masz

- funkcję search_similar_chunks() dla zapisanych embeddingów;
- rankingi dla trzech pytań i różnych wartości top_k;
- opis trafności i kompletności zwracanych źródeł.

Zanim zaczniesz

Najpierw wykonaj notebook 00. Otwórz wskazany starter, wybierz MODE i uruchom Przygotowanie. Pola UZUPEŁNIJ wypełnij samodzielnie. SOURCE wybiera własne R2 albo checkpoint; nie ma automatycznej zamiany wejścia.

```
notebooki/M12_M3_R4_ranking_top_k_start.ipynb
```

Materiał	Do czego służy
chunks_with_embeddings.jsonl z R2	Ten sam wybrany zbiór co w R3.
embedding_manifest.json z R2	Konfiguracja potrzebna do embeddingu pytań.
Własne pytania	Pytanie proste, parafraza i pytanie bez odpowiedzi w bazie.

Gdzie wrócić w prezentacji

W R4 odszukaj „Funkcja wyszukująca najbliższe fragmenty” oraz „Przykład pełnego rankingu”. Funkcję cosine_similarity() odtwórz z R3, ze slajdu „Funkcja podobieństwa w NumPy”.

Praca krok po kroku

Krok 1 Określ oczekiwania — Przeczytaj źródła i przygotuj trzy pytania: proste, jego parafrazę i pytanie bez odpowiedzi w bazie.

- Uzupełnij case_type, expected_chunk_ids i uzasadnienia dla trzech przygotowanych pytań.
- Brak odpowiedzi określ na podstawie treści całej bazy.

Krok 2 Przeanalizuj wyszukiwanie — Przeczytaj search_similar_chunks(query_embedding, chunks, top_k=3) w kod/m3_tools.py.

- Wskaż obliczenie kosinusa, sortowanie malejące i wybór top_k.
- Sprawdź płaski wynik: chunk_id, text, metadata, score.

Krok 3 Sprawdź granice — Przetestuj poprawność top_k i zachowanie dla małej lub pustej bazy.

- top_k to dodatnia liczba całkowita; odrzuć także True i False.
- Dla top_k większego od liczby rekordów zwróć wszystkie. Dla pustej bazy zwróć pustą listę.

Krok 4 Wykonaj porównanie — Utwórz embedding każdego pytania tym samym modelem i wykonaj wyszukiwanie.

- Porównaj top_k równe 1, 3 i 5 na tej samej bazie.
- Zapisz pełne wyniki, pytanie, top_k i identyfikator wybranej próby R2.

Krok 5 Oceń trafność — Przeczytaj zwrócone teksty i porównaj je z oczekiwaniami.

- Pokaż score, chunk_id, document_id, source i dostępną lokalizację fragmentu.
- Dla pytania bez odpowiedzi wyjaśnij, dlaczego ranking nadal może zwracać wyniki.

Krok 6 Zapisz przebieg — Zapisz ranking_results.jsonl i krótki opis wniosków.

- Jeden wiersz JSONL opisuje jedno pytanie i jedną wartość top_k wraz z listą results.
- Nie przedstawiaj score jako prawdopodobieństwa poprawnej odpowiedzi.

Zapis próby: komórka Przygotowanie automatycznie tworzy nowy folder wyniki/R4/<proba_id>/. Zachowaj wyświetloną ścieżkę.

Punkty kontrolne pracy

Kiedy	Co sprawdzasz	Co zrobić przy problemie
Kontrakt wyniku	Każdy wynik ma cztery wymagane pola.	Zachowaj całe metadata zamiast wybranych etykiet.
Kolejność	Wyniki są uporządkowane malejąco po score.	Sprawdź sortowanie i odcięcie do top_k.
Interpretacja	Odróżniasz znalezienie sąsiada od znalezienia odpowiedzi.	Przeczytaj tekst i sprawdź źródło.

Nazewnictwo i zapis wyników

Notebook zapisuje ranking_results.jsonl oraz wnioski.md w wyniki/R4/<proba_id>/. Zachowaj notebook z funkcjami i kontrolami.

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

- □ Zadeklarowałem oczekiwania przed rankingiem.
- □ Sprawdziłem top_k i pusty zbiór.
- □ Porównałem trzy pytania oraz top_k 1, 3 i 5.
- □ Każdy wynik zawiera tekst i źródło.
- □ Opisałem wynik pytania bez odpowiedzi.
- □ Zapisałem rankingi i wnioski.

Gdy pojawi się problem

Objaw	Co sprawdzić
Wyniki są w złej kolejności	Sortuj malejąco po liczbowym score.
Brakuje źródła	Sprawdź, czy kopiujesz całe metadata z rekordu R2.
Wysoki score bez odpowiedzi	Oceń zawartość tekstu; samo podobieństwo nie potwierdza odpowiedzi.

Lista wyników jest wejściem do dalszej pracy z kontekstem. W tym module nie generujesz jeszcze odpowiedzi modelem językowym. Zapisz także uzupełniony notebook.