KARTA PRACY UCZESTNIKA MODUŁ 2 ROZDZIAŁ 4

Podział dokumentów na fragmenty

Porównanie strategii i zachowanie pełnego kontekstu

Czas	Pracujesz z	Rezultat
około 30 min	notebookiem, dokumentami i wynikami wcześniejszych etapów	chunks.json

Porównasz podział według sekcji, akapitów i liczby znaków. Sprawdzisz, czy warunek wykonania polecenia pozostaje razem z procedurą i oczekiwanym wynikiem.

Cel i rezultat rozdziału

Dobierzesz i uzasadnisz strategię fragmentacji dla własnego zbioru tekstów.

Po zakończeniu masz

- porównanie trzech strategii na tych samych danych;
- wybrane fragmenty z identyfikatorem, źródłem i położeniem;
- opis parametrów, uzasadnienie wyboru i ograniczenia.

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_R4_fragmenty_start.ipynb`

Materiał wejściowy	Do czego służy
wyniki/<RUN_ID>/R3	Oczyszczone rekordy z polem clean_text.
dane/zrodla/instrukcja.txt	Procedura do kontroli granic fragmentów.

Ćwiczenie bez LLM: nie potrzebujesz Ollama ani modelu embeddingowego. Najpierw wykonaj test środowiska. Prezentacja M12-M2-v4: slajdy 71–95.

Praca krok po kroku

Krok 1 Wczytaj clean_text — Uruchom Przygotowanie i porównanie strategii.

- Użyj wyniku R3 z tej samej próby.
- Warianty porównawcze mają limit 420 znaków i overlap 60.

Krok 2 Porównaj granice — Obejrzyj fragmenty procedury w trzech wariantach.

- Znajdź warunek, polecenie, nazwę usługi i wynik active.
- Zapisz konkretne przykłady utraty lub zachowania kontekstu.

Krok 3 Oceń overlap — Sprawdź części wspólne sąsiednich fragmentów.

- Nakładanie tekstu nie jest automatycznie duplikatem błędnym.
- Zwiększenie overlapu nie naprawia każdej przeciętej procedury.

Krok 4 Wybierz parametry — Ustaw STRATEGIA, ROZMIAR i OVERLAP.

- Start: sections, 900 znaków, overlap 150; możesz je zmienić.
- Warunek techniczny: 0 ≤ OVERLAP < ROZMIAR.

Krok 5 Uzasadnij wybór — Uzupełnij porównanie, uzasadnienie i ograniczenia.

- Sprawdź wybrane fragmenty, także początek i koniec strony PDF.
- Kod nie łączy stron PDF; ten przypadek wymaga oceny ręcznej.

Krok 6 Zapisz strategię — Uruchom zapis chunks.json i strategia.json.

- Sprawdź źródło, położenie i unikalność identyfikatorów.
- Zachowaj notebook oraz porownanie_strategii.json.

Ważne: zmiana tekstu lub parametrów podziału wymaga ponownego przygotowania metadanych i audytu. Zachowaj wcześniejszą próbę pod innym RUN_ID.

Punkty kontrolne pracy

Po etapie	Masz sprawdzić	Gdy wynik nie pasuje
Po porównaniu	Warianty wykorzystują te same rekordy.	Sprawdź plik R3 i parametry porównania.
Po wyborze	Procedura nie utraciła warunku ani wyniku.	Zmień granice lub rozmiar i oceń ponownie.
Po zapisie	Strategia opisuje zapisany zestaw fragmentów.	Porównaj parametry w JSON i notebooku.

Nazewnictwo i zapis wyników

Zapisz w wyniki/<RUN_ID>/R4/. Pliki: chunks.json, strategia.json, porownanie_strategii.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
Podział przecina procedurę	Sprawdź granice sekcji, limit i overlap.
Nie ma nagłówka	Sprawdź clean_text i strategię sekcji.
Zmieniły się chunk_id	Po zmianie strategii odtwórz również R5 i R6.
Fragmenty są podobne	Odróżnij overlap od powielonego źródła.

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