KARTA PRACY UCZESTNIKA MODUŁ 2 ROZDZIAŁ 3

Czyszczenie i normalizacja tekstu

Usunięcie szumu przy zachowaniu znaczenia dokumentu

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

Oczyścisz tekst z R2 i porównasz go ze źródłami. Szczególną uwagę poświęcisz poleceniom, warunkom, polskim znakom i jednostkom.

Cel i rezultat rozdziału

Zastosujesz jawne reguły czyszczenia i udokumentujesz ich wpływ na treść.

Po zakończeniu masz

- tekst surowy i oczyszczony przy każdym rekordzie;
- rejestr operacji oraz porównanie co najmniej trzech rekordów;
- ocenę zachowania znaczenia i miejsce do ręcznej kontroli.

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_R3_czyszczenie_start.ipynb
```

Materiał wejściowy	Do czego służy
wyniki/<RUN_ID>/R2	documents_extracted.json z surowym tekstem.
dane/zrodla	Oryginały do sprawdzenia zmian.

Ćwiczenie bez LLM: nie potrzebujesz Ollama ani modelu embeddingowego. Najpierw wykonaj test środowiska. Prezentacja M12-M2-v4: slajdy 48–68.

Praca krok po kroku

Krok 1 Wczytaj tekst — Uruchom Przygotowanie i Reguły czyszczenia.

- Sprawdź zgodność RUN_ID z R2.
- raw_text ma pozostać kopią odczytu, bez poprawek.

Krok 2 Rozpoznaj szum — Porównaj odczyt PDF i HTML z oryginałami.

- Znajdź powtarzalny nagłówek, numery stron i podzielony wyraz.
- Odróżnij techniczną stopkę od tytułu sekcji.

Krok 3 Dobierz reguły — Uzupełnij USUN_WIERSZE oraz ZAMIANY.

- Wskazuj dokładne wiersze i konkretne zamiany, potem ponów komórkę.
- Nie usuwaj wszystkich liczb, łączników ani wcięć.

Krok 4 Porównaj zmiany — Oceń co najmniej trzy rekordy, w tym PDF.

- Sprawdź warunek restartu, active, docassist-api oraz 30 s.
- Zachowaj polskie znaki i jednostki z dokumentu HTML.

Krok 5 Uzasadnij operacje — Uzupełnij cztery pola słownika ocena.

- Opisz przykłady zmian, rodzaje szumu i zachowane elementy.
- Wskaż przynajmniej jedno miejsce wymagające ręcznej kontroli.

Krok 6 Zapisz oczyszczony zbiór — Uruchom zapis i otwórz documents_cleaned.json.

- Sprawdź raw_text, clean_text, operations i warnings.
- Zapisz notebook; w R4 podzielisz clean_text na fragmenty.

Ważne: nie nadpisuj raw_text ani plików źródłowych. Zachowaj ślad porównania; tekst poprawny językowo może utracić znaczenie.

Punkty kontrolne pracy

Po etapie	Masz sprawdzić	Gdy wynik nie pasuje
Po regułach	Każda zmiana ma uzasadnioną przyczynę.	Ogranicz regułę do konkretnego artefaktu.
Po porównaniu	Nie zniknęły polecenia, warunki i jednostki.	Porównaj clean_text z raw_text i oryginałem.
Po zapisie	Surowy tekst i rejestr zmian są zachowane.	Sprawdź pola zapisanych rekordów.

Nazewnictwo i zapis wyników

Zapisz w wyniki/<RUN_ID>/R3/. Pliki: documents_cleaned.json, ocena_czyszczenia.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
Reguła nic nie zmienia	Porównaj dokładny zapis w repr(raw_text).
Zniknął łącznik w poleceniu	Usuń ogólną regułę i wróć do surowego odczytu.
Pozostał znak zastępczy	Popraw sposób odczytu w R2 zamiast ukrywać znak.
Ocena dotyczy starszych reguł	Ponów czyszczenie i ocenę po każdej zmianie.

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