KARTA PRACY UCZESTNIKA MODUŁ 2 ROZDZIAŁ 2

Wczytywanie treści z różnych formatów

Cztery formaty i wspólny zapis odczytanego tekstu

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

Odczytasz TXT, Markdown, PDF i HTML. Porównasz wynik z oryginałami, ponieważ poprawne wykonanie programu nie gwarantuje kompletnego odczytu.

Cel i rezultat rozdziału

Zachowasz tekst, identyfikator źródła, metodę odczytu oraz rzeczywistą stronę PDF.

Po zakończeniu masz

- rekordy tekstu z czterech formatów, w tym osobne strony PDF;
- własne obserwacje poprawności odczytu każdego rekordu;
- plik documents_extracted.json i kontrolę odczytu.

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_R2_odczyt_start.ipynb`

Materiał wejściowy	Do czego służy
wyniki/<RUN_ID>/R1	Inwentarz źródeł dopuszczonych do odczytu.
dane/zrodla	Oryginały TXT, MD, PDF i HTML.

Ćwiczenie bez LLM: nie potrzebujesz Ollama ani modelu embeddingowego. Najpierw wykonaj test środowiska. Prezentacja M12-M2-v4: slajdy 27–45.

Praca krok po kroku

Krok 1 Wczytaj wybór R1 — Uruchom Przygotowanie z tym samym RUN_ID.

- Sprawdź, czy inwentarz istnieje i obejmuje cztery formaty.
- Funkcja czyta wyłącznie pozycje z decyzją użyć.

Krok 2 Odczytaj dokumenty — Uruchom komórkę Odczyt czterech formatów.

- Odczytaj pierwsze 300 znaków oraz metodę i nazwę źródła.
- PDF tworzy rekord dla każdej strony z tekstem.

Krok 3 Sprawdź znaki i kolejność — Porównaj każdy rekord z oryginałem.

- Sprawdź początek, koniec, polskie litery, wartości i kolejność.
- W HTML sprawdź usunięcie menu i zachowanie nagłówków.

Krok 4 Sprawdź strony — Porównaj rekordy PDF ze stronami 1 i 2.

- Dla TXT, Markdown i HTML pole page_number ma wartość None.
- Pusty odczyt wymaga diagnozy; sam w sobie nie dowodzi skanu.

Krok 5 Uzupełnij kontrolę — Wpisz ocenę i obserwację dla każdego rekordu.

- Zapisz zgodny albo do korekty oraz konkretną podstawę oceny.
- OCR jest rozszerzeniem; pakiet nie rozpoznaje tekstu z obrazów.

Krok 6 Zapisz odczyt — Uruchom Zapis wyniku odczytu i otwórz oba pliki.

- Zachowaj również wykryte problemy; nie ukrywaj pustych stron.
- Zapisz notebook i przejdź do czyszczenia w R3.

Ważne: zachowaj tekst odczytany bez ręcznych poprawek. Korekty wykonasz na osobnej kopii w R3, z opisem operacji.

Punkty kontrolne pracy

Po etapie	Masz sprawdzić	Gdy wynik nie pasuje
Po odczycie	Są cztery formaty i właściwe źródła.	Wróć do decyzji inwentarza R1.
Po porównaniu	Strony PDF i tekst odpowiadają oryginałom.	Sprawdź znaki, kolejność i liczby.
Po zapisie	Kontrola obejmuje każdy rekord.	Uzupełnij pominięte obserwacje.

Nazewnictwo i zapis wyników

Zapisz w wyniki/<RUN_ID>/R2/. Pliki: documents_extracted.json, kontrola_odczytu.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 biblioteki	Wybierz Python M2 i zainstaluj requirements.txt.
Brak czterech formatów	Sprawdź decyzje użyć w inwentarzu R1.
Nieczytelne znaki	Sprawdź kodowanie; nie ignoruj błędów odczytu.
PDF zwraca pusty tekst	Obejrzyj stronę; oznacz ją do OCR lub korekty.

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