KARTA PRACY UCZESTNIKA MODUŁ 1 ROZDZIAŁ 5

Podstawowa idea systemu RAG

Od przygotowania dokumentów do obsługi pytania użytkownika

Czas	Pracujesz z	Rezultat
około 20 min	notebookiem i tekstami ćwiczeniowymi bez LLM	architektura_projektu.md

Zaprojektujesz przepływ informacji w RAG. Oddzielisz przygotowanie bazy od odpowiedzi na pojedyncze pytanie. Na tym etapie opisujesz architekturę; nie implementujesz jeszcze indeksu ani wyszukiwarki.

Cel i rezultat rozdziału

Oddzielisz przygotowanie bazy od obsługi pytania i wyjaśnisz role elementów RAG.

Po zakończeniu masz

- dwa uporządkowane przepływy informacji;
- opis ról modelu embeddingowego, indeksu i LLM;
- architektura_projektu.md oraz uzupełniony notebook R5.

Zanim zaczniesz

Rozpakuj cały pakiet i pracuj na własnej kopii katalogu Student. Zachowaj razem foldery notebooki, dane, kod i wyniki. Ścieżki w tej karcie są liczone od Student; pliki danych pozostają bez zmian.

Otwórz notebook

```
notebooki/M12_M1_R5_mapa_rag_start.ipynb
```

Materiał wejściowy	Do czego służy
Schemat w notebooku R5	Dwie ścieżki: przygotowanie i obsługa pytania.
Formularz projekt w notebooku	Osiem pól opisujących role, automatyzację i ryzyko.

To ćwiczenie nie wywołuje modelu. W zapisie pracy oznacz „analiza bez LLM”. Prezentacja M12-M1-v2: slajdy 68–82 przed podziałem na rozdziały.

Praca krok po kroku

Krok 1 Przeczytaj zakres zadania — Uruchom Przygotowanie i odczytaj schemat w Zaprojektuj przepływ.

- Nie instaluj bazy wektorowej ani dodatkowego modelu. Ćwiczenie nie wymaga LLM.
- Odróżnij przygotowanie dokumentów od czynności wykonywanych po nadejściu pytania.

Krok 2 Uzupełnij przygotowanie źródeł — Wypełnij pole przygotowanie w słowniku projekt.

- Uwzględnij odczyt, podział tekstu i metadane.
- Dla planowanego wariantu semantycznego opisz reprezentację oraz zapis powiązania: wektor, fragment i źródło.

Krok 3 Opisz obsługę pytania — Wypełnij pole obsluga_pytania.

- Pokaż drogę od pytania przez wyszukanie fragmentów do promptu i generowania.
- Uwzględnij pokazanie źródeł i weryfikację. Sam wybór podobnego fragmentu nie gwarantuje prawdziwości odpowiedzi.

Krok 4 Przypisz role elementom — Uzupełnij model_embeddingowy, indeks i llm.

- Wyjaśnij, który element tworzy reprezentacje tekstu, który organizuje wyszukiwanie, a który generuje odpowiedź.
- Wyszukiwanie semantyczne to wariant kursowy. Nie jest jedyną dopuszczalną metodą wyszukiwania w RAG.

Krok 5 Zapisz granice i ryzyko — Wypełnij co_automatyzuje_rag, czego_nie_robi_kontekst oraz ryzyko.

- Wskaż, jak automatyzowany jest wybór i przekazanie kontekstu.
- Opisz jeden błąd oraz miejsce jego wykrycia. Nie przypisuj samemu dołączeniu dokumentu trwałej zmiany wag modelu.

Krok 6 Sprawdź architekturę — Uruchom zapis architektura_projektu.md i przeczytaj raport.

- Każdy etap powinien mieć wejście, działanie i rezultat. Model nie otrzymuje samoczynnie dostępu do plików.
- Zapisz notebook. Jako rozszerzenie wyjaśnij, co trzeba odświeżyć po zmianie dokumentu źródłowego.

Ważne: komunikat completion sprawdza obecność wpisów, a nie ich poprawność merytoryczną. Po poprawieniu oceny ponownie uruchom komórkę zapisującą raport.

Punkty kontrolne pracy

Po etapie	Masz sprawdzić	Gdy wynik nie pasuje
Po schemacie	Przygotowanie i obsługa pytania są rozdzielone.	Przenieś czynności do właściwego przepływu.
Po rolach	Embeddingi, indeks i LLM mają odrębne zadania.	Usuń przypisanie generowania modelowi embeddingowemu.
Po ryzyku	Wskazano błąd i etap jego wykrywania.	Dodaj konkretną kontrolę wejścia lub odpowiedzi.

Nazewnictwo i zapis wyników

Notebook zapisuje pliki w wyniki/R5/<data_id>/. Zapis <data_id> oznacza katalog nadany automatycznie, a nie nazwę do ręcznego utworzenia. Odczytaj rzeczywistą ścieżkę z komunikatu Wyniki. Zachowaj nazwy: architektura_projektu.md.

Zapisz notebook przez Ctrl+S. Nie zastępuj nazw słowami final lub ostateczny. Do przekazania wybierz uzupełniony notebook i właściwy katalog wyników według ustaleń prowadzącego; karta nie dodaje nowych zasad zaliczenia.

Lista kontroli przed zakończeniem

- □ Otworzyłem właściwy notebook i zachowałem dane źródłowe.
- □ Wszystkie pola UZUPEŁNIJ są zastąpione własną analizą.
- □ Wypełniłem osiem pól projektu.
- □ Nie utożsamiam RAG z trenowaniem modelu ani gwarancją poprawności.
- □ Otworzyłem plik wynikowy i sprawdziłem zgodność z analizowaną próbą.
- □ Zapisałem notebook oraz oznaczyłem rzeczywisty sposób pracy.

Gdy pojawi się problem

Objaw	Najpierw sprawdź
Nie można znaleźć pakietu lub pliku	Rozpakuj cały katalog Student. Sprawdź, czy obok notebooki znajdują się dane oraz kod.
W raporcie pozostało UZUPEŁNIJ	Zmień pola w komórce oceny, uruchom ją ponownie i przeczytaj zapisany raport.
Schemat zaczyna się od „LLM sam czyta pliki”	Dodaj etap odczytu i wyszukiwania wykonywany przez aplikację, zanim tekst trafi do modelu.

Zasada końcowa: zachowaj oryginalny wynik i własne uzasadnienie. Każdą decyzję połącz ze źródłem albo wyjaśnieniem roli danego elementu.