kod szkolenia: DBX-AI / PL DL 1d

Databricks: Mosaic AI i RAG na Databricks

Szkolenie DBX-AI to praktyczne wprowadzenie do budowania produkcyjnych rozwiązań generatywnej AI na platformie Databricks z użyciem Mosaic AI. Uczestnicy budują end-to-end pipeline RAG (Retrieval-Augmented Generation): od podziału dokumentów produktowych na fragmenty i generowania embeddingów, przez tworzenie indeksu wektorowego w Vector Search, po uruchomienie chatbotu opartego na modelu językowym z ewaluacją odpowiedzi w MLflow

Odbiorcy szkolenia

Szkolenie przeznaczone jest dla:

- Inżynierów danych i inżynierów ML chcących budować rozwiązania oparte na dużych modelach językowych i wyszukiwaniu semantycznym
- Programistów aplikacji AI zainteresowanych integracją LLM z danymi firmowymi (asystenci dokumentacyjni, chatboty Q&A, wyszukiwarki semantyczne)
- Specjalistów MLOps odpowiedzialnych za wdrażanie i monitorowanie pipeline'ów AI w produkcji

Korzyści

- Projektowanie i implementacja pipeline'u RAG: podział dokumentów, generowanie embeddingów, Vector Search i serwowanie modelu językowego
- Tworzenie i zarządzanie indeksem wektorowym w Databricks Vector Search z automatyczną synchronizacją z tabelą Delta
- Wdrożenie łańcucha RAG na Serverless Model Serving z rejestracją w Unity Catalog Model Registry
- Ewaluacja jakości odpowiedzi RAG za pomocą MLflow Evaluate z metrykami trafności, ugruntowania i recall kontekstu

- Śledzenie wywołań modeli językowych i kroków retrieval w pipeline'ach agentic z MLflow Tracing

Program szkolenia

Moduł 1 – Mosaic AI: ekosystem i architektura RAG

- Komponenty Mosaic AI: Model Serving, Vector Search, AI Playground i Agent Framework – gdzie co stosować i jak współdziałają
- Wybór modelu do zadania: modele podstawowe Databricks (Foundation Models API), modele osadzeń oraz zewnętrzne modele językowe przez Model Serving
- AI Playground: interaktywne testowanie modeli i inżynieria promptów przed wdrożeniem pipeline'u
- Architektura RAG na Databricks: przepływ od dokumentu przez wyszukiwanie wektorowe do odpowiedzi modelu językowego

Moduł 2 – Przygotowanie dokumentów: chunking i embeddingi

- Strategie podziału dokumentów na fragmenty: stały rozmiar, podział zdaniowy i semantyczny z nakładaniem – wpływ na jakość wyszukiwania
- Wzbogacanie fragmentów metadanymi: kategoryzacja, tagi produktowe i atrybuty do filtrowania hybrydowego
- Generowanie embeddingów z Foundation Models API: wybór modelu osadzeń i jego wpływ na wymiarowość i jakość wyszukiwania semantycznego
- Zarządzanie tabelą Delta z fragmentami i wektorami osadzeń jako podstawą indeksu wektorowego

Warsztat 1 – Korpus dokumentów: chunking i generowanie embeddingów

Uczestnicy przetwarzają dokumenty produktowe RetailHub: dzielą je na fragmenty wybraną strategią, wzbogacają metadanymi i generują embeddingi z Foundation Models API. Efektem jest tabela Delta z fragmentami i wektorami gotowa do zaindeksowania w Vector Search.

Moduł 3 – Databricks Vector Search: indeksowanie i wyszukiwanie

- Typy indeksów Vector Search: Delta Sync Index z automatyczną synchronizacją oraz Direct Access Index – kiedy wybrać który
- Tworzenie indeksu wektorowego, zarządzanie stanem synchronizacji i monitorowanie wydajności w UI
- Wyszukiwanie podobieństwa: miara kosinusowa, iloczyn skalarny i przybliżone wyszukiwanie sąsiadów – różnice w zachowaniu
- Filtrowanie hybrydowe: wyszukiwanie wektorowe z warunkami na metadanych fragmentów

Moduł 4 – Pipeline RAG: retrieval, augmentacja i serwowanie modelu

- Budowanie łańcucha RAG: zapytanie użytkownika → retrieval z Vector Search → wzbogacenie kontekstem → generowanie odpowiedzi przez model językowy
- Rejestracja łańcucha RAG w Unity Catalog Model Registry z MLflow Models from Code – wersjonowanie i aliasy
- Wdrożenie pipeline'u RAG na Serverless Model Serving – konfiguracja endpointu i obsługa zapytań REST
- MLflow Tracing: śledzenie wywołań modeli językowych, kroków retrieval i opóźnień w pipeline'ach

agentic

Moduł 5 – Ewaluacja i monitoring RAG

- Budowanie zestawu ewaluacyjnego: pytania, oczekiwane odpowiedzi i konteksty referencyjne do pomiaru jakości odpowiedzi
- Uruchamianie automatycznej ewaluacji RAG z MLflow Evaluate: metryki trafności odpowiedzi, ugruntowania i recall kontekstu
- Inference Tables: automatyczne logowanie zapytań do Delty dla monitorowania dryfu jakości odpowiedzi w produkcji
- Databricks Agent Framework: gotowe elementy do budowania agentów konwersacyjnych i orchestracji wieloetapowego rozumowania

Warsztat 2 – Kompletny pipeline RAG: od dokumentów do chatbotu

Uczestnicy finalizują projekt: tworzą indeks wektorowy z korpusu przygotowanego w Warsztacie 1, budują łańcuch RAG, rejestrują go w Unity Catalog Model Registry, wdrażają na Serverless Model Serving i uruchamiają automatyczną ewaluację z MLflow Evaluate. Efektem jest działający chatbot produktowy z pomiarem jakości odpowiedzi.

Oczekiwane przygotowanie uczestnika

- Ukończone szkolenie DBX-MLPREP (Przygotowanie Danych do Machine Learning) lub równoważna wiedza: Delta Lake, Unity Catalog, MLflow podstawy, Model Serving
- Ogólna znajomość pojęcia modelu językowego (LLM) i embeddingu – nie jest wymagane doświadczenie w ich wdrażaniu
- Podstawowa znajomość PySpark i Python (poziom skryptowy)

Szkolenie obejmuje

* dostęp do portalu słuchacza Altkom Akademii

Uwaga:W trakcie wybranych szkoleń Akademia zapewnia Uczestnikom dostęp do płatnych narzędzi i usług niezbędnych do realizacji programu szkolenia (w szczególności usług Azure, Fabric oraz innych zasobów chmurowych). Uczestnik zobowiązany jest do korzystania z udostępnionych zasobów wyłącznie w zakresie wynikającym z agendy szkolenia oraz zgodnie z poleceniami trenera. Wykorzystywanie tych zasobów w sposób wykraczający poza zakres szkoleniowy, w szczególności do realizacji projektów prywatnych, testowania dodatkowych usług, modyfikowania konfiguracji środowiska lub podejmowania działań generujących koszty niezwiązane z realizacją ćwiczeń – skutkować będzie obciążeniem Uczestnika kosztami przekraczającymi przewidziany limit, na podstawie wydatków poniesionych przez Akademię (https://www.altkomakademia.pl/ogolne-warunki-uczestnictwa-w-szkoleniach/)

Język

Wykład: polski
Materiały: angielski

Czas trwania

1 dni / 6 godzin

Opis egzaminu

Szkolenie nie zawiera formalnego egzaminu. Po ukończeniu kursu zaleca się podejście do egzaminu Databricks Certified Machine Learning Associate lub śledzenie roadmapy Mosaic AI dla dalszego rozwoju.