kod szkolenia: DBX-AI / PL DL 1d

Databricks: Mosaic AI i RAG na Databricks

Szkolenie Databricks: Mosaic AI i RAG na Databricks to praktyczne wprowadzenie do tworzenia produkcyjnych rozwiązań generatywnej AI na platformie Databricks. Program koncentruje się na budowie kompletnego pipeline'u RAG, który pozwala łączyć modele językowe z danymi firmowymi i tworzyć rozwiązania takie jak chatboty Q&A, asystenci dokumentacyjni czy wyszukiwarki semantyczne.

Uczestnicy przechodzą przez cały proces pracy z dokumentami, embeddingami, indeksem wektorowym, modelem językowym, wdrożeniem oraz ewaluacją jakości odpowiedzi. Dzięki temu szkolenie pokazuje nie tylko, jak zbudować rozwiązanie RAG, ale też jak przygotować je do działania w środowisku produkcyjnym z wykorzystaniem Mosaic AI, Vector Search, MLflow, Unity Catalog Model Registry i Serverless Model Serving.

Odbiorcy szkolenia

Szkolenie jest przeznaczone dla osób technicznych, które chcą budować rozwiązania generatywnej AI oparte na danych firmowych, dużych modelach językowych i wyszukiwaniu semantycznym w środowisku Databricks.
W szczególności kurs sprawdzi się dla:

- inżynierów danych rozwijających rozwiązania AI z użyciem danych przedsiębiorstwa,
- inżynierów ML budujących pipeline'y oparte na LLM, embeddingach i wyszukiwaniu wektorowym,
- programistów aplikacji AI tworzących chatboty Q&A, asystentów dokumentacyjnych i wyszukiwarki semantyczne,
- specjalistów MLOps odpowiedzialnych za wdrażanie, wersjonowanie, monitorowanie i ocenę pipeline'ów AI,

• zespołów technologicznych, które chcą wykorzystywać Databricks do budowy produkcyjnych rozwiązań RAG.
Szkolenie będzie szczególnie wartościowe dla osób, które chcą przejść od eksperymentów z modelami językowymi do praktycznego wdrożenia aplikacji AI z kontrolą jakości, monitorowaniem i integracją z danymi firmowymi.

Korzyści

• Projektowanie pipeline'u RAG – poznasz, jak budować przepływ od przygotowania dokumentów i generowania embeddingów po wyszukiwanie kontekstu oraz generowanie odpowiedzi przez model językowy.
• Praca z Mosaic AI – dowiesz się, jak wykorzystywać komponenty ekosystemu Databricks, takie jak Model Serving, Vector Search, AI Playground, Foundation Models API i Agent Framework.
• Tworzenie indeksów wektorowych – nauczysz się budować i zarządzać indeksem w Databricks Vector Search, w tym wykorzystywać synchronizację z tabelą Delta oraz filtrowanie hybrydowe.
• Wdrażanie rozwiązań RAG – poznasz, jak rejestrować łańcuch RAG w Unity Catalog Model Registry, wersjonować go i uruchamiać na Serverless Model Serving.
• Ewaluacja jakości odpowiedzi AI – nauczysz się oceniać rozwiązania RAG w MLflow Evaluate z wykorzystaniem metryk trafności, ugruntowania odpowiedzi i recall kontekstu.
• Monitorowanie pipeline'ów AI – zobaczysz, jak śledzić wywołania modeli językowych, kroki retrieval, opóźnienia oraz działanie pipeline'ów agentic za pomocą MLflow Tracing i Inference Tables.
• Budowa praktycznego chatbotu produktowego – przejdziesz przez proces tworzenia rozwiązania end-to-end: od korpusu dokumentów po działający chatbot z pomiarem jakości odpowiedzi.

Program szkolenia

1. Moduł – 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
2. Moduł – 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
- 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.

3. Moduł – 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
4. Moduł – 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
5. Moduł – 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.