databricks

Databricks 시험 가이드

Databricks Certified Generative AI Engineer Associate

시험 가이드 피드백 제공

이 시험 가이드의 목적

이 시험 가이드의 목적은 시험의 개요를 제공하고 시험에서 다루어질 주요 내용을 안내하여 시험에 대비할 수 있도록 돕기 위한 것입니다. 이 문서는 시험에 변경 사항이 있을 때마다(그리고 해당 변경 사항이 시험에 적용될 때) 업데이트되므로 그에 맞춰 준비할 수 있습니다. 이 버전은 **2024**년 **6**월 **1**일 현재 공개 버전에 적용됩니다. 시험일 **2**주 전에 다시 방문하여 최신 버전을 확인하시기 바랍니다.

대상에 대한 설명

Databricks Certified Generative AI Engineer Associate 인증 시험은 Databricks를 사용하여 LLM 지원 솔루션을 설계하고 구현하는 개인의 능력을 평가합니다. 여기에는 복잡한 요구 사항을 관리 가능한 태스크로 분류하는 문제 분석뿐만 아니라 포괄적인 솔루션 개발을 위해 현재 생성형 AI 환경에서 적절한 모델, 도구 및 접근 방식을 선택하는 능력도 포함됩니다. 또한 의미론적 유사성 검색을 위한 Vector Search, 모델 및 솔루션 배포를 위한 모델 서빙, 솔루션 수명주기 관리를 위한 MLflow, 데이터 거버넌스를 위한 Unity Catalog와 같은 Databricks 관련 도구를 평가합니다. 이 시험을 통과한 개인은 Databricks 및 해당 도구 세트를 최대한 활용하는 고성능 RAG 애플리케이션 및 LLM 체인을 구축하고 배포할 수 있게 됩니다.

시험 정보

- 질문 수: 45개의 객관식 또는 다중 선택식 질문
- 시간 제한: 90분
- 등록비: $200
- 시험 실시 방법: 온라인 감독
- 테스트 보조: 허용되지 않음
- 기본 요건: 없습니다. 강좌를 이수하고 Databricks에서 6개월간 실습하는 것이 좋습니다. 또한 이 문서의 권장 준비 사항을 참조하세요.
- 유효 기간: 2년
- 재인증: 인증 상태를 유지하려면 2년마다 재인증을 받아야 합니다. 재인증을 받으려면 현재 제공 중인 전체 시험에 응시해야 합니다. 시험에 다시 응시할 준비를 하려면 시험 웹페이지의 '시험 준비하기' 섹션을 검토하세요.

권장 준비사항

- 생성형 AI 학습자 역할, 특히 Databricks를 사용한 생성형 AI 엔지니어링과 관련된 모든 현재 Databricks Academy 과정
- 현재 LLM 및 LLM의 기능에 대한 지식
- 프롬프트 엔지니어링, 프롬프트 생성, 평가에 대한 지식
- LangChain, Hugging Face Transformers 등과 같은 현재 관련 온라인 도구 및 서비스에 대한 지식
- RAG 애플리케이션 및 LLM 체인 개발을 지원하는 Python 및 해당 라이브러리에 대한 실무 지식
- 데이터 준비, 모델 체이닝 등을 위한 현재 API에 대한 실무 지식
- 관련 Databricks 설명서 리소스

시험 개요

섹션 **1:** 애플리케이션 설계

- 특정 형식의 응답을 이끌어내는 프롬프트 설계
- 특정 비즈니스 요구사항을 실현할 모델 태스크 선택
- 원하는 모델 입력 및 출력을 위한 체인 구성요소 선택
- 비즈니스 사용 사례 목표를 AI 파이프라인의 원하는 입력 및 출력에 대한 설명으로 전환
- 다단계 추론을 위해 지식을 수집하거나 조치를 취하는 도구 정의 및 주문

섹션 **2:** 데이터 준비

- 특정 문서 구조와 모델 제약 조건에 맞춰 청킹 전략 적용
- RAG 애플리케이션의 품질을 저하시키는 소스 문서의 불필요한 콘텐츠 필터링
- 제공된 소스 데이터 및 형식에서 문서 콘텐츠를 추출하기 위해 적절한 Python 패키지 선택
- Unity Catalog의 Delta Lake 테이블에 특정 청크 텍스트를 작성하기 위한 작업 및 시퀀스 정의
- 특정 RAG 애플리케이션에 필요한 지식과 품질을 제공하는 데 필요한 소스 문서 식별
- 특정 모델 작업에 맞는 프롬프트/응답 쌍 식별
- 도구 및 메트릭을 사용하여 검색 성능 평가

섹션 **3:** 애플리케이션 개발

- 특정 데이터 검색 요구에 필요한 데이터 추출 도구 생성
- 생성형 AI 애플리케이션에 사용할 Langchain/비슷한 도구 선택
- 프롬프트 형식이 모델 출력 및 결과를 어떻게 변경할 수 있는지 식별
- 응답을 정성적으로 평가하여 품질 및 안전과 같은 일반적인 문제 식별
- 모델 및 검색 평가를 기반으로 청킹 전략 선택
- 키 필드, 용어, 의도에 따라 사용자 입력에서 얻은 추가적인 컨텍스트를 기반으로 프롬프트 보강
- 기준선에서 원하는 출력까지 LLM의 응답을 조정하는 프롬프트 생성
- 부정적인 결과를 방지하기 위해 LLM 가드레일 구현
- 환각이나 개인 데이터 유출을 최소화하는 메타프롬프트 작성
- 사용 가능한 기능을 표시하는 에이전트 프롬프트 템플릿 구축
- 개발하려는 애플리케이션의 속성을 바탕으로 최적의 LLM 선택

- 소스 문서, 예상 쿼리 및 최적화 전략을 기반으로 임베딩 모델 컨텍스트 길이 선택
- 모델 메타데이터/모델 카드를 기반으로 하는 태스크를 위해 모델 허브 또는 마켓플레이스에서 모델 선택
- 실험에서 생성된 공통 메트릭을 기반으로 특정 태스크에 가장 적합한 모델 선택

섹션 **4:** 애플리케이션 어셈블링 및 배포
- 사전 및 사후 처리가 포함된 pyfunc 모델을 사용하여 체인 코드화
- 모델 서빙 엔드포인트에서 리소스로의 액세스 제어
- 요구 사항에 따라 간단한 체인 코딩
- LangChain으로 간단한 체인 코딩
- RAG 애플리케이션을 생성하는 데 필요한 기본 요소 선택 모델 표준 형식(flavor), 임베딩 모델, 검색기, 종속성, 입력 예, 모델 서명
- MLflow를 사용하여 모델을 Unity Catalog에 등록
- 기본 RAG 애플리케이션을 위한 엔드포인트를 배포하는 데 필요한 단계의 순서 지정
- 벡터 검색 인덱스 생성 및 쿼리
- Foundation Model API를 활용하는 LLM 애플리케이션을 제공하는 방법을 식별합니다.
- RAG 애플리케이션의 기능을 제공하는 데 필요한 리소스 식별

섹션 **5:** 거버넌스
- 성능 목표를 달성하기 위해 마스킹 기술을 가드레일로 사용
- 악의적인 사용자 입력으로부터 생성형 AI 애플리케이션을 보호하기 위한 가드레일 기술 선택
- RAG 애플리케이션에 제공되는 데이터 소스에서 문제가 있는 텍스트를 줄이기 위한 대안 추천
- 법적 위험을 피하도록 데이터 원본에 대한 법적/라이선스 요구 사항 사용

섹션 **6:** 평가 및 모니터링
- 일련의 정량적 평가 지표를 기반으로 LLM 선택(크기 및 아키텍처)
- 특정 LLM 배포 시나리오를 모니터링할 주요 메트릭 선택
- MLflow를 사용하여 RAG 애플리케이션에서 모델 성능 평가
- 추론 로깅을 사용하여 배포된 RAG 애플리케이션 성능 평가
- Databricks 기능을 사용하여 RAG 애플리케이션의 LLM 비용 제어

샘플 질문

이러한 문제는 실제 문제와 유사하며 이 시험에서 문제가 어떻게 출제되는지를 일반적으로 이해할 수 있도록 제공된 것입니다. 샘플 문제에는 시험 가이드에 명시된 시험 목표가 포함되어 있으며, 목표에 맞는 샘플 문제가 제공됩니다. 시험 가이드에는 시험에서 다룰 수 있는 모든 목표가 나열되어 있습니다. 인증 시험을 준비하는 가장 좋은 방법은 시험 가이드의 시험 개요를 검토하는 것입니다.

질문 **1**

목표: 특정 문서 구조와 모델 제약 조건에 맞춰 청킹 전략 적용

생성형 AI 엔지니어가 최대 1억 개를 수용하는 벡터 데이터베이스에 1억 5천만 개의 임베딩을 로드하고 있습니다.

레코드 수를 줄이기 위해 취할 수 있는 두 가지 조치는 무엇일까요?

A. 문서 청크 크기 늘리기
B. 청크 간의 오버랩 줄이기
C. 문서 청크 크기 줄이기
D. 청크 간의 오버랩 늘리기
E. 더 작은 임베딩 모델 사용

질문 **2**

목표: 특정 *RAG* 애플리케이션에 필요한 지식과 품질을 제공하는 데 필요한 소스 문서 식별

생성형 AI 엔지니어가 자동차 부품 판매를 지원하기 위해 개발 중인 고객 대면 생성형 AI 애플리케이션의 응답을 평가하고 있습니다. 이 애플리케이션은 고객이 account_id와 transaction_id를 명시적으로 입력해야 질문에 답할 수 있습니다. 초기 출시 후 고객 피드백에 따르면 애플리케이션이 주문 및 청구 세부 정보에는 잘 응답했지만 배송 및 예상 도착 날짜 질문에는 정확하게 응답하지 못했다고 합니다.

다음 수신기 중 이러한 질문에 답하는 애플리케이션의 기능을 향상시키는 것은 무엇일까요?

A. 모든 자동차 부품에 대한 회사 배송 정책 및 지불 조건을 포함하는 벡터 스토어 생성
B. 송장 데이터와 예상 배송 날짜로 채워지는 기본 키로 transaction_id를 사용하여 Feature Store 테이블 생성
C. 예상 도착일에 대한 예시 데이터를 튜닝 데이터셋으로 제공한 후 배송 정보가 업데이트되도록 모델을 주기적으로 미세 조정
D. 주문한 시점을 입력하도록 채팅 프롬프트를 수정하고, 배송 방법이 14일을 초과할 것으로 예상되지 않으므로 모델에 14일을 추가하도록 지시

질문 **3**

목표: 제공된 소스 데이터 및 형식에서 문서 콘텐츠를 추출하기 위해 적절한 *Python* 패키지 선택

생성형 AI 엔지니어가 스캔되어 .jpeg 또는 .png와 같은 형식의 이미지 파일로 저장된 소스 문서에서 검색된 컨텍스트를 사용하는 RAG 애플리케이션을 구축하고 있습니다. 이 엔지니어는 최소한의 코드 줄을 사용하여 솔루션을 개발하기를 원합니다.

소스 문서에서 텍스트를 추출하려면 어떤 Python 패키지를 사용해야 할까요?

A. beautifulsoup
B. scrapy
C. pytesseract
D. pyquery

질문 **4**

목표: 소스 문서, 예상 쿼리 및 최적화 전략을 기반으로 임베딩 모델 컨텍스트 길이 선택

생성형 AI 엔지니어가 LLM 기반 애플리케이션을 개발하고 있습니다. 해당 검색기를 위한 문서는 각각 최대 512개의 토큰으로 청크되었습니다. 생성형 AI 엔지니어는 이 애플리케이션의 품질보다 비용과 대기 시간이 더 중요하다는 것을 알고 있습니다. 선택할 수 있는 여러 컨텍스트 길이 수준이 있습니다.

다음 중 어떤 항목이 엔지니어의 요구를 충족시킬까요?

A. 컨텍스트 길이 512: 가장 작은 모델은 0.13GB이고 임베딩 크기는 384
B. 컨텍스트 길이 514: 가장 작은 모델은 0.44GB이고 임베딩 크기는 768
C. 컨텍스트 길이 2048: 가장 작은 모델은 11GB이고 임베딩 크기는 2560
D. 컨텍스트 길이 32768: 가장 작은 모델은 14GB이고 임베딩 크기는 4096

질문 **5**

목표: 개발하려는 애플리케이션의 속성을 바탕으로 최적의 *LLM* 선택

생성형 AI 엔지니어가 약 한 단락 길이의 메모 필드를 한 문장의 요약으로 업데이트할 수 있는 애플리케이션을 구축하려고 합니다. 이때 요약한 문장은 메모 필드의 의도를 드러내지만 애플리케이션 프런트 엔드에 맞아야 합니다.

이 애플리케이션에 대한 잠재적 LLM을 평가해야 하는 자연어 처리 태스크 카테고리는 무엇일까요?

A. text2text 생성
B. Sentencizer
C. 텍스트 분류
D. 요약

답변
질문 1: A, B
질문 2: B
질문 3: C
질문 4: A
질문 5: D