일반적인 RAG 구성은

(1) **데이터 색인**: 문서를 임베딩해 벡터 DB에 저장,

(2) **검색 단계**: 사용자 쿼리를 임베딩하여 벡터 DB에서 유사한 문서를 검색,

(3) **생성 단계**: 검색된 문서를 LLM에 컨텍스트로 제공하여 답변을 생성하는 구조로 이루어진다.

이때 벡터 임베딩 모델(예: SBERT 등)과 벡터 DB(예: Pinecone, Milvus 등), LLM(API 형태) 등의 기술이 사용된다.

검색 단계에서는 주로 코사인 유사도 등 유사도 함수를 통해 관련 문서를 찾으며, 필요 시 추가 도구(웹 검색, 계산기 등)를 연동할 수도 있다.

10. 프롬프트 질문은 RAG 구성에서 어디에 전달되는가?

1. 검색어를 벡터화
2. 검색어 쿼리와 유사한 문서를 벡터DB에서 추출
3. 추출된 문서를 조합하여 최종 답변 생성

RAG 워크플로에서 사용자의 **프롬프트(질문)는 먼저 검색 단계로 전달된다.**

즉, 입력된 질문은 임베딩 모델을 통해 벡터로 변환되어 벡터 DB에 쿼리된다.

RAG 파이프라인은 “사용자 쿼리를 임베딩하여 인덱싱된 문서에 유사도 검색을 수행하고, 가장 유사한 문서를 추출”하는 방식으로 동작한다.

이후 검색된 문서들과 원래 질문이 함께 LLM에 주어져 답변을 생성하게 된다.

11. AI 에이전트란 무엇인가? 계층구조 적용의 장점은?

AI 에이전트는 환경과 상호작용하면서 주어진 목표를 달성하기 위해 필요한 행동을 스스로 계획하고 수행하는 자율 지능 시스템이다.

예를 들어 고객문의 상담을 자동으로 처리하며 추가 정보를 탐색하는 챗봇이 이에 해당한다.

계층형 에이전트(상위/하위 계층)를 적용하면 상위 에이전트가 전체 작업을 작은 과제로 분해하여 하위 에이전트에게 할당할 수 있다.

이러한 구조를 활용하면 각 하위 에이전트가 독립적이고 전문화된 역할(예: 검색, 분석, 행동 등)을 수행할 수 있어 신뢰성과 재사용성이 높아진다.

12. 고객, 직원, 데이터, 시큐리티 에이전트의 역할은?

- **고객 에이전트**: 고객의 문의에 응대하고 요구에 맞는 정보를 제공한다.