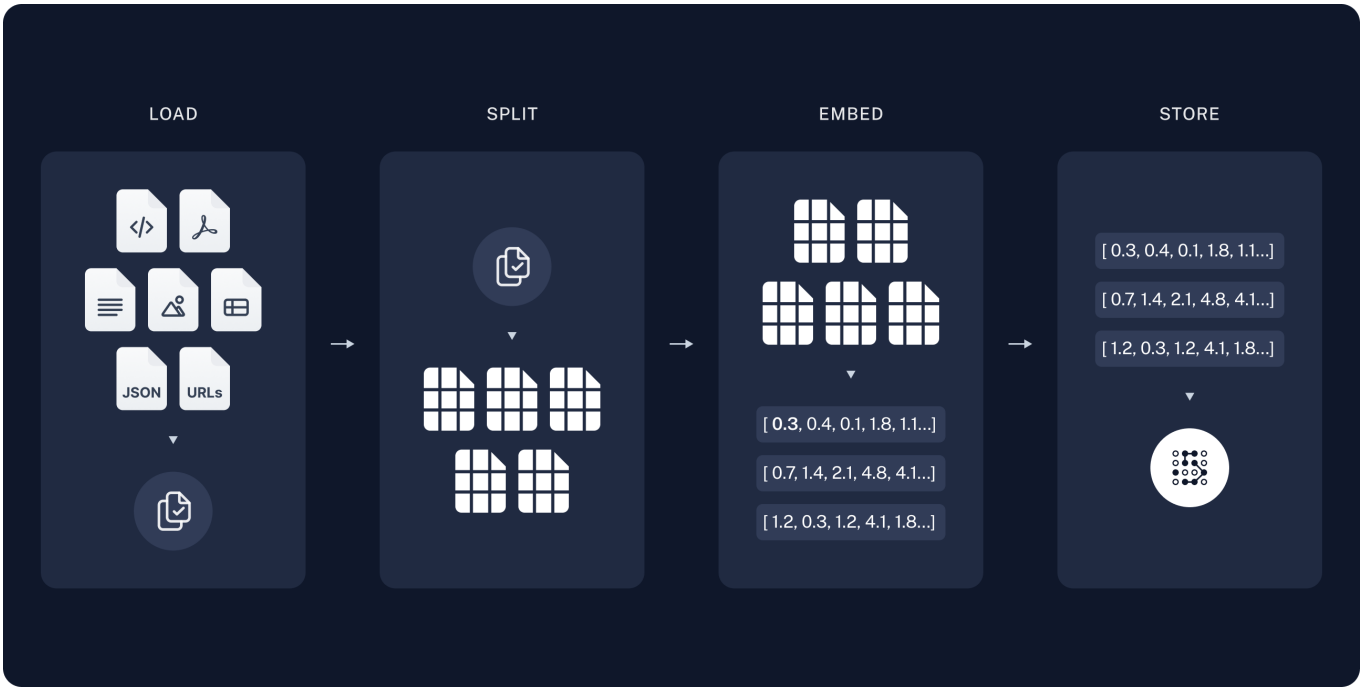
임베딩 (Embedding)

Fast campus



임베딩은 Retrieval-Augmented Generation(RAG) 시스템의 세 번째 단계로, 문서분할 단계에서 생성된 문서 단위들을 **기계가 이해할 수 있는 수치적 형태로 변환**하는 과정입니다. 이 단계는 RAG 시스템의 핵심적인 부분 중 하나로, 문서의 의미를 벡터(숫자의 배열) 형태로 표현함으로써, 사용자가 입력한 질문(Query) 에 대하여 DB 에 저장한 **문서 조각/단락(Chunk) 을 검색**하여 가져올 때 **유사도 계산시 활용**될 수 있습니다.

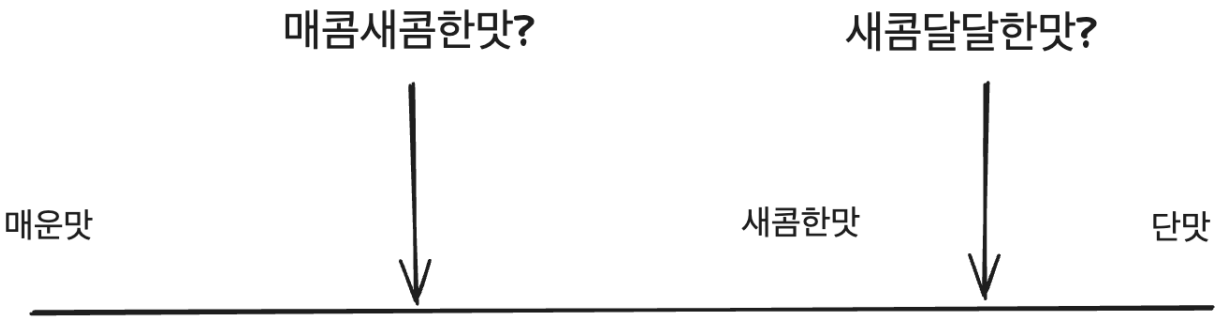
임베딩의 필요성

1. **의미 이해**: 자연 언어는 매우 복잡하고 다양한 의미를 내포하고 있습니다. 임베딩을 통해 이러한 텍스트를 정량화된 형태로 변환함으로써, 컴퓨터가 문서의 내용과 의미를 더 잘 이해하고 처리할 수 있습니다.
2. **정보 검색 향상**: 수치화된 벡터 형태로의 변환은 문서 간의 유사성을 계산하는 데 있어 필수적입니다. 이는 관련 문서를 검색하거나, 질문에 가장 적합한 문

서를 찾는 작업을 용이하게 합니다.

예시

임베딩: 문장을 수치표현으로 변경?



임베딩된 단락 활용 예시

시장조사기관 IDC는 AI 소프트웨어 시장이 2022년 640억 달러에서 2027년 2,510억 달러로 연평균 성장률 31.4%를 기록하며 급성장할 것으로 예상 1

AI 소프트웨어 시장은 AI 플랫폼, AI 애플리케이션, AI 시스템 인프라 소프트웨어(SIS), AI 애플리케이션 개발·배포(AI AD&D) 소프트웨어를 포괄

협업, 콘텐츠 관리, 전사적 자원관리(ERM), 공급망 관리, 생산 및 운영, 엔지니어링, 고객관계관리(CRM)를 포함하는 AI 애플리케이션은 AI 소프트웨어의 최대 시장으로 2023년 전체 매출의 약 3분의 1을 차지하며 2027년까지 21.1%의 연평균 성장률을 기록할 전망 2

AI 비서를 포함한 AI 모델과 애플리케이션의 개발을 뒷받침하는 AI 플랫폼은 두 번째로 시장 규모가 큰 분야로, 2027년까지 35.8%의 연평균 성장률이 예상됨

분석, 비즈니스 인텔리전스, 데이터 관리와 통합을 포함하는 AI SIS는 기존 소프트웨어 시스템과 통합되어 방대한 데이터를 활용한 의사결정과 운영 최적화를 지원하며, 현재 매출 규모는 비교적 작지만 5년간 연평균 성장률은 32.6%로 시장 전체를 웃돌 전망 3

애플리케이션 개발, 소프트웨어 품질과 수명주기 관리 소프트웨어, 애플리케이션 플랫폼을 포함하는 AI AD&D는 향후 5년간 카테고리 중 가장 높은 38.7%의 연평균 성장률이 예상됨

- 1번 단락: [0.1, 0.5, 0.9, ... , 0.1, 0.2]
- 2번 단락: [0.7, 0.1, 0.3, ... , 0.5, 0.6]
- 3번 단락: [0.9, 0.4, 0.5, ... , 0.4, 0.3]

질문: "시장조사기관 IDC 가 예측한 AI 소프트웨어 시장의 연평균 성장률은 어떻게 되나요?"

- [0.1, 0.5, 0.9, ..., 0.2, 0.4]

유사도 계산

- **1번: 80%** -> 선택!
- 2번: 30%
- 3번: 25%

코드

```
from langchain_openai import OpenAIEmbeddings

# 단계 3: 임베딩(Embedding) 생성
embeddings = OpenAIEmbeddings()
```

참고

- 임베딩
- LangChain Text Embeddings