

Deep Learning Application

Chapter 1.

1. LlamaIndex

- LlamaIndex : LLM 기반 애플리케이션 개발에 활용할 수 있는 데이터 프레임워크
 - RAG (Retrieval-Augmented Generation) 시스템 구축에 강점
 - RAG는 LLM이 답변을 생성하기 전에 신뢰할 수 있는 외부 지식 베이스를 참조함으로써 정확성을 높이고, 할루시네이션(잘못된 정보 생성)을 줄이기 위해 개발된 컨텍스트 증강 기술
 - LLM은 방대한 데이터를 기반으로 사전 훈련되지만, 사용자가 보유한 특정 데이터는 학습하지 않았을 가능성.
 - RAG는 이러한 기업 또는 개인의 고유 데이터를 효과적으로 활용하여 LLM과 쉽게 연동할 수 있게 함.
 - 오픈 소스 프로젝트

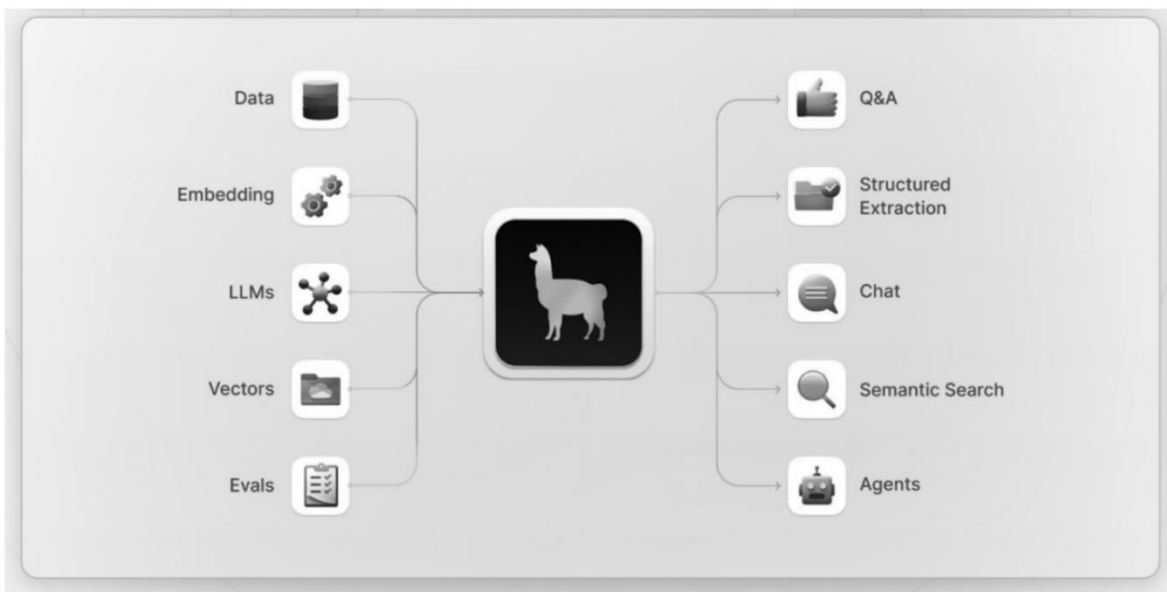


그림 1.1 라마인덱스가 지원하는 작업

1. LlamaIndex

라마인덱스의 주요 특징

- 다양한 데이터 소스 지원: 텍스트, PDF, 웹사이트, AP, 데이터베이스 등 다양한 형식의 데이터를 쉽게 로드
- 데이터 인덱싱: 로드한 데이터를 벡터 임베딩으로 변환하고, 효율적인 검색을 위해 인덱싱
- 관련 문서 검색: 사용자의 쿼리와 가장 연관성이 높은 문서나 데이터를 검색. 다양한 검색 알고리즘을 지원
- 모듈화된 구조: 라마인덱스는 모듈화된 구조로 설계되어 있어 사용자의 요구에 맞게 유연하게 변형해 쓸 수 있음. 각 컴포넌트를 커스터마이징(사용자 맞춤)하거나 교체할 수 있음
- 다양한 백엔드 지원: 여러 종류의 벡터 데이터베이스, 임베딩 모델, LLM 등을 지원
- 쿼리 최적화: 복잡한 쿼리를 자동으로 분해하고 최적화하여, 보다 정확한 응답을 생성.
- 멀티모달 데이터 처리: 텍스트뿐만 아니라 이미지, 오디오 등 다양한 형태의 데이터도 처리.

1. LlamaIndex

1.2. 사용환경 구축

- Python 설치
- Virtual Environment 만들기 (venv vs. pipenv)
 - venv: the built-in way to create an isolated Python environment. It doesn't manage dependencies beyond what pip does.
 - Pipenv: a higher-level project manager that creates venvs for you *and* manages dependencies with a Pipfile/Pipfile.lock for reproducible installs.
- Visual Studio Code 설치
- OpenAI API 키 발급
- 간단한 LlamaIndex 코드 실행 (optional)
- 예제파일 내려받기