PASSION
PIONEER

열정을 가진 자들이 모여 세상을 개척한다

심화반 Week 11.
RAG & LangChain

Presenter: Daehyun Kim
Kakao Tech. Google Developer Groups
MODULABS MODUAI Lab.

Google Developer Groups

PASSION
PIONEER

01 RAG

Google Developer Groups

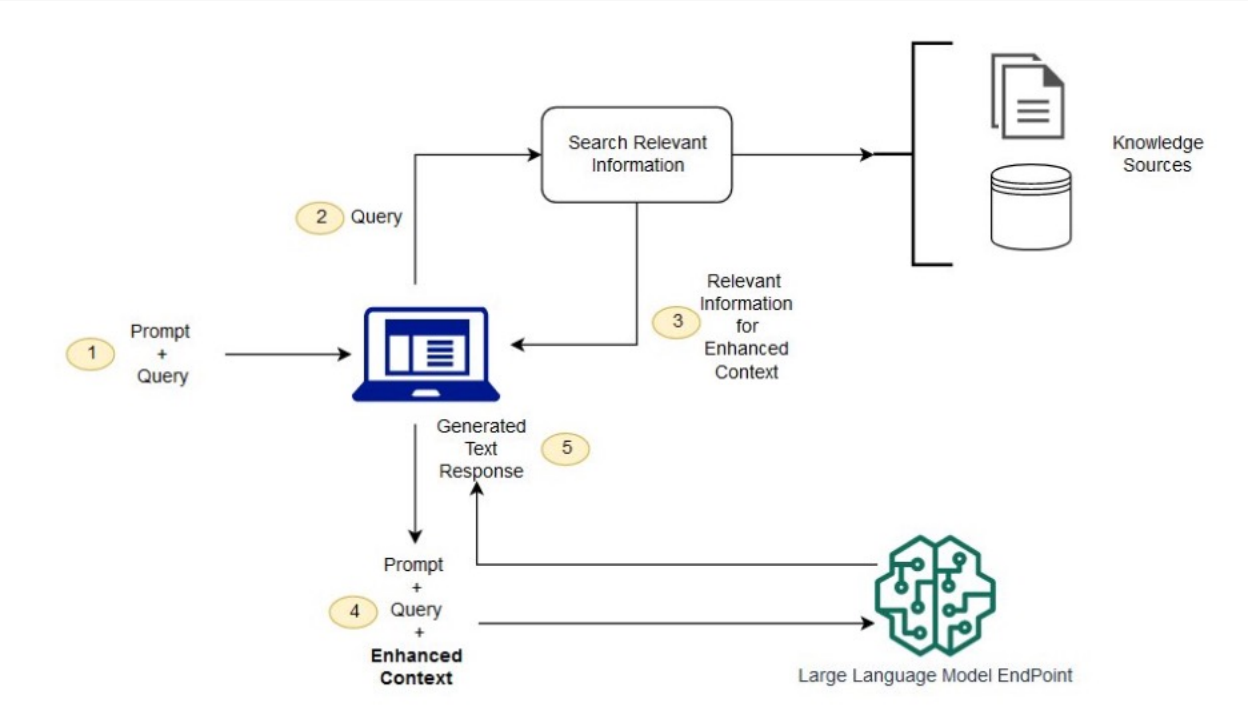
PASSION PIONEER

Retrieve Augment Generation (RAG)

- Retrieve-Augmented Generation (RAG)는 정보 검색(Retrieve)과 텍스트 생성(Generate)을 결합한 하이브리드 언어 모델 아키텍처
- 특정 질의에 대해 외부 지식 베이스에서 관련 정보를 검색한 후, 이 정보를 기반으로 텍스트를 생성하는 방식으로 작동
- 특히 정보가 풍부한 응답을 생성하는 데 유리

PASSION PIONEER

RAG Architecture



PASSION PIONEER

RAG Architecture

RAG의 주요 구조

Retrieve 단계
- 입력 질의(query)와 관련된 문서를 외부 지식 베이스에서 검색
- 쿼리와 문서의 임베딩을 생성하고, 임베딩 간의 유사성을 계산하여 관련 문서를 검색
- 임베딩 알고리즘을 통해 수행 가능

Augmented 단계
- 검색된 문서를 기반으로 입력 질의(query) 강화

Generate 단계
- 입력 질의(query) 또는 검색된 문서등을 기반으로 자연스러운 텍스트를 생성

PASSION PIONEER

RAG Architecture

RAG의 작동 원리

1. 쿼리 임베딩 생성
- 입력된 질의를 임베딩 벡터로 변환
2. 문서 검색
- 쿼리 임베딩을 사용하여 외부 지식 베이스에서 관련 문서들을 검색
3. 문서 임베딩 생성
- 검색된 각 문서를 임베딩 벡터로 변환
4. 문서-쿼리 결합
- 검색된 문서와 쿼리 임베딩을 결합하여 텍스트 생성을 위한 입력으로 사용
5. 텍스트 생성
- 결합된 임베딩을 입력으로 받아 디코더가 새로운 텍스트를 생성

PASSION PIONEER

RAG Architecture

RAG의 장점

- 외부 지식 베이스를 활용하여 더 풍부하고 정확한 정보를 제공
- 다양한 데이터 소스를 결합하여 더 광범위한 지식을 응답에 포함할
- 단순한 생성 모델보다 더 정확하고 정보가 풍부한 응답을 생성
- 내부 LLM모델로 구성하는 경우 보안, 프라이버시 위협에서 어느정도 벗어날 수 있음
- 파인튜닝보다 효율적으로 특화 도메인에 대응 가능
- 검색과 생성을 결합하는 과정이 복잡하며, 모델 학습 및 튜닝이 어려울 수 있
- 임베딩 생성, 문서 검색, 텍스트 생성 등 모든 단계에서 높은 연산 자원이 필요
- 외부 지식 베이스의 품질과 최신성에 크게 의존

PASSION PIONEER

RAG Architecture

RAG 예시
- 질의 응답 시스템 (Question Answering)
: 사용자가 "2022년 FIFA 월드컵 우승 팀은?"이라는 질문을 입력하면, RAG 모델은 최신 뉴스 기사나 공식 발표 자료를 검색 하여 정확한 답변을 제공

- 헬스케어 챗봇
: 사용자가 "고혈압의 일반적인 증상은 무엇인가요?"라고 질문하면, 챗봇은 최신 의학 논문과 건강 정보 웹사이트에서 관련 정보를 검색하여 정확하고 신뢰할 수 있는 답변을 제공합니다. 이는 환자들이 더 신속하고 정확한 건강 정보를 얻는 데 도움이 됩니다

- 학술 논문 검색
: 연구자가 "최근 딥러닝을 활용한 암 진단 연구"에 대해 검색하면, RAG 모델은 최신 학술 논문을 검색하고 주요 내용을 요약하여 제공함으로써 연구자가 필요한 정보를 신속하게 파악할 수 있도록 돕습

- 법률 상담 챗봇
:사용자가 "임대 계약서 작성 시 주의사항은?"이라고 질문하면, 챗봇은 관련 법률 문서와 판례를 검색하여 사용자가 주의해야할 사항들들을 요약하여 제공합니다. 이는 초기 상담 비용을 절감하고, 변호사들이 더 복잡한 문제에 집중할 수 있도록 합니다

PASSION PIONEER

RAG Framework

직접 구현
임베딩
- TF-IDF, 임베딩 모델

정보저장 및 검색
- RDBMS, NoSQL, VectorDB, …

LLM
- Llama, ChatGPT, Phi, …

Langchain
- 대규모 언어 모델(LLM)을 활용한 복잡한 언어 처리 파이프라인을 구축하기 위한 프레임워크
- 다양한 언어 모델과 검색 기법을 결합하여 효율적이고 확장 가능한 자연어 처리(NLP) 시스템을 구축하는 데 중점

LlamaIndex
- 대규모 언어 모델을 기반으로 한 검색 및 생성 작업을 지원하는 프레임워크
- 프레임워크는 대규모 데이터셋을 효율적으로 인덱싱하고, 고성능 검색 기능을 제공하는 데 중점

PASSION PIONEER

RAG Framework

Langchain
- 검색 및 생성을 독립적인 모듈로 구현할 수 있는 구조를 제공
- 따라서 각 구성 요소를 독립적으로 개발, 테스트, 배포 가능
- 다양한 데이터 소스와 검색 방법을 쉽게 통합하는 기능 제공
- 데이터를 처리하고 변환하는 파이프라인을 유연하게 구성 가능
- 사용자가 특정 요구 사항에 맞게 데이터를 전처리하고, 검색 및 생성 단계를 조정 가능
- 통합할 수 있는 환경을 제공합니다. 이는 언어 모델의 학습, 평가, 배포를 위한 도구와 기능을 포함

구조
1. 데이터 소스: 외부 데이터베이스, 웹사이트, 문서 저장소 등 다양한 데이터 소스에서 정보를 검색
2. 검색 모듈: TF-IDF, BM25 등의 검색 알고리즘을 사용해 입력 쿼리와 관련된 문서를 검색하는 기능을 담당
3. 생성 모듈: 언어모델을 사용해 검색된 문서를 바탕으로 자연스러운 텍스트를 생성하는 기능을 담당
4. 파이프라인 관리: 데이터를 전처리하고, 검색 및 생성 단계를 조정하여 최종 응답을 생성

PASSION PIONEER

RAG Framework

LlamaIndex
- The leading data framework for building LLM applications
- 대규모 데이터셋을 효율적으로 인덱싱하고 고성능 검색 기능을 제공하는 데 중점
- 다양한 검색 알고리즘과 언어 모델을 쉽게 통합
- 문서와 쿼리의 임베딩을 생성하고, 유사성을 계산하여 관련 문서를 검색

구조
1. 데이터 인덱싱: 대규모 데이터셋을 효율적으로 인덱싱하여 빠른 검색을 지원하고 그 과정에서 데이터의 구조와 내용을 분석하여 효율적인 검색 가능
2. 쿼리 처리 및 검색: 사용자가 입력한 쿼리를 임베딩 벡터로 변환하고 인덱싱 된 문서를 임베딩 벡터로 변환한 다음 유사도 계산
3. 텍스트 생성: 검색된 문서를 바탕으로 생성에 필요한 입력을 준비
4. 파이프라인 관리: 인덱싱, 검색, 생성을 통합하여 효율적인 작업 흐름을 관리

PASSION PIONEER

RAG Framework

Langchain VS LlamaIndex

Langchain
- 모듈화된 구성 요소: 검색 및 생성 단계를 독립적으로 구현하고 조합할 수 있음
- 확장성: 다양한 데이터 소스와 검색 방법을 쉽게 통합 가능
- 유연한 파이프라인: 데이터를 처리하고 변환하는 파이프라인을 유연하게 구성

LlamaIndex
- 효율적인 인덱싱: 대규모 데이터셋을 구조화된 형태로 저장하여 빠른 검색을 지원
- 강력한 검색 기능: Dense Retrieval 기법을 사용하여 높은 정확도의 검색 결과 제공
- 통합된 생성 기능: 검색된 정보를 바탕으로 자연스러운 텍스트 생성

➔ Langchain은 다양한 응용 분야에서 유연하게 활용할 수 있고,
➔ LlamaIndex는 대규모 데이터셋 기반의 고성능 검색 작업에 적합

PASSION PIONEER

RAG 최신동향

SELF-RAG
- RAG 시스템의 생성 품질과 사실성을 개선하기 위해 SELF-RAG 프레임워크를 제안
- 필요할 때마다 검색하고, 생성한 내용을 자체 반성하여 평가하는 reflection 토큰을 사용
- reflection 토큰을 통해 검색 빈도를 조정하고 사용자 선호도에 맞게 모델 행동 학습

Adaptive RAG
- 쿼리 복잡도에 따라 가장 적절한 전략을 동적으로 선택하는 Adaptive-RAG 프레임워크 개발
- 다양한 복잡도의 쿼리를 처리하기 위해 적응형 RAG 시스템을 제안
- 쿼리 복잡도에 따라 효율성과 정확성을 균형 있게 향상

-> 그외에 새로운 연구/논문들이 계속 나오는 중

PASSION
PIONEER

PNP

Q & A ?

Google Developer Groups