gridone

LLM Course (3)

(주)그리드원

폴리텍대학교 2024. 09. 26

gridone

목차

2일차 수업 내용 요약

2일차 수업 내용 요약

- Ollama 설치
  Ngrok 설치
- 양자화 모델
- 프롬프트 엔지니어링
- 파라미터
- System / User Prompt
- ICL
- N-Shot
- Instruction
- Role
- 구체적인 Task
- 예시 연관정보
- 정규표현식

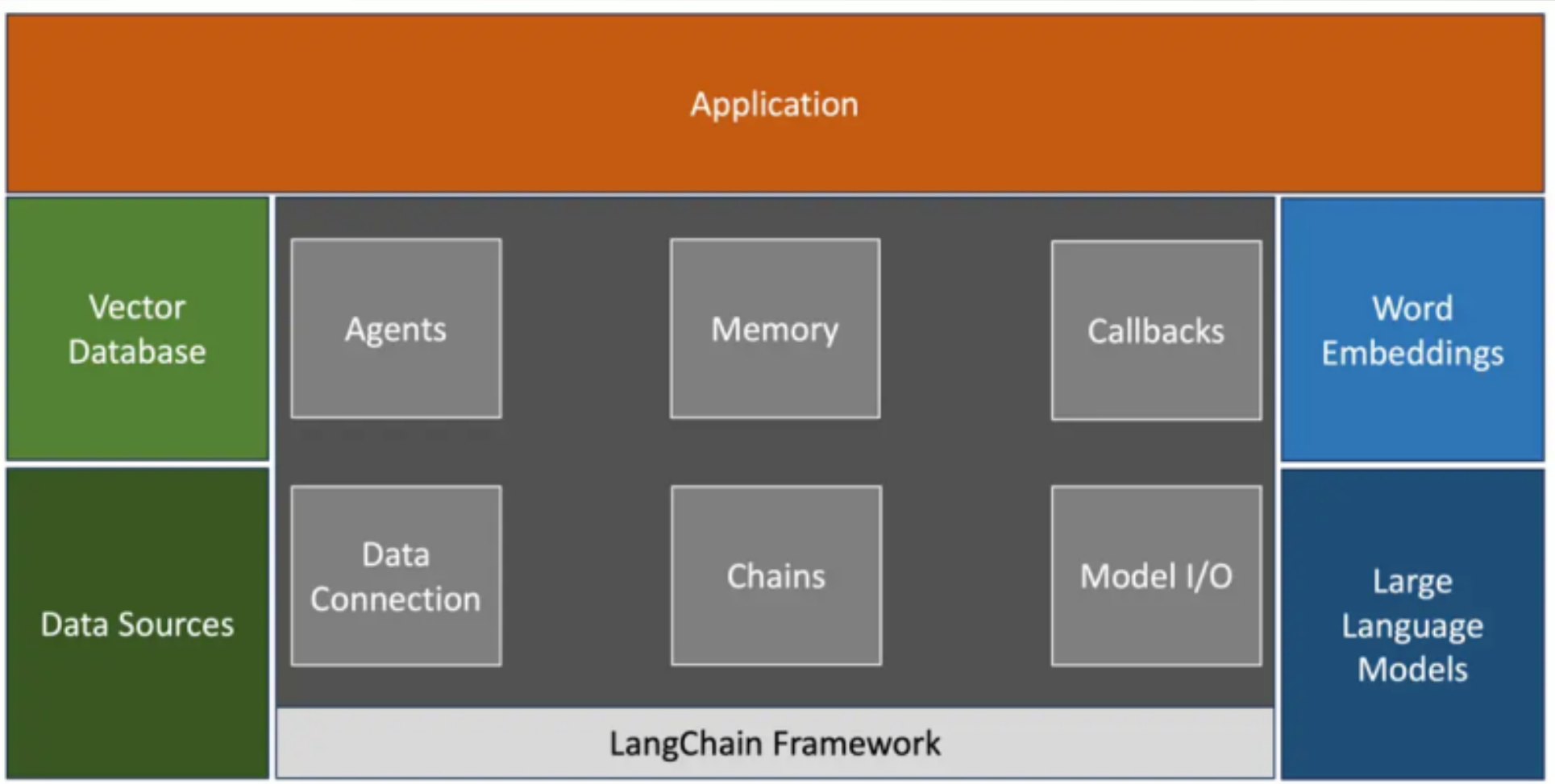
gridone

LangChain

1. LangChain 소개
2. LangChain 구성요소
3. LangChain 실습

LangChain

LangChain



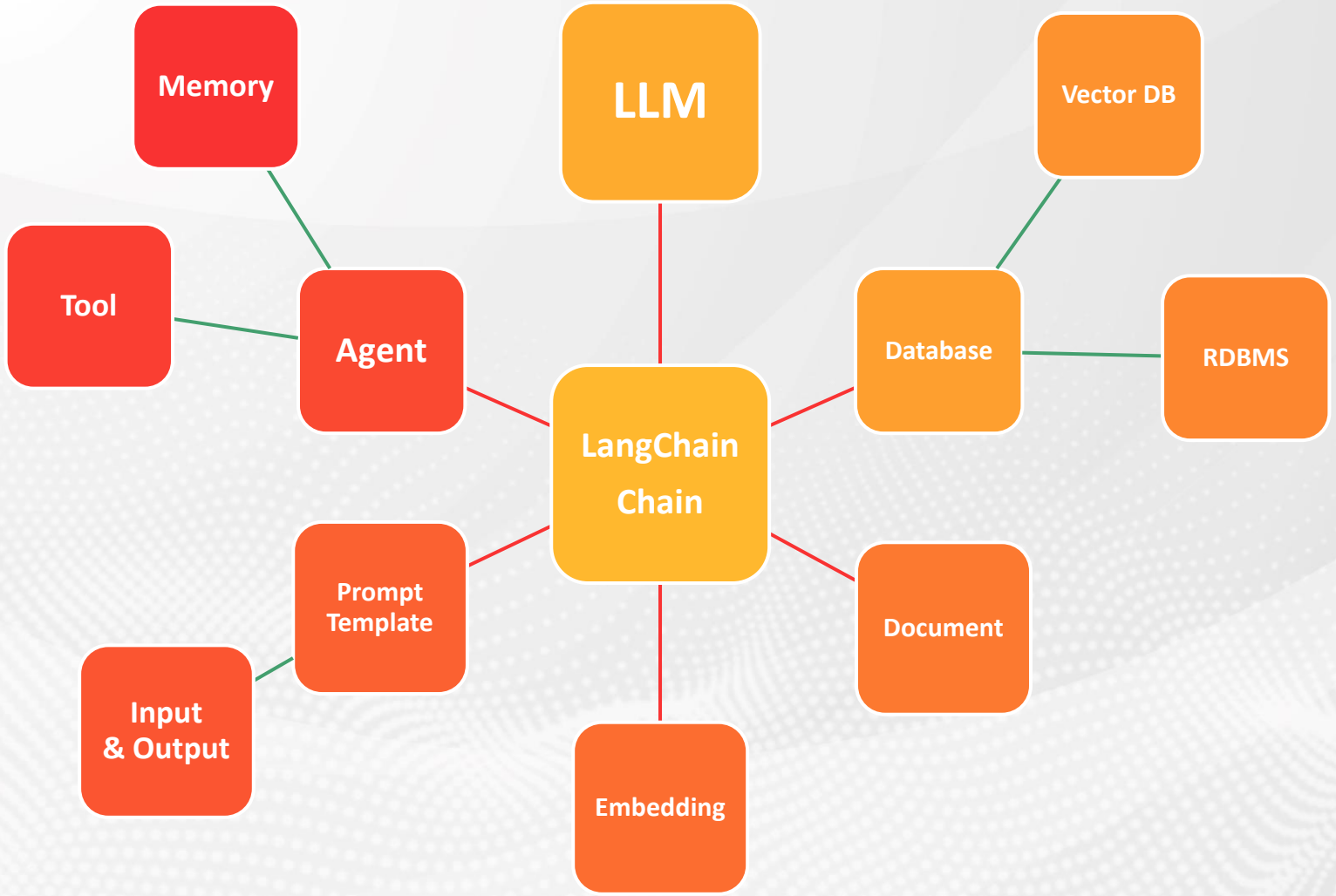
LangChain

LangChain

- 의미
  - LLM을 활용한 애플리케이션 개발을 지원하는 오픈소스 라이브러리
  - LLM과 Application의 용량을 간소화 하도록 설계된 SDK
- 역할
  - LLM이 외부의 '지식' 이나 '계산 능력'을 활용하게 할 수 있도록 하는 것
- 구성요소
  - LLM, 프롬프트 템플릿, 체인, 에이전트, 도구, 메모리
- 효과 – LLM이 학습데이터의 한계를 넘어 LLM의 기능을 확장 시킴

LangChain

LangChain



LangChain

LangChain

- LangChain - LLM
  - 전 세계에서 서비스 하는 LLM(Google, Microsoft, OpenAI 를 포함)과 Local LLM(LLaMa, Mistral, …) 을 호출하는 모듈

- LangChain - PromptTemplate
  - LLM에서 최적의 답변을 도출하기 위한 모듈
  - LLM에 대한 지시, 질문, 답변 예시와 같은 정보를 포함

- LangChain - Chain
  - 모듈과 모듈, 모듈과 함수들을 연결하여 기능을 구현

LangChain

LangChain LLM모델 Python Code

- OpenAI | 🦜🔗 LangChain
- Ollama | 🦜🔗 LangChain
- ChatOllama | 🦜🔗 LangChain
- ChatOpenAI | 🦜🔗 LangChain

- **문장 완성 모델 (Text Completion Model) 과**
  대화 완성 모델(Chat Completion Model)의 차이에 따라
  사용하는 라이브러리 클래스가 다름

LangChain

LangChain

- LangChain - LLM
  - 전 세계에서 서비스 하는 LLM(Google, Microsoft, OpenAI 를 포함)과 Local LLM(LLaMa, Mistral, …) 을 호출하는 모듈

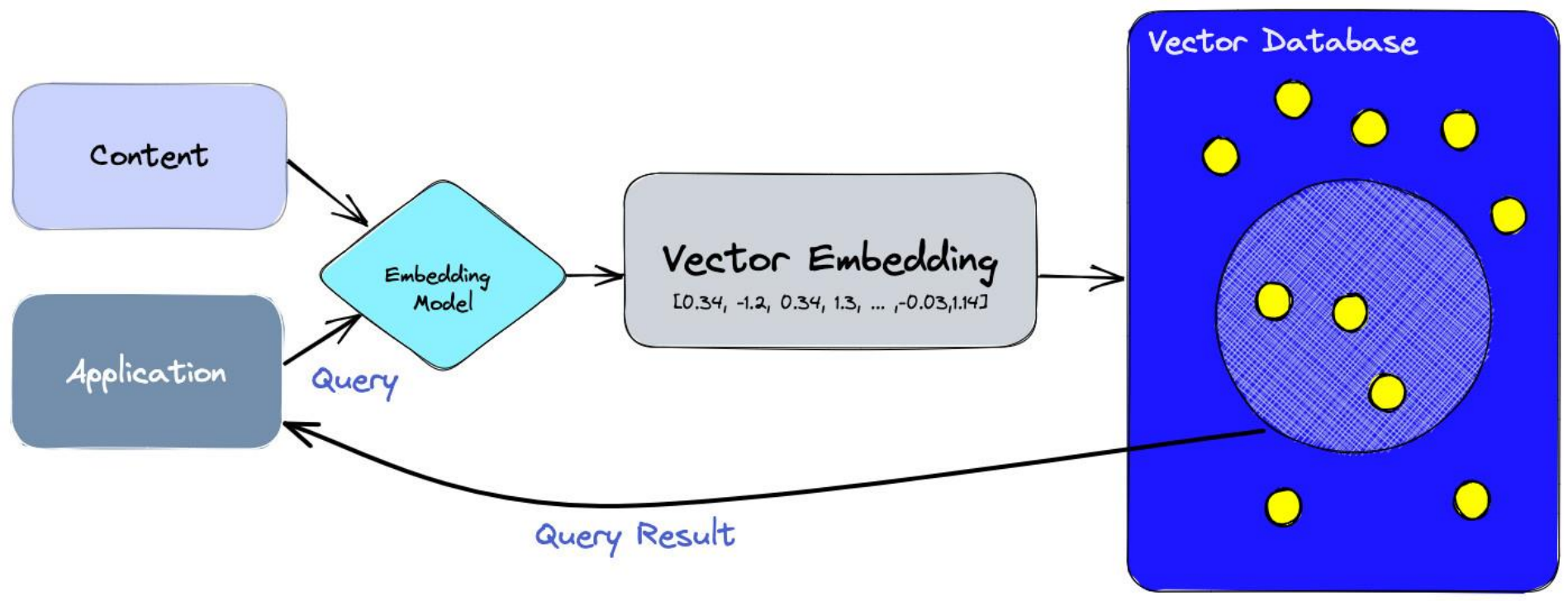
- LangChain - PromptTemplate
  - LLM에서 최적의 답변을 도출하기 위한 모듈
  - LLM에 대한 지시, 질문, 답변 예시와 같은 정보를 포함

- LangChain - Chain
  - 모듈과 모듈, 모듈과 함수들을 연결하여 기능을 구현

LangChain

LangChain

- **Vector DB**



LangChain

LangChain

- Vector DB
  - 벡터 DB를 통해, AI에 시맨틱 정보 검색, 장기 메모리 등의 고급 기능들을 구현할 수 있음.
  - 벡터 임베딩들을 벡터DB에 삽입, 임베딩이 어디에서 생성되었는지 오리지널 콘텐츠에 대한 레퍼런스 포함
  - 어플리케이션이 쿼리를 하면, 같은 임베딩 모델을 이용하여 쿼리에 대한 임베딩을 생성하고, 임베딩으로 DB를 검색하여 비슷한 벡터 임베딩을 찾음

- Vector DB의 장점
  - 데이터 관리 기능 : 데이터 삽입, 삭제, 갱신이 쉬움
  - 메타데이터 저장 및 필터링 : 벡터에 대한 메타데이터 저장이 가능
  - AI도구와의 연동

LangChain 실습

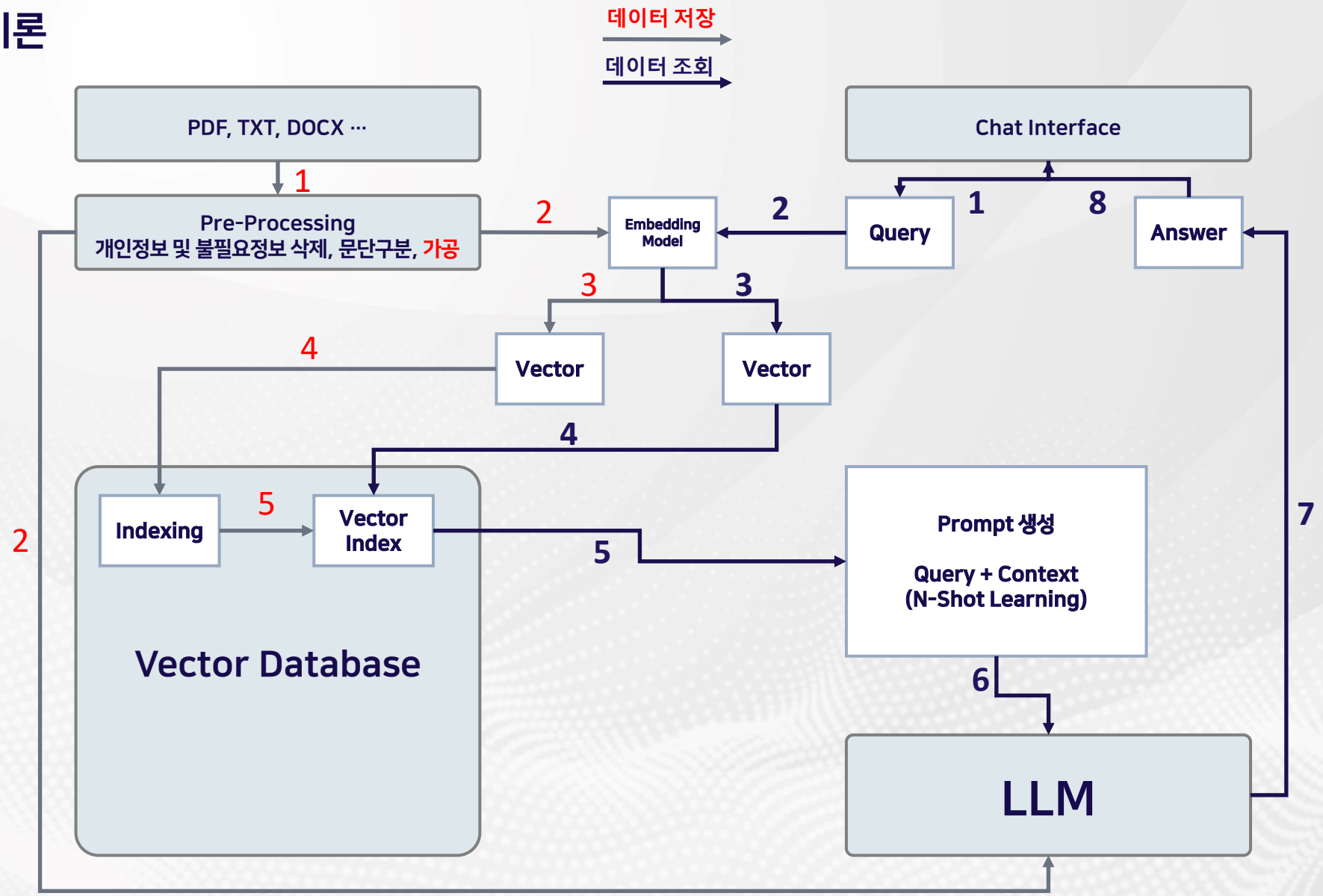
실습 : LangChain 예제 실습

gridone

목차

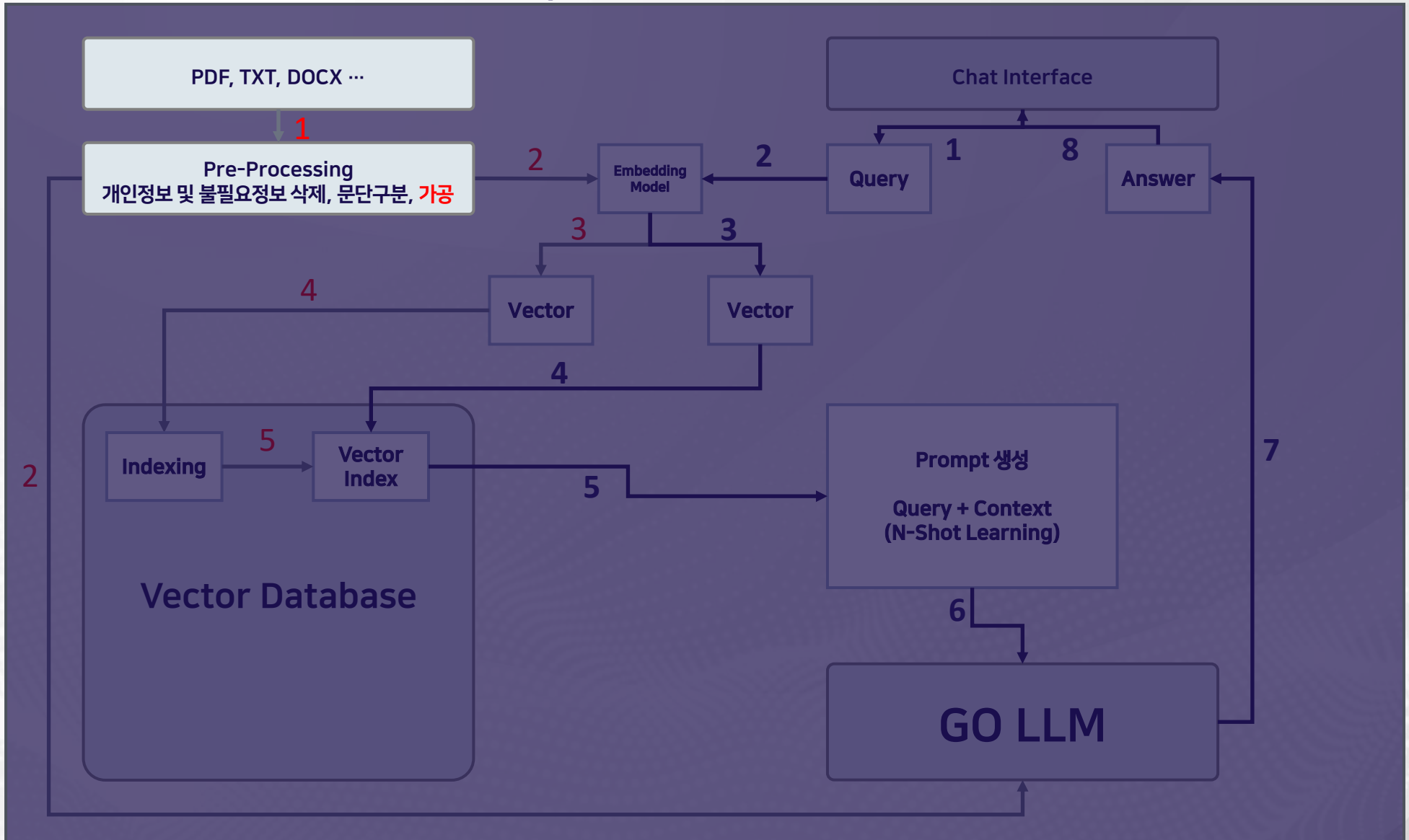
RAG 이론

RAG 이론


데이터 저장
데이터 조회
PDF, TXT, DOCX ···
Chat Interface
1
Pre-Processing
개인정보 및 불필요정보 삭제, 문단구분, 가공
2
Embedding Model
2
Query
1
8
Answer
3
3
Vector
Vector
4
4
Indexing
5
Vector Index
2
Vector Database
5
Prompt 생성
Query + Context
(N-Shot Learning)
7
6
LLM

RAG의 구성요소

RAG의 구성요소 1 : Pre-Processing



RAG의 구성요소

RAG의 구성요소 1 : Pre-Processing

- 전처리

  - 문서의 정보를 분류하고 문서 내 개인정보 혹은 불필요한 정보를 제거
    - 분류 – 텍스트/이미지/표
    - 삭제 – 쪽번호, 반복되는 서명, 장절 제목, 이름, 주민번호, 전화번호 특수문자, HTML태그, 불용어 등

  - 문서를 일관되게 조정하여 필요한 정보를 추출을 용이하게 함.

RAG의 구성요소

RAG의 구성요소 1 : Pre-Processing

- 도큐먼트 청킹
  - 문서의 구조와 내용을 기준으로, 의미단위로 문서를 쪼개는 것.
    - 문장으로 쪼갤 수도 있고, 문단별, 챕터별 등 다양한 쪼개는 방법이 있음
  - Tokenizer가 문장에서 단어를 의미단위로 쪼개듯,
    Chunking은 문서에서 문장을 의미단위로 쪼개는 것
  - 도큐먼트 청킹은 문서의 종류, 도메인, 구조에 따라서 유연하게 쪼갬

RAG의 구성요소

RAG의 구성요소 1 : Pre-Processing 안 한 경우

```
Answer 0
Score:  5.9562817
법제처                                                   40
민법
제469조(제삼자의 변제) ①채무의 변제는 제삼자도 할 수 있다. 그러나 채무의 성질 또는 당사자
의 변제를 허용하지 아니하는 때에는 그러하지 아니하다.
②이해관계없는 제삼자는 채무자의 의사에 반하여 변제하지 못한다.

제470조(채권의 준점유자에 대한 변제) 채권의 준점유자에 대한 변제는 변제자가 선의이며 과실
이 있다.

Answer 1
Score:  5.9562817
못한다. 그러나 그 불법원인이 수익자에게만 있는 때에는 그러하지 아니하다.

제747조(원물반환불능한 경우와 가액반환, 전득자의 책임) ①수익자가 그 받은 목적물을 반환할
액을 반환하여야 한다.
②수익자가 그 이익을 반환할 수 없는 경우에는 수익자로부터 무상으로 그 이익의 목적물을 양수
는 전항의 규정에 의하여 반환할 책임이 있다.
```

RAG의 구성요소

RAG의 구성요소 1 : Pre-Processing

```
민법 [시행 2023. 6. 28.] [법률 제19098호, 2022. 12. 27., 일부개정] 법무부 (법무심의관실) 02-2110-3164

제1편 총칙


제1장 통칙
제1조(법원) 민사에 관하여 법률에 규정이 없으면 관습법에 의하고 관습법이 없으면 조리에 의한다.
제2조(신의성실) ①권리의 행사와 의무의 이행은 신의에 좇아 성실히 하여야 한다. ②권리는 남용하지 못한다.


제2장 인



제1절 능력
제3조(권리능력의 존속기간) 사람은 생존한 동안 권리와 의무의 주체가 된다.
제4조(성년) 사람은 19세로 성년에 이르게 된다. [전문개정 2011. 3. 7.]
제5조(미성년자의 능력) ①미성년자가 법률행위를 함에는 법정대리인의 동의를 얻어야 한다. 그러나 권리만을 얻거나 의무만을 면하는 행위는 그러하지 아니하다.
제6조(처분을 허락한 재산) 법정대리인이 범위를 정하여 처분을 허락한 재산은 미성년자가 임의로 처분할 수 있다.
제7조(동의와 허락의 취소) 법정대리인은 미성년자가 아직 법률행위를 하기 전에는 전2조의 동의와 허락을 취소할 수 있다.
제8조(영업의 허락) ①미성년자가 법정대리인으로부터 허락을 얻은 특정한 영업에 관하여는 성년자와 동일한 행위능력 이 있다. ②법정대리인은 전항의 허락을 취
제9조(성년후견개시의 심판) ① 가정법원은 질병, 장애, 노령, 그 밖의 사유로 인한 정신적 제약으로 사무를 처리할 능력 이 지속적으로 결여된 사람에 대하여 본
제10조(피성년후견인의 행위와 취소) ① 피성년후견인의 법률행위는 취소할 수 있다. ② 제1항에도 불구하고 가정법원은 취소할 수 없는 피성년후견인의 법률행위
제11조(성년후견종료의 심판) 성년후견개시의 원인이 소멸된 경우에는 가정법원은 본인, 배우자, 4촌 이내의 친족, 성 년후견인, 성년후견감독인, 검사 또는 지
제12조(한정후견개시의 심판) ① 가정법원은 질병, 장애, 노령, 그 밖의 사유로 인한 정신적 제약으로 사무를 처리할 능 력이 부족한 사람에 대하여 본인, 배우
```

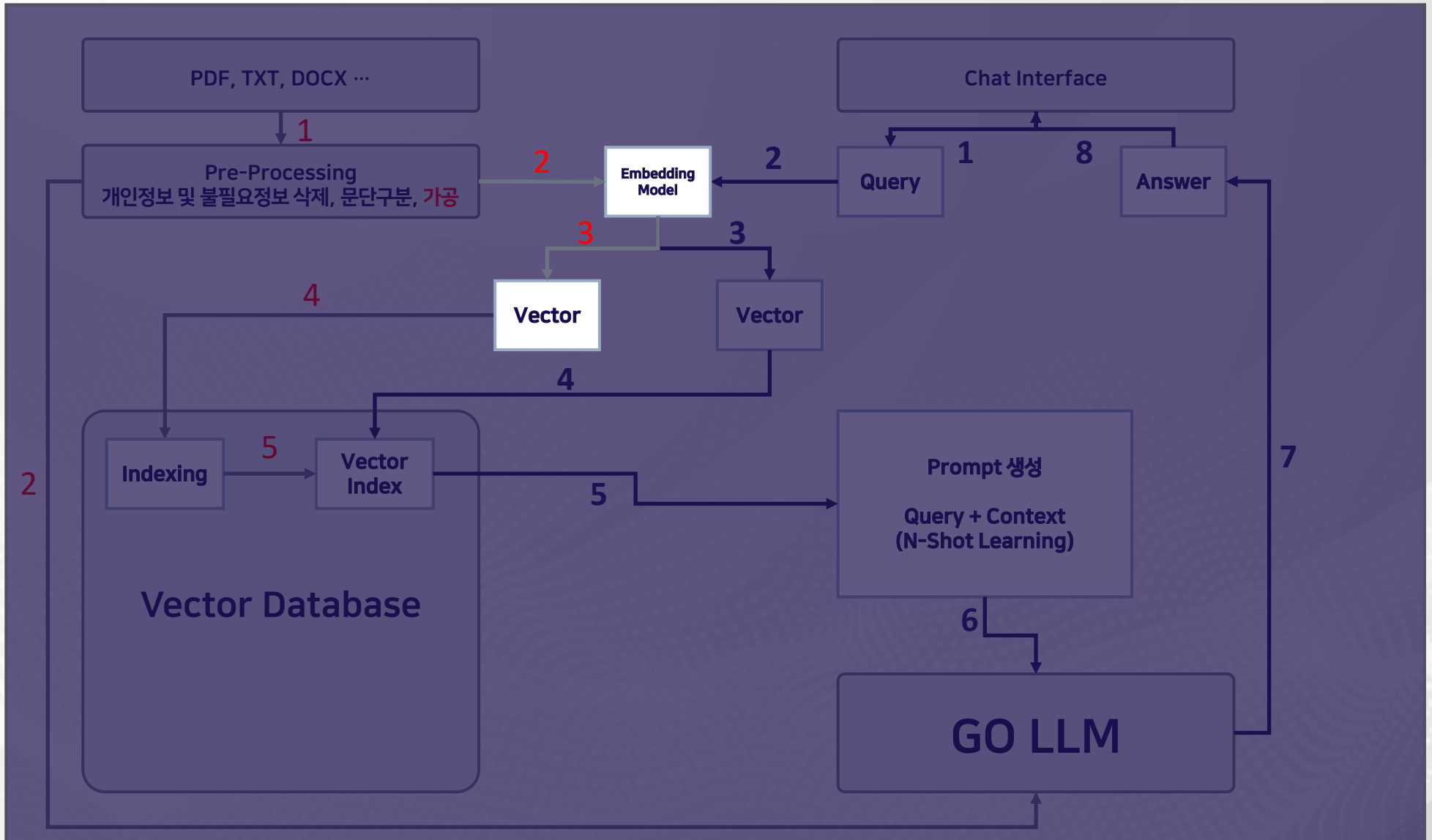
RAG의 구성요소

RAG의 구성요소 1 : Pre-Processing

실습: RAG프로젝트에 적용할 할 PDF문서를 정의하고, 전처리하기

RAG의 구성요소

RAG의 구성요소 2 : Pre-Processing



gridone

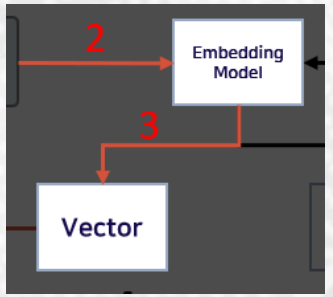
RAG의 구성요소

RAG의 구성요소 2 : Embedding

- **Embedding Model**

  - 전처리된 문서의 텍스트 청크들을 임베딩 모델(인코더)을 통해 벡터로 변환하는 것

  - 벡터로 변환하면서, 단순 수치로 변화하는 것이 아니라, 문장에서 단어와 단어사이의 관계, 문장에서 단어의 순서와 위치 등의 정보를 함께 벡터로 저장



RAG의 구성요소

RAG의 구성요소 2 : Embedding

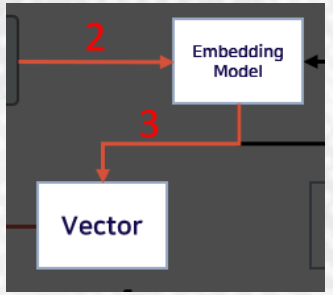
- **Embedding Model**

```
embedder = SentenceTransformer("jhgan/ko-sroberta-multitask")

# Corpus with example sentences
corpus = ['남자','여자','아빠','엄마','누나','왕','여왕','할머니','할아버지','도마뱀']
```

- **'남성' 이라는 단어가 Embedding 모델을 거치면? → Vector**

```
tensor([-1.2694e-01, -2.3778e-01,  6.6165e-01, -1.7557e-01,  6.6660e-01,
        -4.8117e-01, -1.7722e-01,  8.0485e-01, -7.3822e-02, -1.5687e-01,
         5.1065e-02, -2.4846e-01, -5.6627e-01,  1.3569e-01, -3.6579e-02,
         1.3928e-01, -8.4840e-01,  7.7805e-02, -1.8624e-02, -8.9667e-01,
        -7.1099e-01, -3.2162e-01,  7.9521e-01, -2.8384e-01, -4.3439e-01,
         5.2919e-01,  5.2086e-02, -3.5888e-01, -3.0919e-01, -6.1160e-01,
        -3.7066e-01,  6.7306e-01, -4.6369e-01, -1.1722e-01,  1.0218e+00,
        -3.6197e-02, -1.1717e-01, -5.3034e-01, -3.5224e-02, -6.5926e-01,
        -3.4916e-01,  1.6369e-01, -4.0770e-01, -2.6409e-01,  1.8758e-01,
        -3.5828e-01, -6.3495e-01,  2.7272e-01, -1.9814e-01,  6.7971e-01,
         2.8241e-01,  1.1508e-01,  2.9280e-01,  7.1382e-01,  6.4306e-01,
         3.5158e-01,  6.3872e-01, -2.3307e-01,  1.7011e-01,  1.6113e-01,
         1.2837e-01, -5.2138e-02, -1.1354e-01,  4.2775e-01,  6.8378e-01,
        -2.6005e-01, -2.2074e-01, -4.8040e-01, -6.7508e-01,  4.8962e-01,
        -1.1225e-03,  5.2616e-01,  4.5209e-01, -2.1517e-01,  5.8563e-03,
         7.4956e-01, -7.2030e-01,  9.1651e-01,  1.5957e-01, -3.2634e-01,
...
```



gridone

RAG의 구성요소

RAG의 구성요소 2 : Embedding

- **Embedding Model**

```
embedder = SentenceTransformer("jhgan/ko-sroberta-multitask")

# Corpus with example sentences
corpus = ['남자','여자','아빠','엄마','누나','왕','여왕','할머니','할아버지','도마뱀']
```

- **Corpus단어들 중 Query 단어와 유사한 정도**

```
Query: 남성

Top 5 most similar sentences in corpus:
남자 (Score: 0.9134)
아빠 (Score: 0.5391)
할아버지 (Score: 0.4812)
여자 (Score: 0.4573)
왕 (Score: 0.3111)
```

```
Query: 여동생

Top 5 most similar sentences in corpus:
누나 (Score: 0.5975)
여자 (Score: 0.4651)
여왕 (Score: 0.4400)
엄마 (Score: 0.4028)
할머니 (Score: 0.4027)
```

```
Query: 뱀

Top 5 most similar sentences in corpus:
도마뱀 (Score: 0.5833)
할아버지 (Score: 0.3432)
왕 (Score: 0.3404)
남성 (Score: 0.3383)
아빠 (Score: 0.3182)
```

RAG의 구성요소

RAG의 구성요소 2 : Embedding

- **Embedding Model**

```python
# Corpus with example sentences
corpus = ['나는 아침에 밥을 먹었다.',
          '나는 점심에 운동을 한다.',
          '나는 저녁에 책을 읽는다.',
          '나는 자정에 컴퓨터를 한다.',
          '나는 새벽에 잠을 잔다.']
```

- **문장을 가지고서도 가능할까?**

```
Query: 나는 아침식사를 했다.

Top 5 most similar sentences in corpus:
나는 아침에 밥을 먹었다. (Score: 0.9405)
나는 새벽에 잠을 잔다. (Score: 0.4667)
나는 점심에 운동을 한다. (Score: 0.3569)
나는 자정에 컴퓨터를 한다. (Score: 0.2413)
나는 저녁에 책을 읽는다. (Score: 0.2321)
```

```
Query: 나는 조찬을 즐긴다.

Top 5 most similar sentences in corpus:
나는 아침에 밥을 먹었다. (Score: 0.5365)
나는 저녁에 책을 읽는다. (Score: 0.3522)
나는 점심에 운동을 한다. (Score: 0.3090)
나는 새벽에 잠을 잔다. (Score: 0.2934)
나는 자정에 컴퓨터를 한다. (Score: 0.2283)
```

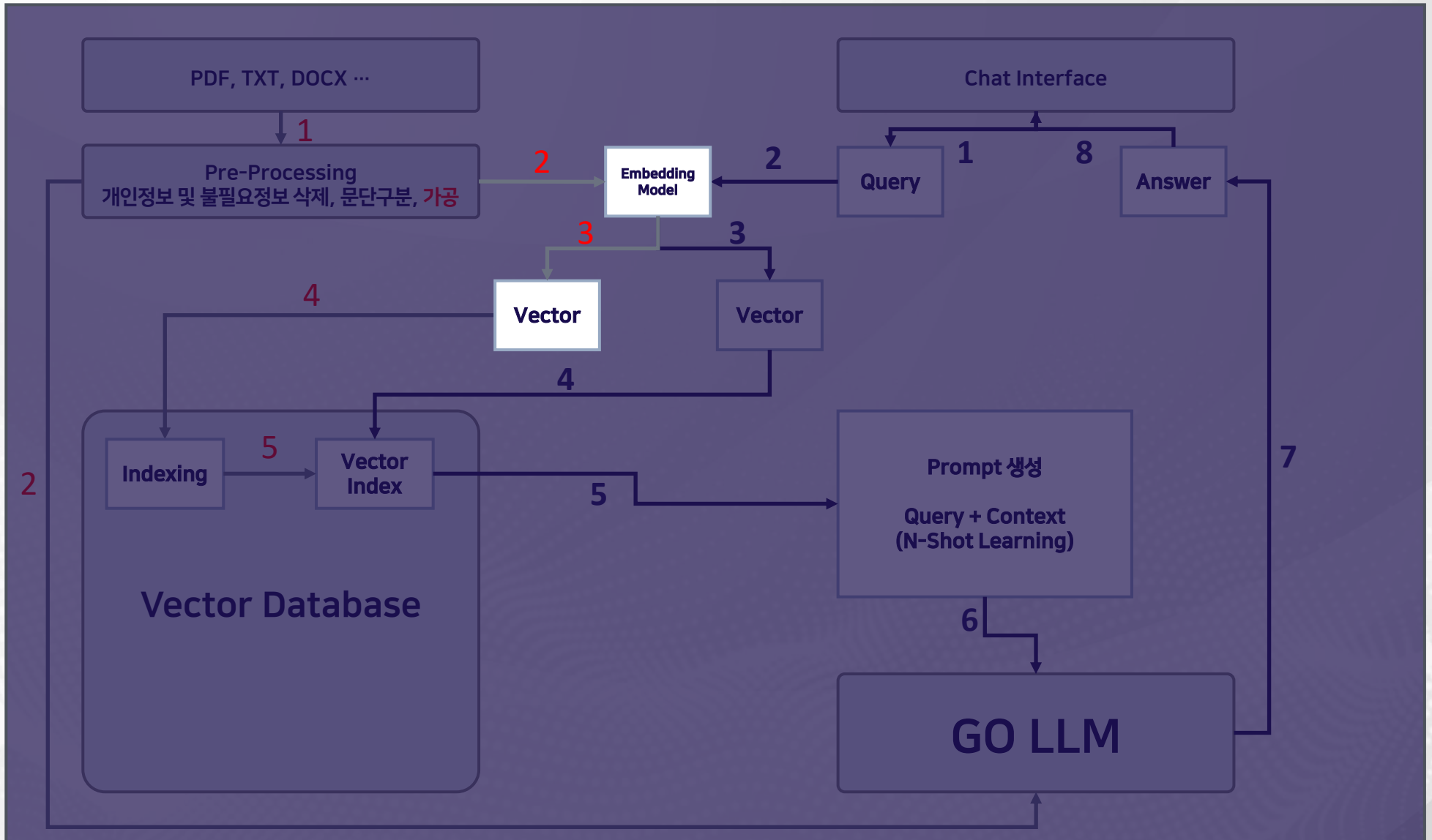
RAG의 구성요소

RAG의 구성요소 2 : Embedding

실습: 임베딩 모델을 가지고서 단어/문장 유사도 확인하기

RAG의 구성요소

RAG의 구성요소 3 : Vector DB 검색(FAISS)



RAG의 구성요소

RAG의 구성요소 3 : Vector DB 검색(FAISS(Vector Index))

- 기존의 키워드 검색
  - 기존의 키워드 검색은 키워드 일치도, 출현빈도 등으로 검색
    - → 남성/남자와 같은 단어를 검색하면 검색이 안되는 문제가 있음
- 벡터DB 검색
  - Embedding을 거친 Vector 정보가 DataBase에 단어와 함께 검색 후 단어의 의미 유사도를 검색할 수 있다면?
    - → 동의어, 유의어, 즉 의미를 중심으로 검색할 수 있게 됨

RAG의 구성요소

RAG의 구성요소 3 : Vector DB 검색(FAISS)

- 벡터DB는 현재 수많은 종류가 있으며, 비용, 오픈소스 여부, 검색 속도, 편의성 등을 고려하여 선택해야 함.

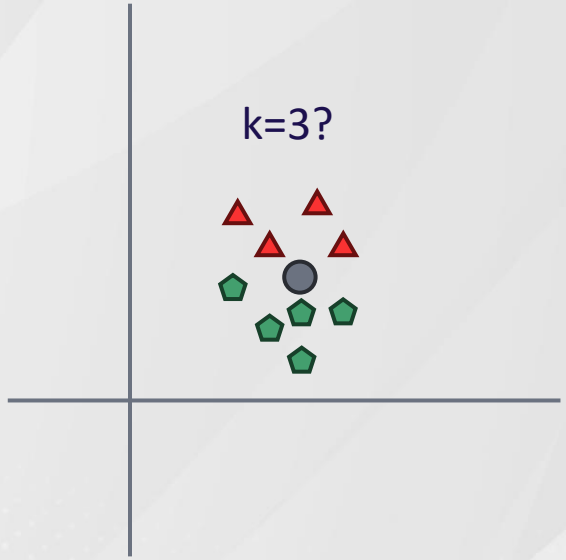
- Picking a vector database: a comparison and guide for 2023 (vectorview.ai)

	Pinecone	Weaviate	Milvus	Qdrant	Chroma	Elasticsearch	PGvector
Is open source	❌	✅	✅	✅	✅	❌	✅
Self-host	❌	✅	✅	✅	✅	✅	✅
Cloud management	✅	✅	✅	✅	❌	✅	(✔)
Purpose-built for Vectors	✅	✅	✅	✅	✅	❌	❌
Developer experience	👍👍👍	👍👍	👍👍	👍👍	👍👍	👍	👍
Community	Community page & events	8k☆ github, 4k slack	23k☆ github, 4k slack	13k☆ github, 3k discord	9k☆ github, 6k discord	23k slack	6k☆ github
Queries per second (using text nytimes-256-angular)	150 *for p2, but more pods can be added	791	2406	326	?	700-100 *from various reports	141

gridone

RAG의 구성요소

RAG의 구성요소 3 : Vector DB 검색(FAISS)



- 벡터DB검색의 주요한 특징
  - 1. k=3, 5, 10, 150 … 을 지정할 수 있음
    - k = k-nearest neighbors (k-최근접 이웃)
      - 가장 유사한 대상의 개수를 지정할 수 있음
      - 전체 1000개의 데이터에서 x를 조회한다고 할 때, k=150으로 지정할 경우, x와 유사한 정도에 따라 150개의 결과를 무조건 반환함
  - → 즉, 가장 가까운 값들을 제외하고는 무관한 데이터가 섞일 가능성이 높음

RAG의 구성요소

RAG의 구성요소 3 : Vector DB 검색(FAISS)

- 벡터DB검색의 주요한 특징

  - 2. 사용자의 잘못된 검색에도 결과를 무조건 도출함
    - 키워드 검색은 검색어와 키워드가 일치 하지 않으면
      검색 결과가 적게 나오거나, 아예 안 나올 수 있음

    - 벡터DB검색은 k=150이라고 지정할 경우, '거리'를 기준으로 검색
      결과를 도출하기 때문에, 충분한 대상들이 있다면
      무조건 검색 결과가 150개가 나옴

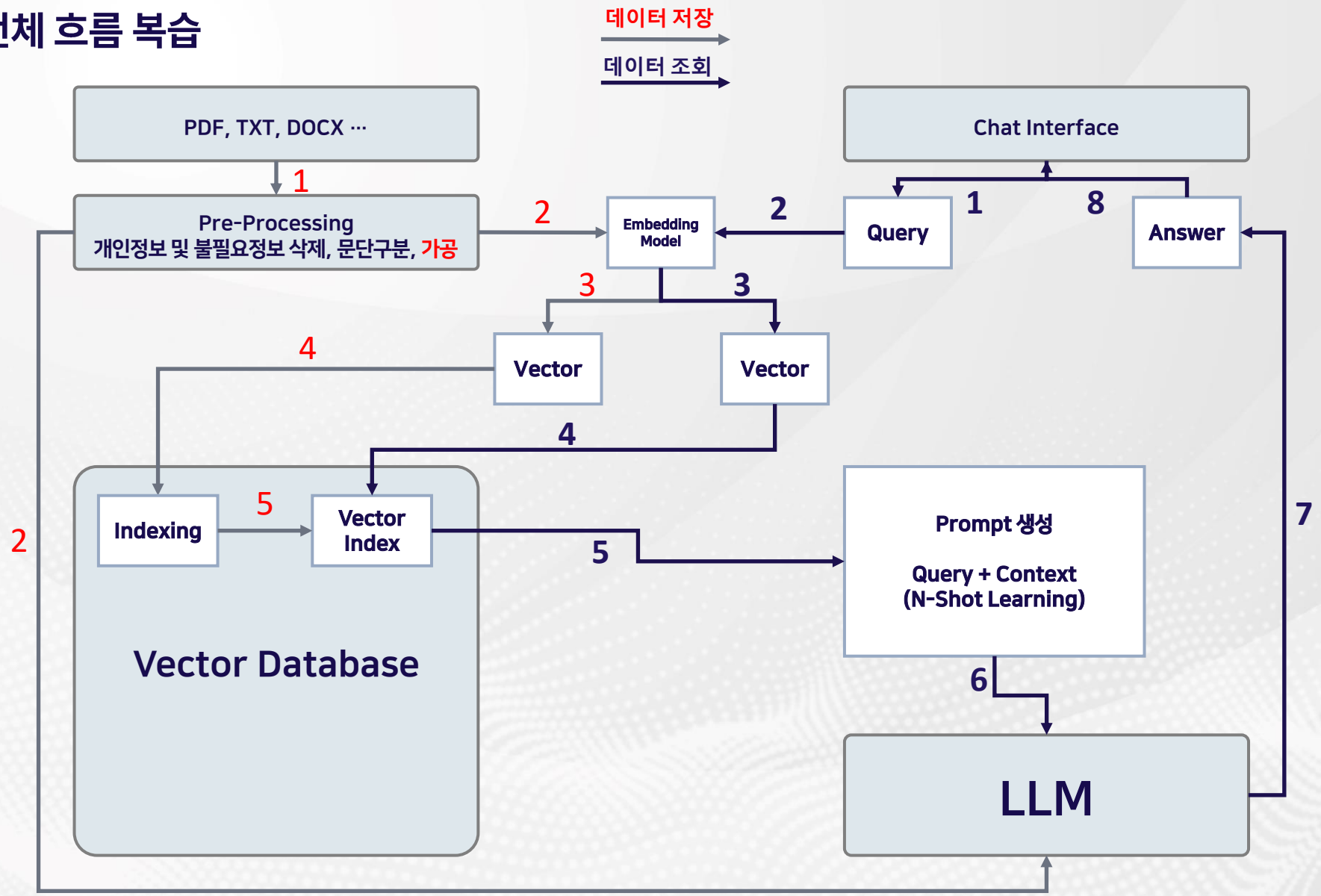
RAG의 구성요소

RAG의 구성요소 3 : Vector DB 검색(FAISS)

실습: PDF문서를 청킹 후 FAISS에 넣어 검색해보기

RAG 이론

RAG 전체 흐름 복습



RAG의 실습

RAG의 실습

실습: Vector DB를 LLM과 Chain하여 결과 얻기

공공데이터 포탈 활용: 미세먼지 안내 챗봇

프로젝트 : 대기오염정보 안내 챗봇

gridone

공공데이터 포탈 활용: 미세먼지 안내 챗봇

프로젝트 : 미세먼지 안내 챗봇

상세기능 | 활용사례 | 추천데이터

상세기능

목록 시도별 실시간 측정정보 조회 조회

시도명을 검색조건으로 하여 시도별 측정소목록에 대한 일반 항목과 CAI최종 실시간 측정값과 지수 정보 조회 기능을 제공하는 시도별 실시간 측정정보 조회

활용승인 절차 개발단계 : 자동승인 / 운영단계 :

신청가능 트래픽 개발계정 : 500 / 운영계정 : 활용사례 등록시 신청하면 트래픽 증가 가능

요청주소 http://apis.data.go.kr/B552584/ArpltnInforInqireSvc/getCtprvnRltmMesureDnsty

서비스URL http://apis.data.go.kr/B552584/ArpltnInforInqireSvc

활용신청

프로젝트 : 미세먼지 안내 챗봇

요청변수(Request Parameter)

항목명(국문)	항목명(영문)	항목크기	항목구분	샘플데이터	항목설명
서비스키	serviceKey	4	필수	-	공공데이터포털에서 받은 인증키
데이터표출방식	returnType	4	옵션	xml	xml 또는 json
한 페이지 결과 수	numOfRows	4	옵션	100	한 페이지 결과 수
페이지 번호	pageNo	4	옵션	1	페이지번호
시도명	sidoName	10	필수	서울	시도 이름(전국, 서울, 부산, 대구, 인천, 광주, 대전, 울산, 경기, 강원, 충북, 충남, 전북, 전남, 경북, 경남, 제주, 세종)
오퍼레이션 버전	ver	4	옵션	1.0	버전별 상세 결과 참고

공공데이터 포탈 활용: 미세먼지 안내 챗봇

프로젝트 : 미세먼지 안내 챗봇

샘플코드

Java | Javascript | C# | PHP | Curl | Objective-C | **Python** | Nodejs | R

```
# Python3 샘플 코드 #


import requests

url = 'http://apis.data.go.kr/B552584/ArpltnInforInqireSvc/getCtprvnRltmMesureDnsty'
params ={'serviceKey' : '서비스키', 'returnType' : 'xml', 'numOfRows' : '100', 'pageNo' : '1', 'sidoName' : '서울', 'ver' : '1.0' }

response = requests.get(url, params=params)
print(response.content)
```

목록

프로젝트 : 미세먼지 안내 챗봇

{"response":{"body":{"totalCount":40,"items":[{"so2Grade":"1","coFlag":null,"khaiValue":"72","so2Value":"0.003","coValue":"0.2","pm25Flag":null,"pm10Flag":null,"o3Grade":"2","pm10Value":"18","khaiGrade":"2","pm25Value":"12","sidoName":"서울","no2Flag":null,"no2Grade":"1","o3Flag":null,"pm25Grade":"1","so2Flag":null,"dataTime":"2024-06-16 11:00","coGrade":"1","no2Value":"0.005","stationName":"중구","pm10Grade":"1","o3Value":"0.056"},

{"so2Grade":"1","coFlag":null,"khaiValue":"65","so2Value":"0.003","coValue":"0.4","pm25Flag":null,"pm10Flag":null,"o3Grade":"2","pm10Value":"25","khaiGrade":"2","pm25Value":"9","sidoName":"서울","no2Flag":null,"no2Grade":"1","o3Flag":null,"pm25Grade":"1","so2Flag":null,"dataTime":"2024-06-16 11:00","coGrade":"1","no2Value":"0.009","stationName":"한강대로","pm10Grade":"1","o3Value":"0.048"},

{"so2Grade":"1","coFlag":null,"khaiValue":"77","so2Value":"0.003","coValue":"0.3","pm25Flag":null,"pm10Flag":null,"o3Grade":"2","pm10Value":"19","khaiGrade":"2","pm25Value":"13","sidoName":"서울","no2Flag":null,"no2Grade":"1","o3Flag":null,"pm25Grade":"1","so2Flag":null,"dataTime":"2024-06-16 11:00","coGrade":"1","no2Value":"0.005","stationName":"종로구","pm10Grade":"1","o3Value":"0.062"},

{"so2Grade":"1","coFlag":null,"khaiValue":"68","so2Value":"0.003","coValue":"0.3","pm25Flag":null,"pm10Flag":null,"o3Grade":"2","pm10Value":"25","khaiGrade":"2","pm25Value":"12","sidoName":"서울","no2Flag":null,"no2Grade":"1","o3Flag":null,"pm25Grade":"1","so2Flag":null,"dataTime":"2024-06-16 11:00","coGrade":"1","no2Value":"0.007","stationName":"청계천로","pm10Grade":"1","o3Value":"0.052"},

RAG의 실습

RAG의 실습

대기질 정보 제공 챗봇

도시 이름을 입력하세요:

서울

궁금한 지역을 입력하세요:

천호대로

지역 선택

DB검색 결과 측정소명: 천호대로, 날짜: 2024-06-16 13:00, 미세먼지농도: 40, 초미세먼지농도: 12, 통합대기환경수치: 63

천호대로의 대기질을 판단하기 위해서는 미세먼지 농도와 초미세먼지 농도를 고려해야 합니다.

- 미세먼지 농도: 40
- 초미세먼지 농도: 12

미세먼지 농도가 31 ~ 80 사이로, 보통 범주에 해당하고 초미세먼지 농도는 좋음 범주(0 ~ 15)에 속해 있습니다. 따라서 천호대로의 대기질은 보통입니다.

또한 통합대기환경수치가 63으로, 이 또한 보통 범주(70 미만)에 해당하므로 천호대로의 전체적인 대기 상태는 보통입니다.

RAG의 실습

RAG의 실습

문제 해결 실습: 대기오염 정보를 Vector DB로 만들어 LLM과 Chain하여 결과 얻기

프로젝트 : 나만의 QA챗봇 만들기

실습: QA챗봇 기획, 데이터 수집, 정제, DB적재 후 QA 챗봇 만들기

- 폴리텍 대학교 학칙 RAG 챗봇

gridone

Find Your Values with AI.
AI와 함께 당신의 가치를 발견하세요.

회사명	(주) 그리드원
Tel	02-2058-2220
E-Mail	gridone@gridone.co.kr
Fax	02-2058-2221
제휴, 제품 문의	02-6412-2339 business@gridone.co.kr
기술 및 교육 문의	02-830-8850 support@gridone.co.kr (기술) edu@gridone.co.kr (교육)
라이선스 요청	license-req@gridone.co.kr
홍보, PR	02-6412-2333