데이터 조회 프로그램

SKN Family AI Camp 14기 : 최종 프로젝트 1팀

프로젝트 주제	LLM 활용 내부 고객 업무 효율성 향상을 위한 구글 API 전문 개발자 지원 AI 기반 문서 검색 시스템
원본 데이터 **(텍스트)**	구글 **API** 공식 문서 (Google for developers) - OAuth 2.0 / Google Identity - People API - Google Drive API - Google Sheets API - Gmail API - API Reference - YouTube Data API - Google Maps - Firestore REST API - Firebase Authentication - BigQuery API 기업 내부 문서 **:** OpenAI 합성 데이터셋(자체 생성)
데이터 전처리 과정	**[구글 API 공식 문서]** OPEN AI API를 사용해서, 구글 API 문서 데이터를 QA셋(jsonl파일)로 변환하였습니다. 구글 API 문서 데이터 원문을 벡터 DB에 넣게 되었을 때, 코드나 문맥이 잘리면서 데이터의 품질이 떨어질 수 있기 때문에 QA셋으로 가공을 하기로 결정하였습니다. 원문 데이터를 벡터 DB에 넣는 작업도 따로 진행 중이므로 QA셋과 원문 데이터를 각각 벡터 DB에 넣은 후 검색 성능을 비교해볼 예정입니다. (1) 각 문서 1개당 최대 10개의 Q&A셋을 생성 ● OPEN AI API를 통해 구글 API 문서 데이터(txt 파일)을 순회 ● 각 문서 1개당 최대 10개의 Q&A셋을 생성하고, 생성된 Q&A셋을 JSONL 파일로 저장 ➔ 문서 길이 등을 고려하지 않고, 한 문서당 최대 10개의 Q&A셋을 만들다 보니, 문서 길이가 긴 경우 Q&A셋으로 만들 때 누락되는 내용이 발생 (2) 페어 단위 Q&A셋 생성 ● 청크 분할 및 오버랩 처리로 긴 문서의 중요 내용이 누락되지 않도록 개선 ● 페어 단위로 처리하여 청크 간 중복 검토를 통해 정보 누락 최소화 ● 질문-답변 생성 시 Q&A가 없으면 생성하지 않도록 하여 불필요한 데이터를 배제 ● Q&A가 없는 페어는 저장하지 않아 최종 결과물의 정확성과 유용성을 확보
DB 사용 용도	**[VectorDB - Chroma]** API 문서 및 내부 문서의 임베딩 벡터 저장 의미 검색 + 키워드 검색 지원 권한 기반 필터링 적용 **[MySQL (관계형 DB)]** 사용자 계정, 직급, 권한 관리 채팅 저장 및 관리 커뮤니티 Q&A, 대화 내역 카드 저장, API 키 관리
사용 데이터	외부 데이터: Google Developers 공식 문서 내부 데이터: 기업 내부 문서 RAG용 전처리 데이터: Google Developers 공식 문서 QA, 기업 내부 문서 QA (말투 포함)

Vector DB 구조	**[구글 API 공식 문서 QA]** C:\Users\Playdata2\Documents\proprecessing>conda activate web_server_env (web_server_env) C:\Users\Playdata2\Documents\proprecessing>python query_chroma.py --query "People API를 사용하려면 어떤 OAuth 2.0 스코프를 요 청해야 하나요?" C:\Users\Playdata2\miniconda3\envs\web_server_env\Lib\site-packages\torch\nn\modules\module.py:1762: FutureWarning: `encoder_attention_mask` is deprecated and will be removed in version 4.55.0 for `XLMRobertaSdpaSelfAttention.forward`. return forward_call(*args, **kwargs) [1] id=c791bbce-2b1e-4567-a082-f2dd5129d23e similarity=71.0% (distance=0.2902) Q: `people.searchDirectoryPeople` API를 호출할 때 필요한 OAuth 범위는 무엇인가요? A: `people.searchDirectoryPeople` API를 호출하기 위해서는 `https://www.googleapis.com/auth/directory.readonly` OAuth 범위가 필요합니다. 이 범위는 인증된 사용자의 도메인 디렉터리 정보를 읽기 위해 필요합니다. -> Q: `people.searchDirectoryPeople` API를 호출할 때 필요한 OAuth 범위는 무엇인가요? -> A: `people.searchDirectoryPeople` API를 호출하기 위해서는 `https://www.googleapis.com/auth/directory.readonly` OAuth 범위가 필요합니다. 이 범위는 인증된 사용자의 도메인 디렉터리 정보를 읽기 위해 필요합니다. -> sources: ["https://developers.google.com/people/api/rest/v1/people/searchDirectoryPeople?hl=ko"] -> tags: people -> last_verified: 2025-08-19 -> source_file: people_v1_troubleshoot-authentication-authorization.txt [2] id=a3f8a0ca-d4da-4a9b-a31f-91650c858586 similarity=67.7% (distance=0.3232) Q: `people.get` 메서드를 사용하여 비공식 데이터에 접근하기 위해 필요한 OAuth 범위는 무엇인가요? A: 비공식 데이터에 접근하기 위해서는 다음과 같은 OAuth 범위 중 하나가 필요합니다: `https://www.googleapis.com/auth/contacts`, `https://www.googleapis.com/auth/userinfo.profile` 등. 이 범위는 요청하는 데이터의 종류에 따라 다릅니다. -> Q: `people.get` 메서드를 사용하여 비공식 데이터에 접근하기 위해 필요한 OAuth 범위는 무엇인가요? -> A: 비공식 데이터에 접근하기 위해서는 다음과 같은 OAuth 범위 중 하나가 필요합니다: `https://www.googleapis.com/auth/contacts`, `https://www.googleapis.com/auth/userinfo.profile` 등. 이 범위는 요청하는 데이터의 종류에 따라 다릅니다. -> sources: ["https://developers.google.com/people/api/rest/v1/people/get?hl=ko"] -> tags: people -> last_verified: 2025-08-19 -> source_file: people_v1_troubleshoot-authentication-authorization.txt [3] id=6fa4ba97-6a77-4503-bb89-4bd4bb893e64 similarity=67.2% (distance=0.3279) Q: People API를 사용하기 위해 OAuth 2.0 클라이언트 ID를 생성하는 방법은 무엇인가요? A: Google Cloud 콘솔에서 메뉴 > 클라이언트로 이동한 후 '클라이언트 만들기'를 클릭하고 애플리케이션 유형으로 '데스크톱 앱'을 선택합니다. 이름을 입력하고 만들기를 클릭하면 OAuth 2.0 클라이언트 ID가 생성됩니다. -> Q: People API를 사용하기 위해 OAuth 2.0 클라이언트 ID를 생성하는 방법은 무엇인가요? -> A: Google Cloud 콘솔에서 메뉴 > 클라이언트로 이동한 후 '클라이언트 만들기'를 클릭하고 애플리케이션 유형으로 '데스크톱 앱'을 선택합니다. 이름을 입력하고 만들기를 클릭하면 OAuth 2.0 클라이언트 ID가 생성됩니다.
Django 내부 **RDB** 구조