S

**SK네트웍스 Family AI 과정 14기  
데이터 수집 및 저장 데이터 조회 프로그램**



| **산출물 단계** | 데이터 수집 및 저장 |
| --- | --- |
| **평가 산출물** | 데이터 조회 프로그램 |
| **제출 일자** | 2025. 08. 22 |
| **깃허브 경로** | https://github.com/skn-ai14-250409/SKN14-Final-2Team |
| **작성 팀원** | 유용환 |

| **원본 데이터**  **(텍스트)** | **[Web scraping 기반 향수 정보 수집]**   1. 자료 출처 : 해외직구 플랫폼 “바이슈코” 2. 수집 방식 : 웹 스크래핑(Web Scraping) 3. 수집항목 :  * 브랜드, 제품명, 용량, 가격, 제품 상세 url,   description (부향률, 메인어코드, 탑/미들/베이스 노트, 향 설명)   1. Output 파일 : perfume.csv |
| --- | --- |
| **데이터 전처리**  **과정** | **[수집데이터 구조 및 전처리 대상]**   1. 전처리 대상 : perfume.csv 2. 처리 방식 : description 텍스트에서 키워드 기반으로 향수 특성 정보를 추출하고, 별도 컬럼으로 분리 3. 추가 생성된 컬럼 :  * 부향률, 메인 어코드, 탑 노트, 미들 노트, 베이스 노트, 향 설명  1. 전처리 된 최종 파일 : after\_prepro\_del\_txt.csv |
| **DB 사용 용도** | **[Vector DB 구축 및 활용 목적]**   1. 사용 목적  * 사용자의 쿼리(Query)에 적합한 향수를 의미기반으로 검색하여 추천 * 자연어 검색 질의(Query)에 대응하기 위해 텍스트를 임베딩하고 벡터 DB에 저장 * 향수 상세정보에 대한 설명 또는 비교 분석 질문에 대응 가능  1. 활용 구조  * 모든 컬럼을 “page\_content + metadata” 구조로 변환 * LangChain + Pinecone 사용하여 데이터 조회 프로그램 구현 |
| **VectorDB 구조** | **<Pinecone Vector DB>**      **<Data\_Search\_Program>** |