

**SK네트웍스 Family AI과정 14기  
 데이터 수집 및 저장 데이터 베이스 설계서**



**□ 개요**

* 산출물 단계 : 데이터 수집 및 저장
* 평가 산출물 : 데이터베이스 설계서
* 제출 일자 : 2025. 09. 25
* 깃허브 경로 : https://github.com/skn-ai14-250409/SKN14-Final-2Team
* 작성 팀원 :

| **개요** | * 소개 * 시스템 개요 |
| --- | --- |
| **구조** | * 시스템 아키텍처 * 요구사항 매트릭스 |
| **설계** | * 데이터 설계 * 설계 근거 * ERD |

| **소개** | - **목적**  - 이 데이터베이스는 대화형 향수 추천 챗봇(ScentPick)이 수집한 향수/브랜드//메인 어코드/용량/가격 정보를 일관되게 저장하고, 사용자 대화 이력과 선호(즐겨찾기·피드백)를 바탕으로 퍼스널라이즈된 추천을 제공하도록 설계되었습니다.  - **범위**  - 사용자 계정/세션/대화 이력 관리   * 향수 메타데이터(브랜드, 라인, 부향률 등) * 노트/메인 어코드 등 매핑(탑/미들/베이스) * 외부 소스 수집(Fragrantica, 리테일러 가격/이미지 등) 및 출처 추적 * 벡터 임베딩(향 설명, 메인 어코드 등)과 RAG 검색 연동 * 사용자 즐겨찾기, 리뷰 |
| --- | --- |
| **시스템 개요** | - **시스템 역할**  **데이터 허브**: 스크래핑으로 수집한 향수·노트·가격·리뷰·이미지 원천을 정규화해 저장  **추천 근거 저장**: 대화 중 사용된 후보/점수/근거 문장(소스 링크) 보존 → 재현 가능성 확보  **개인화**: 사용자 즐겨찾기/리뷰/대화 로그 기반 개인화  - **주요 기능**   * 사용자 인증 및 로그인 시간 기록 * 향수/노트/어코드/브랜드/가격/이미지 등 **도메인 엔터티 관리** * 세션별 **대화 메시지** 저장 * **RAG 검색**을 위한 임베딩/소스 문서 관리 |
| **시스템 아키텍처** | - **데이터베이스 구조**  **관계형 DB (MySQL)**: 기준 마스터(향수/브랜드/어코드), 관계 테이블, 사용자/세션/대화/즐겨찾기/가격  **벡터 저장소 (Pinecone)** : perfumes 향 설명 텍스트, keyword 사전,네이버 리뷰 데이터 임베딩, 인덱스/버전 관리  **적용한 데이터베이스**  - AWS의 RDS(MySQL) 를 기반으로 주요 마스터/관계 데이터  (향수, 브랜드, 어코드, 사용자, 세션, 대화, 즐겨찾기, 가격 등)를 관리함.  - Pinecone 벡터 DB에는 다음 데이터를 임베딩하여 저장함:   * Perfumes: 향 설명 텍스트 임베딩 * Keyword 사전: 노트명·카테고리 매핑 사전 임베딩 * 네이버 리뷰 데이터: 크롤링한 사용자 리뷰를 Full text 그대로 임베딩.   - MySQL RDB에는 Pinecone 벡터의 ID만 참조하며, 원문 텍스트는 규격화 테이블과 함께 관리함. |
| **요구사항 매트릭스** | 요구사항 및 관련 테이블   | 요구사항 | 관련 테이블 | | --- | --- | | 사용자 로그인 정보 저장 | users, socialaccount | | 사용자별 대화내역 기록 | conversations, messages | | 향수/브랜드/노트/어코드 관리 | perfumes, note\_image | | 즐겨찾기, 사용자 피드백 리스트 관리 | favorites,feedback\_events | | 추천 근거 로그 관리 | rec\_runs, rec\_candidates | | 이미지/멀티모달 입력 관리 | messages.chat\_image | | 사용자 프로필 확장 정보 | user.profile\_image\_url | |
| **데이터 설계** | - 테이블 설명  - **users**  - 서비스 회원 기본 정보를 저장(auth\_users 확장)   | 컬럼명 | 데이터 타입 | 설명 | NN  (not null) | | --- | --- | --- | --- | | id | BIGINT | 사용자 ID (PK)  FK -> auth\_user.id | True | | name | VARCHAR(60) | 이름(Unique) |  | | gender | VARCHAR(10) | 성별 |  | | birth\_year | INT | 출생년도 |  | | created\_at | TIMESTAMP | 생성 시각 | True | | updated\_at | TIMESTAMP | 수정 시각 | True | | profile\_image\_url | VARCHAR(200) | 프로필 이미지 url |  |   - **conversations**  - 사용자의 대화목록 정보를 저장   | 컬럼명 | 데이터 타입 | 설명 | NN(not null) | | --- | --- | --- | --- | | id | BIGINT | 대화 ID (PK) | True | | user\_id | BIGINT | FK → auth\_user.id | True | | title | VARCHAR(40) | 사용자 대화 제목 |  | | started\_at | TIMESTAMP | 대화 시작 시각 |  | | upadated\_at | TIMESTAMP | 마지막 업데이트 시각 |  |   - **messages**  - 사용자의 대화내역 정보를 저장   | 컬럼명 | 데이터 타입 | 설명 | NN  (not null) | | --- | --- | --- | --- | | id | BIGINT | 메세지 ID (PK) | True | | conversation\_id | BIGINT | FK → conversations.id | True | | role | VARCHAR(10) | 화자 구분  (user/assistant) | True | | content | TEXT | 메시지 본문 | True | | created\_at | TIMESTAMP | 생성 시각 |  | | chat\_image | VARCHAR(500) | 사용자가 보낸 이미지 |  | | model | varchar | 사용한 gpt model |  |   - **favorites**  - 사용자의 즐겨찾기 내역 저장   | 컬럼명 | 데이터 타입 | 설명 | NN(not null) | | --- | --- | --- | --- | | user\_id | BIGINT | FK → auth\_users.id | True | | perfume\_id | BIGINT | FK → perfumes.id | True | | created\_at | TIMESTAMP | 즐겨찾기 등록 시각 | True |   - **rec\_runs**  - 추천 근거 로깅   | 컬럼명 | 데이터 타입 | 설명 | NN(not null) | | --- | --- | --- | --- | | id | BIGINT | 추천 근거 로깅  ID (PK) | True | | user\_id | BIGINT | FK → auth\_users.id | True | | conversation\_id | BIGINT | FK → conversation.id | True | | request\_msg\_id | BIGINT | FK →  messages.id | True | | query\_text | TEXT | 사용자 원문 질의 | True | | parsed\_slots | JSON | 파싱된 원문 슬롯 | [keyword 파싱] | | agent | TEXT | 사용한 추천 agent명 |  | | model\_version | TEXT | 사용한 모델/인덱스 버전 |  | | created\_at | TIMESTAMP | 생성성 시각 | True |   `  - **note\_images**  - note별 이미지   | 컬럼명 | 데이터 타입 | 설명 | NN(not null) | | --- | --- | --- | --- | | id | BIGINT | 노트 이미지 ID (PK) | True | | category | VARCHAR(50) | 노트가 속한 카테고리 |  | | note\_name | VARCHAR(50) | 노트 이름 |  | | image\_url | VARCHAR(500) | 노트의 이미지 url |  |   - **rec\_candidates**  - 해당 run에서 추천된 각 후보와 근거   | 컬럼명 | 데이터 타입 | 설명 | NN(not null) | | --- | --- | --- | --- | | id | BIGINT | 추천 후보/근거  ID (PK) | True | | run\_rec\_id | BIGINT | FK → rec\_runs.id | True | | perfume\_id | BIGINT | FK → perfumes.id | True | | rank | BIGINT | 해당 run에서의 정렬 순위 | True | | score | FLOAT | 최종 랭킹 점수 | True | | reason\_summary | TEXT | 짧은 추천 이유 문장 |  | | reason\_detail | JSON | 구조화된 근거 |  | | retrieved\_from | TEXT | 후보를 가져온 소스/전략 라벨 |  |       - **socialaccount\_socialaccount**  - 소셜 로그인을 통한 회원가입 정보들을 저장. (django allauth를 통해 생성됨)  - 필요 컬럼만 기술   | 컬럼명 | 데이터 타입 | 설명 | NN  (not null) | | --- | --- | --- | --- | | id | BIGINT | 사용자 ID(PK) | True | | provider | VARCHAR(200) | 소셜 로그인 플랫폼 | True | | uid | VARCHAR(128) | 플랫폼 측 사용자 ID | True | | extra\_data | JSON | 플렛폼에서 제공하는 추가 사용자 정보 |  | | user\_id | INT | FK  →auth\_user.id |  |     - **auth\_user**  - 사용자의 기본정보를 저장. (django allauth를 통해 생성됨)   | 컬럼명 | 데이터 타입 | 설명 | NN  (not null) | | --- | --- | --- | --- | | id | BIGINT | 사용자 ID(PK) | True | | username | VARCHAR(150) | 이름(Unique) | True | | password | VARCHAR(128) | 비밀번호 해시 | True | | email | VARCHAR(254) | 사용자 email | True |   - **feedback\_events**  - 사용자 좋아요/싫어요 이력   | 컬럼명 | 데이터 타입 | 설명 | NN(not null) | | --- | --- | --- | --- | | id | BIGINT | 피드백 이벤트  ID (PK) | True | | user\_id | BIGINT | FK → auth\_users.id | True | | perfume\_id | BIGINT | FK → perfumes.id | True | | rec\_candidate\_id | BIGINT | FK → rec\_candidate.id  특정 추천 리스트에서 노출된 후보와 연결 |  | | source | TEXT | 이벤트가 발생한 맥락 | True | | action | TEXT | 사용자의 좋아요/싫어요 피드백 동작 | True | | context | JSON | 이벤트 발생 시점의 부가 정보(날씨, 월) |  | | created\_at | TIMESTAMP | 생성 시각 |  |     - **perfumes**  - 향수와 관련된 정보들을 저장.   | 컬럼명 | 데이터 타입 | 설명 | NN(not null) | | --- | --- | --- | --- | | id | BIGINT | 향수 ID (PK) | True | | brand | VARCHAR(50) | 브랜드명 | True | | name | VARCHAR(50) | 향수명 | True | | sizes | JSON | 용량 (ml), 리스트 sizes: [30, 50] |  | | detail\_url | VARCHAR(500) | 상세 페이지 URL |  | | description | TEXT | 향수 설명 | True | | concentration | VARCHAR(30) | 부향률(EDP/EDT 등) | True | | gender | VARCHAR(10) | 마케팅/사용 대상으로 설정한 성별 구분(Male, Female, Unisex) | True | | main\_accords | JSON | 메인 어코드 | True | | top\_notes | JSON | 탑 노트 |  | | middle\_notes | JSON | 미들 노트 |  | | base\_notes | JSON | 베이스 노트 |  | | notes\_score | JSON dict | 향수에 해당되는 노트별 점수 |  | | season\_score | JSON dict | 향수의 계절별 연관도 |  | | day\_night  \_score | JSON dict | 향수의 주/야 연관도 |  | | created\_at | TIMESTAMP | 생성 시각 |  | | updated\_at | TIMESTAMP | 수정 시각 |  |   이외 django allauth에서 생성되는 table로 account\_emailadress, account\_emailconfirmation, auth\_group, auth\_group\_permissions, auth\_user\_groups, auth\_user\_user\_permissions, django\_content\_type, django\_migrations, django\_session, django\_site, socialaccount\_socialaccount, socialaccount\_socialapp, socialaccount\_socialapp\_sites, socialaccount\_socialtoken 이 있음. |
|  | - **데이터 무결성 보장**  - 외래 키(FK)를 활용하여 테이블 간 관계를 명확히 정의함.  - **확장성 고려**  - 대화·추천 로그 분리로 장기 보관/분석/ 확장 가능.  - 가격 정보 api로 가져옴으로써 가격 변동 및 상품 변화에 유연하게 대처  - **보안**  - 사용자 민감정보는 최소 수집·암호화, 소셜 로그인 지원. |
| **ERD** |  |