**수집 데이터**

**SKN Family AI Camp 14기 : 최종 프로젝트 2팀**

| **프로젝트 주제** | 사용자와의 대화(이미지, 텍스트)를 통해 향을 읽어내어 맞춤형 향수를 추천하는 RAG 기반 추천 챗봇 | | | | |
| --- | --- | --- | --- | --- | --- |
| **데이터 출처** | 향수 사이트  BYSUCO,  Fragrantica,  Huggingface,  향수사랑카페 | **데이터 형태** | 웹페이지 (크롤링),  csv 파일 | **데이터 정보 및 분량 (사이즈)** | 향수 802종,  허깅페이스 데이터셋(학습: 23442건, 검증: 2613건, 라벨: 17개) |
| **데이터 수집 계기** | 개인의 취향이 확고해지고, 선택의 범위가 광범위해졌으나, 소비자들은 향수 구매 과정에서 여전히 불편함을 겪고 있습니다. 예를 들면, 시향 가능한 매장으로의 접근성이 낮아 직접 향을 경험하기 어렵고, 매장에서 시향을 하더라도 향이 섞이거나 첫 향만으로 판단해 잔향을 충분히 확인하지 못하는 경우가 많습니다.  이러한 한계점은 특히 온라인 구매의 경우 지속력·발향력·성분·계절 적합성 등 구체적인 정보를 얻기 어렵다는 문제로 이어집니다.  또한 수많은 브랜드와 제품 중에서 자신의 취향에 맞는 향을 찾기 어렵고, 전문 용어 위주의 노트 설명은 일반 소비자가 이해하기 어려워 원하는 향이 실제로 어떤 향기인지 파악하기 힘든 실정입니다.  저희 팀은 이러한 문제를 해결하기 위해, LLM 및 멀티모달 챗봇을 통해 사용자가 원하는 향을 분석하고 맞춤형 향수를 추천하는 서비스를 제공하는 것을 목표로 합니다.  기대효과 측면에서 이 시스템은 개인의 취향과 라이프스타일을 반영한 AI 기반 향수 추천 및 분석 시스템으로서 화장품 등 뷰티 코스메틱 분야 전반으로의 확장도 가능할 것으로 보입니다. | | | | |
| **데이터 출처** | * BYSUCO 와 Fragrantica 웹사이트 : 향수명, 브랜드, 주요 노트(Top/Middle/Base), 평점, 리뷰, 주요 어코드 컬럼 등 정보 제공 * 허깅페이스 데이터셋, 향수사랑카페 리뷰 데이터 * <https://www.bysuco.com/> * <https://www.fragrantica.com/> * <https://huggingface.co/> * <https://cafe.naver.com/perfumelove/> | | | | |
| **원본 데이터** |  | | | | |
| **데이터 전처리 시 유의해야 할 점** | * 향수 이름, 향수 브랜드, 향수 종류, 향 설명, 용량, 메인 어코드, 탑/미들/베이스 노트 등 향수 관련 데이터를 DB에 저장하는 과정에서 데이터 정제가 필요. * 노트, 향/재료, 색상/이미지, 계절, 온도/습도, 날씨, 시간대, 장소/상황, 감정/분위기 등 키워드 사전 작성 * 기존 파일에 향수 이름 영어 번역 추가 (잘못 번역된 것 있어서 스크래핑 시 누락되어서 확인 필요) * 노트 이름 영어로 되어 있어서 한국어 번역 추가 * masterpiece, iconic 등의 광고 문구를 제거하고 향 설명과 메인 어코드만 남김 * 크롤링하는 과정에서 too many requests 창이 나타남, 캡챠(로봇이 아님을 체크) 창이 나타남 * 용량에서 ml 제거하기 * 품절이어서 크롤링 안된 제품들이 있음 * 가격 변동 문제 해결 위해 API 사용하기 * 크롤링 데이터에 부향률 용어가 상이해 용어 통일 | | | | |