

외국인 관광객을 위한 피부타입 기반 화장품 추천

Skin Type-Based Cosmetics Recommendation for Foreign Tourists

TEAM MEMBERS 김보성 · 박준석 · 최재성 · 한성민 · 김예원

K-BEAUTY

크롤링

Machine Learning

LLM

Dashboard

I. 문제 정의 & 데이터 파이프라인

문제 정의 및 배경

문제 — 외국인 관광객은 K-뷰티 매장에서 한국어 리뷰와 성분표에 막혀 자신에게 맞는 화장품을 찾기 어렵다.

해결책 — 피부타입 기반 추천과 LLM 다국어 설명으로 구매 장벽을 낮추고, 매장 응대 비용을 절감한다.

데이터 수집 및 전처리

수집 출처 — 올리브영(상세정보 + 리뷰), 식약처 화장품 성분사전 API

378개

수집 상품

161,211건

총 리뷰 (4개 카테고리)

스킨케어 · 마스크팩 · 클렌징 · 선크어

트러블슈팅

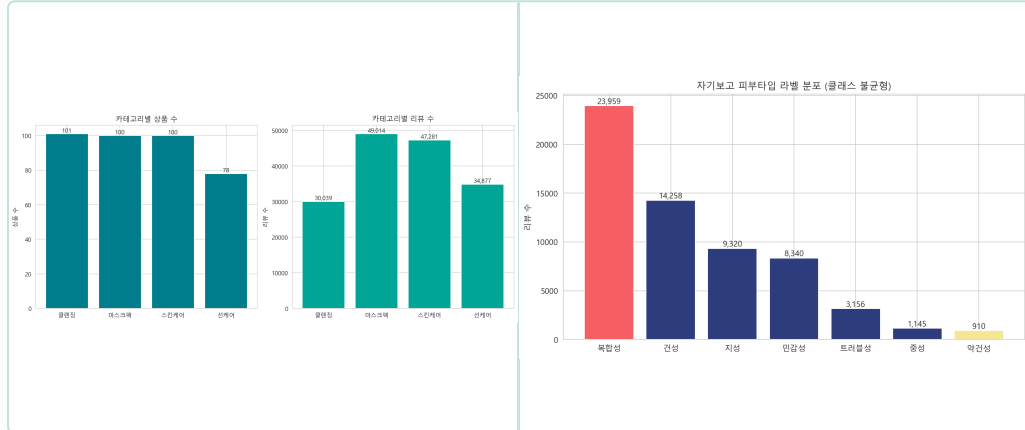
- Shadow DOM 재귀적 탐색 추출 적용 (클렌징)
- Selenium + 체크포인트 자동 저장으로 Cloudflare 403 / 429 차단 우회

전처리

가능성 성분 각주 제거 · 화학명 복원 · 피부타입 태그 정규화 완료

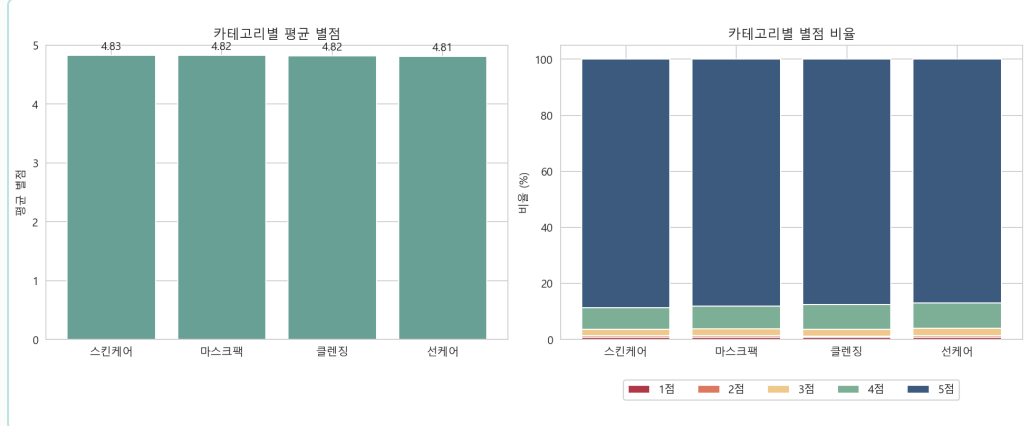
II. 데이터 탐색(EDA) & 주요 인사이트

Insight 1 — 수집 편차 & 클래스 불균형



복합성과 약간성 간 최대 28.6배의 극심한 클래스 불균형 존재.

Insight 2 — 평점의 변별력 상실



전체 리뷰의 87.9%가 5점 만점. 평점만으로는 상품 구별이 불가능해 성분 기반 접근이 필요.

Insight 3 — 주요 성분 분석



정제수 · 글리세린 등은 공통이나, 선크어의 자외선 차단 성분처럼 카테고리별 특화 성분 확인.

III. 머신러닝 · LLM 구현

머신러닝 모델링 — 피부타입 적합도 분류

피처 — 리뷰 텍스트(TF-IDF char n-gram) + 상위 150개 성분(위치 가중치)

0.36

F1-score (macro)

2.5×

기준선(0.14) 대비 향상

RandomForest, 84,111행 학습. **한계 극복** — 28.6배 클래스 불균형으로 인한 소수 클래스 재현율 저하를, 대시보드의 **베이지안 보정(Bayesian Smoothing)** 추천 로직으로 선회해 해결.

LLM 활용 & 다국어 대시보드

생성 — OpenAI gpt-4o-mini로 리뷰 요약 및 성분 기반 추천 이유 생성

5개 국어

한·영·일·중(간체)·베트남어

0건

환각 검증 오류 (140개 조합)

안정성 — 발표 환경 장애 대비, 140개 추천 조합의 다국어 결과를 사전 생성해 **오프라인 캐시 풀백** 구현. 환각 제어 프롬프트로 검증 오류 0건 달성.

시스템 파이프라인 (End-to-End)

