SK네트웍스 Family AI과정 6기

데이터 수집 및 저장 데이터 수집 보고서

▫ 개요

- 산출물 단계 : 데이터 수집 및 저장
- 평가 산출물 : 데이터 수집 보고서
- 제출 일자 : 2025.02.14
- 깃허브 경로 : https://github.com/SKNETWORKS-FAMILY-AICAMP/SKN06-FINAL-3Team.git
- 작성 팀원 : 박서윤, 박유나, 유경상, 장예린

1. 데이터 수집 목적

A. 내부 검색 시스템(sLLM) 구축을 위해 사업 분야, 브랜드 및 제품군, 투자 자료, 경영 전략, 기업 문화 데이터를 수집

B. 화장품 관련 정보 제공을 위해 화장품 성분, 관련 법규(화장품법), 산업 뉴스 및 저널, 경쟁사 제품, 학술 자료 수집

2. 데이터 수집 방법

A. 회사(아모레퍼시픽) 관련 데이터

a. 아모레몰 (자사 제품) 크롤링
   - 데이터 출처: 아모레몰 공식 웹사이트 (https://www.amoremall.com/)
   - 수집 방법: 웹 크롤링을 활용하여 제품 데이터를 JSON 형태로 저장
   - 저장 형식: JSON 파일

b. 아모레퍼시픽 뉴스 크롤링
   - 데이터 출처: 아모레몰 공식 웹사이트 내 뉴스 페이지 (https://www.apgroup.com/int/ko/news/news.html)
   - 수집 방법: 웹 크롤링을 활용하여 뉴스 데이터를 CSV형태로 저장
   - 저장 형식: CSV 파일

c. IR, FA 자료 다운로드
   - 데이터 출처: 아모레몰 공식 웹사이트 내 투자 페이지 (https://www.apgroup.com/int/ko/investors/investors.html#
   - 수집 방법: 제품 데이터 PDF 파일을 저장

- 저장 형식: PDF 파일

B. 화장품 및 성분(특정 도메인) 관련 데이터

a. 올리브영 (자사+타사 제품) 크롤링
   - 데이터 출처: 올리브영 공식 웹사이트 (https://www.oliveyoung.co.kr/)
   - 수집 방법: 웹 크롤링을 활용하여 제품 데이터를 JSON 형태로 저장
   - 저장 형식: JSON 파일

b. 식품의약품안전처 OPENAPI
   - 데이터 출처: 공공데이터포털 (https://www.data.go.kr/data/)
   - 수집 방법: OPENAPI를 활용하여 성분 데이터를 JSON 형태로 저장
   - 저장 형식: JSON 파일

c. 화장품 원료(성분) 정보 크롤링
   - 데이터 출처: ecogolik(https://ecogolik.com/ingredients/)
   - 수집 방법: 웹 크롤링을 활용하여 성분 데이터를 CSV 형태로 저장
   - 저장 형식: CSV 파일

C. 화장품 관련 뉴스/저널 데이터

a. 화장품 관련 뉴스 사이트 크롤링
   - 데이터 출처: 코스인(https://cosinkorea.com/news/article_list_all.html)
   - 수집 방법: 웹 크롤링을 활용하여 뉴스 데이터를 CSV형태로 저장
   - 저장 형식: CSV 파일

b. 잡지사(allure) 다운로드
   - 데이터 출처: 얼루어(https://www.allurekorea.com/)
   - 수집 방법: 웹 크롤링을 활용하여 뉴스 데이터를 CSV형태로 저장
   - 저장 형식: CSV 파일

c. 화장품 관련 학술지 다운로드
   - 데이터 출처: Science ON(https://scienceon.kisti.re.kr/main/mainForm.do)
   - 수집 방법: 대한화장품학회지에서 발표한 자료 내용 검토 후 선별하여 저장
   - 저장 형식: PDF 파일

3. 수집 데이터(요약)

- 정형데이터 위주로 표현하였습니다.

A. 아모레몰 크롤링 데이터

키(key)	설명	키(key)	설명

brand	브랜드 명	product_weight	내용물의 용량 또는 중량
product_id	아모레몰 내 고유 ID	skin_type	피부타입, 색상(호,번)
product_url	제품 URL	expiration_date	사용기한
product_img	제품 이미지 URL	usage_guide	사용방법
product_name	제품 명	product_ingredient	성분
product_price	제품 가격	precaution	사용 시 주의사항
product_info	제품 소개		

B. 아모레퍼시픽 사이트 내 뉴스 크롤링 데이터

키(key)	설명
title	뉴스 제목
Date	뉴스 등록일자
body	뉴스 본문
image_file	이미지 파일 내 파일명
News_URL	해당 뉴스 URL

C. 올리브영 크롤링 데이터

키(key)	설명	키(key)	설명
id	크롤링된 데이터에 부여한 고유 ID	prd_url	해당 제품의 상세 페이지 URL
img_url	제품 이미지(썸네일) URL	file_path	저장된 이미지(썸네일) 파일 경로
img_ext	이미지 파일 확장자 (예: .jpg, .png)	brand	제품 브랜드명
prd_name	제품명	price	정가 (숫자로 변환)
capacity	제품의 용량 또는 중량	specifications	제품 주요 사양

usage_instructions	사용 방법	manufacturer	제조업체 및 책임판매업자 정보
country_of_origin	제조국	ingredients	제품의 모든 성분 목록
fda_approval	기능성 화장품 여부 (식약처 심사 여부)	precautions	사용 시 주의사항
warranty_policy	품질 보증 기준	customer_service_number	소비자 상담 전화번호

D. 공공데이터포털 크롤링 데이터

키(key)	설명	키(key)	설명
INGR_KOR_NAME	한글 성분명	ORIGIN_MAJOR_KOR_NAME	기원 및 정의
INGR_ENG_NAME	영문 성분명	INGR_SYNONYM	이명
CAS_NO	CAS 번호		

E. 아모레퍼시픽 재무 / 회계 데이터 및 내부 보고서 PDF

F. 아모레퍼시픽 및 브랜드 정보(브랜드 스토리, 브랜드 컨셉 등) TXT

G. 화장품 법규 자료 PDF

H. 화장품 동향보고서 PDF

I. 화장품 시장보고서 CSV

J. 화장품 관련 저널 PDF