

데이터 수집 결과 보고서

1. 기본 정보

항목	내용
보고서 제목	선물 추천 챗봇 데이터 수집 결과 보고서
작성일	2025.05.07
프로젝트명	Senpick
수집 목적	선물 상품 및 사용자 리뷰 데이터 확보
활용 예정 모델	GPT-4o
데이터 활용 범위	선물 추천 챗봇 추천용 상품 수집

2. 수집 개요

항목	내용
수집 기간	2025.04.25 ~ 2025.05.07
수집 대상	카카오, 네이버, SSG등의 쇼핑몰에서 카테고리 별로 수집한 상품 정보와 리뷰 데이터
수집 방식	Selenium 기반 웹 크롤링
저장 형식	JSON

3. 수집 데이터 요약

항목	수량
전체 수집 상품 수	17856개
리뷰 수 총합	833347개

누락/결측 데이터	3114건(리뷰데이터 없음)
데이터 용량	217MB

4. 데이터 형식

필드명	데이터 타입	설명
brand_name	문자열 (string)	상품 브랜드명. 예: "메르베"
product_name	문자열 (string)	상품명 또는 제품 상세 이름. 예: "메르베 신생아니삭스..."
category_main	문자열 (string)	대분류 카테고리. 예: "유아동"
category_sub	문자열 (string)	중분류 또는 소분류 카테고리. 예: "액세서리"
options	배열 (array)	상품 옵션(사이즈, 색상 등). 옵션이 없으면 빈 배열 []
product_url	문자열 (string)	상품 상세 페이지 URL
thumbnail_url	문자열 (string)	상품 썸네일 이미지 URL
review_content	배열 (array)	사용자 리뷰 내용 모음. 각 항목은 문자열로 구성됨
price	문자열 (string)	상품 가격. 숫자형으로 변환 가능하며, "4,000" 형태의 문자열로 수집됨

4.JSON 양식 예시

```
{
  "brand": "폴로 랄프 로렌(의류)",
  "product_name": "[단독]코튼 치노 베이스볼 캡 (18 color)",
  "category_main": "패션",
  "category_sub": "브랜드의류",
  "options": [
    "1. 텐 (MAPOHGS0J421242250)",
    "2. 블랙/레드 포니 (MAPOHGS0J421246001)",
    "3. 화이트 (MAPOHGS0J421245100)",
    "5. 뉴포트 네이비/레드 포니 (MAPOHGS0J421237999)",
    "8. 뉴 포레스트 (MAPOHGS0J422104300)",
    "9. 그레이 (MAPOHGS0J420198020)",
    "15. 미네랄 다이 올리브 (MAPOHGS0J422080301)",
    "16. 미네랄 다이 그린스톤 (MAPOHGS0J422079330)",
    "17. 미네랄 다이 로즈 (MAPOHGS0J422084650)",
    "[단독] 18. 퍼플 캐릭터스 (MAPOHGS0J421810500)"
  ],
  "product_url": "https://gift.kakao.com/product/1100651",
  "thumbnail_url":
    "https://img1.kakaocdn.net/thumb/C305x305@2x.q82/?fname=https%3A%2F%2Fst.kak
    aocdn.net%2Fproduct%2Fgift%2Fproduct%2F20200513164955_ed7b9e3918dd412ea73b8c
    55b6a1e4ef.png",
  "review_content": [
    "리뷰1",
    "리뷰2",
    "리뷰3"
  ],
  "price": "89,000원"
}
```