

**SK네트웍스 Family AI과정 13기  
 데이터 수집 및 저장 수집 데이터 보고서**



| **산출물 단계** | 데이터 수집 및 저장 |
| --- | --- |
| **평가 산출물** | 수집 데이터 보고서 |
| **제출 일자** | 2025-07-30 |
| **깃허브 경로** | <https://github.com/SKNETWORKS-FAMILY-AICAMP/SKN13-FINAL-2TEAM> |
| **작성 팀원** | 김지민 |

| **데이터명** | **수집 대상** | **수집 목적** | **사용 예정 기능** | **출처/저작권** |
| --- | --- | --- | --- | --- |
| W컨셉 상품 데이터 | W컨셉 남성, 여성  의류 카테고리  내 상품 상세 페이지 | 플랫폼 특성 분석  및 상품 데이터 확보 | 가격 분포 및  브랜드별 상품 수 분석 | W컨셉 |
| 29CM 상품 데이터 | 29CM 웹사이트의  여성 상의 카테고리  내 상품 상세 페이지 | 감성 셀렉트샵의 상품  큐레이션 및 트렌드 분석 | 리뷰 데이터 기반 상품 평가 모델링 | 29cm |
| 무신사 상품 데이터 | 무신사 웹사이트의  상의, 하의 카테고리 내 상품 상세 페이지 | 국내 최대 패션 플랫폼의  시장 점유율 및 특성 분석 | 상세 데이터를 활용한 사용자 맞춤 추천 시스템 연구 | 무신사 |

1. **수집 데이터 개요**
2. **수집 방법 및 자동화 절차**

각 사이트의 페이지 로딩 방식과 구조에 맞춰 최적화된 수집 절차를 설계했습니다.

**2-1. W컨셉 수집 절차**

W컨셉은 목록 페이지에서 스크롤을 통해 동적으로 상품을 로드 하므로, **'링크 선(先)수집, HTML 후(後)수집'**의 2단계 전략을 사용했습니다.

* **1단계 (상품 링크 대량 수집):** Selenium을 활용해 각 카테고리 페이지의 모든 상품 정보가 로딩되도록 **자동으로 스크롤을 끝까지 내리고**, 페이지 내 모든 상품의 고유 링크를 수집하여 CSV 파일로 1차 저장했습니다.
* **2단계 (상세 페이지 HTML 수집):** 1단계에서 저장한 CSV 파일을 읽어, 상품 링크 목록을 순회하며 Selenium으로 각 링크에 개별적으로 접속해 최종 렌더링된 HTML 소스를 파일로 저장했습니다.

**2-2. 29CM 수집 절차**

29CM는 페이지 이동과 동적 로드가 복합적으로 이루어져, 각 상품 페이지를 직접 방문하고 돌아오는 순회 전략을 채택했습니다.

* **페이지 순회 및 상품 접근:** Selenium을 사용하여 지정된 카테고리의 목록 페이지 URL에 page={page\_num} 파라미터를 변경하며 순차적으로 접근합니다.
* **HTML 수집 및 복귀:** 페이지 끝까지 자동 스크롤하여 로드된 상품 목록에서 각 상품 링크를 클릭하여 상세 페이지로 이동한 후, 로딩이 완료되면 driver.page\_source를 저장하고 driver.back() 명령어로 목록 페이지로 복귀하여 다음 상품 수집을 이어갑니다.

**2-3. 무신사 수집 절차**

무신사는 상품 ID 기반으로 URL 접근이 가능하고, 상세 정보가 탭(아코디언 메뉴) 내에 동적으로 로드되는 특징이 있어 다음과 같이 수집을 진행했습니다.

* **유효 상품 필터링 및 접근:** 상품 ID를 1씩 증가시키는 while 루프를 사용하고, requests.get()으로 **HTTP 상태 코드를 먼저 확인**하여 404 에러 페이지는 사전에 필터링했습니다.
* **동적 컨텐츠 로딩 및 HTML 저장:** Selenium(headless 모드)으로 접속 후, **'상품 고시 정보안내'와 같은 모든 탭 버튼을 찾아 클릭**하여 JavaScript로 숨겨진 상세 정보까지 모두 로드된 최종 driver.page\_source를 저장했습니다.

1. **데이터 설명 및 구성**

향후 각 플랫폼별 데이터의 독립적인 업데이트와 유지보수 편의성을 위해, 데이터는 플랫폼별로 분리된 CSV 파일로 저장하는 방식을 채택했습니다.

**3-1. W컨셉 데이터 구성**

| 파일명 | 필드명 | 데이터 타입 | 설명 | 예시 |
| --- | --- | --- | --- | --- |
| wconcept \_products.csv | 상품명 | string | 상품의 전체 이름 | 페플 패키지 하와유 락스타 맨투맨 4종 2PACK JDMT1435 |
| 링크 | string | 상품 상세 페이지 URL | https://www.wconcept.co.kr/Product/303698473 |
| 브랜드 | string | 상품의 브랜드명 | FP |
| 판매자 | string | 상품 판매자 (브랜드와 동일) | FP |
| 성별 | string | 상품 카테고리의 성별 구분 | 여성 |
| 제품 대분류 | string | 1차 카테고리 | 티셔츠 |
| 제품 소분류 | string | 2차 카테고리 | 스웻 |
| 원래가격 | integer | 할인이 적용되지 않은 정가 | 93600 |
| 할인가 | integer | 할인이 적용된 최종 판매가 | 59800 |
| 좋아요 숫자 | integer | 상품의 ‘좋아요’ 수 | 1 |
| 이미지URL | string | 상품 대표 이미지 원본 URL | https://product-image.wconcept.co.kr/.../303698473\_GO52148.jpg |
| 이미지 파일 경로 | string | 로컬에 저장된 이미지 파일의 경로 | D:/T\_shirts\_W3/product\_images/303698473.jpg |
| 사이즈 | string | 상품 옵션 사이즈 | (정보 없음) |
| 총장 | string | 실측 정보 : 총 기장 | (정보 없음) |
| 어깨 너비 | string | 실측 정보 : 어깨 너비 | (정보 없음) |
| 가슴 단면 | string | 실측 정보 : 가슴 단면 | (정보 없음) |
| 소매 길이 | string | 실측 정보 : 소매 길이 | (정보 없음) |
| 제품 소재 | string | 상품의 소재 정보 | cotton 72%, Polyester 28% |
| 색상 | string | 상품의 색상 정보 | 상품상세설명 참조 |
| 세탁방법 | string | 상품 세탁 및 관리 방법 | 세탁방법 및 취급시 주의사항 1.물에 담궈 두지 마시고... |
| 상품 코드 | string | 상품 정보 고시 내 상품코드 | 303698473 |
| 치수 | string | 상품 정보 고시 내 치수 정보 | 상품상세설명 참조 |
| 상품 ID | string | W컨셉 플랫폼 내 상품 고유 ID | 303698473 |

**3**

**3-2. 29CM 데이터 구성**

| 파일명 | 필드명 | 데이터 타입 | 설명 | 예시 |
| --- | --- | --- | --- | --- |
| 29cm \_products.csv | 파일명 | string | 원본 HTML 파일의 이름 | ...Flower\_Basket\_T-Shirt\_Grey.html |
| 제품이름 | string | <meta property='og:title'> 태그에서 추출 | Flower Basket T-Shirt\_Grey |
| 제품대분류 | string | 제품소분류를 기반으로 '상의' 등으로 자동 분류 | 상의 |
| 제품소분류 | string | 페이지 상단 카테고리(<span>) 태그에서 추출 | 반소매 티셔츠 |
| 원가 | integer | 취소선(line-through) 태그에서 숫자만 추출 | 75000 |
| 할인가 | integer | 상품 최종 판매가(pdp\_product\_price ID)에서 숫자 추출 | 75000 |
| 좋아요 숫자 | integer | '좋아요' 버튼 텍스트에서 숫자만 추출 | 22270 |
| 브랜드 | string | <meta name='keywords'> 태그의 첫 단어 | 더바넷 |
| 택배사 | string | 수집되지 않아 None으로 저장됨 | (정보없음) |
| 상품코드 | string | HTML 전체 텍스트에서 '상품번호' 검색 후 추출 | 3138584 |
| 제품소재 | string | 본문의 소재 관련 키워드(cotton 등) 문장 조합 | Cotton 100% |
| 치수 | string | 본문의 사이즈 관련 키워드(cm, Length 등) 문장 조합 | S: 총장 65 어깨 48 가슴 54 .. |
| 세탁방법 | string | 본문의 세탁 관련 키워드(wash, clean 등) 문장 조합 | Dry clean only Do not wash .. |
| 평점 (전체평균) | float | 채워진 별 아이콘 개수를 세어 평점으로 변환 | 5 |
| 리뷰 개수 | integer | '리뷰 보기' 버튼 텍스트에서 숫자만 추출 | 33 |
| 대표 이미지 | string | 상품 상세설명의 대표 이미지 URL 목록 | https://img.29cm.co.kr/item/... |
| 정보 이미지 | string | 상품 상세설명의 마지막 이미지 URL | https://img.29cm.co.kr/item/... |

**3-3. 무신사 데이터 구성**

무신사는 리뷰 데이터의 양이 방대하고 구조화가 잘 되어 있어, 상품 정보와 리뷰 정보를 별도의 파일로 분리하여 저장했습니다.

[상품 정보]

| 파일명 | 필드명 | 데이터 타입 | 설명 | 예시 |
| --- | --- | --- | --- | --- |
| musinsa \_products.csv | 제품번호 | integer | 상품 고유 ID | 600" |
| 제품이름 | string | 상품명 | "FTC All City Tee Black" |
| 브랜드 | string | 브랜드명 | "에프티씨" |
| 원가 | integer | 정상가 | 49000 |
| 할인가 | integer | 판매가 | 47530 |
| 총 리뷰 개수 | integer | 리뷰 총 개수 | 0 |
| 평균 평점 | float | 리뷰 평균 평점 | 0 |
| 제품 소재 | string | 제품 소재 정보 | "Cotton 100%" |
| 제품 대분류 | string | 제품의 큰 단위의 분류(상의, 하의) | 상의 |
| 제품 소분류 | string | 큰분류에서 한번 더 분류(티셔츠, 청바지) | 청바지 |
| 사진 | string | 이미지url | https:/images/goods\_img/20110527/10011/10011\_1\_500.jpg |
| 성별 | string | 상품 카테고리의 성별 구분 | 남자 |
| 좋아요 수 | string | 사용자들의 좋아요 수 | 4천 |
| 세탁 방법 | string | 상품의 세탁방법 | 세탁시 이염이 될 수 있으니 반드시 단독세탁 하시기 바랍니다. 뜨거운물 세탁시에는 원단 변형이 있을 수 있으니 삼가하시기 바랍니다. 보관시 직사광선과 습한곳을 피하시기 바랍니다. |
| 제품 URL | string | 쇼핑몰의 상품 url | https://www.musinsa.com/products/5253292 |
| 상품 코드 | string | 제품의 고유번호 | 6329501 |
| 실측사이즈정보 | JSON(string) | 실측 사이즈 테이블 정보 | '[{"사이즈명":"M",총장":"72", ...}]' |

[리뷰 정보]

| 파일명 | 필드명 | 데이터 타입 | 설명 | 예시 |
| --- | --- | --- | --- | --- |
| musinsa \_reviews.csv | 제품번호 | integer | 상품 고유 ID | 600 |
| 작성자명 | string | 리뷰 작성자 닉네임 | “stylelover” |
| 리뷰별점 | float | 사용자가 부여한 별점 | 5 |
| 작성자정보 | string | 작성자의 신체 정보 | “180cm, 75kg” |
| 구매옵션 | string | 작성자가 구매한 상품옵션 | “블랙/L” |
| 리뷰내용 | string | 리뷰 본문 텍스트 | “재질이 좋고 핏이 예쁩니다.” |

1. **법적·윤리적 검토**

* **개인정보 포함 여부:** **[✔] 미포함**  
  (무신사 리뷰 데이터의 '작성자명'은 비식별화된 닉네임이므로 개인정보로 간주하지 않음)
* **출처 및 사용권:**
  + **공개 여부:** **[✔] 내부사용 한정**
  + **라이선스 또는 약관 검토 여부:** 각 웹사이트의 robots.txt를 확인하여 수집이 금지된 페이지는 제외하였으며, 수집된 데이터는 상업적 용도가 아닌 분석 목적으로만 활용할 계획입니다.

1. **데이터 품질 및 정합성 관리 방안 (사이트별 상세)**

각 플랫폼의 HTML 구조와 데이터 제공 방식이 달라 사이트별로 최적화된 전처리 전략을 적용했습니다.

**5.1 W컨셉(W Concept) 데이터 전처리**

* **데이터 추출 방식:** BeautifulSoup으로 HTML을 파싱한 후, <script> 태그 내 ep\_page\_brandNM(브랜드명), prodid(상품ID) 등 정보는 정규표현식(re)으로 추출하고, 가격 정보는 div.price\_wrap 클래스 하위 태그를 CSS Selector로 특정하여 수집했습니다.
* **데이터 정제 및 표준화:** <title> 태그에서 추출한 상품명에 포함된 [] 괄호를 제거하고, 가격 필드에서 '원', ',' 문자를 제거 후 integer 타입으로 변환했습니다.

**5.2 29CM 데이터 전처리 (예시)**

* **데이터 추출 방식:** #pdp\_product\_name(상품명), #pdp\_product\_price(가격) 등 고유 id 값을 가진 요소를 중심으로 데이터를 추출하고, '상품정보 보기' 테이블에서 '제품 소재' 정보를 추출하는 로직을 설계했습니다.
* **데이터 정제 및 표준화:** '33개 리뷰 보기'와 같은 텍스트에서 정규표현식으로 숫자만 추출 후 integer로 변환하고, 가격 필드도 숫자형으로 변환했습니다.

**5.3 무신사(MUSINSA) 데이터 전처리**

* **데이터 추출 방식:** <script id="pdp-data"> 태그 안의 JSON 데이터를 직접 파싱하는 방식을 주력으로 사용하여 대부분의 정보를 안정적으로 확보했습니다. 제품 소재, 실측 사이즈 등 JSON에 없는 정보는 HTML 테이블을 직접 파싱하여 보충했습니다.
* **데이터 정제 및 표준화:** HTML 테이블로 된 실측 사이즈 정보를 파싱하여 사이즈별 측정치를 JSON 형태의 문자열로 변환 후 단일 컬럼에 저장했습니다. 리뷰 텍스트의 줄바꿈 문자(\n, \r)를 공백으로 치환하여 데이터 일관성을 확보했습니다.

1. **변경 이력 및 보완 내역**

| 변경일 | 변경자 | 변경내용 | 비고 |
| --- | --- | --- | --- |
|  |  |  |  |
|  |  |  |  |