sb

**SK네트웍스 Family AI 과정 15기  
 데이터 수집 및 저장 수집 데이터 보고서**



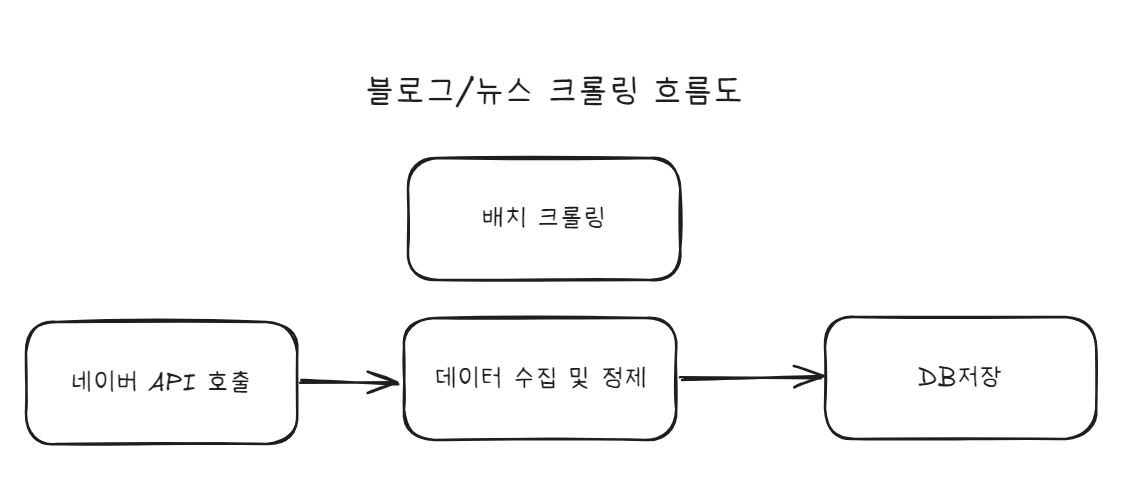
| **산출물 단계** | 데이터 수집 및 저장 |
| --- | --- |
| **평가 산출물** | 수집 데이터 보고서 |
| **제출 일자** | 2025년 10월 2일 |
| **깃허브 경로** | <https://github.com/SKNETWORKS-FAMILY-AICAMP/SKN15-FINAL-4TEAM> |
| **작성 팀원** | 권주연,강민정,기현택,임경원,정민철 |

1. **수집 데이터 개요**

| **데이터명** | **수집 대상** | **수집 목적** | **사용 예정 기능** | **출처/저작권** |
| --- | --- | --- | --- | --- |
| 한샘 가구 카테고리 | 인테리어 디자이너 | 사용자 추천 모델 학습 데이터 | 모델링 | 한샘 |
| 인테리어 카테고리 | 인테리어 디자이너 | 모델 학습 데이터 | 모델링 | 한샘 |
| 인테리어 시공 사례 | 인테리어 디자이너 | 모델 학습 데이터 | 모델링 | 한샘 |
| 인테리어 학술 논문 | 인테리어 디자이너 | 모델 학습 데이터 | 모델링 | RISS |
| 한샘 공식 블로그 | 인테리어  디자이너 | 참고자료 | 레이아웃용 데이터 | 한샘 네이버 블로그 |
| 인플루언서 블로그 | 인테리어  디자이너 | 참고자료 | 레이아웃용 데이터 | 인플루언서 네이버 블로그 |
| 한샘 관련 뉴스 | 인테리어  디자이너 | 참고자료 | 레이아웃용 데이터 | 네이버/언론사 |

1. **수집 방법 및 자동화 절차**

* 수집 방식  
  + 웹 크롤링
  + API 호출
* 수집 도구 또는 스크립트 설명:  
  + Python + BeautifulSoup, requests, Selenium
  + 자동화 여부 및 주기: 자동화 O (블로그/뉴스, 새벽마다)
  + 오류 발생 시 예외 처리 전략: 배치 성공 여부 DB 저장 → 실패 시, 그 시점부터 재수집 or 전날 수집
* 예시 스크립트 또는 흐름도 첨부: (이미지, 순서도 또는 코드)

****

1. **데이터 설명 및 구성**

**3.1 파일 및 필드 설명**

| **파일명 또는 테이블명** | **필드명** | **데이터 타입** | **설명** | **예시** |
| --- | --- | --- | --- | --- |
| Han\_fur.csv | goods\_id | int | 상품번호 | 114578 |
| big\_cat | int | 카테고리 대분류 | 1 |
| cat1,2 | int | 카테고리 중분류 | 101 |
| goods\_name | string | 상품이름 | 시그니처 리노아 붙박이장 |
| price | bigint | 상품가격 | 2449100 |
| image\_url | string | 상품 이미지 링크 | https://image.hanssem |
| Han\_int.csv | goods\_id | int | 상품번호 | 1064295 |
| big\_cat | int | 카테고리 대분류 | 1 |
| cat1,2 | int | 카테고리 중분류 | 101 |
| goods\_name | string | 상품이름 | 유니크 프리미엄 |
| price | string | 상품가격 | 851000 or  견적상담 |
| image\_url | string | 인테리어 이미지 링크 | https://image.hanssem |
| goods\_exp | string | 인테리어  부가 설명 | [현관]펜트리… |
| News, Blog.json | url | string | 본문 링크 | https://n.news.naver… |
| title | string | 제목 | 가구 1위 되찾은 한샘… |
| date | timestamp | 작성 시각 | **2025-09-29 16:45:54** |
| Thesis.csv | title | string | 논문 제목 | 기업의 경쟁… |
| database | string | 논문 출처 | RISS |
| author | string | 논문 저자 | 이OO, OO대학교 |
| journal | string | 논문 학술지 | 대한수학회보 |
| link | string | 논문 링크 | /search/detail/Detail |
| keyword | string | 논문 주요  키워드 | 주거환경 |
| collected\_date | timestamp | 논문 수집 시각 | 2025-09-25 16:42:44 |
| category | string | 논문 카테고리 | 환경 & 주거 |
| seq\_master.csv | seq | int | 사례 고유번호 | 10243 |
| title | string | 사례 제목 | 고급스럽고 세련된… |
| style | string | 적용 스타일 | 4천만원대 |
| space | string | 시공 대상 공간 | 모던 |
| taglist | string | 관련 태그 | 거실확장 |
| seq\_img.csv | seq | int | 사례 고유번호 | 10065 |
| text | string | 사례의 상세설명 | 입구우측 가장끝방입니다. |
| img | string | 사례 이미지 | [https://image…](about:blank)  jpg |
| seq\_list.csv | seq | int | 사례  고유번호 | 10054 |

**3.2 데이터 양**

* 전체 수집 데이터 건수:
  + 가구(4,612개)
  + 인테리어(342개)
  + 시공사례 (11,000여개 + 매핑 23만개)
  + 블로그 (700여개)
  + 뉴스 (200여개)
* 추출된 고품질 데이터 건수 (필터링 후 기준):
  + 가구(4,501개)
  + 인테리어(342개)
  + 시공사례 (11,000여개 + 매핑 23만개)
  + 블로그 (700여개)
  + 블로그 (200여개)

**3.3 저장 위치 및 포맷**

* 저장 경로: AWS S3
* 저장 포맷: CSV / JSON
* 인코딩: UTF-8

1. 법적·윤리적 검토

* 개인정보 포함 여부: X
* 비식별화 조치 여부: X

* 출처 및 사용권:  
  + 공개 여부: [ ] 공개
* 검토자 및 검토 일자: 미정

**4. 데이터 품질 및 정합성 관리 방안**

* 중복 제거 기준: 카테고리별 라벨링 유지, 중복 제거 대신 각 카테고리별로 다중 라벨링 처리
* 정합성 검증 방법:  
  + 스타일이 ‘기타’인 경우 삭제
* Null 처리 및 결측치 전략 : 견적비용이 null 값일 경우 관련 데이터 모두 삭제(약 5500개)
* 표준화 전략:  
  + 텍스트 전처리 (특수문자, \n 제거, &quot 등 제거)

**5. 변경 이력 및 보완 내역**

| **변경일** | **변경자** | **변경 내용** | **비고** |
| --- | --- | --- | --- |
| 25.10.01 | 전원 | 초안작성 | Ver 1.0 |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |