SK네트웍스 **Family AI**과정 **13**기

데이터 수집 및 저장 수집 데이터 보고서

산출물 단계	데이터 수집 및 저장
평가 산출물	수집 데이터 보고서
제출 일자	2025.08.01
깃허브 경로	https://github.com/SKNETWORKS-FAMILY-AICAMP/SKN13-FINAL-3TEAM.git
작성 팀원	기원준, 전진혁, 강지윤, 최호연, 우민규

1. 수집 데이터 개요

대분류	중분류	데이터 종류	데이터 설명	데이터 출처/저작권
sLLM 챗봇 학습 데이터	텍스트 데이터	디자인 가이드 라인 및 규정	브랜드 아이덴티티, 디자인 철학	- Global Autonews (http://www.global-autonews.com/home.php) - 현대 모터스 (https://www.hyundai.com/kr/ko/e) -조택연, 황영지 , 양지제. (2025). 현대자동차 디자인 철학에 내재하는 미의식의 신경학적 해석. 한국자동차공학회논문집, **33**(3), 201-210.
		기술 문서	각 차량 모델의 기술 사양, 공학적 요소 (각 디자인 요소와 관련된 공기역학 계수, 중량, 재료 강도, 생산성 지표 등)	- Global Autonews (http://www.global-autonews.com/home.php)
		자동차 산업 뉴스 및 분석 리포트	시장 동향, 경쟁사 분석, 신차 출시 정보.	- Global Autonews (http://www.global-autonews.com/home.php)
		자동차 전문 포럼 및 리뷰	사용자 경험, 선호도, 불만 사항, 디자인에 대한 전문가 관점. 역대 현대자동차별 발전사 정보	Global Autonews (http://www.global-autonews.com/home.php)

		학술 논문 및 연구 자료	특정 모델에 대한(혹은 역대) 디자인 이론, 공학적인 요소.	- 서상열, 심준엽, 최태현 (2010). 자동차 개발단계에서의 인간공학의 역할. 대한인간공학회지, **29**(1), 7-16. - 김정민(2020). 자동차 차체 형태 디자인이 공기역학 성능에 미치는 영향에 대한 연구. 문화기술의 융합, **6**(1), 501-506. - 권재수, 이준하, 홍성준, 조희욱, 홍병권, 홍존희(2011). 학술 논문 : 차량 개발 단계의 친환경설계 프로세스에 대한 연구. 한국전과정평가학회지, **12**(1), 84-91.
	이미지 데이터	고품질 자동차 렌더링 이미지	다양한 각도, 조명, 배경, 컨셉의 렌더링 이미지(실사 x)	- Alibaba Cloud (https://tianchi.aliyun.com/dataset/98063)
		역대 생산 차량 실사 이미지	다양한 모델, 연식, 트림의 실제 차량 사진.	- 다나와자동차 (https://auto.danawa.com/auto/?Work=brand&Brand=303)
		브랜드 아이덴티티 시각 자료	로고, 색상 팔레트, 폰트 등 시각적 요소.	- 현대 모터스 (https://www.hyundai.com/kr/ko/e)
		현대 컨셉카	차별화 디자인의 프로토타입 뽑아내기위해 이런 디자인 넣기	-NetCarshow (https://www.netcarshow.com/)
	영상 데이터	자동차의 실제 주행 비디오	통일된 현대자동차 모델의 실제 주행 영상	- Youtube https://youtu.be/wwGJ6yMVdOI?si=RzRUL5q6WiOgQsFF https://youtu.be/7uIriNxxGmw?si=j8dGiUOIfPkIAkM_ https://youtu.be/z73biD_V-kU?si=FMhGJqHhNTnkhHCx https://youtu.be/z73biD_V-kU?si=9Egqa-fAe7qnuYgh 이외 10개 영상 추출)
RAG VectorDB 데이터	텍스트 데이터	역대 디자인 프로젝트 문서	컨셉 보고서, 성과 보고서.	- 현대 모터스 (https://www.hyundai.com/kr/ko/e) - Global Autonews (http://www.global-autonews.com/home.php)

		사용자 피드백 데이터	디자이너가 생성된 이미지/답변에 대해 제공하는 정성적/정량적 피드백. (예: "이 디자인은 너무 공격적이다", "측면 라인이 아쉽다", "이 재료는 규정에 맞지 않는다" 등)	- 현대자동차 구매닷컴(https://www.hyundai.com/kr/ko/purchase-event/vehicles-review)
		공학적 요소	디자인 요소와 관련된 공기역학 계수, 중량, 재료 강도, 생산성 지표 등.	- 현대자동차 공홈 https://www.hyundai.com/kr/ko/brand/brandstory/design.html - 현대 자동차 공홈(https://www.hyundai.co.kr/story/CONT0000000000005130) -디자인방향성(https://www.hyundai.com/worldwide/ko/company/innovation/design) - 자동차 인테리어 설계 원리(https://rkd991108.com/entry/자동차-인테리어-설계-원리-인체공학-UX디자인-공간) - 차체 및 구조의 연관성(https://chascow.tistory.com/22) - 차량 제원 (다나와 자동차 현대 / 제네시스) (https://auto.danawa.com/auto/?Work=brand&Brand=303 / 304)

2. 수집 방법 및 자동화 절차

- 수집 방식 (해당 항목에 체크) → 크롤링 시 특이사항 / 구체적으로 어떤 함수들이 어떤 역할
    - ✅웹 크롤링
    - ✅문서 파일 업로드 (PDF, CSV 등)

- 수집 도구 또는 스크립트 설명:
    - 사용한 언어/라이브러리: (예: Python + BeautifulSoup, requests 등) Python + BeautifulSoup, Selenium, requests, Pandas, / 전처리) remgb, PIL, json

- 예시 스크립트 또는 흐름도 첨부: (이미지, 순서도 또는 코드)

3. 데이터 설명 및 구성

3.1 파일 및 필드 설명

- 텍스트 관련 데이터 **(**내용이 어떻게 구성되어있는지에 대한 필드명과 타입을 각각 설명 **ex) txt**데이터는 **ID-title-content** 라면 각 필드에 대한 이름과 타입 설명**)**

파일명 또는 테이블명	필드명	데이터 타입	설명 **(**용도 및 필드에 대한 설명 추가**)**	예시
News_articles.txt Interview_articles.txt preview_articles.txt	asset_library	text	자동차 뉴스 분석 차량 관련 기사	현대자동차는 21일(현지 시각) 미국 로스앤젤레스 골드스테인 하우스에서 대형 전동화 SUV '아이오닉 9'을 세계 최초로 공개했다…
- 현대 모터스튜디오_디자인 관련 문서.pdf - 현대 디자인 모토.txt - 현대자동차 디자인 철학에 내재하는 미의식의 신경학적 해석.pdf	design_material	text	자동차 디자인 관련 문서	…이를 위해 현대자동차는 '센슈어스 스포티니스'라는 동일한 디자인 철학 아래, 각 차량의 독창성 역시 존중하는 디자인 방식을 추구해 나갈 계획입니다. …
- Car_specs.zip - 자동차 차체 형태 디자인이 공기역학 성능에 미치는 영향에 대한 연구.pdf - 자동차 개발단계에서의 인간공학의 역할.pdf	engineering_spec	varchar	디자인에 반영될 수 있는 공학적 요소	전장,"5,135 mm" 전폭,"1,925 mm"
	sales_stat			

hyundai_car_reviews.json	user_review	text	현대 자동차 고객 후기	{"review": "기존 투싼과의 9년 동행을 마무리하며 새로운 차량을 고민하던 중에…}
chat_logs.csv	prompt_text	string	사용자 질문	“오늘 날씨 어때?” “좋아~”
	timestamp	datetime	수집 시간	2025-05-10 15:01:22
generated_result	text_result	text	AI가 생성한 결과	{“ai_response” : 답변에 관한 생성 결과로..}

- 이미지 및 영상 관련 데이터

파일명 또는 테이블명	필드명	데이터 타입	설명	예시
hyundai_car_history_img.zip hyundai_images_nobg.zip	car_images	images	역대 생산 차량 실사 이미지	(이미지 파일)
Hyundai Motor Company Identity Design Guide Book.pdf	brand_identity	text+images	브랜드 아이덴티티 시각 자료	(이미지 파일)
/hyundai_concept_car_nogb.zip	concept_car_images	images	현대자동차 컨셉카	(이미지 파일)

3.2 데이터 양

- 전체 수집 데이터 건수:
  이미지 데이터) 현대 자동차 (2020-2025) 이미지 1988개 + 역대 컨셉차 이미지 56개 **(17.0GB)**
  텍스트 데이터) 자동차 디자인 관련 논문 4개, 자동차 디자인 관련 문서 4개, 현대 자동차 관련 기사 12개, 리뷰 관련 문서 1개, 현대 자동차 역사 관련 문서 1개 **(40.0MB)**
- 추출된 고품질 데이터 건수 (필터링 후 기준): 자동차 제외 배경삭제 작업 -> 2044개 **(11.7GB)**

3.3 저장 위치 및 포맷

- 저장 경로: SKN_project/final_project/data
- 저장 포맷: CSV / JSON / PDF 등 다양한 형태
- 인코딩: UTF-8 / 기타

4. 법적·윤리적 검토

- 개인정보 포함 여부:
    - [☑️] 포함 [ ] 미포함
    - 포함된 경우 필드: (예: 이름, 전화번호 등)
- 비식별화 조치 여부:
    - 이름 마스킹 [☑️] 고유키 대체 [ ] 기타
- 출처 및 사용권:
    - 공개 여부: [ ] 공개 [ ] 내부사용 한정
    - 라이선스 또는 약관 검토 여부:
- 검토자 및 검토 일자:

5. 데이터 품질 및 정합성 관리 방안

- 중복 제거 기준: (예: user_id + timestamp 기준 중복 제거)
- 정합성 검증 방법:
    - 예: 잘못된 날짜 형식 필터링, 필수 필드 누락 제거
- Null 처리 및 결측치 전략:
- 표준화 전략:
    - 텍스트 전처리 여부 (예: 소문자 변환, 특수문자 제거 등)

6. 변경 이력 및 보완 내역

변경일	변경자	변경 내용	비고
2025-00-00		전화번호 필드 삭제 및 익명화 처리	개인정보 보호 강화 조치