SK네트웍스 Family AI과정 10기

데이터 수집 및 저장 수집 데이터 보고서

산출물 단계	데이터 수집 및 저장
평가 산출물	수집 데이터 보고서
제출 일자	2025.07.11
깃허브 경로	https://github.com/SKNETWORKS-FAMILY-AICAMP/SKN12-FINAL-6TEAM
작성 팀원	조성지

1. 수집 방법 및 자동화 절차

A. HTP(House-Tree-Person) 데이터

데이터명	수집 대상	수집 목적	사용 예정 기능	출처/저작권
어린이의 HTP 검사 반응에 대한 해석과 특징연구	HTP 검사 기반 아동 심리 연구 논문	HTP 검사 해석의 특징 및 전통적 분석 방법론 이해	HTP 그림 분석 모델의 이론적 배경 데이터로 활용	건국대학교 교육대학원 / 안정애
HTP평가기준 개발 -리커트 척도-	리커트 척도 기반 HTP 평가 기준 개발 연구 논문	HTP 검사의 정량적, 객관적 평가 지표 및 방법론 연구	HTP 그림 분석 모델의 이론적 배경 데이터로 활용	동의대학교 대학원 / 박성진
아동 미술심리검사를 위한 AI 기반 그림 데이터 분석 모델 연구	AI를 활용한 HTP 그림 분석 연구 논문	AI 모델(LLM, YOLOv5)을 HTP 분석에 적용한 선행 사례 및 한계점 파악	AI 분석 모델의 기능 정의 및 고도화 방향성 수립	동의대학교 대학원 / 박희진
HTP (House-Tree-Person) 검사 해석체계 구축 및 타당성 제고	HTP 검사 해석체계	HTP 그림 검사 채점표 및 해석 비교	HTP 그림 분석 모델의 이론적 배경 데이터로 활용	삼육대학교 대학원 / 백원대

2. 수집 방법 및 자동화 절차

A. HTP 데이터

- 수집 방식
    - HTP 데이터는 논문 PDF 수동으로 다운로드한 후 시스템 업로드 방식으로 수집

- 수집 도구 또는 스크립트 설명
    - 사용한 언어/라이브러리: Python
    - 자동화 여부 및 주기:
        - 수집: 수동(논문 자료 수집)
        - 전처리: HTP 이미지와 설명 부분만 찾아 수동으로 처리
        - 자동화: 비자동화(일회성 수집)

- 데이터 통합 및 정제
    - 마크다운 형식으로 정제 데이터 통합

- 데이터 수동 수집 및 수동 전처리 이유
    - 공신력 있는 데이터 확보 필요성(웹크롤링은 근거가 없어 학술적 타당성을 증명하기 어려움)
    - 데이터의 비표준성(논문 PDF 내 데이터가 이미지, 표 등 다양한 형태로 되어있어 자동화 추출이 어려움)
    - GPT 등을 이용한 데이터 1차 정제 이후 추가 수작업 필요성(GPT 등을 활용해 논문 내용을 정리해도, 요소/조건/이미지 등으로 구성되는 구조화된 데이터셋을 만들기 위해, 형식에 따라 정제하는 추가적인 수작업이 필수적임)
    - 불필요한 정보 수집의 비효율성(논문의 서론 부분은 필요하지 않으므로 불필요한 데이터를 수집하는 데에 드는 리소스를 줄이고자 함)

3. 데이터 설명 및 구성

A. HTP 데이터

3.1 파일 및 필드 설명

파일명	필드명	데이터 타입	설명
rag_doc_house.md, rag_doc_tree.md, rag_doc_person.md	drawing_element	TEXT	그림에서 분석 대상이 된 구성 요소 이름 (예: 창문, 문, 지붕 등)
	element_condition	TEXT	해당 요소의 상태 또는 특징 (예: 크다, 삐뚤어짐, 없음 등)
	emotion_keyword	TEXT	감정 키워드(예: 불안정감, 폐쇄성 등)
	Interpretation _sentence	TEXT	해당 조건에 대한 정성적 해석 문장
	reference_image_url	VARCHAR (2048)	참고 이미지 경로 또는 URL(선택)

- 1-A 항목에 있는 논문의 데이터를 각각 정제하여 하나의 마크다운 파일로 통합
- 하나의 그림을 받아 YOLO를 통해 객체를 탐지 후 해석모델에 넣을 것이므로 HTP 그림검사 항목인 집(House)/나무(Tree)/사람(Person)으로 분할
    - 필드명은 파일마다 동일

3.2 데이터 양

- 전체 수집 데이터 건수: 약 250건 (논문 기반 분석 요소 + 시뮬레이션된 HTP 그림 분석 결과)
- 추출된 고품질 데이터 건수 (필터링 후 기준): 약 180건(이미지 누락, 중복 응답, 감정 키워드 미도출 등의 예외 건 제거 후)

3.3 저장 위치 및 포맷

- 저장 경로: data/raw/htp_interpretation
- 저장 포맷: CSV / SQLite
- 인코딩: UTF-8

4. 법적·윤리적 검토

- 개인정보 포함 여부:
    - [] 포함 [✔️] 미포함
- 비식별화 조치 여부:
    - 해당사항 없음
- 출처 및 사용권:
    - 해당 사항 없음
    - 라이선스 또는 약관 검토 여부: 논문 내 인용된 사례 및 텍스트 기반 시뮬레이션 데이터를 기반으로 작성한 데이터로, 별도 사용 제한 없음

5. 데이터 품질 및 정합성 관리 방안

- 중복 제거 기준
    - user_id + submitted_at 조합을 기준으로 DrawingTest 테이블에서 중복 업로드 그림 제거
    - 또는 session_id + timestamp 기준으로 ChatMessage 테이블에서 중복 메시지 제거
- 정합성 검증 방법
    - timestamp, created_at 등 날짜 필드가 미래 시간이거나 비정상적인 형식인 경우 필터링
    - 외래 키 필드(user_id, session_id 등)가 존재하지 않는 경우 제거
    - 필수 항목(NN 필드) 누락 여부 점검
- Null 처리 및 결측치 전략
    - deleted_at, comment, description 등 비필수 필드는 NULL 허용
    - email, image_url 등 의미 있는 값이 빠진 경우 NULL로 처리하되, 분석 단계에서 제외
- 표준화 전략
    - 사용자 입력 message_text, summary_text 등에 대해 소문자 변환, 불필요한

공백 제거, 이모지 및 특수문자 필터링

- 텍스트 기반 감정 분석 및 요약 모델 입력을 위한 전처리 수행 예정

6. 변경 이력 및 보완 내역

변경일	변경자	변경 내용	비고
2025.07.11	조성지	최초 내용 작성	
2025.07.13	김승학	내용 수정 및 보완	