

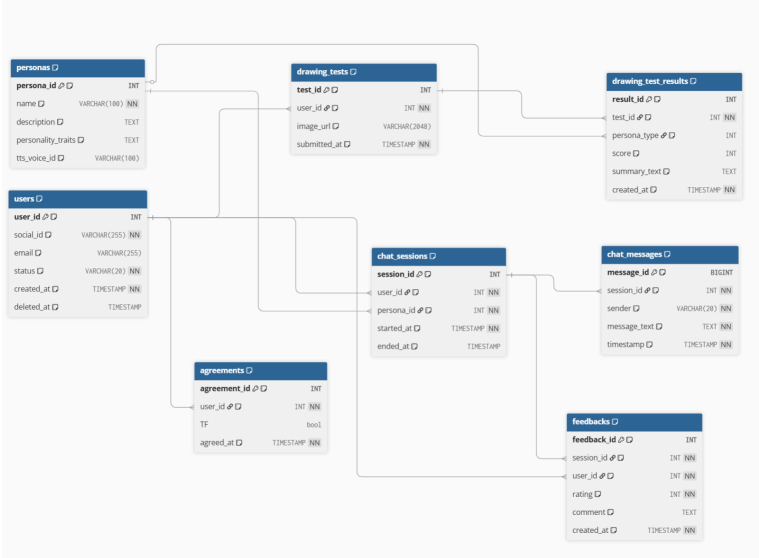
SK네트웍스 Family AI 과정 12기

데이터 수집 및 저장 데이터 조회 프로그램

산출물 단계	데이터 수집 및 저장
평가 산출물	데이터 조회 프로그램
제출 일자	2025. 07.11.
깃허브 경로	https://github.com/SKNETWORKS-FAMILY-AICAMP/SKN12-FINAL-6TEAM
작성 팀원	이정민, 조성지

원본 데이터 (텍스트)	[HTP(House-Tree-Person) 그림 심리검사 관련 논문 PDF] <ul style="list-style-type: none">• 어린이의 HTP 검사 반응에 대한 해석과 특징연구(안정애)• HTP평가기준 개발 -리커트 척도- (박희진)• 아동 미술심리검사를 위한 AI 기반 그림 데이터 분석 모델 연구(박성진)• HTP (House-Tree-Person) 검사 해석체계 구축 및 타당성 제고(백원대)
데이터 전처리 과정	<ul style="list-style-type: none">• 논문 내 서술 내용을 구조화된 필드(요소명, 조건, 감정 키워드, 해석 설명 등)로 가공• GPT API 응답 데이터는 JSON 구조로 저장• 이미지 파일명과 연결하여 정합성 관리
DB 사용 용도	<ul style="list-style-type: none">• 사용자가 제출한 HTP 그림(집, 나무, 사람)에 대한 심리 해석의 정확도를 높이기 위해 RAG(Retrieval-Augmented Generation) 방식 사용• 이를 위해 전문가 심리 해석 사례, 분석 기준 등이 담긴 논문 문서를 전처리한 뒤, 임베딩하여 Vector DB(FAISS)에 저장• GPT 응답 시 해당 DB를 참조함으로써 사용자 그림에 맞춤형 해석 제공 가능 [사용 기술] <ul style="list-style-type: none">• PostgreSQL 기반 관계형 데이터베이스• AWS S3를 통한 이미지 저장• FastAPI와 연동하여 백엔드 API 제공
VectorDB 구조	[Vector DB 기반 흐름 구조]

	<table><tr><th>단계</th><th>처리 내용</th><th>주요 기술/역할</th></tr><tr><td>1. 해석 생성</td><td>사용자의 HTP 그림을 기반으로 GPT-4o가 summary_text 생성</td><td>Vision + Language 통합 모델 (GPT-4o)</td></tr><tr><td>2. 임베딩</td><td>summary_text를 임베딩 벡터로 변환</td><td>OpenAI embedding 또는 sentence-transformers</td></tr><tr><td>3. 유사 문서 검색 (RAG)</td><td>벡터 쿼리를 통해 VectorDB 내 .md 해석 문서들과 유사도 기반 검색 수행</td><td>FAISS, Pinecone, cosine similarity 기반 Top-k 검색</td></tr><tr><td>4. 감정 키워드 추출</td><td>검색된 문서로부터 감정 키워드 수집 (불안감, 애정 욕구등)</td><td>해석 문서 내 포함된 감정 키워드 필드 활용</td></tr><tr><td>5. 키워드 스코어링</td><td>추출된 감정 키워드를 5가지 상담 유형 기준표와 매핑하여 점수화</td><td>Rule 기반 키워드 매칭 (가중치는 균등 적용)</td></tr><tr><td>6. 상담 유형 분류</td><td>감정 점수 벡터를 ML 분류기에 입력해 상담 유형(1~5) 도출</td><td>SVM 또는 RandomForest 기반 분류기 (실험 중)</td></tr><tr><td>7. 챗봇 연동</td><td>분류된 유형에 따라 GPT-4o 기반 상담 챗봇과 연결하여 맞춤형 상담 진행</td><td>유형별 RAG 적용된 GPT-4o 챗봇과 실시간 대화 연동</td></tr></table>	단계	처리 내용	주요 기술/역할	1. 해석 생성	사용자의 HTP 그림을 기반으로 GPT-4o가 summary_text 생성	Vision + Language 통합 모델 (GPT-4o)	2. 임베딩	summary_text를 임베딩 벡터로 변환	OpenAI embedding 또는 sentence-transformers	3. 유사 문서 검색 (RAG)	벡터 쿼리를 통해 VectorDB 내 .md 해석 문서들과 유사도 기반 검색 수행	FAISS, Pinecone, cosine similarity 기반 Top-k 검색	4. 감정 키워드 추출	검색된 문서로부터 감정 키워드 수집 (불안감, 애정 욕구등)	해석 문서 내 포함된 감정 키워드 필드 활용	5. 키워드 스코어링	추출된 감정 키워드를 5가지 상담 유형 기준표와 매핑하여 점수화	Rule 기반 키워드 매칭 (가중치는 균등 적용)	6. 상담 유형 분류	감정 점수 벡터를 ML 분류기에 입력해 상담 유형(1~5) 도출	SVM 또는 RandomForest 기반 분류기 (실험 중)	7. 챗봇 연동	분류된 유형에 따라 GPT-4o 기반 상담 챗봇과 연결하여 맞춤형 상담 진행	유형별 RAG 적용된 GPT-4o 챗봇과 실시간 대화 연동
	단계	처리 내용	주요 기술/역할																						
	1. 해석 생성	사용자의 HTP 그림을 기반으로 GPT-4o가 summary_text 생성	Vision + Language 통합 모델 (GPT-4o)																						
	2. 임베딩	summary_text를 임베딩 벡터로 변환	OpenAI embedding 또는 sentence-transformers																						
	3. 유사 문서 검색 (RAG)	벡터 쿼리를 통해 VectorDB 내 .md 해석 문서들과 유사도 기반 검색 수행	FAISS, Pinecone, cosine similarity 기반 Top-k 검색																						
	4. 감정 키워드 추출	검색된 문서로부터 감정 키워드 수집 (불안감, 애정 욕구등)	해석 문서 내 포함된 감정 키워드 필드 활용																						
	5. 키워드 스코어링	추출된 감정 키워드를 5가지 상담 유형 기준표와 매핑하여 점수화	Rule 기반 키워드 매칭 (가중치는 균등 적용)																						
	6. 상담 유형 분류	감정 점수 벡터를 ML 분류기에 입력해 상담 유형(1~5) 도출	SVM 또는 RandomForest 기반 분류기 (실험 중)																						
7. 챗봇 연동	분류된 유형에 따라 GPT-4o 기반 상담 챗봇과 연결하여 맞춤형 상담 진행	유형별 RAG 적용된 GPT-4o 챗봇과 실시간 대화 연동																							
<p>[HTP 그림 심리검사 관련 Vector DB 예시]</p> <table><tr><th>ids</th><th>documents</th></tr><tr><td>02a1e345-c6a2-4a9c-b27e-8d3e</td><td>지붕은 내적인 공상 활동 및 자기 자신의 생각과 같은 인지과정과 관련됩니다.</td></tr><tr><td>03aed191-3b46-4f95-9b9c-6b2d4</td><td>지나치게 큰 그림은 환경에 대한 적의와 공격성이 강함을 나타내며, 긴장과 조급하기 쉬운 사람에게</td></tr><tr><td>04a1b023-455b-4394-b352-7ebf3</td><td>굴뚝에서 피어나는 연기는 가정 내의 따뜻함이나 혹은 심리적 갈등과 같은 정서적 상태를 반영할 수</td></tr><tr><td>0a1e1712-7d69-45ab-8c2d-2253</td><td>태양의 존재는 일반적으로 따뜻함, 희망, 긍정적인 에너지 또는 권위적인 인물(아버지 등)을 상징합니다</td></tr><tr><td>0aae13a4-cc87-4dc4-a48a-4d26f</td><td>나무 그림은 기본적인 자기상을 나타내며, 나무 기둥은 내적 자아 강도에 대한 주관적 느낌을 가지</td></tr></table>			ids	documents	02a1e345-c6a2-4a9c-b27e-8d3e	지붕은 내적인 공상 활동 및 자기 자신의 생각과 같은 인지과정과 관련됩니다.	03aed191-3b46-4f95-9b9c-6b2d4	지나치게 큰 그림은 환경에 대한 적의와 공격성이 강함을 나타내며, 긴장과 조급하기 쉬운 사람에게	04a1b023-455b-4394-b352-7ebf3	굴뚝에서 피어나는 연기는 가정 내의 따뜻함이나 혹은 심리적 갈등과 같은 정서적 상태를 반영할 수	0a1e1712-7d69-45ab-8c2d-2253	태양의 존재는 일반적으로 따뜻함, 희망, 긍정적인 에너지 또는 권위적인 인물(아버지 등)을 상징합니다	0aae13a4-cc87-4dc4-a48a-4d26f	나무 그림은 기본적인 자기상을 나타내며, 나무 기둥은 내적 자아 강도에 대한 주관적 느낌을 가지											
ids	documents																								
02a1e345-c6a2-4a9c-b27e-8d3e	지붕은 내적인 공상 활동 및 자기 자신의 생각과 같은 인지과정과 관련됩니다.																								
03aed191-3b46-4f95-9b9c-6b2d4	지나치게 큰 그림은 환경에 대한 적의와 공격성이 강함을 나타내며, 긴장과 조급하기 쉬운 사람에게																								
04a1b023-455b-4394-b352-7ebf3	굴뚝에서 피어나는 연기는 가정 내의 따뜻함이나 혹은 심리적 갈등과 같은 정서적 상태를 반영할 수																								
0a1e1712-7d69-45ab-8c2d-2253	태양의 존재는 일반적으로 따뜻함, 희망, 긍정적인 에너지 또는 권위적인 인물(아버지 등)을 상징합니다																								
0aae13a4-cc87-4dc4-a48a-4d26f	나무 그림은 기본적인 자기상을 나타내며, 나무 기둥은 내적 자아 강도에 대한 주관적 느낌을 가지																								
<div>Django RDB구조</div> <ul style="list-style-type: none">• PostgreSQL 기반의 RDB를 사용하며, 전체 테이블 구조 및 관계는 아래 ERD(Entity Relationship Diagram)를 통해 정의• 모든 데이터는 사용자 ID(user_id)를 중심으로 연결되어 있으며, 정규화된 구조로 설계되어 무결성과 확장성을 확보																									

	
<div> <div>검증 코드 및 프롬프트 요약</div> </div>	<div> <div> <div> <div>[검증 코드]</div> <div> <pre> from typing import Dict # 1. 감정 키워드 유형 매핑 사전 (정의된 감정 목록 기반, 가중치 +1 방식) emotion_type_map = { "추진형": ("과대평가", "스트레스", "정열", "욕심", "충동성", "자율성", "자기 억제", "정확성", "자신감", "미충족된 욕구"), "관계형": ("애정", "관심 필요", "외로움", "관계 불안", "신뢰", "애정결핍", "친밀감 형성의 어려움", "의존력", "우유부단"), "안정형": ("안정감", "평온함", "조화로움", "인내심", "무력감", "자기 통제", "불안정감", "정서혼란", "스트레스", "우유부단"), "쾌락형": ("즐거움", "공상", "자율성", "자극", "호기심", "창의력", "장난기", "답답함", "상상력", "미충족된 욕구"), "내면형": ("불안", "우울", "내성적", "위축", "강한 자기 억제", "자기 보호", "감정 통제 및 억제", "자아 성찰", "에민성", "자폐") } # 2. 추출된 감정 키워드 예시 extracted_keywords = {"불안", "자기 보호", "정열", "스트레스"} # 3. 유형별 점수 초기화 score_vector: Dict = { "내면형": 0, "안정형": 0, "쾌락형": 0, "관계형": 0, "추진형": 0 } # 4. 감정 키워드에 따른 점수 누적 for keyword in extracted_keywords: for type_name, keywords in emotion_type_map.items(): if keyword in keywords: score_vector[type_name] += 1 # 5. 예측된 상담 유형 도출 predicted_type = max(score_vector, key=score_vector.get) # 6. 검증 출력 print("✅ 감정 키워드:", extracted_keywords) print("✅ 유형 점수:", score_vector) print("✅ 예측된 상담 유형:", predicted_type) # 7. 검증용 Assert assert sum(score_vector.values()) > 0, "❌ 감정 키워드가 어떤 유형에도 매핑되지 않음" assert predicted_type in score_vector, "❌ 예측 결과 오류" </pre> <div> <div> <div>✅ 감정 키워드: {'자기 보호', '불안', '정열', '스트레스'}</div> <div> <div>✅ 유형 점수: {'내면형': 2, '안정형': 1, '쾌락형': 0, '관계형': 0, '추진형': 2}</div> <div>✅ 예측된 상담 유형: 내면형</div> </div> </div> </div> </div> </div> </div> <div> <div>[프롬프트 주요 내용 요약]</div> <div> <div> <div>심리 분석 요소 식별: 그림에 나타난 시각적 요소들을 구체적으로 나열</div> <div>요소별 해석: 각 요소의 상태를 기반으로 심리적 해석을 제시</div> <div>감정 키워드 도출: 주요 감정 상태를 한 줄 요약</div> </div> </div> </div> </div>
<div> <div>검증 결과</div> </div>	<div> <div> <div>감정 키워드는 5가지 상담 유형(내면형, 안정형, 쾌락형, 관계형, 추진형)에 매핑되어 있으며, 등장 시마다 해당 유형 점수를 +1씩 누적하는 Rule-based 방식으로 분류</div> </div> </div>

	<ul style="list-style-type: none"> ● GPT-4o 해석 결과에 포함된 키워드가 사전에 없을 경우 무시되며, 이에 대한 예외 처리 및 RAG 기반 자동 확장 방식 도입이 필요 ● 감정 키워드는 .md 기반 해석 문서에서 RAG를 통해 실시간 추출 및 활용되며, 이를 통해 키워드 사전을 동적으로 보완 가능 ● 감정 키워드 점수의 동점 발생 시, 사전 정의된 유형 우선순위에 따라 분류 결과를 결정할 예정 ● Rule 기반 방식은 임시 사용이며, 머신러닝 기반 상담 유형 분류기(SVM, RandomForest 등)를 도입할 예정이며, 다양한 테스트를 통해 최적 모델을 선정할 계획 ● 유형 분류 예외 상황을 대비한 fallback 유형 지정 로직 및 예외 키워드 기록·관리 체계도 병행 구축할 예정
--	--