수집 데이터

D-Day	마감
상태	완료
마감일	@2026/01/09
작업완료 여부	
작업 유형	기능 요청
설명	다음 릴리스 노트에 제품 업데이트를 포함합니다.

v2. 수집 데이터

▼ 1. 데이터 개요

구분	데이터명	데이터 수집 목적	데이터 형태	데이터 출처	데이터 크기	처리 상태	비고
DATA-001	밈 기본 정보	밈 정의, 기원, 키워드 수집	JSON	나무위키, Google 검색 (Serper API)	110건	테스트 완료/ 자동화 개발하여 저장 계획	
DATA-002	밈 활용 예시	밈 사용 맥락, 패러디 사례	JSON (내장)	블로그, 뉴스 크롤링	밈당 3~5개의 예시	테스트 완료/ 자동화 개발하여 저장 계획	
DATA-003	밈 관련 YouTube 쇼츠 정보	밈 관련 인기 영상 수집하여 영상 및 오디오 분석	JSON (내장)	YouTube Data API v3	밈당 상위 3개 영상	테스트 완료/ 자동화 개발하여 저장 계획	
DATA-004	영상 분석 데이터	밈 구간 탐지, 동작 분석	JSON	Google Gemini API	밈당 1건	테스트 완료/ 자동화 개발하여 저장 계획	
DATA-005	오디오 분석 데이터	음성 운율, TTS 힌트 생성	WAV	YouTube 영상 추출 (yt-dlp)	밈당 1건	테스트 완료/ 자동화 개발하여 저장 계획	
DATA-006	텍스트 파인튜닝 데이터1	밈 기반 캐릭터 대사 생성 및 대화 스타일 학습을 목적으로 한 모델 파인튜닝용 원천 데이터	tsv	korean SmileStyle Dataset	3706개	추가 수집 중/ 후처리 필요	부장 캐릭터: column[azae] 사원 캐릭터: column[gentle]
DATA-007	텍스트 파인튜닝 데이터2	밈 기반 캐릭터 대사 생성 및 대화 스타일 학습을 목적으로 한 모델 파인튜닝용 원천 데이터	JSONL	한국어 대화 데이터셋 - Office 상황 데이터셋	46414개	추가 수집 중/ 후처리 필요	column[system_utterance]
DATA-008	음성 파인튜닝 데이터	음성 생성 모델의 보이스 클로닝	JSON	감성 및 발화스타일 동시 고려 음성 합성 데이터	560개	추가 수집 중/ 후처리 필요	부장 캐릭터: 80개(감정별20개, 4감정) 사원 캐릭터: 80개(감정별20개, 4감정) 여분 : 400개(80개 5세트)

▼ 2. 데이터 수집 및 활용의 적법성 검토

나무위키 robots.txt

YouTube API 서비스 - 개발자 정책

구분	상업 이용 가능	학습 사용 허용	크롤링 이용약관 준수	개인정보 보호	비고	
DATA-001	X	O	O	O	출처 명시 필수	
DATA-002	O	O	O	O	공개 게시물만 수집	
DATA-003	O	O	O	O	YouTube API 정책 준수	
DATA-004	O	O	O	O	2차 가공 데이터	
DATA-005	O	O	O	O	음성 원본 미저장	
DATA-006	O	O	-	O	상업적 사용의 경우 smilegate_ai@smilegate.com 으로 별도 문의	

구분	상업 이용 가능	학습 사용 허용	크롤링 이용약관 준수	개인정보 보호	비고	
DATA-007	O	O	-	O	공개 한국어 대화 데이터셋 기반 분석·가공 데이터로, 원본 라이선스 범위 내에서 2차 가공 후 사용 가능	
DATA-008	O	O	-	O	원본 라이선스 범위 내에서 2차 가공 후 사용 가능	

▼ 3. 데이터 저장 및 관리

▼ 1. DATA-001 (밈 기본 정보)

- 저장 형식: JSON
- 저장 환경: PostgreSQL (memedb 스키마)
- 데이터 구조:

컬럼명	데이터 타입	제약 조건	설명
meme_id	SERIAL	PK, NOT NULL	밈 고유 식별자
meme_name	VARCHAR	UNIQUE, NOT NULL	밈 이름
definition	TEXT	NOT NULL	밈 정의. 밈의 의미와 사용 방법을 3~5문장으로 설명
origin	JSONB	-	밈의 기원 정보 {source, creator, date, platform}
key_phrase	VARCHAR	-	quotable 밈의 핵심 대사
keywords	JSONB	-	검색 키워드 배열
risk_info	JSONB	-	밈의 위험/논란 정보 {controversies, risk_level}
meme_type	VARCHAR	CHECK (`meme_type` IN ('quotable', 'performable'))	밈 유형. 'quotable'(대사 중심) 또는 'performable'(동작 중심).
is_processed	BOOLEAN	DEFAULT FALSE	영상 분석 완료 여부
created_at	TIMESTAMP	DEFAULT CURRENT_TIMESTAMP	생성 일시
updated_at	TIMESTAMP	DEFAULT CURRENT_TIMESTAMP	수정 일시

- 데이터 정제 및 전처리:
  - 나무위키/웹검색에서 정의 추출 후 LLM으로 요약
  - 키워드 자동 추출 및 중복 제거
  - 논란/민감 주제 자동 감지하여 risk_info 생성

▼ 2. DATA-002 (밈 활용 예시)

- 저장 형식: JSON
- 저장 환경: PostgreSQL (memedb 스키마)
- 데이터 구조:

컬럼명	데이터 타입	제약 조건	설명
example_id	SERIAL	PK, NOT NULL	예시 고유 식별자
meme_id	INTEGER	FK, NOT NULL	참조 밈 ID (밈당 3~5개 예시 저장)
source_url	TEXT	NOT NULL	크롤링한 원본 URL
source_title	VARCHAR	-	원본 페이지 제목
context	TEXT	NOT NULL	밈이 사용된 상황 및 맥락 설명
description	TEXT	-	원문 기반 밈 활용 방식 설명 (3~5문장)
original_quote	TEXT	-	원문에서 직접 발췌한 문구
example_type	VARCHAR	-	예시 유형('챌린지','패러디','상황극','짤','댓글')
created_at	TIMESTAMP	DEFAULT CURRENT_TIMESTAMP	생성 일시

- 데이터 정제 및 전처리:
  - 크롤링된 원문에서만 예시 추출 (할루시네이션 방지)
  - source_url, original_quote 필수 검증
  - example_type 자동 분류

▼ 3. DATA-003 (유튜브 영상 메타데이터)

- 저장 형식: JSON
- 저장 환경: PostgreSQL (memedb 스키마)
- 데이터 구조:

컬럼명	데이터 타입	제약 조건	설명
video_id	VARCHAR	PK, NOT NULL	YouTube 영상 ID (11자)
meme_id	INTEGER	FK	참조 밈 ID
video_title	VARCHAR	-	YouTube 원본 영상 제목
youtube_url	TEXT	-	YouTube 영상 URL
view_count	BIGINT	-	조회수
published_at	TIMESTAMP	-	영상 게시 일시

- 데이터 정제 및 전처리:
  - YouTube Data API v3로 조회수 상위 3개 검색
  - Shorts 영상만 필터링 (60초 이하)
  - 중복 video_id 제거

▼ 4. DATA-004 (영상 분석 데이터)

- 저장 형식: JSON
- 저장 환경: PostgreSQL (memedb 스키마)
- 데이터 구조:

컬럼명	데이터 타입	NULL 여부	Default 값	PK	FK	비고
video_id	VARCHAR	NOT NULL	-	O	youtube.video_id	
time_stamp	JSONB	NULL	-	-	-	밈 구간 정보
motion_prompt_hint	TEXT	NULL	-	-	-	50-100 words
motion_sequence	JSONB	NULL	[]	-	-	시간별 동작
body_parts	JSONB	NULL	{}	-	-	부위별 동작
style_keywords	JSONB	NULL	[]	-	-	
movement_analysis	JSONB	NULL	{}	-	-	
negative_prompt	TEXT	NULL	-	-	-	
camera_motion	VARCHAR	NULL	'static'	-	-	
background_suggestion	TEXT	NULL	-	-	-	
duration_seconds	NUMERIC	NULL	5.0	-	-	

컬럼 상세 설명:

- **video_id**: youtube 테이블의 video_id를 참조.
- **time_stamp**: 밈 구간 정보. `{start, end, detected_text}` 구조.
- **motion_prompt_hint**: 영상생성용 상세 프롬프트 (50-100 words).
- **motion_sequence**: 시간별 동작 분해. `[{time, action}]` 배열.
- **body_parts**: 신체 부위별 구체적 동작. `{head, face, arms, hands, torso}` 구조.
- **camera_motion**: 카메라 움직임. `static` | `slow_zoom_in` | `subtle_pan`

- 데이터 정제 및 전처리:
  - Gemini Vision으로 영상 직접 분석
  - 밈 구간 타이밍 자동 탐지 (key_phrase 기반)
  - motion_prompt 자동 생성

▼ 5. DATA-005 (오디오 분석 데이터)

- 저장 형식: JSON
- 저장 환경: PostgreSQL (memedb 스키마)
- 데이터 구조:

컬럼명	데이터 타입	제약 조건	설명
video_id	VARCHAR	PK, FK, NOT NULL	YouTube 영상 ID 참조
audio_file	TEXT	-	추출된 오디오 파일 저장 경로

컬럼명	데이터 타입	제약 조건	설명
audio_json	JSONB	-	음성 분석 결과
audio_json.prosody	-	-	운율 분석 정보 {pitch_mean, speaking_rate}
audio_json.tts_hint	-	-	TTS 생성 힌트 {voice, speed}
audio_json.detected_text	-	-	음성 인식으로 추출된 텍스트
created_at	TIMESTAMP	DEFAULT CURRENT_TIMESTAMP	생성 일시

- 데이터 정제 및 전처리:
  - ffmpeg로 밈 구간 오디오 추출 (time_stamp 기반)
  - librosa로 운율 분석 (피치, 리듬, 강세)
  - LLM 기반 SSML 동적 생성

▼ 6. DATA-006 (텍스트 파인튜닝 데이터 생성용 1)

- 저장 형식: tsv
- 저장 환경: PostgreSQL (memedb 스키마)
- 데이터 구조:

컬럼명	데이터 타입	제약 조건	설명
utterance_id	INTEGER	NOT NULL	원본 데이터 식별자.
speaker_role	VARCHAR(30)	NOT NULL	캐릭터별 발화 텍스트 부장 캐릭터: column[azae] 사원 캐릭터: column[gentle]
utterance	TEXT	NOT NULL	발화 텍스트. 부장님/상사 역할의 기준 발화 텍스트

- 데이터 정제 및 전처리:
  - 기존 성향 컬럼 중 azae, gentle **유형만 유지**
  - 발화 텍스트가 null 또는 공백인 행 제거
  - UTF-8 인코딩 통일

▼ 7. DATA-007 (텍스트 파인튜닝 데이터 생성용 2)

- 저장 형식: JSON
- 저장 환경: PostgreSQL (memedb 스키마)
- 데이터 구조:

컬럼명	데이터 타입	제약 조건	설명
utterance_id	INTEGER	NOT NULL	원본 데이터 식별자.
domain	VARCHAR(30)	NOT NULL	대화 유형 구분 값 (예: daily, task)
system_utterance	TEXT	NOT NULL	시스템 발화 내용.

- 데이터 정제 및 전처리:
  - system_utterance가 존재하는 데이터만 선별
  - 발화 텍스트가 null 또는 공백인 행 제거
  - UTF-8 인코딩 통일

▼ 8. DATA-008 (음성 파인튜닝 데이터)

- 저장 형식: JSON
- 저장 환경: PostgreSQL (memedb 스키마)
- 데이터 구조:

컬럼명	데이터 타입	제약 조건	설명
audio_id	SERIAL	PK	음성 데이터의 고유 식별자
file_root	TEXT	-	오디오 파일 저장 경로
filename	VARCHAR(30)	-	음성 파일 이름
character	VARCHAR(10)	-	캐릭터(부장/신입)
speaker_id	INT	-	성우 id
speaker_age	INT	-	성우 나이
speaker_gender	VARCHAR(10)	-	성우 성별
origin_text	VARCHAR(200)	-	문장 원문

컬럼명	데이터 타입	제약 조건	설명
ptr	VARCHAR(200)	-	발음 전사
tr	VARCHAR(200)	-	철자 전사
emotion	TEXT	-	감정(예:분노)
intensity	INT	-	감정의 강도
style	VARCHAR(10)	-	스타일(예:독백체)
sub_style	VARCHAR(10)	-	스타일 추가 정보(예:남아)
duration	VARCHAR(20)	-	음성 구간 길이
file_duration	VARCHAR(20)	-	음성 파일 구간 길이
votes	JSONB	-	투표 정보
created_at	TIMESTAMP	DEFAULT CURRENT_TIMESTAMP	생성 일시