수집된 데이터 및 데이터 전처리 문서

데이터 설명 및 구성

1. 데이터 개요

구분	데이터명	수집 목적	데이터 형태	데이터 출처	데이터 크기	처리 상태
DATA-001	밈 기본 정보	밈 정의, 기원, 키워드 수집	JSON	나무위키, Google 검색 (Serper API)	110건	테스트 완료 / 자동화 개발 중
DATA-002	밈 활용 예시	밈 사용 맥락, 패러디 사례	JSON	블로그, 뉴스 크롤링	밈당 3~5개	테스트 완료 / 자동화 개발 중
DATA-003	YouTube 쇼츠 정보	밈 관련 인기 영상 수집	JSON	YouTube Data API v3	밈당 상위 3개	테스트 완료 / 자동화 개발 중
DATA-004	영상 분석 데이터	밈 구간 탐지, 동작 분석	JSON	Google Gemini API	밈당 1건	테스트 완료 / 자동화 개발 중
DATA-005	오디오 분석 데이터	음성 운율, TTS 힌트 생성	WAV + JSON	YouTube 영상 추출 (yt-dlp)	밈당 1건	테스트 완료 / 자동화 개발 중

- **전체 수집 데이터 건수**: 밈 110건 기준 약 550건 (밈당 평균 5건)

2. 데이터 필드 상세

DATA-001: 밈 기본 정보

필드명	데이터 타입	설명	예시
meme_id	BIGINT	밈 고유 식별자 (Auto increment)	1, 2, 3
meme_name	VARCHAR	밈 이름 (UNIQUE)	나니가 스키, 랫댄스
definition	TEXT	밈 정의 (3~5문장)	"나니가 스키"는 일본어로 "뭐가 좋아?"라는...
origin	JSONB	기원 정보 {source, creator, date, platform}	{"source": "러브라이브!", "platform": "TikTok"}
key_phrase	TEXT	quotable 밈의 핵심 대사	何が好き?, Absolute Cinema
keywords	JSONB	검색 키워드 배열	["나니가스키", "러브라이브", "틱톡"]
risk_info	JSONB	위험/논란 정보 {controversies, risk_level}	{"risk_level": "low", "controversies": []}
meme_type	VARCHAR	밈 유형 (quotable | performable)	quotable

DATA-002: 밈 활용 예시

필드명	데이터 타입	설명	예시
example_id	INT	예시 고유 식별자	1, 2, 3
meme_id	BIGINT	참조 밈 ID (FK)	1
source_url	TEXT	크롤링 원본 URL	https://blog.naver.com/...
source_title	TEXT	원본 페이지 제목	나니가스키 밈 사용법
context	TEXT	사용 맥락 설명	친구에게 취향을 물을 때 사용
original_quote	TEXT	원문 발췌	"이 밈은 주로 귀여운 상황에서..."
example_type	VARCHAR	예시 유형	챌린지, 패러디, 상황극, 짤, 댓글

DATA-003: YouTube 쇼츠 정보

필드명	데이터 타입	설명	예시
video_id	VARCHAR	YouTube 영상 ID (11자)	ZMQRVtrPAWs
meme_id	BIGINT	참조 밈 ID (FK)	1
video_title	VARCHAR	영상 제목	나니가스키 챌린지 모음
youtube_url	TEXT	YouTube URL	https://youtube.com/shorts/ZMQRVtrPAWs
view_count	BIGINT	조회수	13466259

DATA-004: 영상 분석 데이터

필드명	데이터 타입	설명	예시
video_id	VARCHAR	YouTube 영상 ID (PK, FK)	ZMQRVtrPAWs
time_stamp	JSONB	밈 구간 {start, end, detected_text}	{"start": 2.8, "end": 3.4, "detected_text": {...}}
motion_prompt_hint	TEXT	영상생성용 프롬프트 (50-100 words)	Cute character tilts head, raises hand...
motion_sequence	JSONB	시간별 동작 분해	[{"time": "0-2s", "action": "머리 기울임"}]
body_parts	JSONB	부위별 동작 {head, face, arms, hands, torso}	{"head": "tilts right", "face": "smile"}
style_keywords	JSONB	스타일 키워드 배열	["cute", "kawaii", "anime"]
camera_motion	VARCHAR	카메라 움직임	static, slow_zoom_in, subtle_pan

DATA-005: 오디오 분석 데이터

필드명	데이터 타입	설명	예시
video_id	VARCHAR	YouTube 영상 ID (PK, FK)	ZMQRVtrPAWs
audio_file	TEXT	추출된 오디오 파일 경로	data/raw/audio/ZMQRVtrPAWs.wav
audio_json.prosody	JSONB	운율 분석 {pitch_mean, speaking_rate}	{"pitch_mean": 220.5, "speaking_rate": 4.2}
audio_json.ssml	TEXT	생성된 SSML	`<speak><prosody rate="fast">...</prosody></speak>`
audio_json.detected_text	TEXT	음성 인식 텍스트	何が好き

데이터 수집 및 활용의 적법성 검토

1. 적법성 검토 결과

구분	상업 이용	학습 허용	크롤링 준수	개인정보 보호	비고
DATA-001	X	O	O	O	출처 명시 필수 (https://namu.wiki/robots.txt)
DATA-002	O	O	O	O	공개 게시물만 수집
DATA-003	O	O	O	O	YouTube Developer Policies 준수
DATA-004	O	O	O	O	2차 가공 데이터
DATA-005	O	O	O	O	음성 원본 미저장 (구간만 추출)

2. 준수 사항

- **저작권 준수**: 각 데이터의 이용약관을 검토하여 상업적 이용 및 2차 가공 가능 여부 확인 완료
- **크롤링 이용약관 준수**: robots.txt 규정에 따라 서버 부하를 최소화하는 방식으로 크롤링 수행

- **개인정보 보호**: 수집 과정에서 개인 식별 정보(이름, 연락처 등)는 즉시 제외 또는 비식별 처리

수집 자동화

1. 수집 방식 및 도구

구분	수집 방식	사용 도구	자동화 주기
DATA-001	검색 API + 크롤링	Serper API, BeautifulSoup	매일 00:00 (Phase 1)
DATA-002	웹 크롤링	BeautifulSoup, requests	밈 수집 시 동시 실행
DATA-003	API 호출	YouTube Data API v3	밈 수집 시 동시 실행
DATA-004	영상 분석	Google Gemini API	주간 콘텐츠 생성 시
DATA-005	오디오 추출 + 분석	yt-dlp, ffmpeg, librosa	주간 콘텐츠 생성 시

2. 수집 자동화 코드

```
# Agent 실행 (DATA-001 ~ DATA-005 통합 수집)
uv run python -m scripts.agent.meme_agent_v8 "밈이름"

# 배치 실행 (다중 밈 수집)
uv run python -m scripts.agent.batch_runner
```

3. 데이터 유효성 검증

검증 단계	검증 방법	예외 처리
1차: 필수 필드 검증	meme_name, definition, source_url 존재 확인	누락 시 수집 실패 로그 기록
2차: 중복 검증	meme_name UNIQUE 제약 조건	중복 시 기존 데이터 업데이트
3차: 출처 검증	usage_examples의 source_url 유효성 확인	유효하지 않은 URL 제외
4차: 영상 ID 검증	YouTube video_id 11자 형식 확인	형식 불일치 시 제외
5차: 밈 구간 검증	time_stamp.start < time_stamp.end 확인	역전 시 전체 오디오 분석으로 fallback

전처리 프로세스 개요

1. 전체 흐름도

① 수집 → ② 결측치 처리 → ③ 이상치 탐지 → ④ 정규화 → ⑤ 데이터 분리/저장

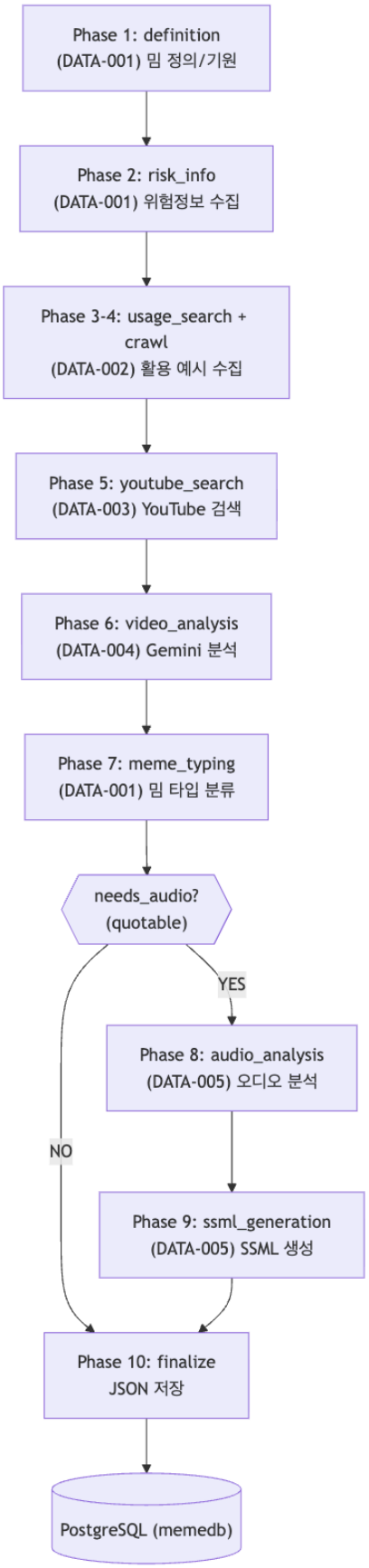
2. 전처리 파이프라인 요약

단계	목적	수행 작업	사용 도구
결측치 처리	누락값 제거/대체	Null 행 제거, 기본값 대체	pandas, PostgreSQL
이상치 처리	비정상 데이터 제거	조회수 0 영상 필터링, 60초 초과 영상 제외	numpy
정규화	텍스트 전처리	중복 키워드 제거, 특수문자 정제	re, nltk
분류	밈 타입 분류	quotable / performable 자동 분류	LLM (GPT-4o)
저장	DB 적재	PostgreSQL memedb 스키마 저장	psycopg2, SQLAlchemy

3. 데이터별 전처리 상세

구분	전처리 작업
DATA-001	나무위키/웹검색 결과 LLM 요약, 키워드 중복 제거, 논란 키워드 자동 감지
DATA-002	크롤링 원문에서만 예시 추출 (할루시네이션 방지), source_url 필수 검증
DATA-003	Shorts 영상만 필터링 (60초 이하), 조회수 상위 3개 선정, 중복 제거
DATA-004	Gemini Vision 분석 결과 JSON 파싱, motion_prompt 50단어 이상 검증
DATA-005	ffmpeg 밈 구간 추출, librosa 운율 분석, LLM SSML 생성

4. 파이프라인 도식화


Phase 1: definition
(DATA-001) 밈 정의/기원
Phase 2: risk_info
(DATA-001) 위험정보 수집
Phase 3-4: usage_search + crawl
(DATA-002) 활용 예시 수집
Phase 5: youtube_search
(DATA-003) YouTube 검색
Phase 6: video_analysis
(DATA-004) Gemini 분석
Phase 7: meme_typing
(DATA-001) 밈 타입 분류
needs_audio?
(quotable)
YES
NO
Phase 8: audio_analysis
(DATA-005) 오디오 분석
Phase 9: ssml_generation
(DATA-005) SSML 생성
Phase 10: finalize
JSON 저장
PostgreSQL (memedb)