SK Networks Family AI Camp 9기

**데이터 전처리 결과서**

가로선

**프로젝트 명: DAIS (Divergent AI with Science)**

**담당자: 조이현**

# <목차>

1. 전처리 목적

2. 전처리 범위

3. 전처리 과정 요약

4. 전처리 결과 요약

# 

# 

# 1. 전처리 목적

원천 유튜브 데이터(raw data)에서 불필요한 HTML 태그·메타정보·특수문자·중복구간

·이모지·URL을 제거하여 학습 데이터의 품질을 확보하기 위함. 따라서 원천 데이터의

자막 텍스트를 토큰화·정규화 및 JSONL 구조로 일관되게 포맷팅하였음. 이를 통하여

모델 파인튜닝 및 RAG 활용 시 토큰 예측 성능을 극대화할 수 있으며, 무결성 체크로

downstream 파이프라인의 자동화·재현성을 보장할 수 있음.

# 2. 전처리 범위

가. 수집 데이터 출처: 지정 12개 유튜브 채널 영상 제목, 공개 자막, 설명

나. 데이터 수집 방식: YouTube Data API v3 + youtube-transcript-api 병행

다. 데이터 수집 기간: 2025년 7월 28일 수요일 - 2025년 07월 29일 목요일

라. 수집 데이터 포맷: UTF-8 인코딩 .jsonl (JSON Lines)

마. 총 데이터 건수: 8,500여 편 영상 → 6,000여 건, 각 레코드 input-output 형태

바. 데이터 특성 요약: 평균 제목 8토큰, 평균 자막 및 설명 1,200토큰

# 

# 

# 

# 

# 3. 전처리 과정 요약

가. 원천 데이터 분석

1) 샘플 수: JSONL 레코드로 환산 시 6,024 건

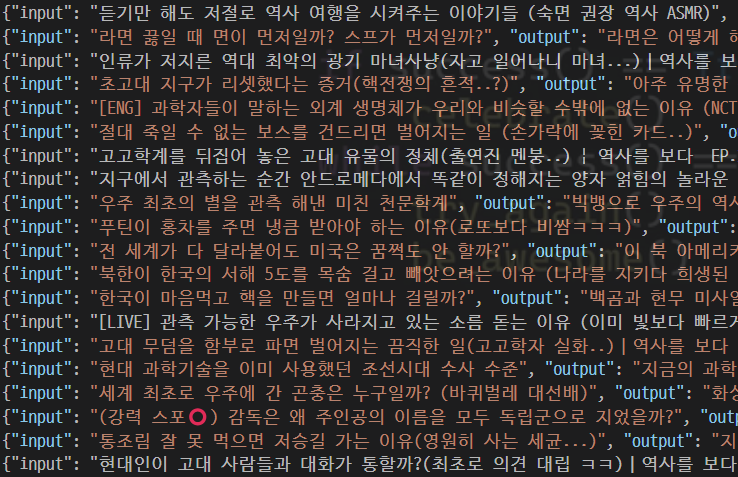
2) 평균 제목 길이: 약 8 토큰 (제목 길이 분포 5 – 80여 토큰)

3) 평균 자막 길이: 약 1,200 토큰 (자막 길이 분포 100 – 3,000여 토큰)

4) 언어: 한국어 자막(자동 생성·수동 업로드 혼합)

5) 유효성 검사: 유튜브 채널 Handle - ID간 매칭 확인

6) 원천 데이터 파일 형식 예시:



나. 전처리 순서

1) 중복 제거: 반복 호출로 인한 중복 추출 및 동일 영상 내 중복 자막 블록 필터링

2) 텍스트 정제: HTML 태그·메타정보·특수문자·중복구간·이모지·URL 제거

3) input 삭제: 시청자의 관심 유도를 위한 자극적 소재로 영상 내용과 매칭 불가

4) JSON 삭제: 900 미만의 length는 정보성 없는 shorts로 판단하여 제거

5) output 삭제: 과학 주제와 연관되지 않은 내용의 스크립트 삭제

6) 언어 변환: 한글이 아닌 문자를 한글로 변환

7) output 정규화: 원천 데이터의 일정하지 않은 길이를 2,000 토큰 이하 요약

8) input 생성 및 정규화: output 기반 48 토큰 이하 요약

9) 텍스트 정제: api 호출로 인한 특수문자 생성 제거

10) 중복 제거: api 호출로 인한 중복 생성 output 제거

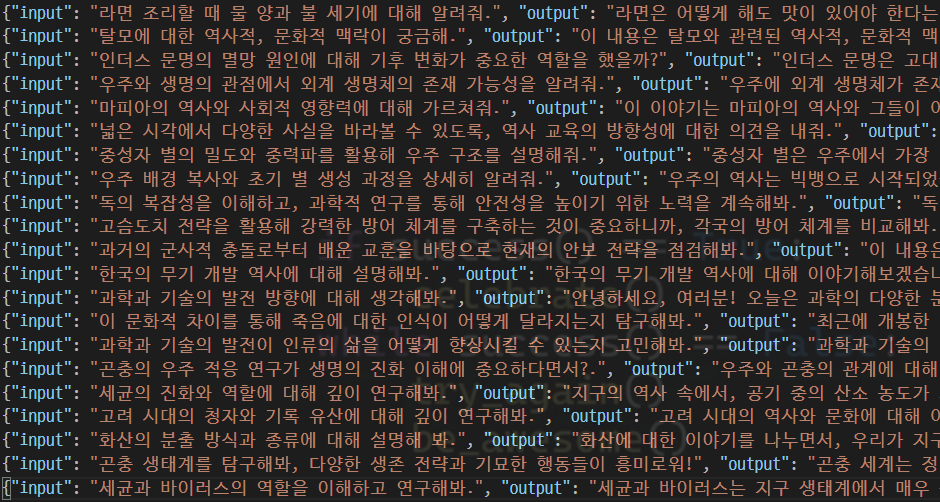
# 4. 전처리 결과 요약

가. 전체 레코드: 6,024건 (연령 제한 및 api 호출 실패 제외)

나. 유효 레코드: 3,697건 (2,327건 제거)

다. 파일 형식: input (최대 48토큰), output (최대 2,000토큰)

라. 파일 형식 예시:



마. 구조적 특징: 토큰 한계 내 정규화를 통한 JSONL 파싱 데이터

마지막으로, 이번 전처리 작업을 통해 HTML 태그·이모지·URL 같은 불필요한 노이즈를 완전히 제거하고, 제목과 자막을 토큰화·정규화(제목 60토큰 이하, 자막 2,000토큰이하)를 적용하여 JSONL 형식으로 일관되게 정리함으로써 전처리를 통한 고품질 데이터셋 확보를 달성하였음. 또한 중복된 레코드 필터링 적용 등을 통해 학습 및 검색 정확도를 높일 수 있는 기반을 마련하였음. 이러한 세부적 정제 과정을 통해 확보된 데이터셋은 향후 모델 파인튜닝 시 과학 AI 인플루언서의 일관된 페르소나 학습과 고성능 예측을 보장하며, RAG 시스템 구축 시에도 안정적이고 재현 가능한 검색·응답 흐름을 지원할 것으로 전망됨.