SK Networks Family AI Camp 9기

**인공지능 데이터 결과서**

가로선

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

**담당자: 조이현**

# <목차>

1. 데이터 수집 목적

2. 데이터 수집 출처

3. 수집 데이터 요약

4. 수집 데이터 형식

5. 전처리 목적

6. 전처리 범위

7. 전처리 과정 요약

8. 전처리 결과 요약

# 1. 데이터 수집 목적

과학 AI 인플루언서 DAIS의 지식 기반 구축을 위해, 유튜브 채널에서 다음과 같은 스크립트(자막 및 설명 텍스트)를 확보하였음. 각 채널 선정 이유는 과학 관련 주제를 친근하고 이해하기 쉽게 풀어주는 특화된 콘텐츠를 제공하며, AI 모델에 다양한 설명 스타일·어투·난이도를 학습시키는 데 최적의 소스라고 판단하였음.

가. 실제 인터뷰·방송 콘텐츠 기반 Q&A 패턴 구축: 방송·인터뷰·온라인 강연 등에서 사용된 질문과 답변을 그대로 재현하여, DAIS의 대화 자연성을 높이고자 함.

나. 과학적 레퍼런스 확보: 과학 콘텐츠에서 핵심 지식을 추출하여 모델의 사실 기반 응답 품질을 강화하기 위함.

다. 다양한 질문 유형·난이도 커버리지 확장: 초보자용 기초 질문부터 심화 학습용 고난이도 해설까지 폭넓은 범위의 Q&A 샘플을 준비하여, 모든 사용자 수준을 만족하기 위함.

라. 컨텍스트 전환 및 대화 흐름 학습: 동일 주제 내 세부 질문 간 맥락 전환 패턴, 다중 턴 대화 예시를 포함하여, 자연스러운 대화 상태 전이를 모델이 익히도록 하고자 함.

마. 페르소나 몰입도 향상: 수집한 언어 표현과 어투 데이터를 통해 ‘친근하면서도 전문적인’ AI 인플루언서 캐릭터를 일관되게 학습시키기 위함.

# 

# 2. 데이터 수집 출처

| 채널명 | 핸들(Username) | 주요 콘텐츠 특징 |
| --- | --- | --- |
| 안될과학 | @Unrealscience | 실험·오류 사례 중심의 ‘실패하는 과학 실험’ 시리즈 |
| 에스오디 | @softdragon | IT·소프트웨어 원리부터 최신 기술 동향까지 폭넓은 과학 이야기 |
| 우주먼지의 현자타임즈 | @wz\_mz | 우주 물리·천문학 주제의 심화 강의 형식 콘텐츠 |
| 지식인미나니 | @iamminani | 일상 속 과학 원리를 만화·애니메이션으로 풀이 |
| 지식은 날리지 | @jisikisknowledge1620 | 퀴즈·문제 풀이를 통해 개념을 점검하는 ‘과학 퀴즈쇼’ 형식 |
| 보다 | @보다BODA | 최신 과학 뉴스·연구 성과를 쉽고 빠르게 정리하는 브리핑 |
| 과학드림 | @ScienceDream | 초중고 교과개념부터 대학 수준 심화까지 단계별 커리큘럼 제공 |
| 한눈에 보는 세상 | @kurzgesagt\_kr | 고퀄리티 애니메이션으로 과학·철학 주제를 시각적으로 설명 |
| 1분과학 | @1minscience | 1분 내외 짧고 강력한 ‘과학 핵심 요약’ 콘텐츠 |
| 과학쿠키 | @snceckie | 생활 밀착형 과학 실험·DIY 콘텐츠 |
| 긱블 | @geekble\_kr | 팟캐스트 스타일 대담과 심층 인터뷰로 과학 이슈를 다룸 |
| 이준석 | @junseoktube | 과학 정책 및 대한민국 기술 트렌드 분석 |

# 

# 3. 수집 데이터 요약

가. 채널 수: 총 12개 채널

나. 영상 수: 채널별 업로드 영상에서 수집된 전체 공개 영상 8,500여 편

다. 스크립트 샘플 수: JSONL 레코드로 환산 시 6,000여 건

라. 평균 제목 길이: 약 8 토큰 (한글 기준 약 20~30자)

마. 평균 자막 길이: 약 1,200 토큰 (한글 기준 약 1,500~2,000자)

바. 언어: 한국어 자막(자동 생성·수동 업로드 혼합)

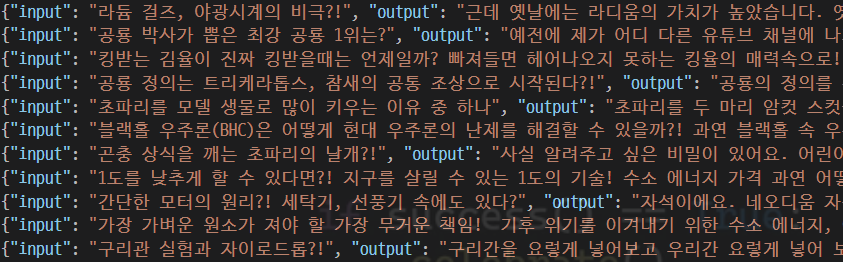
사. 수집 기간: 2025년 7월 28일 수요일 - 2025년 07월 29일 목요일

# 4. 수집 데이터 형식

가. 파일 형식: UTF-8 인코딩된 .jsonl (JSON Lines)

나. 파일 형식 개요: 영상 제목(최대 60토큰), 자막 전체 텍스트(최대 2,000토큰)

다. 파일 형식 예시:



이와 같은 형식으로 학습 데이터셋을 구성할 경우, 파인튜닝 시에는 input/output 페어로 학습 가능하며, RAG 또는 Vector DB 활용으로 용이한 장점이 있음. 이를 통해 두 가지 워크플로우를 모두 지원하는 유연한 구조를 갖추고자 함.

# 5. 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.

# 6. 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

# 

# 

# 

# 

# 7. 전처리 과정 요약

가. 원천 데이터 분석

1) 샘플 수: JSONL 레코드로 환산 시 6,024 건

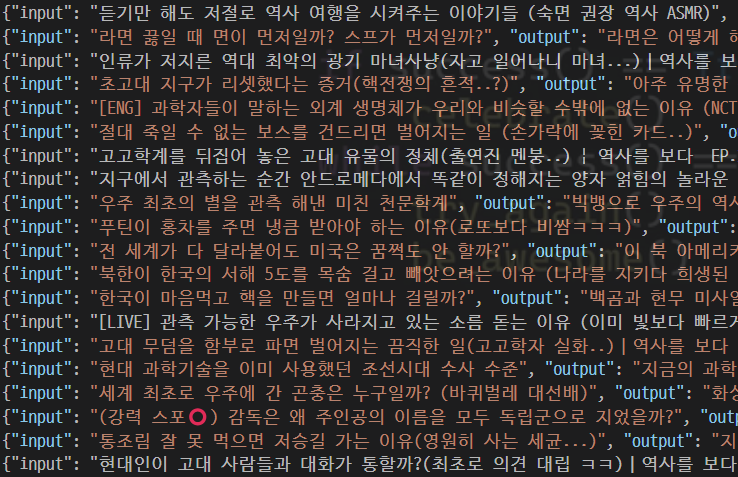
2) 평균 제목 길이: 약 8 토큰 (제목 길이 분포 5 – 80여 토큰)

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

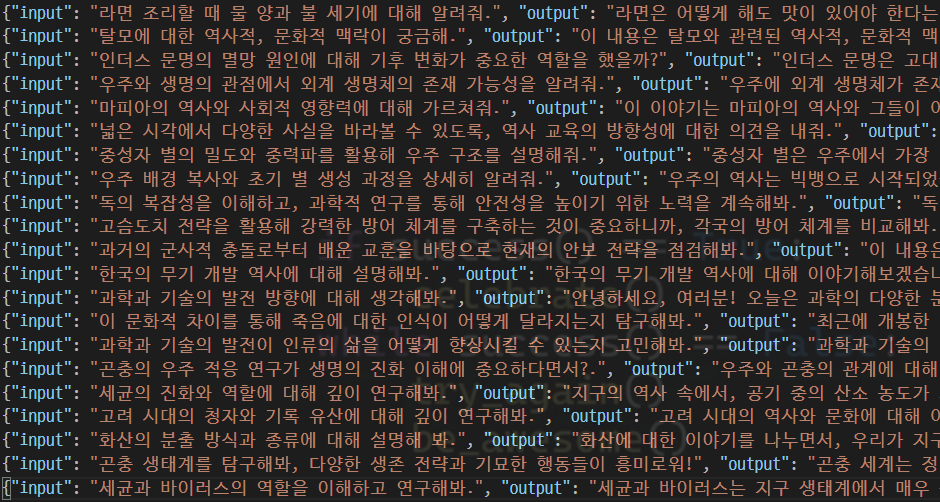
# 8. 전처리 결과 요약

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.