SK네트웍스 Family AI과정 13기

데이터 수집 보고서

산출물 단계	데이터 수집 및 저장
평가 산출물	데이터 수집 보고서
제출 일자	2025-08-01
깃허브 경로	https://github.com/SKNETWORKS-FAMILY-AICAMP/SKN13-FINAL-4TEAM
작성 팀원	박현아, 이재범

1. 개요

1.1 데이터 수집 전략 수립

본 데이터는 '연애 상담 유튜버 콘텐츠 스크립트 데이터'로, 프로젝트 LLM(Love Language Model): 'AI 인플루언서 연애 상담 스트리밍 플랫폼'의 sLLM Fine-tuning 단계에서 사용할 목적으로 수집되었다. 서로 다른 성격과 연애관의 인플루언서가 다양한 시각에서 질문과 사연에 대한 상담을 진행하는 것이 본 프로젝트의 핵심 가치임을 고려했을 때, 여러 sLLM이 고유한 페르소나를 형성할 수 있도록 다양한 톤앤매너 데이터가 요구된다. 이에 4 명의 연애 상담 유튜버의 유튜브 콘텐츠 스크립트 데이터를 수집하였으며, 수집 절차는 다음과 같다.

1) 다양한 페르소나에 대응하는 sLLM을 구축하기 위해, 본 프로젝트에서 요구되는 성격의 유형을 정의하고, YouTube에서 연애 상담을 핵심 콘텐츠로 진행하며 앞서 정의한 성격과 상응하는 유튜버를 탐색한다.
2) 탐색한 유튜버의 콘텐츠에서 음성을 추출하고, 추출된 음성으로 텍스트 형태의 자막을 추출한다. 해당 작업은 Python 프로그래밍을 통해 수행한다.

1.2 수집 데이터 개요

- kimdal_scripts/: 유튜버 "김달(Moon)"의 영상 스크립트 추출 파일
  - 듣기 싫은 말을 부드럽게 해주는 형/오빠
  - 김달 (Moon) - YouTube
- hongcha_scripts/: 유튜버 "홍차 HONGCHA"의 영상 스크립트 추출 파일
  - 직설적으로 현실적인 조언을 해주는 친한 언니/누나
  - 홍차 HONGCHA - YouTube
- shoohee_scripts/: 유튜버 "슈히Shoohee"의 영상 스크립트 추출 파일
  - 상처받은 마음에 공감하면서도 현실을 짚어주며 정신 차리게 해주는 단단한 사람
  - 슈히Shoohee - YouTube
- omar_scripts/: 유튜버 "오마르의 삶"의 영상 스크립트 추출 파일
  - 감정에 휩쓸리기보다 현실적으로 담백하게 통찰하며 위로해주는 차분한 현실주의자
  - 오마르의 삶 - YouTube

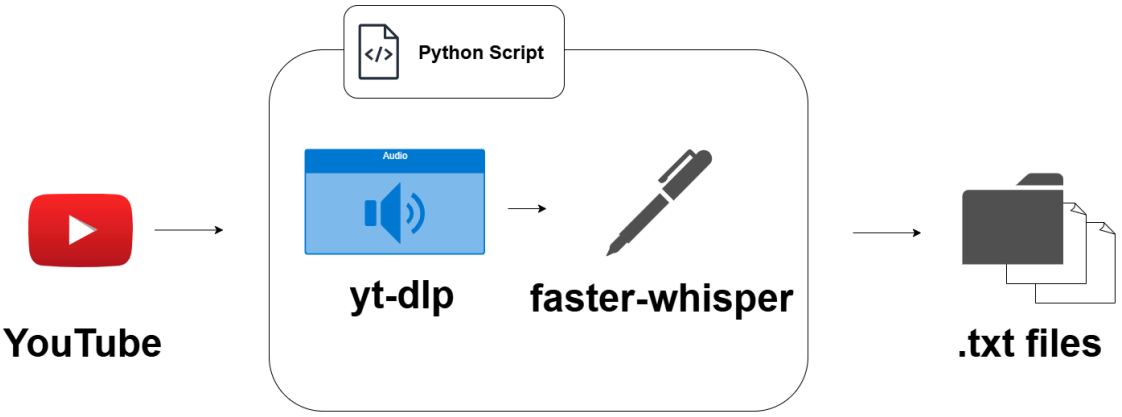
2. 수집 방법 및 자동화 절차

- 수집 방식 (해당 항목에 체크)
    - ○ 웹 크롤링
    - ✔ API 호출
    - ○ 사용자 입력
    - ○ 문서 파일 업로드 (PDF, CSV 등)
    - ○ 기타: _____
- 수집 도구 또는 스크립트 설명:
    - ○ 사용한 언어/라이브러리: python + yt-dlp & faster_whisper
    - ○ 오류 발생 시 예외 처리 전략:
        - ■ Google Colab T4(무료) GPU의 경우 약 4시간의 일일 사용량 제한이 있음. 이에 익일 데이터 수집을 재개해도 이미 수집한 데이터를 중복 수집하지 않는 처리.

```
for i, url in enumerate(video_urls[start_index:], start=start_index + 1):
    print(f"₩n--- [{i}/{len(video_urls)}] 영상 처리 시작: {url} ---")

    downloaded_audio_path = None
    try:
        video_title, video_id = get_video_info(url)
        if not video_title or not video_id:
            continue

        print(f"영상 제목: {video_title}")
        sanitized_title = sanitize_filename(video_title)
        output_filepath = os.path.join(OUTPUT_DIR, f"{sanitized_title}.txt")

        if os.path.exists(output_filepath):
            print(f"'{sanitized_title}.txt' 파일이 이미 존재하므로 건너뜁니다.
            continue
```

        - ■ 톤앤매너 학습 데이터를 구축하기 때문에 2인 이상의 목소리가 섞인 영상은 일관된 톤앤매너를 형성하는 데 노이즈로 작용할 수 있음. 이에 유튜브 채널 선정 단계에서 monologue 콘텐츠 위주의 채널을 선정한 뒤, 2인 이상의 dialogue 콘텐츠는 human annotation 방식으로 제거.

- 예시 스크립트 또는 흐름도 첨부: (이미지, 순서도 또는 코드)



  - yt-dlp
    - 특정 유튜브 채널 url에서 모든 영상 조회
    - 개별 영상에 대한 임시 오디오 파일 추출
  - faster-whisper
    - 추출한 임시 오디오 파일에서 텍스트 추출(STT, Speech-to-Text)
    - 추출한 텍스트를 .txt 파일로 저장
    - 본 단계에서 임시 오디오 파일 제거

3. 데이터 설명 및 구성

3.1 데이터 양

- 전체 수집 데이터 건수: 3558개의 영상

  1020개의 유튜버 홍차의 영상, 1020개의 김달의 영상, 429개의 슈히의 영상, 1180개의 오마르의 영상

- 추출된 고품질 데이터 건수 (필터링 후 기준):

  899개의 유튜버 홍차의 텍스트 파일, 1003개의 유튜버 김달의 텍스트 파일, 354개의 유튜버 슈히의 텍스트 파일, 1110개의 유튜버 오마르의 텍스트 파일

3.2 저장 위치 및 포맷

- 저장 경로: Google Drive: “/content/drive/MyDrive/SKN13-Final-4Team/crawled_data/”
- 저장 포맷: .txt
- 인코딩: UTF-8
- 예시:
    - 파일명: 혹시 금사빠세요 그럼 보고 가세요.txt
    - 내용: 자, 오늘의 할 이야기는 금사빠의 주의사항이라는 이야기예요. 자, 저 같은 금사빠 빨리 손들어. 아니, 저 같은 경우도 사실 진짜 엄청 금사빠예요. 저는 제 취향의 남자만 있으면 가만히 있지를 못해. 제 취향의 남자만 있으면 말을 걸어야 되고 여자친구 있냐고 물어봐야 돼, 내가. 이게 사람을 보는 눈이 길러지는 데까지는 꽤나 오랜 시간이 걸리잖아요. 근데 이 금사빠라는 거는 참 그게 일단 중요하지 않다. 하하하하 일단은 금사빠 본인들 알고 있어야 돼요. 왜 금사빠일까? 그냥 좋으니까. 근데 여러 가지가 있어요… (생략)

4. 데이터 품질 및 정합성 관리 방안

- 표준화 전략:
    - 상기 전략으로 수집된 텍스트 파일 데이터를를 sLLM fine-tuning을 위한 QA 데이터셋으로 전처리 예정.
        - 각 유튜버의 스크립트 일부를 OPENAI: GPT-4.1에게 예시로 제공하여 해당 유튜버의 톤앤매너가 반영된 QA 데이터셋을 만들어내도록 Few-shot Prompting
        - Q: GPT-4.1이 실제 입력으로 받을 스크립트를 기반으로 시청자가 할 만한 질문
        - A: GPT-4.1이 실제 입력으로 받을 스크립트를 기반으로 스트리머가 질문에 대해 답변할 만한 내용. 스트리머의 톤앤 매너를 담고 있어야 함.
    - 데이터셋의 저장 포맷은 .json을 사용하여 일관적인 형식으로 관리할 예정

- 데이터 품질 관리 및 평가 전략:
    - 평가 항목 1. Rationality: 생성된 데이터셋이 실제로 해당 유튜버의 성격과 연애관을 합리적으로 반영하고 있는가?
    - 평가 항목 2. Error Rate: 생성된 데이터셋에 오탈자 혹은 구문론적 오류가 존재하는가?
    - 생성된 데이터셋의 구체적인 평가 방법은 추후 논의를 통해 이끌어 낼 예정.