SK네트웍스 Family AI 과정 10기

데이터 전처리 인공지능 데이터 전처리 결과서

산출물 단계	데이터 전처리
평가 산출물	인공지능 데이터 전처리 결과서
제출 일자	2025-08-10
깃허브 경로	https://github.com/SKNETWORKS-FAMILY-AICAMP/SKN13-FINAL-4TEAM
작성 팀원	박현아

1. 문서 개요

- 프로젝트명: LLM(Love Language Model)
- 전처리 목적: 특정 유튜버(페르소나)의 말투, 화법만을 sLLM이 학습할 수 있도록 1인 화자가 등장하는 데이터만 남기는 것
- 문제 정의: 영상 스크립트에는 해당 유튜버 외에도 다른 화자가 등장할 수 있으며, 이 경우 sLLM 학습 시 톤앤매너가 혼재되어 모델 성능이 저하될 우려가 있음. 따라서, 여러 명의 사람이 등장하는 영상의 스크립트는 전량 삭제하여, 순수하게 1인 화자(페르소나)의 말투만 남기는 방식으로 전처리를 수행함.

2. 데이터셋 개요

- 데이터 출처 및 수집 방법:

  유튜버의 영상 다수를 수집하고, 각 영상을 mp3 등 음성 파일로 변환한 뒤, faster-whisper 라이브러리를 사용하여 자동 전사

- 데이터 구성:

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)

- 원본 데이터 샘플(5~10건 첨부):
  (스크린샷 또는 테이블 형태)

아니 요즘 감사하게도 내 영상 보고 위로벗었다고 해주는 사람들이 많아서 오늘은 내 이별 얘기 한 번 해보려고 내 이별은 생각보다 오래되진 않았어 작년 초였거든? 오랜 시간을 함께한 사람이었는데 둘이 진짜 펑펑을 면서 눈물 가득한 긴 대화를 끝으로 그 사람은 여기로? 나는 여기로 그렇게 갈게를 가게 되었어 진짜 정말 많이 울었어 너무 울어서 그 기간에 찍은 영상이나 사진 보면 눈이 부어있어 와 근데 진짜 너무 힘들더라 근데 나는 유한함의 힘을 믿거든 이 관계도 끝이 있었듯이 이 고통도 끝이 있을 거라고 믿었어 근데 정말 시간이 지나니까 괜찮은 지도라구 인생 살면서 이런 저란 이별을 겪으면서 듣는 생각은 이별은 확된 게 아니라는 거야 난 그 이별 덕에 더 성장할 수 있었고 더 많은 도정도 할 수 있었어 물론 자주 겪고 싶지 않아 그러니까 혹시 작년이 나처럼 아파하고 있다면 진짜 괜찮아질 거라고 그 이별이 너에게 축복 같은 일이 될 거라고 말해주고 싶어 친구야 괜찮아질 거야 진짜로 괜찮아질 거야

그 상대방이 나랑 안 맞는 거지 내가 모든 사람과 안 맞는 게 아니잖아요 고백 타이밍은 언제라고 생각하시나요 고백 타이밍은 저번에도 말씀드렸는데 고백을 100% 성공하는 방법이 있어요 근데 그거를 몰라서 못 하는 게 아니라 안 돼서 못 하는 거거든요 고백을 해서 100% 성공할 수 있는 방법은요 여러분들 스스로 본인 스스로가 내가 상대에게 고백을 했을 때 이 사람은 무조건 승락을 할 거야라고 오는 그 느낌 있어요 초기 그때 고백하면은 99%도 아니고 100% 200% 성공해요 근데 대부분 그거에 대한 자신감도 없고 확신이 없는 거예요 상대가 어떻게 생각하는지에 대해서 내 스스로서 확신이 없는 거예요 그러니까 고백을 할까요 말까요 아니면은 고백을 하면 될까요 안될까요 이런 질문들이 나오는 이유가 내가 고백하면 니가 당연히 받아야지 이런 생각을 가지고 있으면 돼요 근데 그렇게 불확실한 마음을 가지고 어떡하지 어떡하지 이러고 있으니까 고백이 안 되는 거예요 그러니까 고백을 할 때 가장 중요한 거는 고백을 하기 전에 여러분들의 행동 그리고 상대방이 여러분들한테 대하는 태도 그걸 보고 스스로가 확신에 차 있어야 돼요 그리고 고백을 나는 조만간 할 거니까 내가 고백을 했을 때 받아주려면 내가 그만큼 더 최선을 다해서 노력을 해

29살 남자입니다. 약 5년간의 연애 끝에 이별했고, 돌아선 사람의 마음을 이해합니다. 달림 영상 매일 보고 있습니다. 말씀하신 대로 이 다음번 연애는 후회 없이 할 수 있겠죠. 근데 이 사람이 아니라는 게 너무 마음이 아프고 미련이 남습니다. 꼭 극복하고 싶습니다. 제가 뭘 해야 할까요? 영상에서 하신 말들 다 이해가지만 지금 저에게 더 현실적인 조언이 필요합니다. 정신 차리고 싶습니다. 마지막에 영상에서 하신 말들 다 이해가 가지만 저에게 더 현실적인 조언이 필요합니다. 라고 하셨는데 이거는 솔직하게 얘기하면 영상에서 했던 얘기가 다 이해가는 것 까진 맞는데 지금 저에게 더 현실적인 조언이 필요합니다. 이 말은 그 이상의 조언을 구하고 싶은 것 보다는 지금 당장에 나한테 내 얼굴 보고 지금 이 타이밍에 니 입으로 직접 좀 얘기해줘 약간 이런 게 더 큰 거죠. 형이 나한테 얘기해준다면 그게 위로가 될 것 같아요. 약간 이런 거. 자 쉽게 얘기할게요. 만났어 5년 만났어 이까지가 이제 5년의 기간이라고 볼게요. 5년 동안 잘 만났어 여기서 헤어졌다 그죠. 이 헤어진 시점에서 얘기를 하는 겁니다. 여기서 이 시점에서 내 상황은 어떠냐면 이게 아니고 이거에요. 이해가 되요? 보통 일반적인 사람이 그냥 5년을 연애 안 하고 그냥 살아왔어 그러면 그 사람은 일로 가요. 일로 간다고 방향이 그냥 내가 이렇게 살아왔구나. 앞으로는 이렇게 가

그 사람이 이상형을 얘기할 때 혹시나 장난식으로라도 어? 나네? 이 지랄하는 남자가 있다 그런 사람들도 오늘 준비해온 내용은 남자들이 관심있는 여자 앞에서 하는 행동 6가지 정도를 말씀드리려고 준비를 좀 해왔습니다 자주 보시는 분들은 늘 느끼셨겠지만 제일 중요한 한 가지는 맨 마지막에 말씀드린다는 점 건너뛰기 하셔도 괜찮습니다 편하신 대로 하시고 아니면 정리되어 있는 거 보고 그냥 나가셔도 됩니다 단 세부적인 내용은 못 들었으니 후회하는 건 책임지지 않습니다 굉장히 잣게 여성분들께서 이런 질문을 해요 어떤 남자를 알게 됐는데 차를 태워주고 차문도 열어주고 계산도 자기가 하고 이 정도면 호감 아닌가요? 이런 식으로 물어보시는 분들이 계시는데 물론 호감일 수 있죠 가끔씩 방송률 말씀드리는 게 기본적인 매너에 혹하지 마라 이런 얘기를 하거든요 관심이 없어도 모든 관계에서 밉보이는 걸 싫어하는 남자들은 그냥 진짜 단순 매너라고 생각을 해서 아무 의미 없이 몸에 베어서 그냥 해주는 경우도 정말 많습니다 이런 얘기들은 흔하게 들을 수 있는 것들이잖아 그래서 오늘은 제가 진짜 확실한 것들만 위주로 관심있는 여자 앞에서 하는 남자의 행동을 여섯 가지 말씀을 드리겠습니다 자 첫 번째 제일 확실한 거 카톡 메시지가 세 줄 이상 온다 100%입니다 예시를 들어 드릴게요 둘이 약속이 있어서 만나게 됐어 헤어졌네

멍청한 눈 그 눈부터 갈아 끼우세요 여러분들 원할머니 보고쌈님이 말씀을 해주신게 저게 맞는 말이에요 여러분들 그 딱 한가지만 거슬리고 신경쓰이면 만나지 말라는 거지 그러니까 여러분들이 연애를 시작하실때요 중요하게 알고 계셔야 되는게 대부분 경험이 부족할 때는 모든 사람이 다 잘 완벽하게 맞을 수 없으니까 어느 하나는 포기하고 감당하고 만나야 된다라고 생각을 해요 근데 그 말도 일리는 있어요 맞는 말이에요 틀린 말은 아니거든요 근데 중요한 부분이 뭐냐면 여러분들이 생각을 못하시는 것 중에 하나가 연애를 시작하면은 감당해야 되는 부분이 있는 건 사실이에요 근데 그게 감당해야 될 부분이 감당할 만한 가치가 있는 부분이냐를 생각해야 돼요 근데 그냥 싸그리 통틀어서 그냥 어떤 문제라고 그냥 특정 한가지라고 생각하고 이건 내가 감당해야지 이러고 가니까 그걸로 자꾸 트러블이 생기는 거예요 한번 생각을 해보세요 이게 감당할 만한 사회적 통염으로 봤을 때 가치가 있는 부분인가 아니면 누가 봐도 아닌 것 같은데 그냥 그 한 가지가 안 맞는 건가 이 두 가지 중에 한 가지를 생각을 해보세요 그러면 대부분 여러분들은 가치 없는 걸 가지고 자꾸 감당하려고 해요 방금 전에 말씀하셨던 여사친이 많은데 남사친이 많은데

3. 전처리 프로세스 개요

- 전체 흐름도:

1. 페르소나로 선정한 유튜버들의 영상을 크롤링하여 음성 파일로 변환.
2. 변환한 음성 파일을 faster-whisper 를 사용하여 스크립트로 전사
3. 유튜버의 영상 중 해당 유튜버 이외의 인물이 등장하는 영상은 톤앤매너 학습 시 노이즈로 작용할 우려가 있을 수 있어 해당 영상의 스크립트 삭제
4. 해당 유튜버의 말투나 화법 등을 페르소나로 정의하고, few shot prompting을 사용하여 QA 데이터셋으로 변환

4. 세부 전처리 단계

4.1 이상치 처리

- 정의한 이상치 기준: 페르소나로 정의한 유튜버 이외의 인물이 등장하는 영상

- 처리 방식 및 영향:

  sLLM 파인튜닝 과정에서, 설정한 페르소나 외의 말투나 화자가 포함될 경우 학습 노이즈가 발생할 가능성 존재.
  특히, 데이터셋 내에서 여러 사람이 등장하거나 말투가 혼재되는 경우, 페르소나 일관성이 저해될 우려가 있음.
  따라서, 페르소나 일관성 유지를 최우선으로 하여, 화자가 혼재된 스크립트나 타인의 말투가 섞인 구간은 데이터셋에서 제외.
  최종적으로, 단일 페르소나의 말투가 유지되는 선별 스크립트를 기반으로 QA 데이터셋 구축 및 학습 데이터 정제.

4.2 표준화

수집된 텍스트 파일 데이터를 sLLM fine-tuning을 위한 QA 데이터셋으로 전처리
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.

사용한 라이브러리: json.JSONDecodeError, openai

4.3 데이터 변환 및 생성

데이터셋의 저장 포맷은 .json 형식으로 변환

5. 학습/검증 데이터 분리

첫 번째 검증 방법(test dataset 1) – 유사도 평가용

모든 파일에서 추출한 QA 데이터셋 전체를 학습 데이터로 사용한 뒤, 별도의 test dataset 두 개를 구성

첫 번째 test dataset의 목적은 실제 서비스 상황을 반영하여 모델이 생성한 답변과 기준 정답(A)의 유사도를 평가

1. 주요 서비스 상황과 사용 목적에 맞춘 시나리오 정의
   예시: "연애 고민 실시간 상담", "일상 생활 속 고민" 등
2. 각 시나리오별로 대표적인 질문 유형과 예상 대화 흐름을 기획
3. 시나리오별로 적용할 페르소나(유튜버 스타일, 공감·조언 중심, 위트·공감 혼합 등)의 말투와 어투를 세부 정의
4. 정의된 시나리오, 페르소나 말투 예시, 그리고 **학습 데이터와 중복되지 않는 QA 데이터 일부**를 포함하여 Few-shot Prompting 프롬프트를 설계
5. OpenAI GPT-4.1 API를 활용하여 Few-shot Prompting으로 test QA 데이터셋을 생성
6. 생성된 QA 쌍에 대해 자연스러움, 다양성, 실제 사용 환경 반영 여부를 검토하고 필요 시 추가 보정
7. 최종 test QA 데이터셋을 JSON 파일 포맷으로 저장·관리

두 번째 검증 방법(test dataset 2) – 답변 차이성 평가용

두 번째 test dataset의 목적은 동일한 질문에 대해 서로 다른 모델이 생성한 답변의 차이를 평가이를 위해 연애 상황을 '만남', '과정', '이별', '재회' 4단계로 정의하고, 각 단계별 5개의 시나리오를 작성

1. 각 단계별로 시나리오를 정의
   예시 :

{ "relationship_stage": "동아리에서 처음 만난 대학 새내기들, 서로에게 호감이 있지만 아직 연락을 주고받는 사이까진 가지 못한 상태",
"conflict_situation": "상대방이 나에게 관심이 있는지 확신이 서지 않아 먼저 다가가도 되는지 망설여지는 상황",
"question_tone": "설렘과 기대, 그리고 상대방의 마음을 알 수 없어 약간 불안한 톤",
"format_condition": "상대방과 자연스럽게 친해질 수 있는 대화 주제와 접근 방법을 구체적으로 예시로 제시"
}

2. 각 시나리오에서 대표적인 질문 1개를 생성
3. 총 20개의 질문을 JSON 파일 포맷으로 저장·관리

이렇게 구성한 데이터셋을 활용해 동일 질문에 대해 파인튜닝한 sLLM들이 생성하는 답변이 얼마나 상이한지를 평가

6. 전처리 결과 요약 및 평가

- 전처리 후 전체 건수: 3,639개의 텍스트 파일 -> 3,373개의 텍스트 파일
    - 이상치 제거: 총 266건의 텍스트 파일 삭제

- 총 3373개의 텍스트 파일 -> **14,441쌍**의 QA 데이터셋

파일명	**QA 개수**
hongcha_text_preprocessing	3,715
kimdal_text_preprocessing	4,967
moogonghae_text_preprocessing	55
omar_text_preprocessing	4,569
shoohee_text_preprocessing	1,135

- 향후 활용 방안:

    - sLLM의 톤앤매너 일관성을 학습시키기 위한 QA 데이터셋 구축에 사용
    - 이렇게 만들어진 QA 데이터셋을 활용하면, sLLM이 오로지 유튜버 한 명의 스타일만 반영해 일관된 어투와 개성 학습 가능

7. **변경 이력**

변경일	변경자	변경 내용	비고
2025-05-10	홍길동	tag 결측치 '기타' 대체	전처리 안정성 확보 목적