

**SK네트웍스 Family AI 과정 14기  
 데이터 전처리 인공지능 데이터 전처리 결과서**



| **산출물 단계** | 데이터 전처리 |
| --- | --- |
| **평가 산출물** | 인공지능 데이터 전처리 결과서 |
| **제출 일자** | 2025.08.29 |
| **깃허브 경로** | [4팀\_깃허브](https://github.com/SKNETWORKS-FAMILY-AICAMP/SKN14-FINAL-4Team) |
| **작성 팀원** | 공지환 |

1. **문서 개요**

* 프로젝트명 : 4-seed[포-시드]
* 전처리 목적 : 챗봇에 적용될 TTS를 파인튜닝하여 특정 모델의 화법과 억양, 어조 등을 원할하게 TTS에 구현하기 위함
* 문제 정의 : 챗봇에 적용되는 TTS 모델의 파인튜닝을 위한 음성의 metadata 데이터셋 구축

1. **데이터셋 개요**

* 데이터 출처 및 수집 방법:

1. 유튜브에서 영상 추출(y2mate)

2. 영상의 말을 적절한 길이의 세그먼트로 분할

* 데이터 구성:

| **항목명** | **설명** | **예시** |
| --- | --- | --- |
| \_id | 발화자 식별자 | 0001 |
| spk | 발화자명 | Hong |
| language | 언어코드 | KR |
| text | 발화 텍스트 | 이러네. 물에 담가 놔야 돼. 이 소금기를 빼는 거예요. 엄마가 담가 |
| phones | 음소 | ᄋ ᅵ ᄅ ᅥ ᄂ ᅦ . ᄆ ᅮ ᄅ ᅦ ᄃ ᅡ ᆷ ᄀ ᅡ ᄂ ᅪ ᄋ ᅣ ᄃ ᅫ . ᄋ ᅵ ᄉ ᅩ ᄀ ᅳ ᆷ ᄀ ᅵ ᄅ ᅳ ᆯ ᄈ ᅢ ᄂ ᅳ ᆫ ᄀ ᅥ ᄋ ᅨ ᄋ ᅭ . ᄋ ᅥ ᆷ ᄆ ᅡ ᄀ ᅡ ᄃ ᅡ ᆷ ᄀ ᅡ |
| tone | 성조/억양 정보 | 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 |
| word2ph | 음소 단위 매핑 | 1 3 2 7 6 5 5 7 6 1 3 3 1 4 5 2 2 2 1 2 5 5 5 6 1 7 5 1 2 1 1 |

* 원본 데이터 샘플:



1. **전처리 프로세스 개요**

* 전체 흐름도 : ① 수집 → ② 결측치 처리 → ③ 이상치 탐지 → ④ 데이터 분리
* 전처리 파이프라인 요약:

| **단계** | **목적** | **수행 작업** | **사용 도구/라이브러리** |
| --- | --- | --- | --- |
| 결측치 처리 | 누락값 제거 | Null 행 제거, 특수값 대체 | pandas |
| 이상치 처리 | 비정상 데이터 제거 | 단어 수 기준, 이상시간 필터링 | numpy |
| 텍스트 정규화 | 한글 발음으로 변경 | 영문/숫자/기호 등을 한글로 인식할 수 있는 수치로 치환 | MeloTTS의 cleaner.clean\_text() |
| 분리 | 학습/검증 분할 | train:val = 9:1 | 추출한 dataframe에서 분리 |

1. **세부 전처리 단계**

* 결측치 존재 여부: 있음
* 결측 컬럼 및 비율:

| **컬럼명** | **결측률** | **처리 방법** |
| --- | --- | --- |
| text | 5% | 제거/직접 타이핑 |

* 코드 예시 (Python):

| df = df.dropna(subset=['text’']) df.loc[df["\_id"] == "0025", "text"] = "김치에 마늘을 넣어서 이거 조금 끓으면 두부 넣을게요" |
| --- |

1. **이상치 처리**

* 정의한 이상치 기준 : Text의 인코딩 오류(한글/영어로 작성되지 않은 경우)
* 처리 방식 및 영향:

| **항목** | **기준** | **처리 방식** | **처리 수** |
| --- | --- | --- | --- |
| text | 인코딩 오류 | 제거/직접 타이핑 | 2건 |

1. **정규화 및 표준화**

* 텍스트 정규화:  
  + **텍스트를 한글로 변경** : “outdoor” → “아웃도어”
  + **한글을 한글 발음으로 변경** : “있잖아” → “읻짜나”
  + **성조 파악** : 0, 1 …
  + **단어 -> 음소 매핑** : “너무 간단하고 산뜻하게…” -> “1 2 4 4 …”
* 사용 라이브러리 : math, pandas, MeloTTS.text.cleaner

1. **학습/검증 데이터 분리**

* 분리 기준 및 방법 :  
  + 기준 : 무작위 분할
  + 비율 : Train 90% / Val 10%
* 분리 코드:

| val\_n = max(1, math.floor(len(df) \* 0.1))  df\_val = df.sample(val\_n, random\_state=42)  df\_tr = df.drop(df\_val.index) |
| --- |

* 분리 후 건수:

| 구분 | 데이터 수 |
| --- | --- |
| 학습 데이터 | 43건 |
| 테스트 데이터 | 4건 |

1. **전처리 결과 요약 및 평가**

* 전처리 후 전체 건수: 43건 → 41건
* 품질 향상 지표:  
  + 결측값 제거: 2건 정제
  + 이상치 수정: 2건
* 향후 활용 방안:  
  + 챗봇의 TTS 파인튜닝에 사용