# 인공지능 데이터 수집 및 전처리 결과서

**2025**년 **11**월 **17**일

## **1.** 문서 개요

본 문서는 "음성에서 문서로, *AI*가 만드는 회의 자동화 시스템" 구축을 위해 인공지능 학습에 필요한 데이터 수집과 전처리 과정을 정리하기 위해 작성되었다.

본 프로젝트는 음성 인식(Whisper), 화자 분리(Pyannote)를 이용해서 회의 전문을 텍스트로 반환, 요약, 태스크 추출, 이슈 및 안건 추출을 통해 최종적으로 회의록 반환을 목표로 한다. 이를 위해서 한국어 IT 도메인에 특화된 고품질 텍스트 데이터를 이용한 파인튜닝이 선행되어야한다. 한국어 IT 도메인 특화 데이터를 구축하기 위해 수행된 수집·정제·가공 과정을 문서화하였다.

## **2.** 데이터 사용 목적

- 파인튜닝 대상 모델: ***skt/A.X-4.0-Light***
- 선정 이유: 한국어 업무 환경에서의 문맥 이해 능력이 우수하고, 도메인 적응이 용이한 경량 LLM
- 적용 목적:
    1. 복잡한 회의 내용을 도메인 특화 언어모델이 안정적으로 이해하도록 하기 위함
    2. 소프트웨어 개발·기획·기술 PM 등 IT 실무 용어를 정확하게 해석하도록 모델을 조정

## **3.** 원천 데이터 서치

***skt/A.X-4.0-Light*** 모델의 도메인 특화를 위해 아래 네가지 파인튜닝 텍스트 데이터를 수집하였다.

| 구분 | 데이터 유형 | 출처 | 수집량 | 용량 | 비고 |
|---|---|---|---|---|---|
| 텍스트 | 파인튜닝용 용어 사전 | (한국정보통신기술협회) | 41,967개 | 5 MB | IT 용어 크롤링 |
| 텍스트 | 파인튜닝용 용어 사전 | (한국정보통신기술협회) | 79개 | 19 KB | IT 용어 pdf 추출 |
| 텍스트 | 파인튜닝용 용어 사전 | OpenAI GPT 모델 생성 | 744개 | 123 KB | 판교어 정의쌍 |
| 텍스트 | 파인튜닝용 용어 사전 | 크몽 - 사회초년생 및 IT 입문자를 위한 판교어 300선 | 300개 | 317 KB | 판교어 pdf 추출 |

## 4. 전처리

### - 텍스트 정제 (중복 제거 및 hallucination 검증)

IT 도메인 특화 텍스트 데이터를 다음 세 가지 방법을 통해 중복 또는 비의미 항목 제거하였다.

1) 중복 용어 발생 시 대표 용어 하나만 선택
2) 맥락상 IT 도메인과 일치하지 않는 항목 제외
3) OpenAI GPT와 Gemini API 교차 검증을 통한 hallucination 검증 (존재 여부 / 정의 확인)

- 확인 코드 및 결과

```
print("중복된 '첫 단어' 용어:")
for term, occ in term_map.items():
    if len(occ) > 1:
        print(f"\n[용어] {term}  ({len(occ)}개)")
        for idx, q in occ:
            print(f"  - index {idx}: {q}")
```

확인 코드

```
[용어] 스펙  (2개)
  - index 5: 스펙(Spec)이란 무엇인가?
  - index 271: 스펙 아웃(Spec Out)이란 무엇인가?

[용어] 스코프  (4개)
  - index 6: 스코프(Scope)란 무엇인가?
  - index 76: 스코프 크리프란 무엇인가?
  - index 283: 스코프 다운이란 무엇인가?
  - index 284: 스코프 업이란 무엇인가?

[용어] 레거시  (2개)
  - index 19: 레거시(Legacy)란 무엇인가?
  - index 713: 레거시 시스템이란 무엇인가?

[용어] 릴리즈  (4개)
  - index 23: 릴리즈(Release)란 무엇인가?
  - index 335: 릴리즈 노트란 무엇인가?
  - index 554: 릴리즈 캔디데이트(RC)란 무엇인가?
  - index 580: 릴리즈 트레인이란 무엇인가?

[용어] 버퍼  (2개)
  - index 25: 버퍼(Buffer)란 무엇인가?
  - index 556: 버퍼 타임이란 무엇인가?
```

결과

결과에서 확인 가능한 '스펙', '스펙 아웃'과 같이 반복되는 단어 제외, 단어와 의미가 정확하게 일치하는 경우만 제거하였다. 추가로, 똑같은 의미가 중복되지만 한글과 영어로 다르게 표시되는 경우도 확인 후 제거하였다.

### - QA 템플릿 변환

파인튜닝 데이터로 활용하기 위해 질문-답변(QA) 형태로 재구성하였다.

- *"OO*란 무엇인가*?"* 형식의 일관된 질문 생성
- 파인튜닝 데이터셋으로 사용 가능한 형태로 정규화

아래는 각 데이터의 템플릿 변환 전, 후 예시이다.

**1. TTA** 용어사전 웹 크롤링 데이터

- JSON 형태 (용어, 정의)

1) 전처리 전

{"라다 방식": "몇 개의 주파수와 그 주파수의 펄스 시간 위치를 잘 맞추어 정한 상대국만이 교신할 수 있는 무선 통신 방식의 한 가지. 비밀도가 높고, 주파수의 이용도가 좋다."}

2) 전처리 후

{"question": "라다 방식이란 무엇인가", "answer": "몇 개의 주파수와 그 주파수의 펄스 시간 위치를 잘 맞추어 정한 상대국만이 교신할 수 있는 무선 통신 방식의 한 가지. 비밀도가 높고, 주파수의 이용도가 좋다."}

**2. PDF** 용어집 추출 데이터

- JSON 형태 예시 (용어, 정의)

{"question": "검색 증강 생성란 무엇인가", "answer": "대규모 언어 모델(LLM)에 쌓인 데이터와 별개의 외부 데이터를 이용해 답변 정확도를 높여주는 기술"}

```python
if term and definition:
    question = f"{term}란 무엇인가?"
    qa_list.append({
        "page": page_idx,
        "question": question,
        "answer": definition
    })
    print(f"[{page_idx}쪽] {term} → 추출 완료")
```

**3.** 합성 데이터**(OpenAI API) - IT** 기업 실무 기술 용어 **/** 판교어

- JSON 형태 예시 (용어, 정의 - 예문 포함)

1) 전처리 전

{"question": "탭핑이란 무엇인가", "answer": "'간 본다'는 의미로, 본격 진행 전 유관부서의 반응을 살피는 행위. 가볍게 의견을 묻거나 확인한다는 의미이다.", "example": "개발팀에 먼저 탭핑해볼게요."}

2) 전처리 후

{"question": "탭핑이란 무엇인가", "answer": "'간 본다'는 의미로, 본격 진행 전 유관부서의 반응을 살피는 행위. 가볍게 의견을 묻거나 확인한다는 의미이다. 예: '개발팀에 먼저 탭핑해볼게요.'"}

## **5.** 데이터 품질 검증

| 검증 항목 | 검증 내용 | 결과 | 비고 |
|---|---|---|---|
| 필드 완전성 | - 용어 누락 여부<br>- 정의 누락 여부<br>- JSON 구조 검사 | 용어 누락 0건<br>정의 누락 0건<br>JSON 파싱 오류 0건 | 자동 스크립트로 검증 |
| 문장 구조 유효성 | - 단문/무의미 문자 여부<br>- HTML 태그 문자 포함 여부 | 단문 0건<br>무의미 문자 0건<br>HTML 0건 | 정규식 기반 필터링 |
| 중복 체크 | - 동일 용어명 중복 여부 | 용어 중복 0건 | RapidFuzz |