인공지능 학습용 데이터셋 구축 안내서

2021. 2

과학기술정보통신부

NIA 지능정보원
한국지능정보사회진흥원

〈일 러 두 기〉

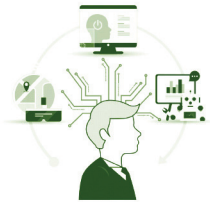
○ ‘인공지능 학습용 데이터셋 구축 안내서’(이하 구축 안내서)는 인공지능 학습용 데이터 구축 사업을 추진하는 수행 및 참여기관이 사업 초기에 수립해야 할 유형별 데이터 구축 가이드라인 작성을 돕기위한 목적으로 가이드라인 작성에 필수적으로 포함되어야 할 내용을 중심으로 안내 함

○ 본 구축 안내서는 텍스트, 음성, OCR 이미지, 영상(동적/정적 이미지) 등 4개 데이터 유형별로 제공하여, 학습 데이터 구축 특성에 맞게 적용할 수 있도록 함

○ 본 구축 안내서와 부록1. 인공지능 학습용 데이터셋 구축 공통참조기준을 기반으로 ‘인공지능 학습용 데이터셋 구축계획서’ 작성 및 구축 업무에 활용해야 함

Ⅰ. 텍스트 데이터

Ⅱ. 음성 데이터



Content

Ⅳ. 영상(동적/정적 이미지) 데이터

Ⅴ. 부 록

I 텍스트 데이터

목 차

표 목 차

그림 목 차

제1장 | 개 요

1 작성 배경

- 인공지능 학습용 데이터 구축 사업 확대에 따라 다양한 역량의 수행 및 참여기관 참여로 사업 진척도 및 데이터 품질의 편차가 발생
- 인공지능 학습용 데이터 품질 향상 및 성공적인 사업 추진을 위해 수행 및 참여기관을 대상으로 데이터 구축 기준, 절차 등 노하우 공유 필요

2 작성 목적

- 인공지능 학습용 데이터 구축에 보편적으로 적용되는 데이터 유형별로 데이터 구축에 필요한 절차 및 구성요소를 제시하여 데이터 구축 과정에서의 시행착오를 줄이고 체계적인 계획 수립을 지원한다.
- 국내 인공지능 학습용 데이터 구축 시 활용된 다양한 가이드라인 사례를 제시하여, 향후 수행 및 참여기관에서 수립해야할 데이터 구축 가이드라인 작성에 참조할 수 있도록 한다.
- 향후 인공지능 학습용 데이터 구축 사업 추진 시 본 구축 안내서를 배포하여 다양한 수행 및 참여 기관의 역량 향상 및 성공적인 사업 수행을 지원한다.

- 궁극적으로 양질의 인공지능 학습용 데이터 구축 및 개방을 통해 국내 인공지능 산업 활성화 및 발전에 기여한다.

3 작성 범위

- 인공지능 학습용 데이터 구축에 필수 공정단계인 데이터 획득·정제·라벨링·검사 단계를 본 구축 안내서의 작성 범위로 한다.
- 텍스트 타입의 원시데이터를 클래스 분류 또는 텍스트 타입으로 라벨링하는 데이터 구축을 본 구축 안내서의 작성 범위로 한다.
- 기존 TTA 인공지능 학습용 데이터 구축 가이드라인, '20년 1차·추경(2차) 인공지능 학습용 데이터 구축사업 수행 및 참여기관의 데이터 구축 가이드라인, 국내 주요 인공지능 전문기업이 인공지능 학습용 데이터 구축을 위해 제작한 인공지능 데이터 구축 지침·가이드 등의 자료들을 검토하여 인공지능 학습용 데이터 구축 시 공통적으로 고려해야 할 사항들을 도출하여 구축 안내서에 반영한다.

4 용어 정의

- 데이터 획득 (Data Acquisition)
  - 인공지능의 기계학습에 필요한 데이터를 현실 세계에서 직접 수집 또는 생성하거나, 이미 보유하고 있는 조직이나 시스템 등으로부터 법률적 제약이 없도록 '원시데이터'를 확보하는 활동

- 데이터 정제 (Data Refinement)
  - 획득한 원시데이터를 기계학습에 필요한 형식으로 맞추거나 불필요한 중복을 제거하며, 개인정보를 비식별화하여 처리하는 등 일련의 전처리 과정을 통해 '원천데이터'를 확보하는 활동

- 데이터 라벨링 (Data Labeling)
  - 인공지능이 기계학습에 활용할 수 있도록 기능이나 목적에 부합하는 정보를 원천데이터에 부착하는 활동

- 라벨링데이터 (Labeled Data)
  - 원천데이터에 부여한 '참값', 파일형식이나 해상도 등의 속성, 그리고 설명이나 주석 등이 포함된 '어노테이션'의 집합

- 원시데이터 (Raw Data)
  - 기계학습을 목적으로 획득 단계에서 수집 또는 생성한 음성, 이미지, 영상, 텍스트 등의 데이터

- 원천데이터 (Source Data, Unlabeled Data)
  - 원시데이터를 라벨링 공정에 투입하기 위해 필요한 전처리 등 정제 작업을 수행한 데이터로 라벨링데이터가 부여되지 않은 상태의 데이터

- 인공지능 학습용 데이터 구축
  - 임무정의, 데이터 획득, 데이터 정제, 데이터 라벨링 등 인공지능 학습용 데이터를 구축하는 일련의 활동

- 참값 (Ground Truth)
  - 인공지능의 기계학습 목적에 따라 원천데이터에 라벨링된 정확한 값이나 사실의 의미적 표현

- 어노테이션 (Annotation)
  - 데이터 라벨링 시 원천데이터에 주석을 표시하는 작업을 의미하며, 추가 부착되는 설명정보 데이터는 기능 목적에 따라 다양한 형태로 표현될 수 있으며 이러한 설명정보 표현방식을 지칭

    ※ 용어사용 예 : 사물 바운딩박스 어노테이션, 클래스 라벨링 어노테이션 등

- 광학문자인식 (OCR, Optical Character Recognition)
  - 사람이 쓰거나 기계로 인쇄한 문자의 영상을 기계가 읽을 수 있는 문자로 변환하는 것

    ※ 자세한 용어 정의는 '인공지능 학습용 데이터 품질관리 가이드라인 Ⅴ.부록- 용어 정의'를 참조

제2장 | 구축 가이드라인 작성 방법

1 데이터 구축 목적 정의

- 데이터 구축 목적 정의
  - 인공지능 학습용 데이터 구축 목적은 단순한 데이터 수집, 모음이 아닌 구축된 데이터를 인공지능 학습 모델에 적용하여 의미있는 수준의 정확도를 확보하고 서비스 등에 유용하게 활용되는 것을 목표로 정의한다.
  - 목적 정의에는 데이터의 구축 배경 또는 필요성, 구축되는 데이터에 대한 명확한 정의, 구축 방향 및 활용(예상) 분야 등을 포함한다.
  - 구축될 학습용 데이터가 실제로 어떤 산업, 서비스, 연구분야에서 활용될 수 있는 지, 정의하여 데이터 구축 방향에 대한 타당성을 재확인한다.
  - 데이터의 저장, 기록이나 해석에서 오류의 가능성이 없도록 명확한 단어, 어휘체계를 사용하여 정의한다.

【참고】 데이터 구축 목적 정의 예시

〈예시 1 – 문서요약 텍스트 구축〉

- 데이터 구축 목적
  - 다양한 한국어 원문 데이터로부터 정제된 추출 및 생성 요약문을 도출하여 검증된 한국어 문서 요약 AI 데이터셋 구축 및 배포
  - 구축된 한국어 문서 요약 AI 데이터셋을 기반으로 추출/생성 요약 AI 알고리즘 개발 및 배포
  - 추출/생성 요약 AI 알고리즘을 이용한 문서 요약 관련 서비스 구현 및 API 배포
- 데이터 구축 필요성
  - 인공지능이 텍스트를 이해하고 핵심 내용을 요약적으로 전달하기 위해서는 AI SW가 해당 텍스트의 주요 내용이 무엇인지를 이해할 수 있는 형태로 가공된, 다양한 유형의 대규모 요약 텍스트 데이터 구축이 필요
  - 국내 인공지능 기반 요약 기술 개발과 관련된 다수의 연구들에서는 해당 텍스트의 제목을 본문의 요약문으로 가정하거나 뉴스 기사의 제목 혹은 첫 문장을 전체 기사의 요약문으로 가정하여 학습데이터로 활용함에 따라 본문 전체의 핵심 내용이나 의미 전달을 온전히 포함하지 못하는 한계점을 내포함
  - 특정 채널에 편향되지 않는 요약기술 개발을 위해서는 채널별로 균형 있는 데이터 원문 수집과 함께, 텍스트 성격에 따라 핵심내용에 영향을 미치지 않는 부분들에 대한 정제 작업이 필수로 요구됨 (중략)
- 데이터 구축 방향성
  - 다양한 한국어 원문 데이터로부터 정제된 추출 및 생성 요약문을 도출하여 검증된 한국어 문서 요약 인공지능 데이터셋 구축
  - 특정 채널에 편향되지 않는 인공지능 요약기술 개발을 위해서는 채널별로 균형 있는 데이터 원문 수집과 함께, 텍스트 성격에 따라 핵심내용에 영향을 미치지 않도록 데이터를 구축
- 데이터 활용 분야
  - 원문 요약 API : 원문을 입력하면 원문의 요약문과 시각화 결과를 제공하는 서비스
  - 고품질 뉴스 기사 판단 및 요약 서비스 : 콘텐츠 품질지표, 구조 명확성, 정보량 등 지표를 기반으로 기사의 품질을 평가하는 서비스

〈예시 2 – 한국어 SNS 텍스트 데이터 구축〉

- 데이터 구축 목적
  - 한국어 구어체 텍스트 기반의 정보검색, 대화분석, 질의응답, 명령어 이해, 언어모델 학습 등의 자연어 처리 AI 기술 개발을 위한 한국인의 일상대화 메신저 채팅 데이터 구축
  - 범용 모바일 메신저인 카카오톡 메신저 대화 원문 수집
- 데이터 구축 필요성
  - 텍스트 자동 요약은 인공지능을 활용한 고차원의 자연어 처리 기술이며 정보의 소스와 채널 다각화와 대량화, 전송의 신속화, 정보 소비 패턴의 변화 등으로 인해 요약이 매우 중요한 서비스로 등장함
  - 코로나19 팬데믹으로 많은 산업 분야가 위축되었으나 언택트(untact, 비대면) 소비가 일상화되면서 교육, 공공 등의 분야에서의 챗봇 도입이 활성화되고 있으나 국내의 챗봇 사례는 대부분 단순 패턴 매칭에 의한 챗봇 서비스(1단계), 지능형 비서 서비스(2단계)에 머물러 있음
  - 챗봇 서비스가 감성 비서 서비스(3단계)에 도달하여 공감 능력을 갖추고 각종 서비스에 선제 대응하기 위해서는 한국인의 구어체 일상대화 내용 요약기술이 필수적으로 요구됨
- 데이터 활용분야
  - 연구 분야: 한국인들이 일상생활 속 메신 저를 통한 텍스트 커뮤니케이션에서 사용하는 대화 방식과 표현 및 어휘를 처리할 수 있는 언어모델 연구 – 정보검색, 대화 엔진, 질의응답, 명령어 이해 등
  - 산업 분야: AI 상담센터, 챗봇, AI 스피커, 개인비서, 스마트홈 등 한국어 구어 자연어 처리 엔진이 필요한 산업

2 데이터 구축 시 고려사항

- 데이터 종류 및 규모
  - 획득해야 할 데이터의 규모를 설정한다. 이때 대상으로 하는 산업 분야 및 서비스에서 요구되는 수준과 사업기간과 획득에 드는 시간과 비용을 종합적으로 고려하여 구축 규모를 선정한다.

- 어노테이션 타입
  - 데이터 활용 분야를 고려하여 구축되는 데이터의 어노테이션 타입을 정의한다.

표 2-1. 텍스트 데이터 어노테이션 타입 및 용도

어노테이션 타입	주요 활용 용도
• 클래스 라벨(단일, 다중)	• 텍스트 분류(Text Classification) ※ 감성, 주제 등
• 단어(구문) 라벨	• 명명된 엔티티(용어, 단어) 인식(Named Entity Recognition)
• 텍스트 라벨	• 문장 번역 • 문장 요약
• 단어(구문) 라벨링 및 두 단어 사이의 관계	• 관계-의존성 정의 (Relation-Dependencies)
• 기타	• 그 밖의 용도

- 데이터 구축 프로세스 정의
  - 데이터 구축 목적 정의, 데이터 획득, 데이터 정제, 데이터 라벨링, 데이터 검사에 이르는 일련의 데이터 구축 프로세스를 사전에 정의하고, 각 프로세스에 따르는 이슈 및 검토사항 등을 도출한다.
  - 데이터 구축 프로세스는 구축 단계별 주요 작업에 대해 서술하나, 순서도·표 등을 활용해 구조화하여 구축 관계자 및 작업자들이 쉽게 이해할 수 있도록 한다.

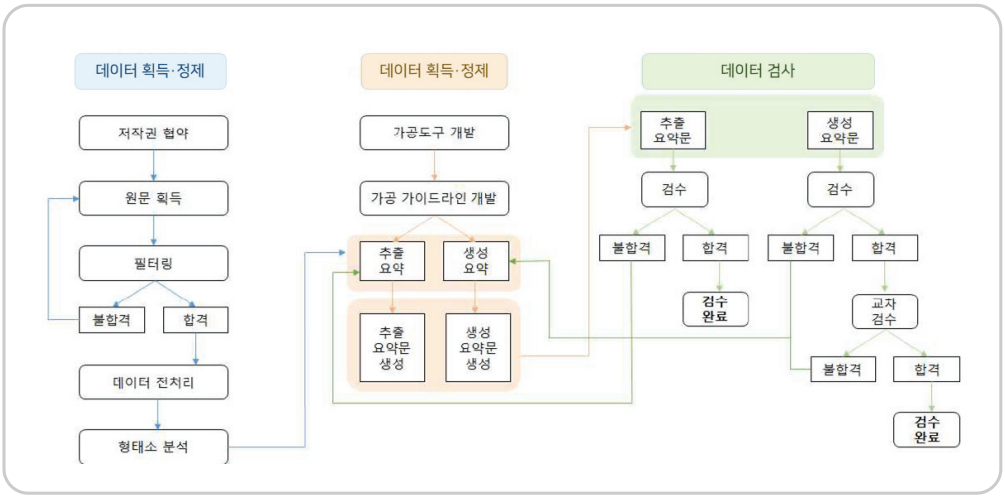


그림 2-1. 데이터 구축 프로세스(순서도 형식) 정의 예시(문서요약 텍스트)

【참고】 데이터 구축 프로세스 정의 예시(한국어 대화 요약 데이터)

- 준비 단계
  - 데이터 제공자와 대화 참여자를 모집함
  - 모집된 데이터 제공자와 대화 참여자에게 개인정보 이용 동의와 저작권 활용 계약 체결을 진행함
  - 데이터 제공을 할 수 있는 온라인 설문 도구(구글 폼)의 접속점을 안내함
- 데이터 획득 단계
  - 데이터 제공자들이 온라인 설문 도구(구글 폼)를 통해 원시데이터를 제공함
- 데이터 정제 단계
  - 데이터 제공자들이 온라인 스프레드시트(구글 시트)를 이용하여 개인정보 비식별화 작업을 수행함
  - 기계적인 절차에 의해 비유효 데이터의 필터링, 화자정보 부착, 형식 변환이 이루어짐
- 데이터 라벨링 단계
  - 데이터 제공자들이 대화 데이터에 대한 주제, 유형의 분류, 요약문 생성 등의 데이터 라벨링 작업을 수행함
- 데이터 검사 단계
  - 교차 검사 방식으로 원시데이터의 형식 요건과 개인정보 비식별화, 요약문을 대상으로 검사를 수행함

[데이터 구축 프로세스(표 형식) 정의]

단계	세부 절차	설 명	산출물
준비	작업 환경 구축	• 작업 도구 선정	
	작업 대상 선정	• 획득할 데이터의 규격 및 조건 선정	
	데이터 제공 기관 검토	• 작업자 모집 기관을 검토	
	작업자 확정	• 원시데이터 작업자 및 제공자와 계약 체결	개인정보수집 및 이용 동의서, 근로계약서, 저작권활용계약서
	작업 지침서 작성	• 작업 지침서 및 가이드 동영상 제작	작업 지침서 가이드 동영상
획득	원시데이터 획득	• 카카오톡 대화문(텍스트) 형태의 원시데이터 획득	원시데이터
정제	부적합 데이터 선별	• 데이터 수집 요건 미충족 대화 제외 • 중복 데이터 제외	요건 미충족 및 중복 제외 데이터
	데이터 비식별화	• 개인정보 마스킹 및 비식별화 • 민감정보 등의 삭제	비식별화 데이터
라벨링	작업 인력 교육	• 데이터 라벨링 작업 교육	
	요약 데이터 구축	• 비식별화 데이터의 요약 작업	요약 데이터
	유형 및 주제 구분	• 요약 데이터의 유형 및 주제 구분 작업	라벨링데이터
검사	요약문 검사	• 요약문 검사 기준 부합 여부 확인	검사 완료 데이터
	외부 기관 품질 인증	• 관련 외부 기관의 품질 인증	품질 인증서

- 데이터 품질 수준
  - 데이터 제작을 의뢰하는 고객이 있는 경우 고객이 요구하는 데이터 품질수준을 기본으로 하며, 세부적인 사항은 협의하여 결정한다.
  - 특정 고객 없이 범용적으로 활용할 수 있는 데이터를 제작하는 경우에도 해당 산업 및 서비스 분야에서 요구되는 품질 수준을 갖

추기 위해, 해당 분야 산업 관계자 및 전문가 등의 검토를 통해 적절한 품질 수준을 설정한다.

- 데이터 활용 목적에 맞는 데이터를 구축하는지 구축 데이터의 시간대, 주제, 비효율성, 인과관계 등을 검토한다.
- 구축되는 데이터는 모집단 및 프로세스에 대한 충분한 정보를 얻을 수 있는 지, 구축되는 도메인의 모집단 또는 프로세스를 대표할 수 있는지 검토한다.
- 품질 수준 및 측정 방법은 논문, 연구, 유사 사업(사례) 등을 통해 객관적이고 명확하게 제시되어야 하며, 기존에 구축된 인공지능 학습용 데이터와 최소한 동일 품질 수준 이상을 갖추는 것을 목표로 해야 한다.

● 데이터 제작 도구

- 구축 대상 데이터가 수행 및 참여기관에서 보유한 구축 도구(소프트웨어)로 목표로 하는 수준으로 제작할 수 있는지 검토한다.
- 제작 도구는 자체 개발한 솔루션을 활용하거나, 타사의 상용 솔루션, 또는 오픈소스 도구를 활용할 수 있으며 이 중 적합한 방법을 선택한다.
- 구축할 데이터의 특성에 맞게 구축 도구에 관한 환경설정을 진행한다.

● 가이드라인 수정 및 이력관리

- 인공지능 학습용 데이터 구축 진행 중 발생되는 예외상황(edge case), 애매모호한 상황 등 데이터 구축 설계단계에서 제작한 가이드라인에서 변경이 필요한 사항이 발생 시, 가이드라인을 업데이트 하고 작업자에게 신속히 배포할 수 있는 방안을 마련한다.
- 구축 과정에서 큰 영향을 미치는 작업방법, 라벨링 세부조건 등에 관한 변경사항 발생 시 고객사 또는 각 산업 전문가 및 관계자 등의 검토·협의를 통해 적합한 방법을 도출할 수 있도록 변경 검토 절차를 마련한다.

※ 변경사항에 대한 검토 및 배포가 제대로 이루어지지 않을 때, 최악의 경우 라벨링을 다시해야 할 수 있음

- 작업 및 검사 인력 운영 방식
  - 대량의 데이터를 구축해야 하는 인공지능 학습용 데이터 특성 상 필요한 작업자 수와 수행 및 참여기관의 시설·노동 환경을 고려하여 내부조직, 아웃소싱, 크라우드소싱 또는 혼합 방식 등 적합한 작업자 운영 방식을 선정한다.
  - 작업자 관점에서 데이터 제작 과정에서 발생할 수 있는 다양한 유사사례 및 예시 등을 포함하여 매뉴얼을 제작한다.

【참고】 크라우드소싱 작업인력 운영 방식 수립 사례

- 기본 원칙
  - 입사 후 7일간 수습 신분으로 Half time (주20시간) 근무
  - 성실성 등 근무태도 평가 후 단기계약 가능
  - 계약 후 한달 근무 완료 시 아래의 세 가지 근무 타입 중 선택

[크라우드소싱 작업자 근무 타입별 운영방식 및 특성]

업무 장소	근무 유형	운영방식 및 고려사항
사무실 (인하우스)	하프타임	• 13시~18시 근무 • 단시간 집중도 높은 근무 • 전체 근무자 중 64.8% 비중
	풀 타임	• 09시~18시 근무 • 작업 이해도 및 작업 역량 단기간 내 확보 가능 • 매니저, 리더 포함 근무자 중 35.2% 비중
재택	자율근무	• 일정 퀄리티 이상의 작업 결과에 대해 건별로 보수 지급 • 작업 내용에 대한 피드백 지연 발생 • 인하우스 근로자 대비 퇴사율이 높아 계약 베이스로 운영하지 않을 경우 리소스 예측과 관리가 어려움

표 2-2. 작업자 운영 방식 특성 비교

구분	내부조직	아웃소싱	크라우드소싱 (crowd-sourcing)
특징	• 품질에 대한 상시 교육 및 피드백 가능 • 작업환경을 위한 운영비, 작업공간 및 인프라 필요(전기·통신시설, 컴퓨터 등)	• 높은 업무 전문성 및 경험 보유 • 요구사항 정의 및 기준 합의에 많은 시간 소요	• 높은 업무 접근성(장소 제약없음) • 품질 교육 및 피드백에 한계가 있음 • 크라우드소싱 대가산정에 대한 명확한 기준 마련이 어려움
적합 용도	• 머신러닝 훈련에 대한 높은 수준의 이해가 필요한 작업 • 라벨링 결과에 대한 공정간(수집, 정제, 라벨링, 학습모델 등) 긴밀한 피드백을 요구하는 작업	• 데이터 구축에 전문적인 지식과 숙련도가 요구되는 작업	• 단기간에 대량의 데이터를 처리해야 하는 작업 • 작업 난이도가 비교적 낮고, 데이터 보안수준이 낮은 작업

- 작업자 대상 매뉴얼 작성
  - 데이터 획득·정제·라벨링·검사 단계에 참여하는 작업자들이 인공지능 학습용 데이터셋 구축 취지에 부응하여 데이터 제작이 이루어질 수 있도록 작업자들이 직접 활용하는 매뉴얼을 제작한다.
  - 작업자 대상 매뉴얼에는 구축 목적·정의, 제작 절차, 제작 도구 활용방법과 작업 기준, 작업 결과 처리·저장 방법 등의 내용을 포함한다.
  - 작업자 관점에서 데이터 제작 과정에서 발생할 수 있는 다양한 케이스를 포함하여 매뉴얼을 제작한다.

3 데이터 획득 및 정제 방법

3.1 데이터 정의

- 원시데이터 정의
  - 인공지능 학습용 데이터 구축에 필요한 원시데이터 항목을 검토하고, 각 항목 별로 데이터 획득에 필요한 정보(데이터 획득정보, 획득방법, 획득 단계에서 필요한 요건 등)들을 검토하여 문서화 한다.
  - 원시데이터 대상 및 획득방법을 아래와 같이 육하원칙에 따라 정의할 수 있다.

표 3-1. 원시데이터 획득 시 검토사항 및 예시

5W1H	항 목	예 시
What	• 측정대상 • 획득 시 포함되어야 할 변수들	• 사회적으로 많이 활용·언급되는 상식, 자식 텍스트 • 시간별, 주제별, 지역별. 매체별 검토 (필요시 도메인 전문가, 인공지능 전문가 협의 후 대상 객체를 명확히 함)
When	• 획득 기간 (From, To)	• 뉴스 데이터(1주, 11.14 ~ 11.20), 인터넷 커뮤니티 데이터(2주, 11.21~12.4), 법률 데이터(1주, 12.5~12.11)
Where	• 획득장소 / 프로세스	• 00 주식회사 데이터팀 내 데이터 수집 서버
Who	• 획득 담당자 / 획득하는 사람	• 00 주식회사 데이터팀 데이터 수집 담당자
How	• 획득 방법, 측정주기, 샘플 크기, • 데이터 양식	• 뉴스, 커뮤니티, 법률 분야별 데이터를 제공하는 기관(기업)의 데이터 수집 API 신청 후 수집 서버와 연결하여 수집하며, json 파일에서 메타데이터는 DB로, 본문은 txt 파일로 변환하여 저장
Why	• 측정 목적 / 기대 결과	• 목적에 맞는 획득 데이터 이해와 프로세스 능력의 파악 / 추세분석

  - 획득할 원시데이터 내역에 대한 정의 및 현황정보 등의 사항을 정리한다.

표 3-2. 원시데이터 현황정보 작성 예시(문서요약 텍스트)

원시데이터 종류	신문	기고문	잡지	법률	...
자료 형태	뉴스 텍스트	오피니언 텍스트	웹진 기사 텍스트	법원 판결문, 뉴스 텍스트	...
자료 규모	30만 건	3만 건	1만 건	3만 건	...
목적	기본요약 알고리즘 확보	의견제시 요약 확보	장문 요약 및 문법적 다양성 확보	실용문 핵심 사실관계 추출	...
확보방안	언론사 저작권 협의	-	-	법률뉴스, 법원 사이트 등	...

- 원시데이터 포맷
  - 원시데이터의 파일 형식은 특정 수집 장비 및 처리 도구에 종속되지 않으며, 보편적으로 통용되는 포맷을 활용한다.
    ※ hwp, docx 등 특정 워드프로세서에서만 호환되거나 pdf 등 기계 가독이 어려운 포맷은 배제
    ※ 텍스트 인코딩은 특정 OS, 특정 프로그램이 아닌 보편적으로 활용되는 UTF-8 인코딩을 준수

- 원시데이터 획득 규모
  - 원시데이터 획득 후 정제, 라벨링, 검사 과정에서 기준 미충족으로 버려지는 데이터 양을 고려하여 구축 목표치 이상의 데이터를 획득하도록 계획한다.
    ※ 구체적인 목표치 대비 획득량은 데이터 구축 공정 난이도 및 구축 기간 등을 고려하여 설정

3.2 획득 데이터 특성 분석

- 원시데이터 획득 관련 이슈사항 도출
  - 획득할 원시데이터의 범위 및 방법을 명확히 하기 위해 데이터 규모·획득범위·수집처 등에 대한 세부 이슈사항을 도출하여 가이드라인에 기술한다.

【참고】 원시데이터 특성 분석 예시(문서 요약 데이터)

〈예시 1 – 문서 요약 데이터〉

- 이용 기간의 제한이 없이 영구적으로 저작권이 확보된 40만 건의 원문데이터를 확보
  - 신문기사
    1) 신문기사는 요약 알고리즘의 핵심적인 데이터로서 10개 언론사로부터 30만 건을 획득
    2) 현재 226종 신문 지면, 67종의 잡지 지면, 2,457종의 온라인 매체로부터 뉴스 텍스트를 수집하여 원문 데이터로 구축하고 있으며, 이중 종합면 30%, 정치 20%, 경제 20%, 사회 20%, 문화 및 스포츠 기타 10%의 비율로 원문 기사를 확보함
  - 기고문
    1) 기고문은 사실관계의 전달에 중점을 둔 일반적인 뉴스 텍스트와 달리 개별적인 주장을 담고 있는 형태의 문서로서 신문의 오피니언 면을 통해 확보
    2) 다양한 형태의 주장이 나올 수 있기 때문에 특정 매체에 대한 집중도를 줄이고 정치/경제/사회/문화/과학 등 다양한 주제를 균등하게 배분하여 학습 데이터를 구축

〈예시 2 – 한국어 대화 요약 데이터〉

- 부적절하거나 자연스럽지 않은 내용 제거
  - 혐오, 차별적 내용, 선정적인 내용, 반사회적 등을 포함한 대화는 데이터셋에 포함하지 않음
  - 데이터 구축 가이드라인에서 대화 진행 시 대화의 목적이 본 과제를 위한 데이터 구축임을 의식하지 않도록 안내하더라도 해당 내용이 대화에 포함될 경우 해당 내용을 삭제하고 해당 대화가 대화 요건을 충족하면 데이터셋에 포함함
- 요건 비충족 또는 불필요 발화 제거
  - 이모티콘으로만 구성되었거나 ㅋㅋㅋ, ㅎㅎㅎ 등이 과도하게 사용된 발화 삭제
  - 대화방 출입, 이미지 등의 콘텐츠 공유, 광고 메시지 등으로 이루어진 발화 삭제
  - 위 기준에 따라 삭제된 내용이 포함된 대화의 요약 모델링에 제약이 있을 수 있음
- 개인정보 비식별화
  - 대화에 포함된 인명 등의 개인정보는 모두 비식별화되므로 개인정보가 포함된 대화의 요약 모델링에 제약이 발생할 수 있음

- 이모지의 처리
  - 메신저를 이용한 대화에는 텍스트뿐만 아니라 사진 등의 이미지, 이모지 등이 광범위하게 사용됨
  - 대부분의 이모지는 대화를 텍스트로 변환시 소실되며 최종 데이터에서 대화에 포함된 이모지를 플레이스홀더(place holder) 문자열로 대치하므로 대화 모델링에 있어 제약이 발생할 수 있음

- 원시데이터 적합성 검토
  - 원시데이터 항목별 데이터 획득 방법, 법적문제 발생가능여부 등을 검토하여 실제로 인공지능 학습용 데이터 구축에 활용할 수 있는 데이터를 선정한다.

- 원시데이터 선정
  - 데이터 품질, 획득 가능성(가능여부 및 획득량), 획득 비용, 수행 및 참여기관의 기술수준, 법적 요건 등을 검토하여 획득할 데이터를 최종 선정한다.
  - 선정된 원시데이터를 획득하기 위해 필요한 정보, 또는 원시데이터 획득현황을 파악하기 위한 데이터 명세서 또는 정의서를 작성하여 데이터 획득 기준으로 활용한다.

표 3-3. 원시데이터 명세서 작성 예시(문서 요약 텍스트)

데이터명		문서 요약 텍스트 AI 데이터
데이터 포맷		txt(텍스트 파일)
활용 분야		뉴스기사 요약, 법률분서 요약, 사업보고서 요약 등 핵심내용을 신속하고 정확하게 파악할 수 있는 인공지능 요약기술 개발에 활용
데이터 요약		다양한 한국어 원문데이터로부터 정제된 추출 및 생성 요약문을 도출하고 검증한 한국어 문서요약 인공지능 데이터셋
데이터 출처		전국종합일간, 지역종합일간, 경제일간, 스포츠일간, 전문일간, 전문주간 등 60개 언론매체로부터 신문 (30만 건), 기고문 (3만 건), 잡지 (1만 건), 법률 (3만 건), 논문 (3만 건)의 원시데이터 확보
데이터 이력	**배포버전**	TextSummaryAIDataSet_ver1.
	개정이력	신규
	작성자/배포자	수행기관(000)
데이터 통계	**데이터 구축 규모**	원문 총 40만 건, 요약문 총 80만 건 (추출요약 40만 건/생성요약 40만 건)
	데이터 분포	매체별 분포 : 전국종합일간(45%), 지역종합일간(12.5%), 경제일간(12.5%), 전문일간(12.5%), 잡지(2.5%), 판결 해설문(7.5%), 논문(7.5%) 주제별 분포 : 신문-종합(22.5%), 신문-정치(9.3%), 신문-경제(9.3%)...(중략)... , 기고문(7.5%), 잡지-시사(1.25%), 문화예술(0.75%)...(중략)...
기타 정보	**대표성**	별도문서 참고
	독립성	
	유의사항	
	관련 연구	해당사항 없음

3.3 획득 절차 및 항목

- 데이터 획득·정제 절차 수립
  - 원시데이터 획득 및 정제 절차 수립 시 데이터 획득 방법별로 명확하게 획득·정제 절차가 정의될 수 있도록 한다.

    1) 원시데이터 직접 제작

    ※ 녹취, 필사 등

    2) 수행 및 참여기관 내·외부에 있는 데이터 수집

    ※ API, 크롤링, 직접수령 등
  - 데이터 관점 뿐만 아니라, 기관간 역할, 작업자 업무, 작업자-관리자 간 관계, 행정요소 등 사람 관점에서 실질적인 구축작업에 필요한 사항을 종합적으로 고려하여 절차를 수립한다.

표 3-4. 데이터 획득 방안 정의 예시

	데이터 획득 형태	수집장비	데이터 형식	수집처(장소)	담당 인원
1	API	수집 서버	JSON → TXT	000공사 API	과제 인력 및 크라우드소싱 인력
2	웹 크롤링	크롤링 서버	여러 텍스트 포맷 → TXT	00 포털 사이트 (oo.com)	000사 크롤링 담당자

- 텍스트 데이터 획득방법·절차 수립
  - 데이터 직접 제작 시* 데이터 제작 범위, 제작과정 및 지침, 등록 및 저장방법 등을 중심으로 획득방법·절차를 수립한다.

    ※ 크라우드소싱 방식으로 데이터를 수집하는 경우에도 해당
  - 외부 데이터 수집을 통한 데이터 획득 시 데이터 수집 범주, 수집량, 수집처별 수집방법, 저장방법 등을 중심으로 획득방법·절차를 수립한다.

【참고】 원시데이터 획득 절차 예시(한국어 대화 텍스트(크라우드소싱))

- 원시데이터의 획득 방안
  - 데이터 획득 협력 기관의 회원들을 대상으로 크라우드소싱에 의해 크라우드 워커를 모집
  - 크라우드 워커 모집 단계에서 개인정보 수집 및 이용 동의와 저작권 활용 계약을 체결함
  - 대화 요약 데이터 지침에 따라 대화 데이터의 기준, 대화 진행 방법, 데이터 제출 방법 등의 교육을 진행
  - 대화 참여자 전원의 개인정보 수집 및 이용 동의 및 저작권 활용 계약 체결(아래에서 상술)이 이루어졌으며 대화의 규격 및 조건이 충족되는 경우 과거 대화 데이터에 대화 주제 분류 부가하여 제출하는 것을 허용
  - 신규 대화를 생성한 크라우드 워커는 대화 주제 분류와 함께 추출된 텍스트 형식의 대화 데이터를 데이터 수집 관리자에게 제출
  - 과제 초기 원시데이터의 수집은 구글폼(Google Forms)을 활용
  - 데이터 수집 관리자는 제출된 대화에 참여한 대화 참여자의 개인정보 수집 및 이용 동의, 저작권 활용 계약 체결, 화자 정보 제공 여부 등을 확인

- 인공지능 알고리즘 편향 방지 및 다양성 확보를 위한 데이터 획득 방안
  - 대화의 주제 분류에는 개인, 주거, 여가, 여행, 상거래, 공공서비스, 전문지식 등 16개 분야가 포함됨
  - 본 과제에서는 대화 주제별 편향을 방지하기 위해 상기 16개 분야가 전체 데이터 세트의 3% ~ 9% 범위가 되도록 할당
  - 과적합 방지를 위해 크라우드 워커별 작업 할당량 상한선을 정함
  - 데이터 검사 과정에서 화자 정보에 대한 정량 지표(참여자 수, 참여자 성별 및 연령대)를 활용하여 화자 특성에 따른 편향을 제어함

- 구글폼(Google Forms)과 구글 스프레드시트를 활용한 획득
  - 크라우드 워커가 구글폼으로 작성된 대화문 제공 신청서에서 이름, 연락처, 전자우편 주소, 계좌번호, 발화자 정보를 입력하여 제출함.
  - 신청서 작성을 완료한 크라우드 워커에게 개인정보 수집 및 이용 동의 계약서, 근로 계약서 및 저작권 활용 계약서를 온라인 서명 플랫폼을 통해 발송함
  - 계약서 서명을 완료한 크라우드 워커는 카카오톡 대화문 등록 구글폼 페이지에서 이름, 연락처, 전자우편 주소, 성별, 연령, 거주지를 입력한 후 원시데이터에 해당하는 대화 원문을 제출함
  - 발화자의 카카오톡 대화명과 연락처를 함께 제출하게 하여 어노테이션에 활용함
  - 구글폼에 입력한 정보들은 실시간으로 구글 스프레드시트에 응답 결과로 저장되어 관리자가 확인 및 데이터 가공이 가능함

【참고】 원시데이터 획득 절차 예시(문서요약 텍스트(외부 데이터 수집))

- 신문기사
  - 신문기사는 요약 알고리즘의 핵심적인 데이터로서 10개 언론사로부터 30만건을 획득
  - 현재 226종 신문 지면, 67종의 잡지 지면, 2,457종의 온라인 매체로부터 뉴스 텍스트를 수집하여 원문 데이터로 구축하고 있으며, 이 중 종합면 30%, 정치 20%, 경제 20%, 사회 20%, 문화 및 스포츠 기타 10%의 비율로 원문 기사를 확보함
- 기고문
  - 기고문은 사실관계의 전달에 중점을 둔 일반적인 뉴스 텍스트와 달리 개별적인 주장을 담고 있는 형태의 문서로서 신문의 오피니언 면을 통해 확보
  - 다양한 형태의 주장이 나올 수 있기 때문에 특정 매체에 대한 집중도를 줄이고 정치/경제/사회/문화/과학 등 다양한 주제를 균등하게 배분하여 학습데이터를 구축
- 잡지
  - 잡지의 경우 전문성이 뚜렷하고 텍스트의 길이가 긴 편이므로 1차 학습데이터 구축에 있어서 1만 건 수준으로 제한
  - 시사/경제, 공학/기술, 문화/라이프, 예술/엔터테인먼트, 요리/건강, 취비/레포츠, 컴퓨터/인터넷 등 7개의 카테고리로 구분
- 법률
  - 판결문은 방대한 분량과 만연체적 특성으로 인해 요약에 어려움이 있으므로, 기자를 통해 판결의 요지가 정확히 제시된 뉴스 형태 및 실제 법원이 판결 취지를 요약해서 제공하는 공보성 게시글 (각 법원의 '우리법원의 주요판결'과 공보판사가 작성한 '보도자료' 등)을 원문으로 수집
  - 이러한 자료들은 공공에 공개되는 자료로서 저작권 활용에 제약이 없음
  - 각 법원의 판결 데이터를 전자적으로 크롤링하여 1차 가공 후 개별 로컬 DB에 구축
  - 각 판결문마다 해당 법률 조항을 메타정보로 포함시킴으로써 향후 요약 내용 및 판결 차이의 원인 분석 등에 활용
- 논문
  - 논문의 경우 '초록 (abstract)'을 원문데이터로 하여 요약문을 추출하는 방식으로 진행
  - 논문 초록의 저작권은 출처를 명시하는 경우 자유롭게 사용 가능하므로 저작권의 이용 권한 내에서 활용하며, 논문 저작권을 유통하는 한국학술정보원과 협의를 통해 논문의 초록 데이터를 API로 제공받아 사용
  - 수식이나 용어가 난해한 자연과학 분야가 아닌 인문/사회과학이나 예술 계통의 논문을 1차 학습데이터 구축 대상으로 선정

- 데이터 획득항목 정의
  - 획득단계에서 텍스트 문장과 함께 확보해야할 정보를 정의한다.
    1) 텍스트 메타데이터 : 제목, 텍스트 길이, 생성일 등
    2) 도메인 정보 : 주제, 매체유형, 획득처, 문장유형 등
  - 텍스트 데이터 획득 시 수집 및 저장할 정보는 **'부록1. 인공지능 학습용 데이터셋 구축 공통참조기준'**을 준용하여 정의한다.
    1) 라벨링 공통참조항목은 텍스트 데이터 획득 시 공통적으로 기록해야할 정보이다.

표 3-5. 텍스트 데이터 획득 시 라벨링 공통참조항목

No.	속성명	항목 설명	Type	필수여부	작성예시
1	Dataset.identifier	데이터셋 식별자	string	필수	TEXT_QnA_LAW_01 (데이터유형_목적_분야_순번)
2	Dataset.name	데이터셋 이름	string	필수	법률 관련 인공지능 질의응답 학습용 데이터 셋
3	Dataset.src_path	데이터셋 폴더 위치	string	필수	/dataSet/text/
4	Dataset.label_path	데이터셋 레이블 폴더 위치	string	필수	/dataSet/text/
5	Dataset.category	데이터셋 카테고리	number	필수	0: 텍스트 분류, 1: 문서요약, 2:질의응답, 3: 기계번역 등
6	Dataset.type	데이터셋 타입	number	필수	0: 텍스트, 1: 이미지, 2:영상, 3: 음성 등

2) 라벨링 선택항목은 '문서요약' 및 '문서분류' 등 원시데이터 출처 정보가 중요한 경우 선택적으로 기록할 수 있는 정보이다.

표 3-6. 텍스트 데이터 획득 시 라벨링 선택항목

No.	속성명	항목 설명	Type	필수여부	작성예시
1	info.filename	원시데이터 파일명	string	선택	NEWS_000001 (매체유형_순번)
2	info.title	원시데이터 제목	string	선택	이스라엘 75세 남성 화이자 백신 접종 후 사망… "백신 연관성 없는 듯"
3	info.mediatype	매체유형	string	선택	뉴스, 블로그, SNS 등
4	info.medianame	매체명	string	선택	중앙일보
5	info.category	원시데이터 카테고리	string	선택	정치, 경제, 연예, 스포츠 등
6	info.size	원시데이터 크기 (글자수)	string	선택	270
7	info.date	발행일자	string	선택	2020.12.29 12:40:23 (yyyy.MM.dd HH:mm:ss)

3) 라벨링 선택항목(저작권 정보)은 데이터 획득 시 저작권 정보가 필요한 경우 기록하는 정보이다.

표 3-7. 텍스트 데이터 획득 시 라벨링 선택항목(저작권 정보)

No.	속성명	항목 설명	Type	필수여부	작성예시
1	licenses.id	라이센스 고유 번호	string	선택	http://www.apache.org/licenses/LICENSE-1.0
2	licenses.name	라이센스 이름	string	선택	Apache License 1.0
3	licenses.url	문서 식별자	string	선택	NEWS_000001

- 획득 데이터 저장 및 관리
  - 획득 파일에 대한 저장, 전송, 백업 등 관리 절차 및 방안을 수립한다.

【참고】 획득 데이터 저장 방안 수립 예시

- API로 수집한 데이터는 수집 서버, 외장하드, 클라우드에 3중 백업 진행
  - 외장하드 고장에 대비하기 위해 NAS* 등의 추가 장비를 활용하여 주기적으로 백업
- 촬영자 관리 및 촬영분 일일 관리를 위해 직접 촬영 데이터를 “촬영일자”_“촬영자” 기재한 폴더 형태로 구성하여 원시데이터 저장

  * NAS(Network Attached Storage) : 네트워크 결합 스토리지

  - 획득한 파일을 체계적으로 분류하기 위해 데이터 종류 및 분류에 따른 라벨링데이터 파일 명명법과 파일 저장구조를 정의하고, 정의된 내용에 맞게 파일을 저장한다.

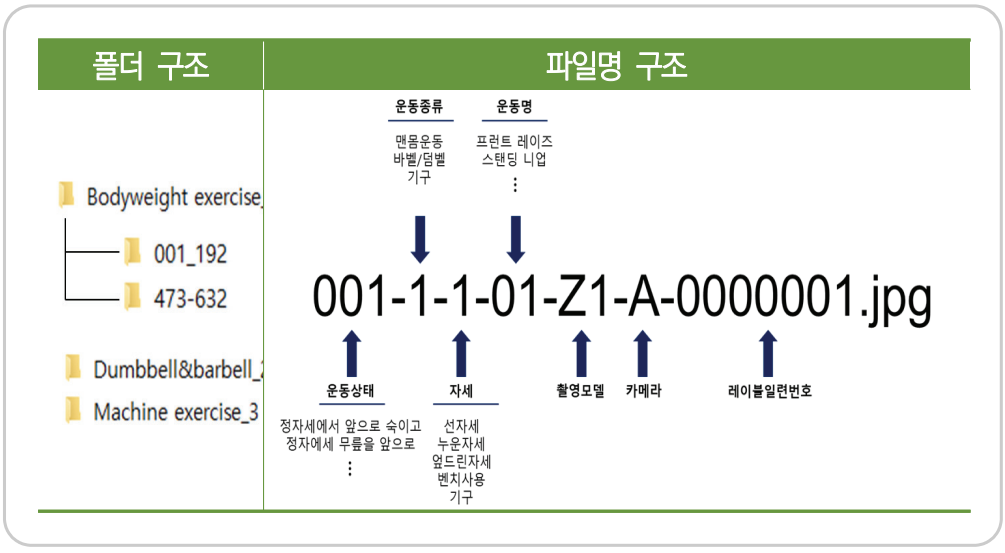


그림 3-1. 획득 데이터 폴더구조 및 파일명 코드화 예시

3.4 획득 데이터 정제 방식

- 정제 프로세스 수립
  - 어노테이션 단계에 들어가기 전에 학습용 데이터로 적합한 데이터를 선별하고 처리하는 정제 프로세스를 획득방법별로 수립한다.
  - 데이터 정제는 도구(소프트웨어)를 활용하여 정해진 규칙에 따라 제외 또는 변환하는 방법, 작업자가 직접 눈으로 확인하여 검사하는 방법 등을 적용할 수 있다.

- 정제 기준 수립
  - 데이터 구축 목적, 데이터 유형, 도메인 특성에 따른 데이터 정제 기준을 수립한다.
  - 텍스트 분량, 텍스트 문법의 정확성, 텍스트 내용의 적절성, 획득 주제와의 연관성 등을 고려하여 부적절한 데이터를 필터링하거나 라벨링하기 적합한 형태 및 내용으로 수정한다.
  - 데이터 라벨링에 포함하지 않아야 할 개인정보 등을 필터링하는 정제 기준을 마련한다.

【참고】 텍스트 데이터 정제 기준 및 정제 방법 예시 (문서요약 텍스트)

데이터 정제 기준	내 용
문장 분리	• 문장분리 기술을 활용하여 3줄 요약으로 문장을 분리 • 분리된 문장은 한 줄씩 개행하여 라벨링 작업자의 작업 효율을 증대
문장구분 오류	• 오픈소스 문장 분리기 (Koala NLP)와 참여기관이 자체 보유한 한국어 어휘사전 기반 문장 분리기를 결합하여 99% 이상 정확도를 갖는 문장 분리 수행
오탈자 수정	• 참여기관이 자체 보유한 한국어 어휘사전을 활용하여 초성-중성-종성 간의 관계를 고려하여 오탈자 판단 및 수정
수식어 하이라이트	• 기 보유하고 있는 통합사전과 추가로 보완할 도메인 사전을 바탕으로 형태소 분석을 통해 품사를 구분함 • 하이라이팅된 수식어는 라벨링 작업자들이 요약 시 내용을 파악하는데 방해되는 수식어를 쉽게 구분할 수 있도록 함

【참고】 텍스트 데이터 개인정보 정제 기준 예시 (한국어 대화 텍스트)

범주	항 목	변 환	예 시
이름	실명 실명(변형) 특수 애칭, 별명, 대화명, 필명	#@이름#	(수정 전) 소연이 넘 고마오♥ (수정 후) #@이름# 넘 고마오♥
	일반 애칭, 별명 공인 실명		자기야, 여보 등 김연아, 빌 게이츠 등
온라인	아이디 이메일 주소 URL	#@계정#	(수정 전) sample@sample.com 으로 보내 (수정 후) #@계정#으로 보내
각종 번호 및 비밀 번호	고유 식별 번호 (주민번호, 학번, 사번 등)	#@신원#	(수정 전) 응 학번은 200101-1234567 (수정 후) 응 학번은 #@신원#
	전화번호	#@전번#	(수정 전) 언니 번호 010-1234-5678이야 (수정 후) 언니 번호 #@전번#이야
	금융 번호 (계좌, 카드번호 등)	#@금융#	(수정 전) 신한 110-234-456-789 김연아 (수정 후) #@금융#
	일련번호 (구매자) 식별 번호 사업자 등록 번호 비밀번호	#@번호#	(수정 전) 사업자등록번호 123-45-67890 (수정 후) 사업자등록번호 #@번호#
장소	상세 주소(동 이하) 거주 아파트 및 건물명	#@주소#	(수정 전) 배송지는 서구 연희동 123 로요 (수정 후) 배송지는 서구 #@주소# 로요
	거주지 역명 (지하철역, 기차역 등) 방문 장소(비정기적) 상호명		도곡역 3번출구로 오세요 연세대 앞 정류장이야 롯데리아에서 만날래?
출신 및 소속	출신 및 소속 학교 출신 및 소속 직장 출신 및 소속 부대	#@소속#	(수정 전) 한국대학교에 재학중입니다 (수정 후) #@소속#에 재학중입니다

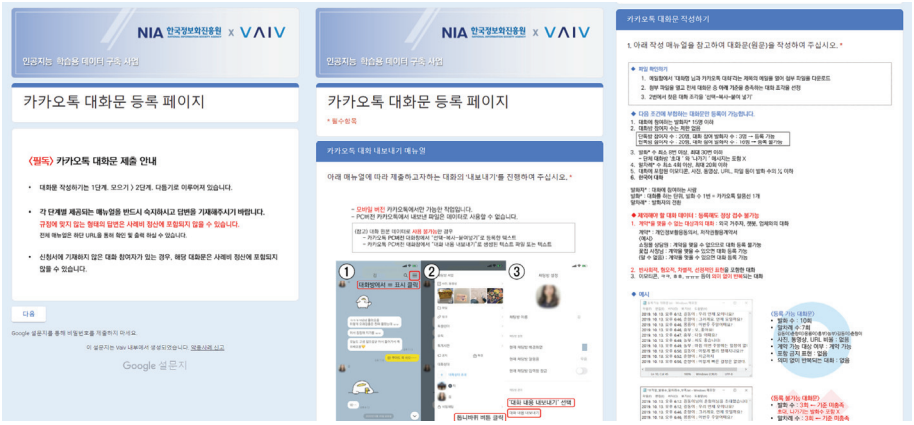
3.5 획득 도구 및 정제 도구

- 획득 및 정제도구
  - 데이터 획득 및 정제도구 개발 및 활용 계획을 작성한다.
  - 텍스트 자동 획득으로 API 또는 크롤링 방식으로 자동화한 텍스트 수집기를 개발·활용할 수 있다.
  - 데이터 구축 목적에 맞게 텍스트를 Parsing*하는 방법을 수립하여 정제도구로 개발한다.

    * Parsing(파싱) : 문장을 해부(분할)하는 것 또는 낱말의 품사, 문법적 관계를 설명하는 것
  - 데이터 획득, 정제도구를 자체적으로 개발하기 어려운 경우, 시중의 제작 도구 또는 그와 유사한 역할을 할 수 있는 서비스·애플리케이션을 활용할 수 있다.

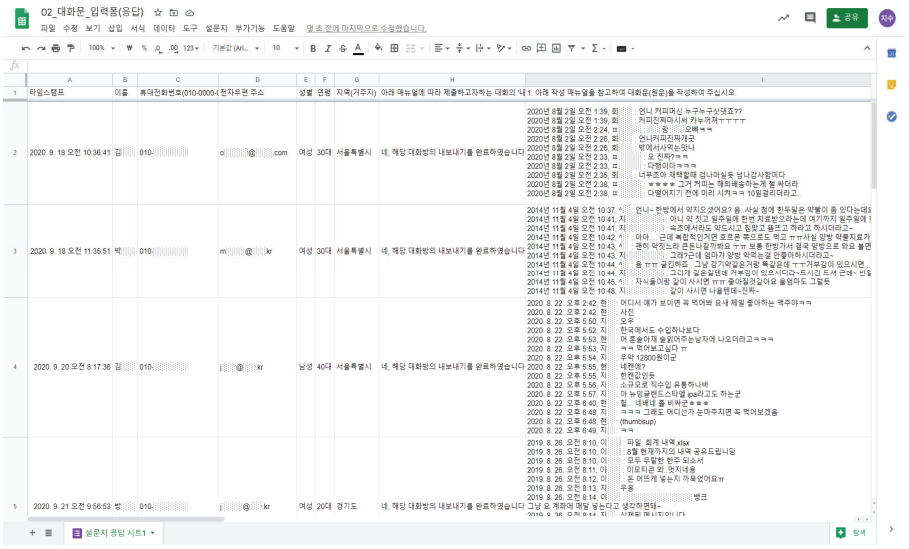
【참고】 데이터 획득·정제도구 구성 예시(기존 플랫폼 활용 사례)

- 구글폼(Google Forms)을 활용한 데이터 획득 및 저작 도구 도입
  - 본 과제에서 구축하는 데이터는 일반 텍스트 형식으로 구조화되고 유지하므로 온라인 공동 작업과 버전 유지가 이루어지는 구글폼(Google Forms) 시스템으로 데이터의 수집이 가능
  - 다양한 용도로 폭넓게 사용되고 있는 구글폼(Google Forms)은 실시간 응답 관리와 스프레드시트와 연동하여 모든 정보를 확인 가능하므로 데이터 활용도를 높일 수 있음
  - 공개된 클라우드 시스템 활용으로 데이터 수집 및 저작 도구의 개발 시간을 단축



[구글폼 활용 데이터 획득 도구 구성]

- 데이터 관리 및 정제
  - 구글폼(Google Forms)과 연계된 구글 스프레드시트에서 응답 데이터의 실시간 확인 및 가공이 가능함



[획득 데이터 관리(스프레드시트 활용)]

3.6 획득 시 고려사항

- 법·제도 준수
  - 데이터 획득 대상, 획득방법이 법·제도를 저촉하거나 또는 사회 윤리에 어긋나지 않도록 한다.
  - 개인정보 및 사생활 보호가 필요한 항목 획득 시, 개인정보보호법 등에 따라 적절한 법적, 기술적 절차를 거친 데이터를 활용하며, 그렇지 않은 데이터는 정제 과정에서 처리될 수 있도록 한다.

※ 법적 절차 : 개인정보 활용 동의, 초상 활용 동의, 명예훼손 가능성 여부 검토 등
※ 기술적 절차 : 데이터 유형별로 적용할 수 있는 익명처리 기법 적용
① 수치형 데이터 : 데이터 범주화 등
② 텍스트 데이터 : 이름, 민감정보 키워드 데이터 변환 등
③ 이미지·동영상 데이터 : 모자이크·블러 처리, 크롭(자르기) 등
④ 음성 데이터 : 크롭(자르기) 등

- 데이터가 3자 제공 및 대중에 개방에 문제가 없도록 법적요건 및 동의서 내용 등을 검토한다.
- 저작권 보호 대상인 데이터 획득 시 법에 저촉되지 않는 범위 내에서 획득할 수 있는 방안을 마련하며, 저작권 보호 대상 저작물 활용 필요 시 가급적 동의서, 계약서 등을 활용한 서면 자료 확보를 권장한다.

  ※ 예) 텍스트 내 특정 기업, 제품명 등 노출 가능 여부

  ※ 예) 뉴스, 출판물 등 저작권 보호 대상 저작물 활용 시 관련 당사자·기업·기관과 협의방안

- 개인정보활용동의서 및 저작물 활용 동의서 등 법적 요건을 준수하기 위한 관리방안을 마련한다.

  ※ '인공지능 학습용 데이터 품질관리 가이드라인'의 Ⅲ.품질관리기준, 3.1.5 체계준수성-보안준수의 【참고】개인정보보호 및 보안관련 법령·고시·권고 참조

【참고】 온라인 서명 플랫폼 활용 예시

- 원시데이터 제공자와 개인정보 수집 및 이용 동의 및 저작권 활용 계약을 전자서명 계약 기반으로 체결함으로써 민감한 정보 제공에 따른 위험 부담을 최소화함
- 저작권 이용 허락 계약 체결은 온라인 서명 플랫폼을 활용해 계약서를 업로드하여 법적으로 유효한 온라인 서명 기능을 구현

- 데이터 다양성 확보
  - 인공지능 학습모델이 현실을 잘 반영하고 본래의 구축 목적을 달성할 수 있도록, 획득하는 데이터가 일부 범주에만 치우치지 않고 가능한 다양한 시간·공간·집단·수준이 포함될 수 있도록 한다.

【참고】 데이터 다양성 확보/미확보 예시

- (다양성 미확보) 전문분야 한-영 말뭉치 학습 데이터 획득 시, 'IT 기술', '국제스포츠행사', '축제/행사' 3개 분야로 한정된 말뭉치만 획득함
- (다양성 확보) 획득 분야와 획득 수량을 미리 정의하고 통계치로 관리하여 데이터 다양성을 확보함

분야	대상 자료	획득 문장 수
국제 스포츠	스포츠경향/월드/서울 등 최근 3년 신문(중복 문장 빈도 고려)	50만
금융 · 증시	머니투데이, 이데일리, 한국경제, 헤럴드 경제 등 최근 1년 신문	40만
가정통신문	전국 공립 유치원, 초중 교 홈페이지의 가정통신문 전체 신문	20만
축제/행사	대한민국 구석구석, 한국 관광공사, 여행신문사 등 전체 신문	40만
IT	전자신문 최근 5년 신문	30만
의료/보건	일간보사, 메디칼타임즈, 메디컬투데이, 데일리메디 등 전체 신문	30만
대법원 판례	국가법령센터 판례 정보 전체 판례	40만
향토문화	전국 90개 지역의 한국향토문화전자대전 홈페이지 전체 자료 전국 시군구 지자체 홈페이지	30만
음식	한식(전체), 한국음식문화, 식품외식경제, 식품저널 전체 자료	30만
합계		310만

- 데이터 편향 방지 및 윤리 준수
  - 인공지능 학습모델이 인간의 비윤리 또는 편견을 학습하지 않고 사회적 윤리를 준수할 수 있도록 비윤리적 내용, 편견·편향된 데이터의 획득은 지양한다.

    ※ 딥페이크 분류, 가짜뉴스 분류, 비속어 필터링 등 비윤리·편향·왜곡된 정보 특성을 학습하는 것을 목적으로 구축하는 데이터는 예외로 할 수 있다.

【참고】 인공지능 학습용 데이터의 윤리성 검토

- 국내 모 업체에서 개발한 챗봇은 커플들의 메신저 대화를 학습데이터로 활용하였으나, 성차별성·인종차별성 표현, 성소수자·장애인들에 대한 혐오표현들을 여과없이 표현하여 사회적 문제가 된 결과 서비스 일시중단에 이르는 결과가 발생하였음

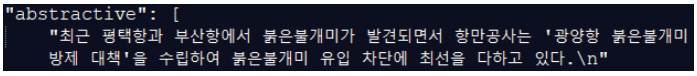
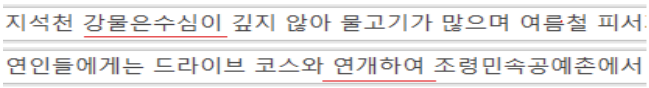
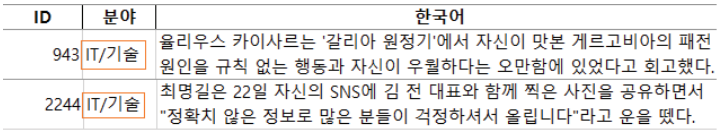
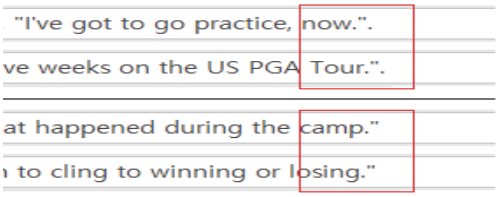
- 사업계획서 및 데이터 구축 요건 일치
  - 사업계획 당시 정의한 데이터 구축 기준에 맞춰 데이터를 획득·정제하도록 구축 현황을 모니터링한다.

【참고】 크라우드소싱 작업인력 운영 방식 수립 사례

1. 사업계획 시 txt 포맷으로 획득하기로 했으나, 실제 구축되는 데이터에 txt, json 파일이 혼재
2. 사업계획 시 2인 이상 화자의 자연어 대화를 획득하기로 하였으나, 1인 화자의 발표내용 문장이 포함됨
3. 구축 목적(노인 돌봄을 위한 감성 대화 서비스 구현) 대비 획득한 데이터의 범위가 너무 좁아서 구축 목적을 달성하기 불가능함 (의식주, 건강 관련 대화만 획득하는 등)

- 기타 텍스트 데이터 획득 시 품질 고려사항
  - 아래의 사례를 참고하여 텍스트 데이터 획득·정제에 필요한 사항을 가이드라인에 반영한다.

표 3-8. 텍스트 데이터 획득 시 품질 고려사항

항 목	내 용
불필요한 철자 제거	• 문장 내에 정상적인 문장과 무관한 기호, 접두(접미)어, 코드 등 철자를 제거한다. - 예) 문장 끝에 있는 '\n'는 실제 문장에 무관한 요소이므로 제거해야 함 
틀린 맞춤법 정제	• 띄어쓰기 및 맞춤법이 올바르지 않은 텍스트 획득 시, 설명 및 정제하는 방법을 가이드라인에 포함한다. 
획득 목적 관련성	• 구축 목적과 관련성이 낮은 텍스트 획득 시 이를 정제하는 방안을 마련한다. 
문장 부호	• 문장이 대화나 인용문으로 끝나는 등 따옴표, 큰따옴표와 마침표가 연이어 나타나는 경우 인용구 내 마침표가 제대로 사용되었는지 확인한다. 
기계 생성 문장	• 사람이 직접 작성한 문장 대신, 기계가 자동생성한 문장을 획득하려고 할 경우 인공지능 학습용 데이터셋 구축에 활용하는데 적합한지, 그리고 구축 취지에 부합하는지 검토한다. ※ 기계가 자동생성한 문장은 기존 인공지능 학습 알고리즘을 통해 도출된 '결과물'이므로, 이를 검사, 구축하는 것은 학습용 데이터 구축이 아닌 인공지능 모델 개발의 영역으로 볼 수 있다.

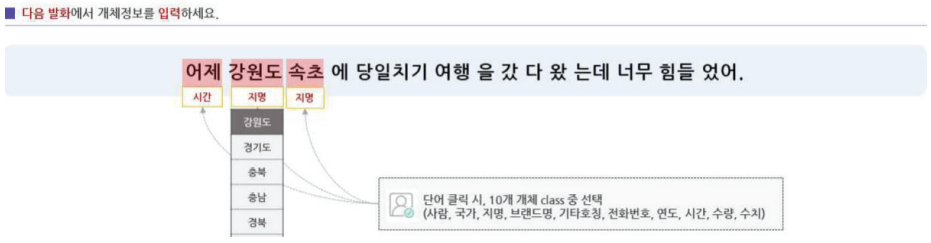
4 데이터 라벨링 작업

4.1 데이터 특성 식별 분류 체계 및 고려 사항

- 라벨링 작업 대상 및 범위 정의
  - 원천데이터 내에서 어떤 항목들을 라벨링 해야 하는지 대상과 범주를 정의한다.
  - 원천데이터 내에서 데이터 구축 목적에 부합하는 내용을 최대한 반영할 수 있는 정보를 라벨링할 수 있도록 라벨링 대상 범위를 정의하며, 데이터 품질 및 구축 목적과 무관한 내용을 불필요하게 라벨링하는 사항의 존재 여부 등을 검토한다.

【참고】 잘못된 라벨링 대상 설정 예시(대화 텍스트 감성 분류)

- 대화의 감성 분류를 목적으로 인공지능 학습용 데이터셋을 구축할 때, '시간', '지명' 키워드를 라벨링하는 것이 본래의 구축 목적 달성에 필요한 것인지 합리적인 근거가 없음



  - 특히 텍스트 전체에 대한 라벨링이 아닌, 하나의 텍스트 안에 특정 키워드, 문장 등을 라벨링해야 하는 경우 작업자들이 어떤 대상을 라벨링해야 하는지 판단할 수 있도록 세부적인 기준을 마련한다.
  - 원천데이터에 포함된 개인정보는 라벨링 대상에서 제외하거나, 익명처리 등 비식별화를 통해 개인정보를 알아볼 수 없게 라벨링한다.

※ 개인정보 활용 및 3자제공 동의를 받은 경우 동의 범위 내에서 개인정보를 라벨링 데이터로 활용할 수 있음

【참고】 개인정보에 대한 라벨링 익명처리 예시

- 대화자들의 신분 보장을 위해 이름, 주민등록번호, 카드 번호, 전화번호 등 개인정보와 관련된 사항은 그대로 전사하지 않고 아래와 같이 영문 기호로 적는다.
- 사람 이름은 '&name&'로 전사한다. 정치인, 연예인 등 유명인 이름은 그대로 전사한다.

(발화) 저는 김철수입니다.
(전사) 저는 &name&입니다.

- 여러 이름이 나올 때는 '&name1&', '&name2&' 등 name 뒤에 숫자를 붙여 구별한다.

(발화) 그때 철수랑 민수랑 너랑 나랑 갔잖아. 철수도 알고 있지?
(전사) 그때 &name1&이랑 &name2&이랑 너랑 나랑 갔잖아. &name1&도 알고 있지?

- 집 주소 등 개인정보에 해당되는 주소는 동 단위 까지만 전사하고, 그 이하 구체적인 주소는 영문 기호로 적는다.

(발화) 홍제동 한양아파트 205동 304호로 배달해주세요.
(전사) 홍제동 &address&로 배달해주세요.

- 클래스 정의 및 관리
  - 원천데이터의 특성을 바탕으로 부여할 수 있는 클래스 리스트 또는 클래스의 범주를 정의한다.
  - 클래스를 정의할 때는 원천 데이터 내에 존재하는 다양한 값들을 모두 커버할 수 있도록 정의하고, 클래스 이름이 중복되거나 모호한 의미를 갖지 않도록 한다.
  - 클래스 이름은 의미를 바르고 명확하게 나타낼 수 있도록 적절한 어휘를 선택한다.

【참고】 잘못된 클래스 정의 예시 (텍스트 주제 및 감성 분류)

- 클래스 간 의미 중복이 있어 구분이 애매함
  - (학업) vs (학교폭력) : 학업의 개념 안에 학교폭력을 포함
  - (학업 및 진로) vs (진로/취업/직장) : 전자와 후자가 서로 교집합이 있음
  - (상처) vs (슬픔) : 전자와 후자의 경계를 나누기 모호함
- 라벨링 표기법에 일관성이 없음
  - (청소년) vs (청소년(10대)) : 전자와 후자를 같은 의미로 라벨링에 활용하였으나 표기법에 일관성이 없음
  - (여성) vs (FEMALE) : 전자와 후자를 같은 의미로 라벨링에 활용하였으나 언어·문자가 일관성이 없음
- 범위와 주제에 맞지 않는 클래스 존재
  - 성별 구분에 (기타) 클래스 존재 : (남성), (여성) 외에 (기타) 성별을 정의할 수 있는지, 또는 의미가 무엇인지 불명확

- 라벨링 진행 중에 이전에 정의되지 못했거나 새롭게 정의가 필요한 클래스가 발견될 경우 클래스 항목 업데이트 방안을 마련한다.
- 클래스를 정해진 목록에서 선택하지 않는 경우에도, 작업자마다 일관된 기준 및 규칙에 따라 속성값을 부여할 수 있도록 하는 기준을 마련한다.

※ 주로 OCR 이미지, 텍스트, 음성 전사 등 문장, 텍스트로 값을 부여하는 데이터가 해당

4.2 데이터 라벨링 방법 및 절차

- 개요
  - 획득→정제 과정을 통해 도출된 원천데이터를 라벨링하여 학습데이터를 생성하기 위한 과정 및 고려사항을 작성한다.
  - 라벨링 지원 도구를 활용하며, 용어 및 분류체계를 준수하여 라벨링한다.

- 라벨링 작업 방식
  - 라벨링할 정보의 특성에 따라 자동, 반자동, 수동 방식을 결정한다. 원천데이터로부터 추출하는 방식이 정형화되어있고 자동화할 수

있는 사항인 경우 자동 방법을 고려할 수 있으며, 기계가 판단하기 어려운 사항은 반자동 또는 수동 방식이 적절하다. 반자동 방식은 자동으로 라벨링한 이후 사람이 다시 확인하여 수정하는 방식으로 작동된다.

- 작업 배분
  - 획득된 데이터를 라벨링 작업자에게 배분하고 라벨링 결과를 다시 저장하는 파일 저장체계 및 프로세스를 정의한다.

- 라벨링 작업 기준
  - 데이터별 어노테이션 기준, 라벨링 기준 등을 상세히 기술하며, 구체적인 예시를 들어 설명하여 작업자들이 혼동없이 명확한 기준을 갖고 빠르게 작업할 수 있도록 한다.

    ※ 레이블 범주, 레이블 부여기준(ground truth) 제시, 레이블 부여 예시, 애매한 내용이 나올 경우의 처리 기준, 자주 실수하는 예시, 검사 기준 등

【참고】 라벨링 작업 안내 예시 (문서요약 텍스트)

1. 메타데이터 확인 : 라벨링도구에서 제공되는 원문에 대한 정보 (메타데이터)를 확인하며, 메타 데이터의 종류는 아래와 같음 (중략)
2. 키워드 확인 : 제공된 문서의 제목을 통해 키워드를 확인하며, 키워드는 고유명사 (이름, 지명)를 먼저 선택한 뒤 이후에 관련된 동사를 확인함 (중략)
3. 원문 정독 : 기사의 구조를 파악하고 키워드가 포함된 문장 (리드문)을 찾되, 기사의 구조 형태는 아래와 같음 (중략)
4. 생성요약문 작성 : 생성요약은 추출요약으로 형성된 문장, 제목의 키워드, 동의어 등을 활용하여 기사의 내용을 한 문장으로 요약함을 원칙으로 함
   - 원문 내 육하원칙의 내용이 모두 포함되어야 함
   - 기술할 때 함축을 위해 1) 집단화, 2) 추상화, 3) 동의어 등을 활용함
     ※ 집단화 : 원문에 포함된 다양한 객체 등을 하나의 그룹으로 묶어 집단을 대표하여 요약하는 것을 의미함
     ※ 추상화 : 원문에 포함된 구체적 내용을 대표할 수 있는 개념으로 치환하여 수정하는 것을 의미함
     ※ 동의어 변환 : 글의 내용 변화를 최소화하면서 같은 의미의 단어로 수정하는 것을 의미함
   - 요약은 원문 글자수의 10% 내외로 함을 원칙으로 함

- 텍스트 데이터 라벨링 작업
  - 인공지능 학습 데이터 구축 목적, 도메인, 활용 분야(번역, 문서 요약 등, 대화형 챗봇 등)를 고려하여 텍스트 입력 절차 및 기준을 수립한다.

【참고】 라벨링 작업 안내 예시 (문서요약 텍스트)

[텍스트 데이터 라벨링 절차]

절 차	내 용
전처리데이터(입력)	수집 및 정제 후 전처리된 원천데이터(문서 형태)를 라벨링 도구를 통해 제공
가이드라인 확인	라벨링 기준 및 검사 기준으로 활용되는 문서요약 가이드라인 확인
정제된 문서의 사전 정보 확인	수집 및 정제, 전처리 과정에서 생성된 문서 정보(카테고리, 글자 수 등)을 확인
핵심 키워드 선정	제목 등 사전 정보를 통해 핵심 키워드 확인
정제된 문서 정독	원천데이터 문장 확인
문장 선정	5문장 내외의 중요 문장 확인
주요문장 선정 (라벨링 1단계)	선정된 중요 문장 중 3문장 선택
3문장 추출 요약문 생성 (라벨링 산출물 1)	원문 데이터 내 중요도에 따라 순차적으로 입력
생성요약문 작성을 위한 어군 선택	원문 데이터에서 확인된 주요 내용을 함축적으로 표현할 수 있는 단어들 선택(집단화, 추상화, 유의어)
선정된 어군 내 선택 (라벨링 2단계)	원문 데이터에서 수정 불가 어군 및 수정 가능 어군을 선택한 후, 수정 가능. 어군을 수정
생성요약문 생성 (라벨링 산출물 2)	생성 요약문 입력

[텍스트 데이터 라벨링 안내]

	구 분	내 용
입력	원천데이터 (정제)	수도가 헷갈리는 나라 중 하나인 캐나다. 잠시 머뭇거렸다면? 오타와가 정답이다. ... (중략)... 서쪽의 오대호와 동쪽의 세인트로렌스강 중간에 위치한 오타와는 물길을 따라 유럽 대륙에 비버 모피와 오크 목재를 수출한 역사가 있다. 오타와의 랜드마크를 꼽는다면 신고딕 양식으로 지어진 국회의사당과 리도 운하다. 둘은 위치상 맞붙어 있어 이 주변은 늘 관광객으로 북적거린다. 리도 운하는 오타와 시내 중심부에서 시작해 킹스턴 온타리오 호수까지 202㎞나 이어진다. 영국의 지배를 받던 시절, 미국의 침략에 대비하기 위해 군사 물자를 실어 나르는 통로로 건설된 리도 운하는 1826년부터 공사를 시작해 6년 후 완공됐다. 전쟁에 한 번도 사용된 적은 없지만 운하가 생기면서 교통의 중심으로 떠오른 오타와는 비약적으로 발전했다. 리도 운하는 19세기 초 아메리카 대륙의 역사와 건설 기술을 담은 사례로 가치를 인정받아 2007년 유네스코 세계문화유산에 등재됐다. ... (이하 생략)
출력	추출요약	오타와의 랜드마크를 꼽는다면 신고딕 양식으로 지어진 국회의사당과 리도 운하다. 리도 운하는 오타와 시내 중심부에서 시작해 킹스턴 온타리오 호수까지 202km나 이어진다. 리도 운하는 19세기 초 아메리카 대륙의 역사와 건설 기술을 담은 사례로 가치를 인정받아 2007년 유네스코 세계문화유산에 등재됐다.
	생성요약	오타와 시내 중심부에서 202km나 연결된 오타와의 랜드마크인 리도 운하는 19세기 초 아메리카 대륙의 역사와 건설 기술을 담은 사례로 가치를 인정받아 2007년 유네스코 세계 문화유산에 등재됐다.

【참고】 라벨링 작업 안내 예시 (한국어 대화 텍스트)

• (예시 1) 육하원칙 6가지 요소 포함하여 한 문장으로 요약

구 분	내 용
예시 대화문	2019. 5. 9. 오후 2:16, 춘향 : 향단쓰~ 나 주말에 쇼핑가는데 같이 갈래? 2019. 5. 9. 오후 2:16, 춘향 : 아마 용산? 근처 갈듯 ㅎㅎㅎ 2019. 5. 9. 오후 2:17, 향단 : 네넹! 조아요!! 저도 코트 보러 가야하는데 잘됐네요 2019. 5. 9. 오후 2:17, 향단 : 용산이면 집에서도 멀지 않고 좋아용ㅋㅋ 2019. 5. 9. 오후 2:22, 춘향 : 오~ 좋앗어!! 나도 옷 사러 가는 거였어 토요일이 좋아? 2019. 5. 9. 오후 2:23, 춘향 : 아님 일요일 점심 같이 먹고 볼까? 2019. 5. 9. 오후 2:24, 향단 : ㅎㅎ네네 아, 전 일욜이 더 나아요. 괜찮으세요? 2019. 5. 9. 오후 2:24, 춘향 : 오케이 그럼 이번주 일욜 12시에 용산역에서 만나자 2019. 5. 9. 오후 2:45, 향단 : 네네!
육하원칙	이번주 일요일에 / 용산역에서 / 언니와 동생이 / 옷을 / 사려고 / 만나기로 했다 언제 / 어디서 / 누가 / 무엇을 / 왜 / 어떻게
한 문장 요약	이번주 일요일에 용산역에서 언니와 동생이 옷을 사려고 만나기로 했다

• (예시 2) 대화문에 등장하는 키워드 2개 이상 포함하여 한 문장으로 요약

구 분	내 용
예시 대화문	2019. 10. 13. 오후 7:12, 길동이 : 우리 오늘 저녁에 뭐 먹으러 갈까요? 2019. 10. 13. 오후 7:46, 춘향이 : 저는 곱창 먹고싶어요! 2019. 10. 13. 오후 7:46, 춘향이 : 유투브에서 맛집을 봤거든요! 2019. 10. 13. 오후 7:46, 흥부 : 오.. 좋아요! 저도 곱창 좋아요! 2019. 10. 13. 오후 7:47, 길동이 : 곱창! 2019. 10. 13. 오후 7:49, 길동이 : 야채곱창인가요?
키워드 포함	오늘 저녁에 유튜브 맛집으로 곱창을 먹으러 가기로 했다 키워드 / 키워드
한문장 요약	오늘 저녁으로 유투브 맛집에서 곱창을 먹기로 했다.

4.3 데이터 어노테이션 포맷과 형식 정의 및 입력

- 어노테이션 포맷 및 저장 형식
  - 텍스트 데이터는 고정된 필드나 스키마가 존재하지 않는 비정형 데이터이기 때문에, 학습용 데이터로서 가치를 부여하는 어노테이션 정보를 저장할 수 있는 별도의 데이터 구조와 파일 포맷을 정의한다.
  - 어노테이션 파일 포맷은 특정 소프트웨어에 종속되지 않고 쉽게 열고 편집할 수 있는 포맷으로 선택하며, 구조화된 어노테이션 정보를 저장하기 적합한 포맷을 선택한다.

    ※ json, xml 등

- 어노테이션 정보 저장 구조
  - 어노테이션 정보(라벨링데이터)에 포함되어야 할 사항을 데이터 유형별(텍스트, 이미지, 동영상, 음성 등) 라벨링 참조 기준과 구축 목적에 따라 필요한 항목을 종합적으로 고려하여 정의한다.
  - 텍스트 데이터 어노테이션 정보 구조는 '부록1. 인공지능 학습용 데이터셋 구축 공통참조기준'을 준용하여 정의한다.

    1) 자연어 처리
    - 어노테이션 공통자연어 처리를 통해 라벨링 시 공통적으로 기록해야할 정보이다.

표 4-1. 자연어 처리 어노테이션 공통항목

No.	속성명	항목 설명	Type	필수여부	작성예시
1	annotations[].morp.id	형태소 식별자 (출현 순서)	string	선택	NNG_00001 (형태소태그_순번)
2	annotations[].morp.lemma	형태소	string	선택	일반명사
3	annotations[].morp.type	형태소 태그	string	선택	NNG, NNP, NNB 등
4	annotations[].morp.position	문장 내 위치	number	선택	231 (형태소 위치)
5	annotations[].morp.weight	형태소 분석 결과 신뢰도	number	선택	0.92 (0 ~ 1)

- 자연어 의미 분석을 위해 단어 또는 구문 단위 분석을 위한 어노테이션 공통항목은 아래와 같다.

표 4-2. 단어(구문) 분석 어노테이션 항목

No.	속성명	항목 설명	Type	필수여부	작성예시
1	annotations[].text_id	원문 텍스트 식별자	string	선택	TXT_0001 (분류_순번)
2	annotations[].word[].string	단어	string	선택	네이버
3	annotations[].word[].label	단어 레이블	string	선택	기업
4	annotations[].word[].start	원문 내 단어 시작 지점	number	선택	100
5	annotations[].word[].end	원문 내 단어 종료 지점	number	선택	102

- 자연어 의미분석을 위해 문장 단위로 어휘의미 분석 시 어노테이션 할 항목은 아래와 같다.

표 4-3. 문장 단위 어휘의미 분석 어노테이션 항목

No.	속성명	항목 설명	Type	필수여부	작성예시
1	annotations[].WSD.id	어휘의미 식별자 (출현 순서)	string	선택	WSD_0001 (분류_순번)
2	annotations[].WSD.text	어휘 텍스트	string	선택	배
3	annotations[].WSD.weight	어휘의미 분석 결과 신뢰도	number	선택	0.92 (0 ~ 1)
4	annotations[].WSD.position	문장 내 위치	number	선택	2310
5	annotations[].WSD.begin	어휘의 첫 형태소 식별자	string	선택	LC_0012
6	annotations[].WSD.end	어휘의 끝 형태소 식별자	string	선택	LC_0017

2) 클래스 분류

- 텍스트 데이터를 클래스로 라벨링할 때 어노테이션 공통항목은 아래와 같다.

표 4-4. 클래스 분류 어노테이션 항목

No.	속성명	항목 설명	Type	필수여부	작성예시
1	annotations[].id	어노테이션 식별자	string	선택	CL_0001 (분류_순번)
2	annotations[].class	클래스 분류 (클래스 정의 필요)	number	선택	0: 정치, 1: 사회, 2: 연예, 등

3) 개체명 인식

- 텍스트 내 개체명 인식을 위한 어노테이션 항목은 아래와 같다.

표 4-5. 텍스트 내 개체명 인식 어노테이션 항목

No.	속성명	항목 설명	Type	필수여부	작성예시
1	annotations[].NE.id	개체명 식별자	string	선택	LC_0012 (개체명분류_순번)
2	annotations[].NE.text	개체명 텍스트	string	선택	광화문
3	annotations[].NE.type	개체명 타입	string	선택	관광명소 (LC_TOUR)
4	annotations[].NE.begin	객체명 구성 첫 형태소 식별자	string	선택	LC_0009
5	annotations[].NE.end	객체명 구성 끝 형태소 식별자	string	선택	LC_0015
6	annotations[].NE.weight	개체명 인식 결과 신뢰도	number	선택	0.92 (0 ~ 1)

4) 관계-의존성 어노테이션

- 텍스트 데이터 내 단어(구문) 간 관계-의존성을 라벨링할 때 어노테이션 공통항목은 아래와 같다.

표 4-6. 관계-의존성 어노테이션 공통항목

No.	속성명	항목 설명	Type	필수여부	작성예시
1	annotations[].dependency[].id	어절 ID (출현 순서)	string	선택	DEF_0021 (분류_순번)
2	annotations[].dependency[].text	의존구문 텍스트	string	선택	안녕하세요. 좋은 아침입니다.
3	annotations[].dependency[].head	부모 어절의 ID	string	선택	DEF_0020
4	annotations[].dependency[].label	의존관계 레이블	string	선택	-
5	annotations[].dependency[].mod[]	자식 어절들의 ID	string	선택	-
6	annotations[].dependency[].weight	의존구문 분석 결과 신뢰도	number	선택	0~1

- 어노테이션 정보(라벨링데이터)가 어떤 원천데이터와 매칭되는 지 확인할 수 있도록 어노테이션 구조 및 내용을 정의한다.

※ 학습용 데이터는 원천데이터 + 라벨링데이터로 구성됨을 고려

표 4-7. 텍스트 데이터 어노테이션 구조 정의 및 구축 사례

항 목	설 명	json 포맷 구축 형태
name	파일명	{ "name": "문서요약 프로젝트", "delivery_date": "2020-09-08 11:36:41", "documents": [{ "id": "343753195", "category": "정치", "media_type": "online", "media_sub_type": "지역지", "media_name": "광주매일신문", "size": "small", "char_count": "814", "publish_date": "2019-05-02 18:32:00", "title": "군 공항 이전사업 홍보 영상물 만든다", "text": [[{ "index": 0, "sentence": "市, 국방부·전남도 협의 통해 이전 후보지서 필요성 설명", "highlight_indices": "" },], ...], "document_quality_scores": { "readable": 4, "accurate": 4, "informative": 3, "trustworthy": 4 }, "extractive": [2, 3, 4], "abstractive": [" 광주시에 따르면 국방부·전남도와 협의해 후보지를 직접 찾아가 주민들에게 사업의 필요성을 설명할 사업 소개와 함께 후보지로 선정되면 지원되는 다양한 사업이 상세하게 담긴 군 공항 이전사업 관련 홍보 영상물을 제작 중이다."] }, ...] }
documents	문서 배열	
문서정보		
id	문서 아이디	
category	카테고리	
media_type	미디어 유형 (ex: online)	
media_sub_type	미디어 유형 (ex: 중앙지)	
media_name	미디어 명 (ex: 국민일보)	
size	길이 (ex: small)	
char_count	본문길이	
publish_date	게시시간 (yyyy-MM-dd hh:mm:ss)	
title	제목	
text	본문(문단/문장)	
index	순번	
sentence	문장	
원문 평가정보		
readable	가독성	
accurate	정확성	
informative	정보성	
trustworthy	신뢰성	
라벨링 결과		
extractive	추출요약문 정보	
abstractive	생성요약문 정보	

4.4 데이터 라벨링 완료 후 관리 방법

- 데이터 관리 기본사항
  - 목적에 맞는 데이터 어노테이션 기준을 수립하고 데이터 사용 목적에 맞게 관리한다.
  - 데이터의 사용 목적에 맞는 일관된 자료인지 확인한다.
  - 데이터들의 편향성을 확인 후 필요에 따라 데이터 추가한다.
  - 보존 일정 및 규정 준수 요구 사항에 따라 데이터 보관, 관리한다.

- 데이터 저장 관리
  - 원천데이터에 추가된 라벨링 정보를 저장하고 관리하는 기준을 수립한다. 파일을 체계적으로 분류하기 위해 데이터 종류 및 분류에 따른 라벨링데이터 파일 명명법과 파일 저장구조를 정의한다. 정의된 내용에 맞게 파일을 저장하도록 작업자에게 안내한다.
  - 작업자들이 원천데이터 및 라벨링 정보 저장 구조에 맞게 저장할 수 있도록 저장 절차를 정의하고, 작업자를 대상으로 배포한다.

- 데이터 백업 관리
  - 원천데이터 및 라벨링데이터의 훼손 및 멸실을 방지하기 위해 안전한 보관방법 및 백업방안(백업 시스템 및 프로세스 구축, 관리 절차 등)을 마련한다.

- 데이터 관리 조직 운영 방안
  - 데이터셋 제작 책임자는 품질관리 책임자로서 획득되는 데이터의 품질을 주기적으로 검사 및 관리한다.
  - 주기적인 실무협의체와의 미팅을 통해 데이터 품질에 대한 피드백을 공유하고 논의한다.
  - 데이터 품질 제고를 위해 데이터 라벨링 방안에 대하여 전문 컨설턴트 등 외부 기관의 조언을 받을 수 있다.

4.5 데이터 라벨링 방식에 적합한 도구 선정

- 라벨링 도구 선정
  - 데이터 구축 목적 달성을 위해 원천데이터 형태, 구축 목적에 부합하는 라벨링 도구를 선정한다.
  - 기존의 도구를 가지고 인공지능 학습용 데이터 구축 목표 달성이 어려울 경우, 기존 라벨링 도구의 기능을 추가하거나 완전히 새로 개발하는 방법을 고려한다.

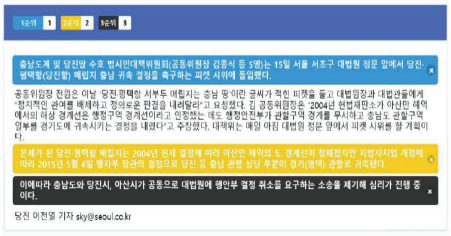
표 4-8. 국내외 주요 텍스트 데이터 라벨링 도구

No	도구명	요금	설 명
1	SELECTSTAR	유료	• 크라우드소싱 기반의 데이터 수집 및 라벨링 도구 • 자연어 처리, TTS(Text-to-Speach) 기능 • 사이트 주소 : https://selectstar.ai/
2	AIWORKS	유료	• 크라우드소싱 기반의 데이터 수집 및 라벨링 도구 • 텍스트 분류, 엔티티 태깅, 질의응답, TTS(Text-to-Speach) 기능 • 사이트 주소 : https://aiworks.co.kr/
3	Amazon Sagemaker Groundtruth	유료	• Amazon AWS의 머신러닝 모델 제작도구인 SageMaker의 부속 서비스 • 라벨링 작업의 일괄처리 및 자동화할 수 있는 API 기능 • 텍스트 분류 및 엔티티 태깅 기능 • 사이트 주소 :https://aws.amazon.com/ko/sagemaker/groundtruth/

- 라벨링 도구 활용 매뉴얼 작성
  - 작업자가 활용할 도구의 사용법에 대한 매뉴얼을 작성한다. 매뉴얼은 Step-by-Step 형태로 쉽게 따라할 수 있도록 작성하며, 이미지, 동영상 등을 활용하여 이해를 도울 수 있다.

【참고】 라벨링 도구 활용 매뉴얼 작성 예시

- 추출요약 도구기능 : 지문 내 의미 있는 구/절/문장 등 특정 범주에 대해 라벨링하는 기법으로, 문서요약 과제에서는 지문 내 문장의 중요도에 따라 3개의 문장을 추출하여 라벨링을 진행함
- 생성요약 도구기능 : 하나의 범주 (지문)에 대하여 새로운 범주 (문장)으로 set을 이뤄 어노테이션하는 기법으로서, 문서요약 과제에서는 지문에서 추출한 3개의 문장을 기반으로 1개의 새로운 문장을 만듦으로써 어노테이션을 진행함



[추출요약 (Sequence Labeling) 형식]

[생성요약 (Sequence-to-Sequence) 형식]

- 라벨링 도구 선정 및 개발 시 고려사항
  - 표준적인 어노테이션 및 라벨링 작업 가능 여부, 표준 파일 포맷 지원 여부 등을 고려한다.
  - 한국어 및 다국어를 입력할 때 깨짐현상이 없는 인코딩을 지원하는지 고려한다.
  - 특히 크라우드소싱 방식을 적용하는 경우, 다양한 작업환경(컴퓨터 성능, OS, 네트워크 등)에서의 실행 가능 여부를 확인한다.
  - 라벨링 결과를 효과적으로 관리하고 작업배분을 할 수 있는 관리 기능이 충실한 지 확인한다.

- 작업자 또는 검사자가 어노테이션 결과를 눈으로 바로 확인할 수 있는 시각화 기능이 있는 지 고려한다.

【참고】 어노테이션 시각화 예시

- 라벨링된 값, 어노테이션 영역 등 어노테이션 정보를 파일(json 등) 뿐만 아니라 눈으로 확인할 수 있는 기능도 함께 제공한다.

5 처리 데이터 검사

5.1 검사 절차 정의

- 개요
  - 인공지능 학습용 데이터 구축을 위한 품질 검사 절차·방법은 데이터 유형, 도메인, 목표 서비스에 따라 달라질 수 있으며 사업 기간 및 예산 등 현실적인 여건을 고려하여 수립한다.
  - 데이터 검사 절차 및 규격은 데이터 구축 목적 정의 단계에서 수립한 데이터 활용 분야·목적을 달성할 수 있도록 정의한다.

- 검사 절차 정의
  - 다량의 데이터를 한정된 시간 내에 최적의 품질로 검사할 수 있도록 하는 검사 단계 및 절차를 수립한다.
  - 검사 프로세스는 학습용 데이터 구축 공정(획득, 정제, 라벨링) 각 단계별로 검사가 수행되는 형태를 기본으로 하며, 데이터 구축 공정 및 데이터 특성을 반영하여 적합한 절차를 수립할 수 있다.

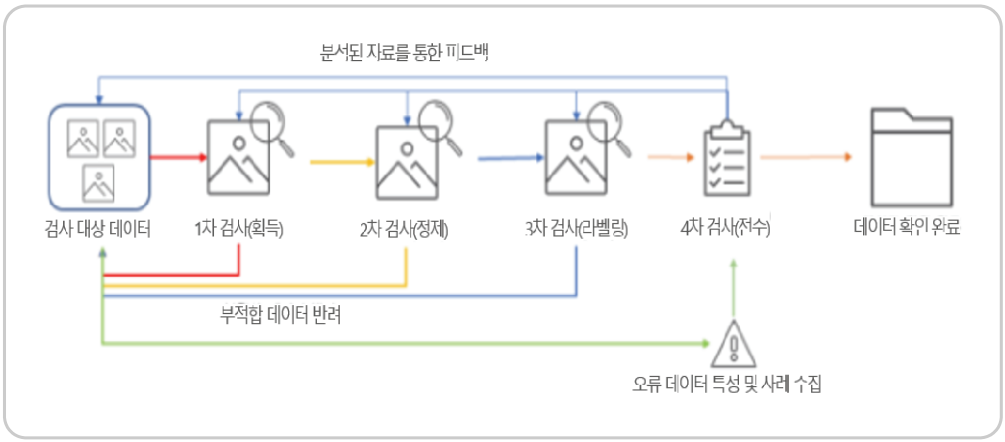


그림 5-1. 데이터 검사 절차 정의

- 검사 규모
  - 데이터 구축 설계 단계에서 구축될 데이터에 대한 품질 수준을 미리 정의하고, 품질 검사를 위한 검사 규모 및 방법*을 설정한다.
    * 전수 검사, 샘플링(00%), 다단계 샘플링 등
  - 전수 검사가 아닌 샘플링 방법으로 데이터를 검사할 경우, 검사 대상 데이터가 편향되지 않으면서 무작위로 추출될 수 있도록 한다.
    ※ 각 클래스별 동일한 비율로 추출되도록 함(층화추출)
    ※ 데이터가 구축된 순서 등이 특정 타이밍에 집중되지 않도록 함(파일명 정렬 후 무작위 추출 방법 등 적용)

5.2 검사 방식

- 검사 항목 정의
  - 구축 공정(획득, 정제, 라벨링)별로 공통적으로 적용할 수 있는 검사 요구사항을 고려하여 검사 항목을 정의할 수 있다.
  - 검사 항목은 데이터 및 절차 측면에서 적합성·정확성·유효성, 준비성·완전성·유용성 지표를 측정할 수 있도록 한다.
    ※ 자세한 절차 및 내용은 'Ⅳ. 품질검사 방법' 내용을 참고하며 아래 표 내용 참고 가능

표 5-1. 구축 공정별 주요 검사 항목

검사 절차	검사 항목	요구사항
1차 검사 (획득)	법·제도 준수	원시데이터 획득 시 관련 법·제도적 규정 등을 반드시 준수해야 함
	사실적인 획득 환경 구성	원시데이터를 인위적인 환경과 조건 하에 획득해야 하는 경우 사실적인 획득 환경을 구성하여야함
	데이터 동기화	다중 데이터 소스 간 정교한 동기화를 위한 절차를 마련하여야 함
	편향성 방지	데이터 편향을 방지하기 위한 절차를 마련하여야함
2차 검사 (정제)	정제 기준의 명확성	데이터 사용 목적에 적합한 정제 기준 수립 여부
	중복성 방지	데이터 정제 후 정보 비교 후 중복도 여부
	정제 작업 매뉴얼	정제 작업을 위한 매뉴얼 작성 및 관리 여부
	정제 도구	정제 작업에 사용될 SW 도구를 확보 및 사용 방법을 숙지

검사 절차	검사 항목	요구사항
	정제 작업 방식	데이터 특성 및 활용 목적에 맞는 적절한 정제 방식 선정 여부 및 선정 기준 타당성 여부
3차 검사 (라벨링)	라벨링 가이드	목적에 맞게 작성된 라벨링 가이드에 대한 타당성 여부를 검사 후 라벨링 작업자들에게 내용 가이드 전달
	어노테이션 항목	목적에 맞는 어노테이션 구성인지 여부를 검사 후 확인된 내용을 포함하도록 작업자들에게 전달
	라벨링 검사 도구	자동화 도구를 통해 검사 후 검사자가 육안으로 부적합 데이터 여부 2차 확인과 촬영된 영상(동적/정적) 이미지의 누락, 번짐 및 조건 오류를 전수 검사
4차 검사 (전수)	부적합 판정 데이터 분포 확인	데이터의 오류율, 특성 분포 확인을 통한 데이터 수집, 정제, 라벨링, 부문 최적화
	외부 검사자	외부 검사자(TTA 등), 도메인 전문가, 데이터 요청자

- 점검 기준 및 점검표 작성
  - 데이터를 일관된 기준으로 검사하기 위해, 데이터 정확성 및 구축 취지에 부합할 수 있는 참값(ground truth)을 정의하고 이 참값을 기준으로 검사 항목 및 채점 기준(통과 기준)을 정의한다.
  - 검사항목 및 채점 기준(통과 기준)을 검사자가 쉽게 확인하고 적용할 수 있도록 체크리스트 등의 형태로 작성하여 배포한다.

【참고】 검사 기준 및 검사 항목 예시(문서요약 텍스트)

1. 원문 내 육하원칙의 요소가 모두 있는 경우에 생성요약문도 육하원칙의 요소를 모두 포함하였는지 기준으로 검사 진행
2. 생성요약문의 경우 지문 내 중요도에 따라 추출된 3개의 문장을 기반으로 생성된 문장이므로, 내용적 측면에서 생성 요약문이 문서 전체 내용을 얼마나 커버하는지에 대한 정량적 수치화가 어려움
3. 검사 기준은 라벨링시 사용한 가이드라인 준수 여부를 중심으로 확인함
   - 추출요약 : 중요 문장에 대한 추출 여부를 5점 척도를 활용하여 4점 이상을 합격점수로 활용함
   - 생성요약 : 추상적인 내용이 포함되어 있어 질적 요소를 포함하는 5점 척도를 활용하여 4점 이상을 합격 점수로 활용함

[검사 항목 점검표]

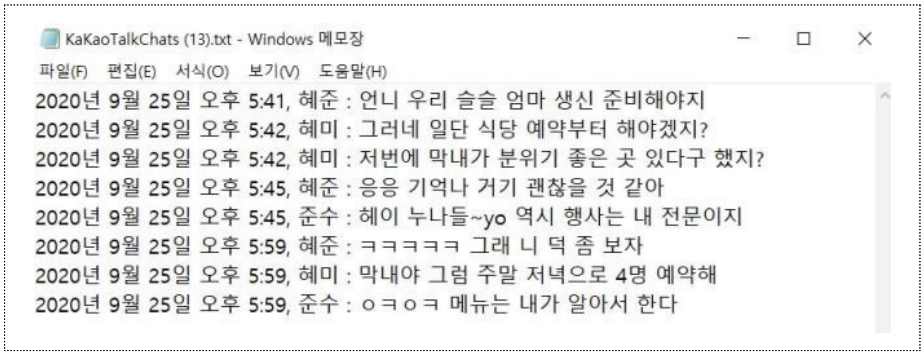
검사 기준	추출요약	생성요약
키워드의 활용	제목에서 제시된 주요 단어가 포함된 문장 선택 여부	제목에서 제시된 주요 단어의 활용 여부
리드문의 활용	가장 중요하게 선택된 문장의 구성 형태 확인	가장 중요하게 선택된 문장 활용 여부
6하원칙 내용의 포함	6하원칙 내용이 포괄된 문장의 선택 여부	생성된 문장에서 6하원칙 내용의 포함 정도
주관적 문장의 포함	주관적 문장 수정 불가	-
의미의 중복	선택된 3개 문장 중 중복된 내용의 수준 - 극단적 중복을 회피	-
구체적 문장의 활용	3개 문장 중 1문장은 구체적 문장이 포함되어야 함	구체적 문장의 수정 여부
문장의 추출	문장 변형 금지	문장 원형 추출 금지
추상화, 집단화, 동의어의 선택	-	문장 수정시 추상화, 집단화, 동의 선택의 정도를 질적으로 판단
비문, 미완성 문장	-	비문, 미완성 문장은 반드시 2점 이하로 입력
문장의 길이	-	전체 문장의 10% 내외로 요약 - 극단적 축약, 극단적인 복문은 요약 실패로 판정

[검사 점수 산정 기준]

기준(숫자가 높을수록 우선적으로 판단)	점수
1. 문장 구성 : 생성요약문이 비문 2. 문장의 완결성 : 오탈자, 띄어쓰기 오류 등이 문장의 50%를 넘는 경우 3. 핵심 키워드 : 생성요약문에 포함되지 않은 경우 4. 요약의 포괄성 : 생성요약문이 글의 일부만 요약하는 경우 5. 동의어 치환 : 고려대상 아님	1점
...	...
1. 문장 구성 : 생성요약문이 정문 2. 문장의 완결성 : 오탈자, 띄어쓰기 오류 등이 발견되지 않음 3. 핵심 키워드 : 생성요약문에 키워드가 포함된 경우 4. 요약의 포괄성 : 생성요약문이 글의 전체를 요약하고 5. 동의어 치환 : 동의어 치환 2번 이상 포함	5점

【참고】 검사 안내 예시(한국어 대화 텍스트)

- 대화의 유형 및 주제 적합/부적합 판정 기준
  - 대화의 유형에는 일상 대화, 토론 대화, 질의응답 대화 셋으로 분류됨
  - 토론 대화는 하나의 토론 주제에 대해 두 개 이상의 대립되는 의견이 있는 대화를 적합한 대화로 정함
  - 질의응답 대화는 3인 이상의 대화에서 대표질문 하나에 대한 답변으로 구성된 대화를 적합한 대화로 정함
  - 대화의 주제 분류에는 개인, 주거, 여가, 여행, 상거래, 공공서비스, 전문지식 등 16개 분야가 포함됨
  - 대화의 주제는 육하원칙을 바탕으로 '무엇을'에 해당하는 분류가 선택되었는지 확인하여 판정함
  - 검사 안내 예시



잘못된 요약	바람직한 요약
가족 행사 준비	주말에 엄마 생신 축하를 위해 가족식사를 할 식당을 예약하기로 했다

- 검사 도구
  - 검사자가 데이터를 빠르게 확인하고 검사할 수 있는 도구를 마련한다. 검사 결과를 작업자에게 즉각적으로 피드백하려는 경우에는 데이터 라벨링 도구와 연계한 시스템으로 검사 도구를 구축할 수 있다.
  - 검사 내용이 단순·반복적인 경우 검사 항목 일부의 자동화 처리를 고려한다.

5.3 검사 결과

- 결과 피드백
  - 각 검사 단계별로 검사자가 검사한 결과를 작업자 및 관리자, 고객에게 피드백하는 체계를 마련한다.
  - 검사결과 집계, 검사내용 확인, 의견사항 전달, 작업자·검사자 재배정, 이슈사항 공유, 피드백 시기 및 주기 등에 대한 절차 및 방법을 수립하여, 검사 현황 및 결과가 빠르게 공유되고 검사를 통과하지 못한 데이터에 대한 재 라벨링이 원활하게 될 수 있도록 한다.

【참고】 검사결과 피드백 프로세스 정의 예시

1. 검사 후의 피드백 프로세스는 다음과 같음
   - 검사 후 불합격 판정 → 검사 의견 제시 (라벨링 결과 문제점을 검사자가 기록) → 라벨링 작업자에게 재전달 → 2차 라벨링 → 2차 검사 → 불합격의 경우 재 라벨링 요청 (검사 의견 기록) → 재 라벨링 (3차 라벨링) → 3차 검사

2. 3차 검사 결과가 불합격인 경우 1) 문서가 매우 어려워 작업자가 라벨링할 수 없음, 2) 작업자의 역량이 해당 원천데이터를 라벨링할 능력이 없음, 3) 원천데이터의 형태가 라벨링이 불가능함 등의 사유로 라벨링 불능 데이터로 판정하고, 3차 검사 불합격 데이터를 별도 저장하여 향후 다른 작업자를 통해 재 라벨할 예정

3. 라벨링시 정확도를 높이기 위해 검사 합격된 문서에만 보상을 명문화하였음

4. 크라우드소싱의 특성상 작업자에 대한 교육이 어려운 관계로, 검사시 검사의견을 최대한 명확하고 구체적으로 전달

- 검사결과 처리
  - 미검사/검사 여부, 검사 통과/미통과 여부를 구분할 수 있도록 파일 관리 체계, 메타데이터 구조 등을 설계한다.
  - 목표 데이터, 원천데이터 대비 검사를 통과한 학습 데이터 현황을 모니터링하는 체계를 구성하여, 일정 내 학습 데이터를 확보할 수 있도록 한다.
  - 검사결과 및 피드백 사항을 데이터 제작 도구 및 시스템 상에서 어떻게 구현할 것인지 설계한다.

※ 데이터별 검사이력 보관방법, 데이터 전송, 작업자 대상 피드백 전달 방법 등

【참고】 데이터 검사결과 저장을 위한 메타데이터 구조 예시

- Amazon Sagemaker Groundtruth의 텍스트 클래스 분류 어노테이션에 대한 검증(verification) 메타데이터 항목

```
"verify-text-classification":"1",
    "verify-text-classification-metadata":
    {
        "class-name": "bad",
        "confidence": 0.93,
        "type": "groundtruth/label-verification",
        "job-name": "verify-text-classification",
        "human-annotated": "yes",
        "creation-date": "2018-11-20T22:18:13.527256",
        "worker-feedback": [
            {"comment": "The class of the sentence can't fit with its meaning."}
        ]
    }
```

참 고 자 료

1. ㈜바이브컴퍼니, 인공지능 데이터 구축·활용 가이드라인-한국어 대화 요약데이터
2. ㈜바이브컴퍼니, 인공지능 데이터 구축·활용 가이드라인-한국어 텍스트 AI 데이터 (한국어 SNS 데이터)
3. ㈜비플라이소프트·㈜위고·㈜테스트웍스·고려대학교·주식회사 에이아이닷엠, 문서 요약 텍스트 구축 가이드라인
4. ㈜슈퍼브에이아이, 데이터 라벨링 사업운영 A to Z
5. ㈜포티투마루, 관광분야 지식베이스 구축 절차서
6. 한국정보통신기술협회(TTA), AI 학습용 데이터 구축사업 공통기준
7. 한국정보통신기술협회(TTA), 2020년 인공지능 학습용 데이터 구축 사업(1차) 중간산출물 20종 검토
8. 한국지능정보사회진흥원(NIA), 2020년 인공지능 학습용 데이터 구축사업(2차) 가이드라인 참조
9. Amazon, AWS Sagemaker Groundtruth 도큐멘테이션 (https://docs.aws.amazon.com/sagemaker/latest/dg/sms.html)

Ⅱ 음성 데이터

목 차

표 목 차

그림 목 차

제1장 | 개 요

1 작성 배경

- 인공지능 학습용 데이터 구축 사업 확대에 따라 다양한 역량의 수행 및 참여기관 참여로 사업 진척도 및 데이터 품질의 편차가 발생
- 인공지능 학습용 데이터 품질 향상 및 성공적인 사업 추진을 위해 수행 및 참여기관을 대상으로 데이터 구축 기준, 절차 등 노하우 공유 필요

2 작성 목적

- 인공지능 학습용 데이터 구축에 보편적으로 적용되는 데이터 유형별로 데이터 구축에 필요한 절차 및 구성요소를 제시하여 데이터 구축 과정에서의 시행착오를 줄이고 체계적인 계획 수립을 지원한다.
- 국내 인공지능 학습용 데이터 구축 시 활용된 다양한 가이드라인 사례를 제시하여, 향후 수행 및 참여기관에서 수립해야할 데이터 구축 가이드라인 작성에 참조할 수 있도록 한다.
- 향후 인공지능 학습용 데이터 구축 사업 추진 시 본 구축 안내서를 배포하여 다양한 수행 및 참여기관의 역량 향상 및 성공적인 사업 수행을 지원한다.

- 궁극적으로 양질의 인공지능 학습용 데이터 구축 및 개방을 통해 국내 인공지능 산업 활성화 및 발전에 기여한다.

3 작성 범위

- 인공지능 학습용 데이터 구축에 필수 공정단계인 데이터 획득·정제·라벨링·검사 단계를 본 구축 안내서의 작성 범위로 한다.
- 음성(소리) 타입의 원시데이터를 클래스 분류 또는 텍스트 타입으로 라벨링하는 데이터 구축을 본 구축 안내서의 작성 범위로 한다.
  ※ 음성(소리) 타입으로 라벨링하는 데이터 구축은 제외
- 기존 TTA 인공지능 학습용 데이터 구축 가이드라인, '20년 1차·추경(2차) 인공지능 학습용 데이터 구축사업 수행 및 참여기관의 데이터 구축 가이드라인, 국내 주요 인공지능 전문기업이 인공지능 학습용 데이터 구축을 위해 제작한 인공지능 데이터 구축 지침·가이드 등의 자료들을 검토하여 인공지능 학습용 데이터 구축 시 공통적으로 고려해야 할 사항들을 도출하여 구축 안내서에 반영한다.

4 용어 정의

- 데이터 획득 (Data Acquisition)
  - 인공지능의 기계학습에 필요한 데이터를 현실 세계에서 직접 수집 또는 생성하거나, 이미 보유하고 있는 조직이나 시스템 등으로부터 법률적 제약이 없도록 '원시데이터'를 확보하는 활동

- 데이터 정제 (Data Refinement)
  - 획득한 원시데이터를 기계학습에 필요한 형식으로 맞추거나 불필요한 중복을 제거하며, 개인정보를 비식별화하여 처리하는 등 일련의 전처리 과정을 통해 '원천데이터'를 확보하는 활동

- 데이터 라벨링 (Data Labeling)
  - 인공지능이 기계학습에 활용할 수 있도록 기능이나 목적에 부합하는 정보를 원천데이터에 부착하는 활동

- 라벨링데이터 (Labeled Data)
  - 원천데이터에 부여한 '참값', 파일형식이나 해상도 등의 속성, 그리고 설명이나 주석 등이 포함된 '어노테이션'의 집합

- 원시데이터 (Raw Data)
  - 기계학습을 목적으로 획득 단계에서 수집 또는 생성한 음성, 이미지, 영상, 텍스트 등의 데이터

- 원천데이터 (Source Data, Unlabeled Data)
  - 원시데이터를 라벨링 공정에 투입하기 위해 필요한 전처리 등 정제 작업을 수행한 데이터로 라벨링데이터가 부여되지 않은 상태의 데이터

- 인공지능 학습용 데이터 구축
  - 임무정의, 데이터 획득, 데이터 정제, 데이터 라벨링 등 인공지능 학습용 데이터를 구축하는 일련의 활동

- 참값 (Ground Truth)
  - 인공지능의 기계학습 목적에 따라 원천데이터에 라벨링된 정확한 값이나 사실의 의미적 표현

- 어노테이션 (Annotation)
  - 데이터 라벨링 시 원천데이터에 주석을 표시하는 작업을 의미하며, 추가 부착되는 설명정보 데이터는 기능 목적에 따라 다양한 형태로 표현될 수 있으며 이러한 설명정보 표현방식을 지칭

    ※ 용어사용 예 : 사물 바운딩박스 어노테이션, 클래스 라벨링 어노테이션 등

- 광학문자인식 (OCR, Optical Character Recognition)
  - 사람이 쓰거나 기계로 인쇄한 문자의 영상을 기계가 읽을 수 있는 문자로 변환하는 것

    ※ 자세한 용어 정의는 '인공지능 학습용 데이터 품질관리 가이드라인 Ⅴ.부록-1 용어정의'를 참조

제2장 | 구축 가이드라인 작성 방법

1 데이터 구축 목적 정의

- 데이터 구축 목적 정의
  - 인공지능 학습용 데이터 구축 목적은 단순한 데이터 수집, 모음이 아닌 구축된 데이터를 인공지능 학습 모델에 적용하여 의미있는 수준의 정확도를 확보하고 서비스 등에 유용하게 활용되는 것을 목표로 정의한다.
  - 목적 정의에는 데이터의 구축 배경 또는 필요성, 구축되는 데이터에 대한 명확한 정의, 구축 방향 및 활용(예상) 분야 등을 포함한다.
  - 구축될 학습용 데이터가 실제로 어떤 산업, 서비스, 연구분야에서 활용될 수 있는 지 정의하여 데이터 구축 방향에 대한 타당성을 재확인한다.
  - 데이터의 저장, 기록이나 해석에서 오류의 가능성이 없도록 명확한 단어, 어휘체계를 사용하여 정의한다.

【참고】 데이터 구축 목적 정의 예시(한국어 상담음성 데이터)

- 데이터 구축 목적
  - 한국인의 음성을 문자로 바꾸어 주고, 문맥을 이해하는 한국어 음성언어처리 기술 개발을 위한 AI 학습용 한국어 음성 DB 구축을 목표로 유무선, 웹 기반 등 다양한 방식으로 상담센터에 연락하여 상담하는 내용을 녹음한 음성 데이터를 구축
  - 구축되는 상담 데이터는 특정 도메인에 국한되지 않고, 다양성을 확보할 수 있도록 3개 이상의 도메인으로 데이터셋을 구성해야 하며, 음성인식 성능을 높일 수 있도록 1000명 이상의 녹취 인원을 확보
- 데이터 구축 필요성
  - 스타트업 기업은 다양한 고객 대응 서비스를 위한 콜센터를 운영이 어려우며, 개인 휴대전화로 주문, 상담, 환불, 배송조회 등의 콜센터 업무까지 직접 수행하는 실정
  - 일반 콜센터의 업무 시간이 18시로 제한되어 있어서 적극적인 대 고객 서비스를 제공하지 못해 매출 신장 및 소비자의 만족도 향상에 한계가 있음
  - 24시간, 365일 대 고객 서비스를 제공할 수 있는 AI 상담센터의 필요성이 대두되며, 이러한 AI 콜센터 구축을 위한 데이터셋을 구축이 필요
- 데이터 활용분야
  - AI 상담센터를 위한 음성상담 음성인식기술 및 언어이해, 언어생성 연구 및 서비스 개발에 활용할 수 있음

2 데이터 구축 시 고려사항

- 데이터 종류 및 규모
  - 획득해야할 데이터의 규모를 설정한다. 이때 대상으로 하는 산업분야 및 서비스에서 요구되는 수준과, 사업기간과 획득에 드는 시간과 비용을 종합적으로 고려하여 구축 규모를 선정한다.

- 어노테이션 타입
  - 데이터 활용 분야를 고려하여 구축되는 데이터의 어노테이션 타입을 정의한다.

표 2-1. 음성 데이터 어노테이션 타입 및 용도

어노테이션 타입	주요 활용 용도
• 클래스 라벨	• 오디오 분류(Audio Classification) • 오디오 세그멘테이션(Audio Segmentation)
• 텍스트 전사	• 음성인식(음성→텍스트 변환)((Speech to Text)

- 데이터 구축 프로세스 정의
  - 데이터 구축 목적 정의, 데이터 획득, 데이터 정제, 데이터 라벨링, 데이터 검사에 이르는 일련의 데이터 구축 프로세스를 사전에 정의하고, 각 프로세스에 따르는 이슈 및 검토사항 등을 도출한다.
  - 데이터 구축 프로세스는 구축 단계별 주요 작업에 대해 서술하나, 순서도·표 등을 활용해 구조화하여 구축 관계자 및 작업자들이 쉽게 이해할 수 있도록 한다.

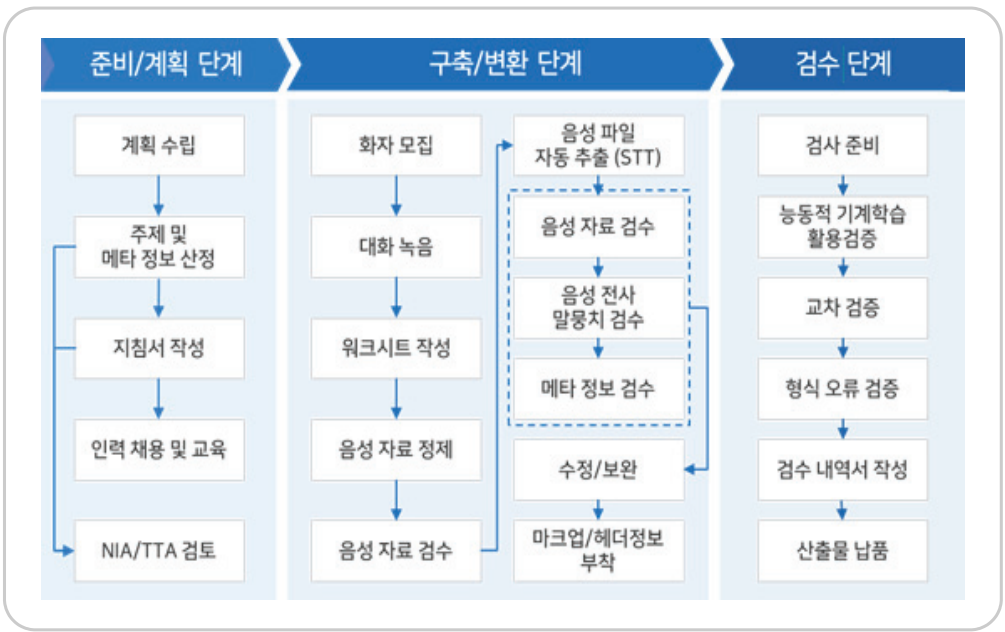


그림 2-1. 데이터 구축 프로세스(순서도 형식) 정의 예시(한국어 방언 음성 데이터)

표 2-2. 데이터 구축 프로세스(표 형식) 정의 예시(한국어 상담 음성 데이터)

구축 단계	세부 절차	설 명
수집	수집 도메인 선정	구축하려고 하는 서비스 업종 등의 적용 분야를 선정
	서비스 시나리오 작성	제공하려는 서비스 시나리오를 수립하여 작성
	시나리오별 스크립트 작성	하나의 시나리오에서 다양한 경우의 문장들을 작성하고 상품명 같은 엔티티를 다양하게 적용하여 작성
	크라우드소싱을 이용한 녹취	콜센터의 특징상 전화 녹취가 될 수 있는 환경을 제시해야 하며, 동일 문장의 녹취 인원이 최대 50명이 넘지 않도록 함.
정제	원천데이터 검사	스크립트와 다른 녹취, 과도한 소음이 포함, 녹취자의 목소리 너무 낮은 음원 등을 제거하는 과정을 통해 데이터를 정제
라벨링	라벨링 인력 교육	인공지능 학습용 데이터 라벨링에 필요한 작업 교육과 훈련 수행
	데이터 라벨링	음원의 녹취와 스크립트 상의 불일치 부분에 대해 스크립트 수정과 음원을 문장 단위로 분리 저장
검사	음성 모델 학습	구축된 데이터셋의 음원으로 음성모델 생성
	전수 검사	생성된 음성 모델을 통해 자동 전사된 텍스트와 스크립트와 텍스트 비교

- 데이터 품질 수준
  - 데이터 제작을 의뢰하는 고객이 있는 경우 고객이 요구하는 데이터 품질수준을 기본으로 하며, 세부적인 사항은 협의하여 결정한다.
  - 특정 고객 없이 범용적으로 활용할 수 있는 데이터를 제작하는 경우에도 해당 산업 및 서비스 분야에서 요구되는 품질 수준을 갖추기 위해, 해당 분야 산업 관계자 및 전문가 등의 검토를 통해 적절한 품질 수준을 설정한다.
  - 데이터 활용 목적에 맞는 데이터를 구축하는지 구축 데이터의 시간대, 주제, 비효율성, 인과관계 등을 검토한다.

- 구축되는 데이터는 모집단 및 프로세스에 대한 충분한 정보를 얻을 수 있는 지, 구축되는 도메인의 모집단 또는 프로세스를 대표할 수 있는지 검토한다.
- 품질 수준 및 측정 방법은 논문, 연구, 유사 사업(사례) 등을 통해 객관적이고 명확하게 제시되어야 하며, 기존에 구축된 인공지능 학습용 데이터와 최소한 동일 품질 수준 이상을 갖추는 것을 목표로 해야 한다.

● 데이터 제작 도구
- 구축 대상 데이터가 수행 및 참여기관에서 보유한 구축 도구(소프트웨어)로 목표로 하는 수준으로 제작할 수 있는지 검토한다.
- 제작 도구는 자체 개발한 솔루션을 활용하거나, 타사의 상용 솔루션, 또는 오픈소스 도구를 활용할 수 있으며 이 중 적합한 방법을 선택한다.
- 구축할 데이터의 특성에 맞게 구축 도구에 관한 환경설정을 진행한다.

● 가이드라인 수정 및 이력관리
- 인공지능 학습용 데이터 구축 진행 중 발생되는 예외상황(edge case), 애매모호한 상황 등 데이터 구축 설계단계에서 제작한 가이드라인에서 변경이 필요한 사항이 발생 시, 가이드라인을 업데이트하고 작업자에게 신속히 배포할 수 있는 방안을 마련한다.
- 구축 과정에서 큰 영향을 미치는 작업방법, 라벨링 세부조건 등에 관한 변경사항 발생 시 고객사 또는 각 산업 전문가 및 관계자 등의 검토·협의를 통해 적합한 방법을 도출할 수 있도록 변경 검토 절차를 마련한다.

※ 변경사항에 대한 검토 및 배포가 제대로 이루어지지 않을 때, 최악의 경우 라벨링을 다시해야 할 수 있음

- 작업 및 검사 인력 운영 방식
  - 대량의 데이터를 구축해야 하는 인공지능 학습용 데이터 특성 상 필요한 작업자 수와 수행 및 참여기관의 시설·노동 환경을 고려하여 내부조직, 아웃소싱, 크라우드소싱 또는 혼합 방식 등 적합한 작업자 운영 방식을 선정한다.
  - 작업자의 성과를 측정하는 기준을 마련하고, 특히 크라우드소싱 작업자 등 정확한 노동시간을 측정하기 어려운 작업자에 대한 임금 또는 보상 지급기준을 정의한다.

【참고】 크라우드소싱 작업인력 운영 방식 수립 사례

- 기본 원칙
  - 입사 후 7일간 수습 신분으로 Half time (주20시간) 근무
  - 성실성 등 근무태도 평가 후 단기계약 가능
  - 계약 후 한달 근무 완료 시 아래의 세 가지 근무 타입 중 선택

[크라우드소싱 작업자 근무 타입별 운영방식 및 특성]

업무 장소	근무 유형	운영방식 및 고려사항
사무실 (인하우스)	하프타임	• 13시~18시 근무 • 단시간 집중도 높은 근무 • 전체 근무자 중 64.8% 비중
	풀 타임	• 09시~18시 근무 • 작업 이해도 및 작업 역량 단기간 내 확보 가능 • 매니저, 리더 포함 근무자 중 35.2% 비중
재택	자율근무	• 일정 퀄리티 이상의 작업 결과에 대해 건별로 보수 지급 • 작업 내용에 대한 피드백 지연 발생 • 인하우스 근로자 대비 퇴사율이 높아 계약 베이스로 운영하지 않을 경우 리소스 예측과 관리가 어려움

표 2-3. 작업자 운영 방식 특성 비교

구분	내부조직	아웃소싱	크라우드소싱 (crowd-sourcing)
특징	• 품질에 대한 상시 교육 및 피드백 가능 • 작업환경을 위한 운영비, 작업공간 및 인프라 필요(전기·통신시설, 컴퓨터 등)	• 높은 업무 전문성 및 경험 보유 • 요구사항 정의 및 기준 합의에 많은 시간 소요	• 높은 업무 접근성 (장소제약없음) • 품질 교육 및 피드백에 한계가 있음 • 크라우드소싱 대가 산정에 대한 명확한 기준 마련이 어려움
적합 용도	• 머신러닝 훈련에 대한 높은 수준의 이해가 필요한 작업 • 라벨링 결과에 대한 공정간(수집, 정제, 라벨링, 학습모델 등) 긴밀한 피드백을 요구하는 작업	• 데이터 구축에 전문적인 지식과 숙련도가 요구되는 작업	• 단기간에 대량의 데이터를 처리해야 하는 작업 • 작업 난이도가 비교적 낮고, 데이터 보안수준이 낮은 작업

- 작업자 대상 매뉴얼 작성
  - 데이터 획득·정제·라벨링·검사 단계에 참여하는 작업자들이 인공지능 학습용 데이터셋 구축 취지에 부응하여 데이터 제작이 이루어질 수 있도록 작업자들이 직접 활용하는 매뉴얼(예 : 수집 매뉴얼, 라벨링 매뉴얼 등)을 제작한다.
  - 작업자 대상 매뉴얼에는 구축 목적·정의, 제작 절차, 제작 도구 활용방법과 작업 기준, 작업 결과 처리·저장 방법 등의 내용을 포함한다.
  - 작업자 관점에서 데이터 제작 과정에서 발생할 수 있는 다양한 유사사례 및 예시 등을 포함하여 매뉴얼을 제작한다.

3 데이터 획득 및 정제 방법

3.1 데이터 정의

- 원시데이터 분석
  - 인공지능 학습용 데이터 구축에 필요한 원시데이터 항목을 검토하고, 각 항목 별로 데이터 획득에 필요한 정보(데이터 획득정보, 획득방법, 획득 단계에서 필요한 요건 등)들을 검토하여 문서화한다.
  - 원시데이터 대상 및 획득 방법을 아래와 같이 육하원칙에 따라 정의할 수 있다.

표 3-1. 원시데이터 획득 시 검토사항 및 예시

5W1H	항 목	예 시
What	• 측정대상 • 획득 시 포함되어야 할 변수들	• 일반인의 일상적 대화 주제별 음성 • 장비별, 객체별, 시간별, 종류별, 사람별, 지역별 검토 (필요시 도메인 전문가, 인공지능 전문가 협의 후 대상 객체를 명확히 함)
When	• 획득 기간 (From, To)	• 2주간(11.14~11.28), 평일 7시간(9:00~12:00, 13:00~17:00)
Where	• 획득장소 / 프로세스	• 00동 00 스튜디오 내 녹음실
Who	• 획득 담당자 / 획득하는 사람	• 00 주식회사 미디어센터 내 데이터 수집 담당 • 그 외 크라우드 소싱 인력 20명
How	• 획득 방법, 측정주기, 샘플 크기, • 데이터 양식	• 녹음인원 1팀(2명)씩 입장후 정해진 스크립트를 녹음, 녹음 시 화자의 발음, 톤 모니터링, 1팀당 최대 20분 분량 녹음, 녹음 종료 후 녹음결과 리뷰
Why	• 측정 목적 / 기대 결과	• 목적에 맞는 획득 데이터 이해와 프로세스 능력의 파악 / 추세분석

- 획득할 원시데이터 내역에 대한 정의 및 현황정보 등의 사항을 정리한다.

표 3-2. 원시데이터 획득 대상 및 규모 정의(한국어 방언 음성 데이터)

지역	1그룹(10대~20대)	2그룹(30대~40대)	3그룹(50대~70대)	합계
강원도	800	800	400	2,000
경상도	800	800	400	2,000
전라도	800	800	400	2,000
제주도	800	800	400	2,000
충청도	800	800	400	2,000
합계	4,000	4,000	2,000	10,000

- 음성 데이터는 일반적으로 사람의 음성과 사물·동물의 소리로 구분되며, 그 중에서 사람 음성은 화자와 발화되는 텍스트로 구성된다. 데이터 구축 목적에 따라 발화된 텍스트 내용과 함께 획득해야할 추가적인 정보들을 정의한다.
- 화자의 특성이 반영되어야 하는 학습용 데이터의 경우 화자의 성별, 연령, 지역 등 세부 정보를 확보해야 하며, 화자 특성별 그룹을 나누는 근거를 제시한다.

표 3-3. 음성 데이터 주요 추가 정보

특 성	분 류
화자 특성	• 성별, 연령, 화자 수 등
텍스트 특성	• 주제, 분야, 문장 형태, 문맥, 화자 수 등
상황 특성	• 일반 상황, 긴급 상황, 소음이 심한 상황 등

- 원시데이터 포맷
  - 원시데이터의 파일 형식은 특정 획득 장비 및 처리 도구에 종속되지 않으며, 보편적으로 통용되는 포맷을 활용한다.

※ wav, mp3 등

※ 펄스부호변조(pcm, Pulse-code modulation) 타입으로 학습용 데이터를 구축하는 케이스는 본 구축 안내서에서 다루지 않음

- 원시데이터 획득 규모
  - 원시데이터 획득 후 정제, 라벨링, 검사 과정에서 기준 미충족으로 버려지는 데이터 양을 고려하여 구축 목표치 이상의 데이터를 획득하도록 계획한다.

    ※ 구체적인 목표치 대비 획득량은 데이터 구축 공정 난이도 및 구축기간 등을 고려하여 설정

3.2 획득 데이터 특성 분석

- 원시데이터 획득 관련 이슈사항 도출
  - 획득할 원시데이터의 범위 및 방법을 명확히 하기 위해 데이터 규모·획득범위·수집처 등에 대한 세부 이슈사항을 도출하여 가이드라인에 기술한다.
  - 녹음 환경에 대한 구체적인 정보(녹음기기, 녹음상태, 성능, 환경정보)를 파악 및 분석한다.

【참고】 원시데이터 특성 분석 예시(한국어 방언 음성 데이터)

- 녹음 환경 구성
  - 두 명의 화자가 편안하게 이야기할 수 있는 사무실 환경 마련
  - 녹음실은 외부와 차단된 상태로 대화에 참여한 두 명만이 대화할 수 있도록 구성
  - 화자는 각각 헤드셋 마이크를 착용하고 발화
  - 상대방의 목소리가 들어가지 않도록 적정거리 유지
- 녹음 화자 모집
  - 특정 성별, 연령, 지역 등이 편중되지 않도록 사전 협의하여 진행
  - 한 화자당 최대 녹음시간은 가능한 약 30분으로 하고 동일 화자가 중복 참여하지 않도록 제한하나, 동일 주제가 아닐 경우에는 허용
  - 녹음 화자 모집 시 최초 2인 1조로 신청자를 최우선으로 하며, 1인이 개별 신청했을 경우 비슷한 연령대 및 관심사를 구분하여 조 편성
  - 주제에 따라 1인 녹음, 3인 이상 녹음을 허용
- 연령별 그룹 기준으로 분류
  - 총 3개의 그룹으로 구성되며 1그룹은 10대~20대, 2그룹은 30대~40대, 3그룹은 50대~70대로 정하였으며 녹음이 어렵다고 판단되는 0~9세 / 80세 이상의 대상자는 제외이나 녹음이 가능할 경우 3그룹에 포함하여 진행

- 원시데이터 적합성 검토
  - 원시데이터 항목별 데이터 획득 방법, 법적문제 발생가능여부 등을 검토하여 실제로 인공지능 학습용 데이터 구축에 활용할 수 있는 데이터를 선정한다.

- 원시데이터 선정
  - 데이터 품질, 획득 가능성(가능여부 및 획득량), 획득 비용, 수행 및 참여기관의 기술수준, 법적 요건 등을 검토하여 획득할 데이터를 최종 선정한다.
  - 선정된 원시데이터를 획득하기 위해 필요한 정보, 또는 원시데이터 획득현황을 파악하기 위한 데이터 명세서 또는 정의서를 작성하여 데이터 획득 기준으로 활용한다.

표 3-4. 원시데이터 명세서 작성 예시(한국어 대화 음성)

데이터 명		한국어 대화 음성 AI 데이터
데이터 포맷		음원: ***. PCM, 전사: ***.txt, 메타정보: json, xml
데이터 요약		한국인의 일상 대화를 인식하고 음성을 문자로 실시간 변환하는 AI기술 개발을 위한 대화음성 데이터 셋 구축 - 특정 상황(대화주제), 말씨나 말투 등 환경에 국한되지 않고 대화·음성 추출, 발화자 연령·특성·단위 분류를 통해 활용성 확보
데이터 출처		크라우드소싱 업체, 춘천 MBC, EBS
데이터 이력	**배포버전**	koreaspeech0000000.txt v1.0
	개정이력	신규
	작성자/배포자	수행기관(000)
데이터 통계	**데이터 구축 규모**	총 4,000시간 1TB 이내 - 연령별(1,000H), 지역별(1,000H), 방송콘텐츠(2,000H)
	데이터 분포	연령별 : 1그룹(10대~20대)(33.3%), 2그룹(30대~40대)(33.3%), 3그룹(50대~70대)(33.3%) 지역별 : 수도권(30%), 강원도(10%), 충청도(10%), 경상도(20%), 전라도(20%), 제주도(10%)
기타 정보	**대표성**	수도권/강원/충청/전라/경상/제주 6개 지역
	독립성	원시데이터는 라벨링데이터와 별개로 데이터셋으로 제공하므로 독립성 유지
	유의사항	인공지능 학습데이터를 활용한 다양한 알고리즘 도출
	관련 연구	-

3.3 획득 절차 및 항목

- 데이터 획득·정제 절차 수립
  - 원시데이터 획득 및 정제 절차 수립 시 데이터 획득 방법별로 명확하게 획득·정제 절차가 정의될 수 있도록 한다.

    1) 원시데이터 직접 제작

    ※ 녹음, 녹취 등

    2) 수행 및 참여기관 내·외부에 있는 데이터 수집

    ※ API, 크롤링, 직접수령 등
  - 데이터 관점 뿐만 아니라, 기관간 역할, 작업자 업무, 작업자-관리자 간 관계, 행정요소 등 사람 관점에서 실질적인 구축작업에 필요한 사항을 종합적으로 고려하여 절차를 수립한다.

표 3-5. 데이터 획득 방안 정의 예시

	데이터 획득 형태	수집장비	데이터 형식	수집처(장소)	담당 인원
1	스튜디오 녹음	스탠드 콘덴서 마이크	WAV / MP3	000 스튜디오 (00시 00구 00동)	과제 인력 및 크라우드소싱 인력
2	동영상 크롤링	크롤링 서버	동영상 포맷 → MP3	00 동영상 사이트 (oo.com)	000사 크롤링 담당자

- 음성 데이터 획득 방법·절차 수립
  - 데이터 직접 제작 시* 데이터 제작 범위, 제작과정 및 지침, 등록 및 저장방법 등을 중심으로 획득 방법·절차를 수립한다.

    ※ 크라우드소싱 방식으로 데이터를 수집하는 경우에도 해당
  - 외부 데이터 수집을 통한 데이터 획득 시 데이터 수집 범주, 수집량, 수집처별 수집방법, 저장방법 등을 중심으로 획득 방법·절차를 수립한다.

【참고】 음성 데이터 획득 환경 구성 예시

- 수집(녹음) 환경은 사람의 대화, 독백, 사물 및 동물에 대한 소리 수집 환경에 맞게 환경을 구성 또는 조건에 맞는 환경일 때 수집한다.
- 사람의 대화 수집의 경우 외부 잡음 및 조용한 곳에서 진행하고 울림 현상이 없는 곳에서 진행하도록 하여야 한다. (예: 전용녹음실, 울림이 없는 회의실, 조용한 가정, 스터디카페 미팅룸 등)
- 부득이하게 울림이 심한 곳에서 녹음을 진행할 경우 벽 또는 창문에 커튼을 이용하여 울림 현상을 최소화하여야 한다.
- 동물 또는 사물의 소리를 수집할 경우 다른 외부 잡음이 들어가지 않도록 주의하여야 한다. (예: 사람의 말소리, 자동차 등 주변 소음 등)

- 데이터 구축 목적과 녹음 환경에 따라 알맞은 녹음 장비 및 음질을 선정한다.

【참고】 음성 데이터 녹음 장비 검토 예시

- 수집(녹음) 장비 : 아래의 장비들을 활용해 사전 녹음을 해보고 목적에 가장 적합한 장비유형을 선택한다.
  - 휴대폰 음성 녹음기 : 누구나 녹음 가능
  - 핀마이크 : 옷에 부착하여 휴대성 우수
  - 붐마이크 : 고품질 녹음이 가능
- 음질 : 소리의 샘플링 주파수(8K ~ 44.1K, 48K)를 다르게 녹음하지만, 샘플링 주파수에 정확한 규격이 없는 경우 고품질 샘플링으로 녹음 진행을 기본으로 한다.

※ 낮은 샘플링 주파수로 녹음할 경우 요구사항에 맞는 샘플링 주파수로 변경이 발생할 경우 문제가 발생할 수 있음

- 사람 음성이 아닌 사물·동물 소리 녹음 시 녹음 방법을 검토한다.

【참고】 사물·동물 소리 녹음 시 검토 예시

- 기타 사물 또는 동물의 소리를 녹음할 경우는 동영상 장비를 이용하여 촬영 후 소리만 발취하여 사용할 수도 있다.
- 필요한 소리를 수집하고자 할 경우 사람처럼 소통이 불가능하기 때문에 동영상으로 계속 촬영 후 필요한 부분만 편집하여 소리만 분리하여 사용하도록 한다. 소리에 대한 라벨링 작업은 영상을 확인하여 메타 데이터(라벨링) 작업을 하도록 한다.

- 음성 녹음 절차 및 주의사항을 수립하고 녹음자 대상 교육·안내를 실시한다.

【참고】 음성 데이터 녹음 규칙 예시

1. 데이터 수집(녹음) 진행 시 대화 형태의 녹음인 경우 구성원은 진행요원 1명, 대화를 하는 화자 2~3명으로 구성하되 가능하면 화자 3명이 녹음을 진행할 수 있도록 한다.
2. 2명이 녹음하기로 계획할 때 녹음 신청 지원자 결원이 발생할 경우 녹음 진행을 할 수 없는 문제가 발생된다.
3. 녹음을 신청한 화자의 개인정보를 확인하고 개인정보 이용 및 제3자 제공 동의서, 저작권 이용허락 계약서를 체결한다. 작성이 완료되면 녹음 시 유의 사항들을 충분히 설명하고 녹음을 진행한다.
4. 데이터셋 녹음 시간이 10분 단위 기준으로 녹음할 경우 최소 13분에서 최대 15분으로 녹음을 진행한다. 녹음 후 정제를 진행하면서 묵음 구간 잡음 구간 등 편집을 하게 되면 필요한 데이터셋 기준에 필요한 10분의 시간이 부족한 경우가 자주 발생하기 때문이다.
5. 동일한 화자가 동일한 주제로 중복 대화하지 않는 것을 원칙으로 한다.
6. 사람간 대화일 경우 원활한 녹음을 위하여 주제를 제시하고 대화를 나누는 방법으로 진행한다. 주제가 없는 대화일 경우 대화를 계속해서 이어가기가 어려운 문제가 발생된다.

※ 주의사항

① 녹음 중 웃지 말 것.
(어쩔 수 없이 웃음이 날 경우 고개 돌리고 소리는 내지 말 것)
② 절대 말을 끼어들지 말고 아, 네, 음 등의 추임새도 넣지 말 것
③ 발언 시 말끝을 흐리지 말고 명확하게 마무리 말 것
④ 발언 시 고개를 돌리지 말고 마이크 방향으로 대고 말할 것
⑤ 발언권을 얻었을 시 말을 급하게 시작하지 말 것
⑥ 절대 말을 서로 주고받지 말고 발언 순서 지킬 것
⑦ 질문을 받고 단답형으로 말을 끝내지 말 것
⑧ 기침은 고개를 돌리고 할 것
⑨ 휴대폰은 진동으로 하고 책상 위에 두지 말 것
⑩ 상대방의 질문이 끝나면 2초 후에 답변할 것
⑪ 진행요원이 녹음 종료하기 전까지 프로그램을 끄는 중에 말하지 말 것
(진행요원이 끝났다고 안내해주기 전까지는 침묵)

- 녹음할 대상에 따라 녹음 기준 및 내용을 정의한다.

표 3-6. 녹음 기준 수립 예시

항 목	녹음 기준
장소	• 방음 스튜디오 or 일반적인 조용한 실내
녹음시간	• 2단계 (최대 15초 / 최대 30초)
화자 수	• 2단계 (2명 / 3명)
주제 및 언어	• 한국어(전문용어 및 일상적으로 사용되는 외국어는 일부 포함 가능) • 일상적인 대화

- 데이터 획득항목 정의
  - 획득단계에서 음성 파일과 함께 확보해야할 정보를 정의한다.
    1) 음성 메타데이터 : 녹음일시, 길이, 녹음자, 화자, 화자 수 등
    2) 도메인 정보 : 클래스 정보, 주제 등
  - 음성 데이터 획득 시 수집 및 저장할 정보는 아래의 항목들을 정의한다.
    1) 라벨링 공통참조항목은 음성 데이터 획득 시 공통적으로 적용될 수 있는 메타정보이다.

표 3-7. 음성 데이터 획득 시 라벨링 참조항목

No.	속성명	항목 설명	Type	작성예시
1	Dataset.identifier	데이터셋 식별자	string	SPEECH_QnA_COMMON_01 (데이터유형_목적_분야_순번)
2	Dataset.name	데이터셋 이름	string	상담 관련 인공지능 질의응답 학습용 데이터 셋
3	Dataset.src_path	데이터셋 폴더 위치	string	/dataSet/speech/
4	Dataset.label_path	데이터셋 레이블 폴더 위치	string	/dataSet/speech/
5	Dataset.category	데이터셋 카테고리	number	0: 음성 분류, 1: 음성전사, 2:질의응답, 등
6	Dataset.type	데이터셋 타임	number	0: 텍스트, 1: 이미지, 2:영상, 3: 음성 등

2) 라벨링 선택항목은 원시데이터 출처 정보가 중요한 경우 선택적으로 기록할 수 있는 정보이다.

표 3-8. 음성 데이터 획득 시 라벨링 선택항목

No.	속성명	항목 설명	Type	작성예시
1	info.filename	원시데이터 파일명	string	NEWS_000001 (매체유형_순번)
2	info.date	원시데이터 생성일시	string	2020-12-10 17:00
3	info.category	원천데이터 카테고리	string	일상대화, 발표, 강의, 상담 등
4	info.mediatype	매체유형	string	직접 녹음, 뉴스, 블로그, SNS 등
5	info.speakers	화자 목록 및 화자 특성	object	화자1 {유형 : 상담사, 성별 : 남성} 화자2 {유형 : 고객, 성별 : 여성}
6	info.size	원시데이터 길이(시간)	number	15(초)

3) 라벨링 선택항목(저작권 정보)은 데이터 획득 시 저작권 정보가 필요한 경우 기록하는 정보이다.

표 3-9. 음성 데이터 획득 시 라벨링 선택항목(저작권 정보)

No.	속성명	항목 설명	Type	작성예시
1	licenses.id	라이센스 고유 번호	string	http://www.apache.org/ licenses/LICENSE-1.0
2	licenses.name	라이센스 이름	string	Apache License 1.0
3	licenses.url	문서 식별자	string	NEWS_000001

- 획득 데이터 저장 및 관리
  - 획득 파일에 대한 저장, 전송, 백업 등 관리 절차 및 방안을 수립한다.

【참고】 획득 데이터 저장 방안 수립 예시

- 녹음 진행 후 명동에 위치한 사무실에 복귀하여 녹음 데이터를 데이터 보관 서버, 클라우드 및 외장하드에 3중 백업 진행
  - 외장하드 고장에 대비하기 위해 NAS* 등의 추가 장비를 활용하여 주기적으로 백업
- 녹음 진행상황에 대한 일일 관리를 위해 “녹음일자”_“녹음자” 기재한 폴더 형태로 구성하여 원시데이터 저장

* NAS(Network Attached Storage) : 네트워크 결합 스토리지

  - 획득한 파일을 체계적으로 분류하기 위해 데이터 종류 및 분류에 따른 라벨링데이터 파일 명명법과 파일 저장구조를 정의하고, 정의된 내용에 맞게 파일을 저장한다.

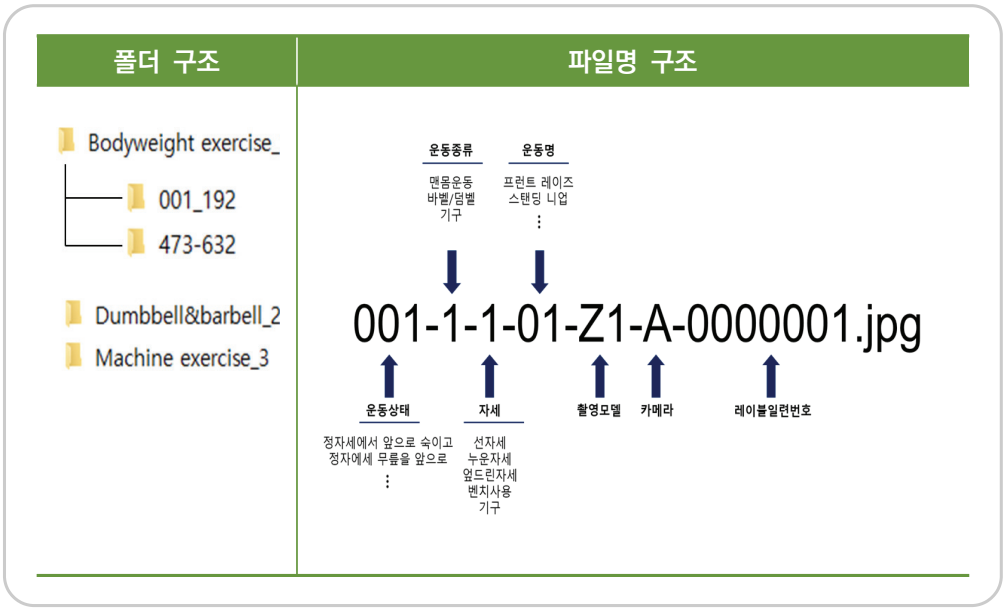


그림 3-1. 획득 데이터 폴더구조 및 파일명 코드화 예시

3.4 획득 데이터 정제 방식

- 정제 프로세스 수립
  - 어노테이션 단계에 들어가기 전에 학습용 데이터로 적합한 데이터를 선별하고 처리하는 정제 프로세스를 획득방법별로 수립한다.
  - 데이터 정제는 도구(소프트웨어)를 활용하여 정해진 규칙에 따라 제외 또는 변환하는 방법, 작업자가 직접 눈으로 확인하여 검사하는 방법 등을 적용할 수 있다.

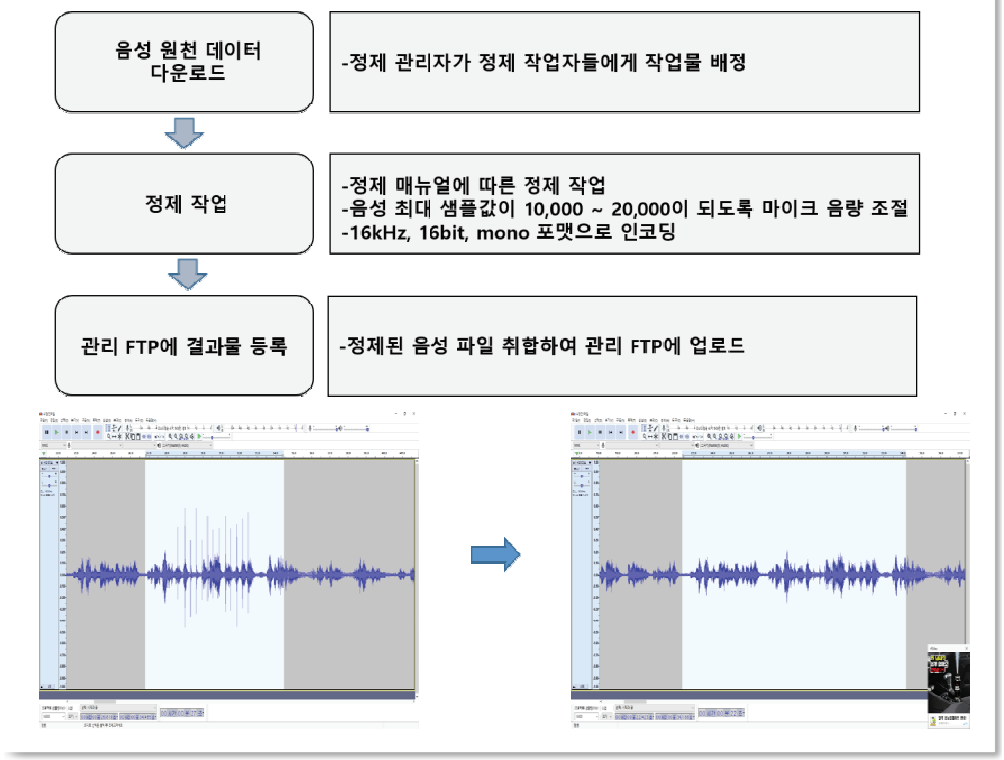
- 정제 기준 수립
  - 데이터 구축 목적, 데이터 유형, 도메인 특성에 따른 데이터 정제 기준을 수립한다.
  - 음성 데이터 정제 기준으로 음량, 발음의 정확성, 소음 및 잡음, 개인정보, 저작권, 그 밖의 데이터 구축목적 부합성 등의 요소를 고려하여 부적절한 데이터를 필터링하거나 라벨링하기 적합한 형태 및 내용으로 수정한다.
  - 데이터 라벨링에 포함하지 않아야 할 개인정보 등을 필터링하는 정제 기준을 마련한다.

표 3-10. 음성 녹음 시 주요 정제기준

기 준	고 려 사 항
음량	• 음량이 너무 크거나 작을 때 허용 범위
발음	• 화자의 발음이 불분명할 때 허용 범위
소음 및 잡음	• 음성 이외에 소음, 잡음이 심할 때 허용 범위
잘림	• 발화된 문장이 완성되지 않고 끝났을 때 허용 범위
안들림	• 음성이 들리지 않을 때 허용 범위
개인정보처리	• 개인정보보호법 위배 여부
저작권	• 저작권 침해 가능성 여부

【참고】 음성 데이터 정제 절차 예시

- 음성(소리) 데이터 정제는 녹음된 음성 파일을 작업자가 확인 후 제대로 녹음이 되었는지 확인한다.
- 음성(소리)의 처음부터 끝까지 들어보며 잡음, 말 겹침, 소음 구간, 긴 묵음 구간 등을 파악해 해당 부분을 사운드 편집 툴을 이용하여 수정 및 삭제 편집한다.



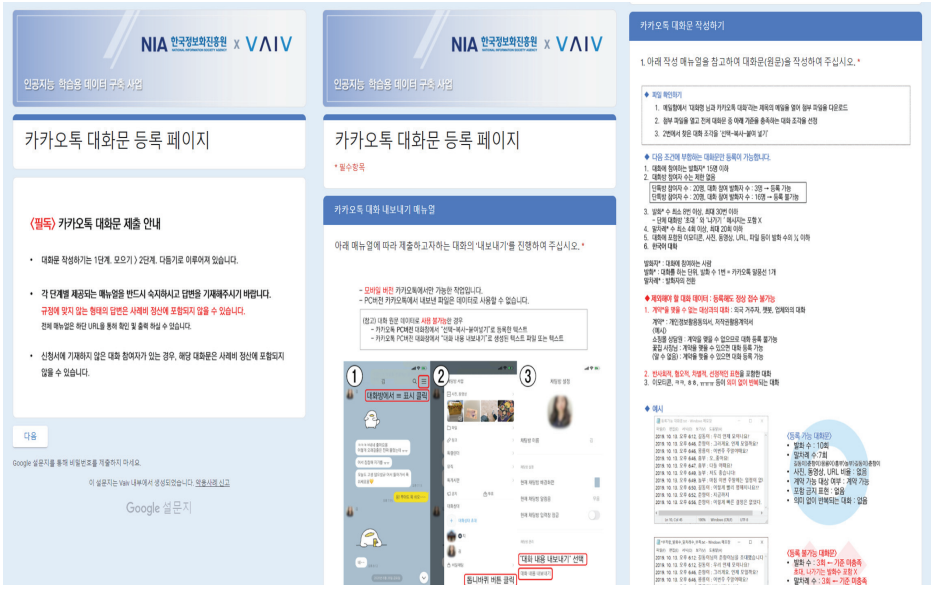
3.5 획득 도구 및 정제 도구

- 획득 및 정제도구
  - 데이터 획득 및 정제도구 개발 및 활용 계획을 작성한다.
  - 특별한 세팅 환경에서의 녹음을 요하는 데이터가 아닌 경우, 데이터 정제 시 단순 자르기-붙여넣기 수준의 작업으로 진행될 경우 시중의 오디오 녹음, 편집 툴을 활용할 수 있다.

- 데이터 획득, 정제도구를 자체적으로 개발하기 어려운 경우, 시중의 제작 도구 또는 그와 유사한 역할을 할 수 있는 서비스·애플리케이션을 활용할 수 있다.

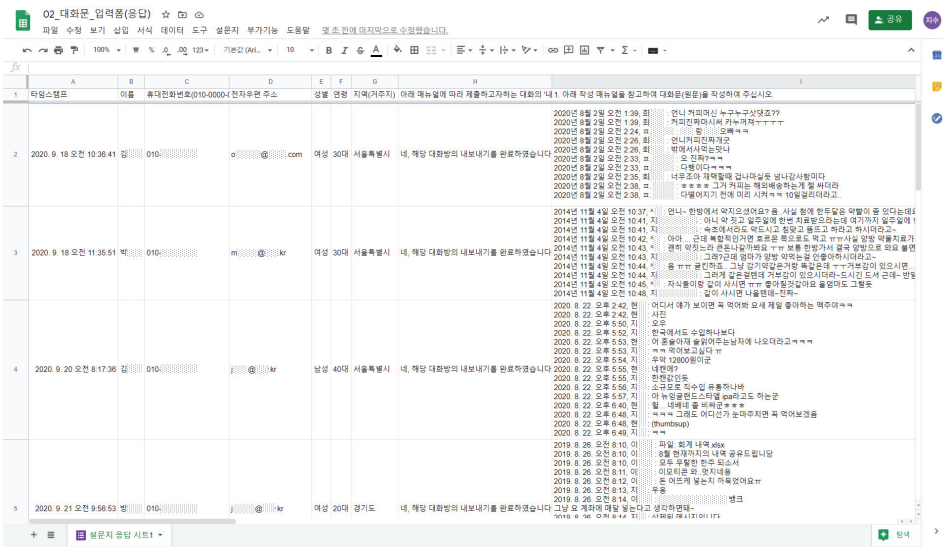
【참고】 데이터 획득·정제도구 구성 예시(기존 플랫폼 활용 사례)

- 구글폼(Google Forms)을 활용한 데이터 획득 및 저작 도구 도입
  - 본 과제에서 구축하는 데이터는 일반 텍스트 형식으로 구조화되고 유지하므로 온라인 공동 작업과 버전 유지가 이루어지는 구글폼(Google Forms) 시스템으로 데이터의 수집이 가능
  - 다양한 용도로 폭넓게 사용되고 있는 구글폼(Google Forms)은 실시간 응답 관리와 스프레드시트와 연동하여 모든 정보를 확인 가능하므로 데이터 활용도를 높일 수 있음
  - 공개된 클라우드 시스템 활용으로 데이터 수집 및 저작 도구의 개발 시간을 단축



[구글폼 활용 데이터 획득 도구 구성]

- 데이터 관리 및 정제
  - 구글폼(Google Forms)과 연계된 구글 스프레드시트에서 응답 데이터의 실시간 확인 및 가공이 가능함



[획득 데이터 관리(스프레드시트 활용)]

3.6 획득 시 고려사항

- 법·제도 준수
  - 데이터 획득 대상, 획득방법이 법·제도를 저촉하거나 또는 사회 윤리에 어긋나지 않도록 한다.
  - 개인정보 및 사생활 보호가 필요한 항목 획득 시, 개인정보보호법 등에 따라 적절한 법적, 기술적 절차를 거친 데이터를 활용하며, 그렇지 않은 데이터는 정제 과정에서 처리될 수 있도록 한다.

※ 법적 절차 : 개인정보 활용 동의, 초상 활용 동의, 명예훼손 가능성 여부 검토 등

※ 기술적 절차 : 데이터 유형별로 적용할 수 있는 익명처리 기법 적용

① 수치형 데이터 : 데이터 범주화 등

② 텍스트 데이터 : 이름, 민감정보 키워드 데이터 변환 등

③ 이미지·동영상 데이터 : 모자이크·블러처리, 크롭(자르기) 등

④ 음성 데이터 : 크롭(자르기), 음성변조 등

- 데이터가 3자 제공 및 대중에 개방에 문제가 없도록 법적요건 및 동의서 내용 등을 검토한다.
- 저작권 보호 대상인 데이터 획득 시 법에 저촉되지 않는 범위 내에서 획득할 수 있는 방안을 마련하며, 저작권 보호 대상 저작물 활용 필요 시 가급적 동의서, 계약서 등을 활용한 서면 자료 확보를 권장한다.

※ 예) 음성 내 특정 기업, 재품명 등 노출 가능 여부

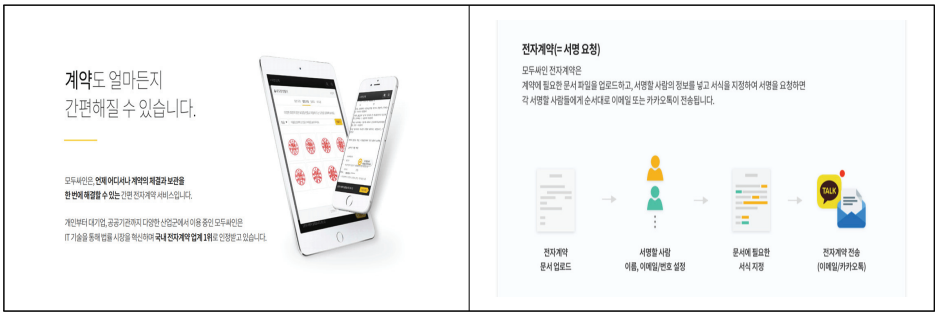
※ 예) 뉴스, 유튜브 영상 등 저작권 보호 대상 저작물 활용 시 관련 당사자·기업·기관과 협의방안

- 개인정보활용동의서 및 저작물 활용 동의서 등 법적 요건을 준수하기 위한 관리방안을 마련한다.

※ '인공지능 학습용 데이터 품질관리 가이드라인'의 Ⅲ.품질관리기준, 3.1.5 체계준수성-보안준수의 【참고】개인정보보호 및 보안관련 법령·고시·권고 참조

【참고】 온라인 서명 플랫폼 활용 예시

- 원시데이터 제공자와 개인정보 수집 및 이용 동의 및 저작권 활용 계약을 전자서명 계약 기반으로 체결함으로써 민감한 정보 제공에 따른 위험 부담을 최소화함
- 저작권 이용 허락 계약 체결은 온라인 서명 플랫폼을 활용해 계약서를 업로드하여 법적으로 유효한 온라인 서명 기능을 구현



● 데이터 다양성 확보

- 인공지능 학습모델이 현실을 잘 반영하고 본래의 구축 목적을 달성할 수 있도록, 획득하는 데이터가 일부 범주에만 치우치지 않고 가능한 다양한 시간·공간·집단·수준이 포함될 수 있도록 한다.

【참고】데이터 다양성 확보/미확보 예시

- (다양성 미확보) 전문분야 음성 학습 데이터 획득 시, 'IT 기술', '의료/보건' 2개 분야로 한정된 음성 데이터만 획득함
- (다양성 확보) 획득 분야와 획득 수량을 미리 정의하고 통계치로 관리하여 데이터 다양성을 확보함

분야	대상 자료	획득 파일 수
스포츠	지상파 3사(KBS·MBC·SBS) 최근 1년 스포츠 뉴스 영상 내 음성	5,000
금융 · 증시	지상파 3사(KBS·MBC·SBS) 및 경제 뉴스(000 등)의 최근 1년 경제분야 뉴스 영상 내 음성	4,000
IT	000 동영상 사이트 내 IT 채널(100개) 영상 내 음성	3,000
의료/보건	지상파 3사(KBS·MBC·SBS) 및 000 동영상 사이트 내 의료·병원 채널(100개) 영상 내 음성	3,000
향토문화	000 방송사 00 프로그램 최근 3년 영상 내 음성	3,000
음식	000 방송사 00 프로그램, 00 최근 3년 영상 내 음성	3,000
합계		21,000

● 데이터 편향 방지 및 윤리 준수

- 인공지능 학습모델이 인간의 비윤리 또는 편견을 학습하지 않고 사회적 윤리를 준수할 수 있도록 비윤리적 내용, 편견·편향된 데이터의 획득은 지양한다.

※ 딥페이크 분류, 가짜뉴스 분류, 비속어 필터링 등 비윤리·편향·왜곡된 정보 특성을 학습하는 것을 목적으로 구축하는 데이터는 예외로 할 수 있다.

- 사업계획서 및 데이터 구축 요건 일치
  - 사업계획 당시 정의한 데이터 구축 기준에 맞춰 데이터를 획득·정제하도록 구축 현황을 모니터링한다.

【참고】 사업계획서 및 데이터 구축 요건과 실제 데이터 획득 간 불일치 사례

1. 사업계획 시 MP3 포맷으로 획득하기로 했으나, 실제 구축되는 데이터에 그 외의 포맷(WMA, AAC 등)이 존재
2. 사업계획 시 128kbps 비트레이트 오디오 파일로 획득하기로 하였으나, 실제 획득한 데이터에 더 낮은 품질(92kbps 이하)의 파일이 포함됨
3. 구축 목적(노인 돌봄을 위한 감성 대화 서비스 구현) 대비 획득한 데이터의 범위가 너무 좁아서 구축 목적을 달성하기 불가능함 (의식주, 건강 관련 대화만 획득하는 등)

- 기타 음성 데이터 획득 시 품질 고려사항
  - 아래의 사례를 참고하여 음성 데이터 획득·정제에 필요한 사항을 가이드라인에 반영한다.

표 3-11. 음성 학습데이터 획득 시 품질 고려사항

항 목	내 용
기계 생성 음성	• 사람이 직접 녹음한 음성 대신, 기계가 자동생성한 음성(TTS)을 획득하려고 할 경우 인공지능 학습용 데이터셋 구축에 활용하는 데 적합한지, 그리고 구축 취지에 부합하는 지 검토한다. ※ 기계가 자동생성한 음성은 기존 인공지능 학습 알고리즘을 통해 도출된 '결과물'이므로, 이를 검사, 구축하는 것은 학습용 데이터 구축이 아닌 인공지능 모델 개발의 영역으로 볼 수 있다.

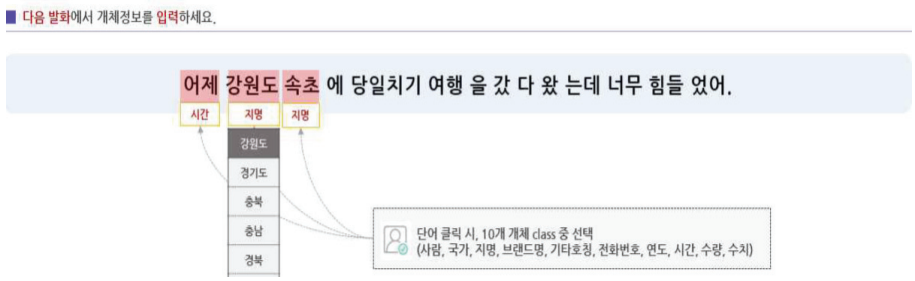
4 데이터 라벨링 작업

4.1 데이터 특성 식별 분류 체계 및 고려 사항

- 라벨링 작업 대상 및 범위 정의
  - 원천데이터 내에서 어떤 항목들을 라벨링 해야 하는지 대상과 범주를 정의한다.
  - 원천데이터 내에서 데이터 구축 목적에 부합하는 내용을 최대한 반영할 수 있는 정보를 라벨링할 수 있도록 라벨링 대상 범위를 정의하며, 데이터 품질 및 구축 목적과 무관한 내용을 불필요하게 라벨링하는 사항의 존재 여부 등을 검토한다.

【참고】 잘못된 라벨링 대상 설정 예시(대화 텍스트 감성 분류)

- 대화의 감성 분류를 목적으로 인공지능 학습용 데이터셋을 구축할 때, '시간', '지명' 키워드를 라벨링하는 것이 본래의 구축 목적 달성에 필요한 것인지 합리적인 근거가 없음



  - 원천데이터에 포함된 개인정보는 라벨링 대상에서 제외하거나, 익명처리 등 비식별화를 통해 개인정보를 알아볼 수 없게 라벨링한다.

※ 개인정보 활용 및 3자제공 동의를 받은 경우 동의 범위 내에서 개인정보를 라벨링 데이터로 활용할 수 있음

【참고】 개인정보에 대한 라벨링 익명처리 예시

- 대화자들의 신분 보장을 위해 이름, 주민등록번호, 카드 번호, 전화번호 등 개인정보와 관련된 사항은 그대로 전사하지 않고 아래와 같이 영문 기호로 적는다.
- 사람 이름은 '&name&'로 전사한다. 정치인, 연예인 등 유명인 이름은 그대로 전사한다.

(발화) 저는 김철수입니다.
(전사) 저는 &name&입니다.

- 여러 이름이 나올 때는 '&name1&', '&name2&' 등 name 뒤에 숫자를 붙여 구별한다.

(발화) 그때 철수랑 민수랑 너랑 나랑 갔잖아. 철수도 알고 있지?
(전사) 그때 &name1&이랑 &name2&이랑 너랑 나랑 갔잖아. &name1&도 알고 있지?

- 집 주소 등 개인정보에 해당되는 주소는 동 단위 까지만 전사하고, 그 이하 구체적인 주소는 영문 기호로 적는다.

(발화) 홍제동 한양아파트 205동 304호로 배달해주세요.
(전사) 홍제동 &address&로 배달해주세요.

- 특히 음성 전체에 대한 라벨링이 아닌, 하나의 음성 내에 특정 키워드, 문장 등을 라벨링하는 경우 작업자들이 어떤 대상을 라벨링해야 하는지 판단할 수 있도록 세부적인 기준을 마련한다.

- 클래스 정의 및 관리
  - 원천데이터의 특성을 바탕으로 부여할 수 있는 클래스 리스트 또는 클래스의 범주를 정의한다.
  - 클래스를 정의할 때는 원천 데이터 내에 존재하는 다양한 값들을 모두 커버할 수 있도록 정의하고, 클래스 이름이 중복되거나 모호한 의미를 갖지 않도록 한다.
  - 클래스 이름은 의미를 바르고 명확하게 나타낼 수 있도록 적절한 어휘를 선택한다.

표 4-1. 클래스 정의 예시

클래스명	클래스 해당 주제 예시
개인 및 관계	이름, 전화번호, 가족, 국적, 고향, 성격, 외모, 개인의 기호(선호), 직업, 종교, 반려동물, 연애(관), 결혼(관), 이상형, 인간 관계, SNS
주거와 생활	숙소, 방, 가구, 침구, 주거비, 생활 편의 시설, 지역, 지리, 가전 제품, 자취, 집안일, 육아, 부동산, 주거시설, 이사, 생활비, 자동차
상거래(쇼핑)	쇼핑 시설 및 장소, 식품, 의복, 가정용품, 물건 및 가격, 택배, 중고거래, 서비스, 교환 및 환불, 구매 후기
식음료	식사, 음식, 음료, 배달, 외식, 맛집, 식사 메뉴, 야식, 디저트, 요리
공공 서비스	우편, 전화, 통신, 휴대전화, 인터넷 서비스, 은행, 관공서
여가와 오락	휴일, 취미, 동아리 및 동호회 활동, 관심사, 방학, 휴가, 행사, 술, 웹서핑
일과 직업	취업, 스펙, 직장 생활, 업무, 회식, 급여, 계약, 협상, 회의
행사 및 모임	초대, 방문, 소개팅, 약속, 가족 및 친척 행사, 공적 행사, 사적 모임(친목 모임)
…	…

【참고】 잘못된 클래스 정의 예시 (텍스트 주제 및 감성 분류)

- 클래스 간 의미 중복이 있어 구분이 애매함
  - (학업) vs (학교폭력) : 학업의 개념 안에 학교폭력을 포함
  - (학업 및 진로) vs (진로/취업/직장) : 전자와 후자가 서로 교집합이 있음
  - (상처) vs (슬픔) : 전자와 후자의 경계를 나누기 모호함
- 라벨링 표기법에 일관성이 없음
  - (청소년) vs (청소년(10대)) : 전자와 후자를 같은 의미로 라벨링에 활용하였으나 표기법에 일관성이 없음
  - (여성) vs (FEMALE) : 전자와 후자를 같은 의미로 라벨링에 활용하였으나 언어·문자가 일관성이 없음
- 범위와 주제에 맞지 않는 클래스 존재
  - 성별 구분에 (기타) 클래스 존재 : (남성), (여성) 외에 (기타) 성별을 정의할 수 있는지, 또는 의미가 무엇인지 불명확

- 라벨링 진행 중에 이전에 정의되지 못했거나 새롭게 정의가 필요한 클래스가 발견될 경우 클래스 항목 업데이트 방안을 마련한다.

- 클래스를 정해진 목록에서 선택하지 않는 경우에도, 작업자마다 일관된 기준 및 규칙에 따라 속성값을 부여할 수 있도록 하는 기준을 마련한다.

※ 주로 OCR 이미지, 텍스트, 음성 전사 등 문장, 텍스트로 값을 부여하는 데이터가 해당

4.2 데이터 라벨링 방법 및 절차

- 개요
  - 획득→정제 과정을 통해 도출된 원천데이터를 라벨링하여 학습 데이터를 생성하기 위한 과정 및 고려사항을 작성한다.
  - 라벨링 지원 도구를 활용하며, 용어 및 분류체계를 준수하여 라벨링한다.

- 라벨링 작업 방식
  - 라벨링할 정보의 특성에 따라 자동, 반자동, 수동 방식을 결정한다. 원천데이터로부터 추출하는 방식이 정형화되어있고 자동화할 수 있는 사항인 경우 자동 방법을 고려할 수 있으며, 기계가 판단하기 어려운 사항은 반자동 또는 수동 방식이 적절하다. 반자동 방식은 자동으로 라벨링한 이후 사람이 다시 확인하여 수정하는 방식으로 작동된다.

- 작업 배분
  - 획득된 데이터를 라벨링에게 작업자에게 배분하고 라벨링 결과를 다시 저장하는 파일 저장체계 및 프로세스를 정의한다.

- 라벨링 작업 기준
  - 데이터별 어노테이션 기준, 라벨링 기준 등을 상세히 기술하며, 구체적인 예시를 들어 설명하여 작업자들이 혼동없이 명확한 기준을 갖고 빠르게 작업할 수 있도록 한다.

※ 레이블 범주, 레이블 부여기준(ground truth) 제시, 레이블 부여 예시, 애매한 내용이 나올 경우의 처리 기준, 자주 실수하는 예시, 검사 기준 등

【참고】 라벨링 작업 안내 예시(한국어 상담음성 데이터)

- 개요
  - 표준발성에서 벗어나거나 같은 전사에 대하여 두 가지 이상 발음이 가능한 경우 발음전사와 철자전사를 병행하며, 이 경우 (철자전사)/(발음전사)로 표기한다 (이 문서에서 향후 이를 '이중전사'라 칭한다). 예) (컴퓨터)/(컴퓨타)
  - 발음전사: 발성된 내용을 소리 값에 최대한 가깝게 표기한다. 이는 음성인식의 음향 모델링을 주된 목적으로 한다.
  - 철자전사: 표준어법에 맞게 표기한다. 이는 음성인식의 언어모델링 등을 주된 목적으로 한다.
  - 숫자, 외래어, 기호, 도량형 및 온도 단위는 발음 전사를 수행하되, 별도의 목록표를 생성하여 발음 전사별로 해당되는 표준 표기를 명시한다 (1.3, 1.7, 1.8절 참조).
  - 이중전사를 할 때, 이중전사의 범위를 표시하기 위해 괄호('(', ')')를 사용한다.
  - 이중전사, 잡음, 중복 발성 등을 나타내기 위한 특수 기호(meta symbol, 예: '/', '(', ')', '*', '+')는 원래의 목적으로만 표기되어야 한다. 특수기호가 실제 발성된 경우에는 발성된 형태를 반영하여 발음전사 한다. 분수 표기도 풀어서 표기한다.
  - 전사 과정에서 삽입되는 모든 기호('(), /')등)는 아스키코드만 사용하도록 한다.
  - 단일 발화 문장은 최소 2개 이상의 어절 혹은 5글자 이상으로 이루어져야 한다.
  - 단일 발화의 음성 길이는 최대 20초를 넘지 않도록 한다.

- 잡음
  - 잡음은 대표 발화자의 음성을 제외한 전사된 다른 모든 소리로 정의한다.
  - 단어의 앞과 뒤에 거의 붙어 발생된 잡음은 단어와 분리하여 표기한다.
  - 잡음이 있는 상황에서 사람에게서 발생하는 잡음은 명확히 구분될 정도로 큰 것만 표기해도 좋다.
  - 다음에 정의된 잡음 이름 뒤에 '/'를 붙여 표기한다.
    예) - b : 숨소리 / - l : 웃음 소리(laugh) / - o : 다른 사람의 말소리가 포함된 경우 문장의 맨 앞에 표기 / - n : 주변의 잡음 (상기에 명시된 요소 이외의 잡음)

- 발화자의 표기
  - 각 발화 문장의 발화자는 모두 표기한다.
  - 발화자의 성별 및 각 성별의 등장 순서에 따라 남성의 경우 'M + 등장순서', 여성의 경우 'F + 등장순서'로 표시한다.

예) 해당 음성의 첫번째로 등장한 남자: ‘M1’, 해당 음성의 첫번째로 등장한 여성: ‘F1’, 해당 음성의 3번째로 등장한 남자: ‘M3’

- 해당 발화자 판별이 어려울 시 ‘etc’로 표현한다.
- 하나의 대화 안에서 같은 사람일 경우 화자 정보를 동일하게 유지한다.
- 발화자 지역, 나이대, 성별, 녹음방식, 잡음환경, 수집디바이스에 대한 정보가 있을 경우 발화자 정보에 추가로 표시한다.
- 두 명 이상의 발화자가 동시에 말을 할 경우 주 발화자(먼저 말을 한 대화의 주체)가 대표 발화자가 되며 이를 기준으로 잡음 처리 규칙을 따른다.

- 음성 데이터 라벨링 작업
  - 음성 데이터를 텍스트로 전사하는 등의 라벨링 작업 시 구축 목적, 도메인, 활용 분야(챗봇 등)를 고려하여 라벨링 절차 및 기준을 수립한다.

표 4-2. 음성 라벨링(전사) 규칙 수립 예시

항목	전사 기준
대화 메타정보 및 화자정보	• 화자 간 대화 및 화자 메타 정보를 부착한다.
전사 입력 원칙	• 띄어쓰기를 준수하며, 한글 이외의 제반 기호와 특수문자 등은 사용하지 않는다. • 일반적인 기호, 전화번호나 카드 번호의 하이픈(-) 등의 특수문자나 괄호 등 구두 기호도 모두 소리나는 대로 적는다. • 영어 알파벳, 단어도 모두 소리 나는 대로 한글로 전사한다.
방언 전사 원칙	• 이중 전사를 원칙으로 하며, 그 형태는 다음의 예시와 같이 “(방언)/(표준어)” 형태로 기재하되 괄호와 빗금 사이에는 공백을 두지 않는다. • 방언형에 대한 표준어 대응쌍은 가급적 음절 및 어절 수를 맞추어 제시한다.
전사 단위	• 기본 원칙은 문장 단위로 하나, 중간에 쉼이 있는 경우에는 단위를 구분한다. • 한 줄에 전사하는 분량이 6초를 넘지 않도록 제한하며, 이때 띄어쓰기 단위는 10개 이내가 된다. • 단위 설정 기준에 대한 판단이 어려운 경우에는 일단 줄의 바꾸어 전사한다.
끊어진 단어	• 끊어진 단어는 발화된 그대로 전사하고 하이픈(-)으로 표시한다. • 불완전하게 발화된 음절이 둘 이상인 경우에도 음절마다 하이픈(-)으로 표시하여 전사한다.

항목	전사 기준
담화 표지 (추임새)	• '이', '그', '저', '어', '아', '에', '음', 등 기존 품사의 의미나 기능을 가지지 않는 것은 추임새로 보고 물결 표시(~)를 사용하여 표시한다.
잘 들리지 않는 부분	• 잘 들리지 않거나 전혀 들리지 않는 부분은 (0)과 같이 전사한다. • 잘 들리지 않지만 추정이 가능한 경우에는 (0) 안에 추정하여 전사한다. ex) (더 힘들어) • 들리지 않는 음절은 그 음절의 수만큼 x를 붙여 다음과 같이 전사한다. ex) (xx해야)
준음성 및 기타 소리	• 웃음, 목청 가다듬는 소리, 박수, 노래 등은 다음과 같이 태그하여 전사한다. ex) @웃음, @목청, @박수, @노래 • 감탄, 놀람 등은 '오', '앗', '어머' 등으로 들리는 대로 전사하며 태그하지 않는다.
익명성 보장을 위한 전사	• 대화자들의 신분 보장을 위해 이름, 주민등록번호, 카드번호, 전화번호 등 개인정보와 관련된 사항은 노출되지 않도록 비식별화한다.
문장 종결	• 한 문장이 끝나면 반드시 문장부호(마침표, 물음표, 느낌표)를 표기한다.

4.3 데이터 어노테이션 포맷과 형식 정의 및 입력

- 어노테이션 포맷 및 저장 형식
  - 음성 데이터는 고정된 필드나 스키마가 존재하지 않는 비정형 데이터이기 때문에, 학습용 데이터로서 가치를 부여하는 어노테이션 정보를 저장할 수 있는 별도의 데이터 구조와 파일 포맷을 정의한다.
  - 어노테이션 파일 포맷은 특정 소프트웨어에 종속되지 않고 쉽게 열고 편집할 수 있는 포맷으로 선택하며, 구조화된 어노테이션 정보를 저장하기 적합한 포맷을 선택한다.

    ※ json, xml 등

- 어노테이션 정보 저장 구조
  - 어노테이션 정보에 포함되어야 할 사항을 데이터 유형별(텍스트, 이미지, 동영상, 음성 등) 라벨링 참조 기준과 구축 목적에 따라 필요한 항목을 종합적으로 고려하여 정의한다.

- 어노테이션 정보(라벨링데이터)가 어떤 원천데이터와 매칭되는 지 확인할 수 있도록 어노테이션 구조 및 내용을 정의한다.

※ 학습용 데이터는 원천데이터 + 라벨링데이터로 구성됨을 고려

표 4-3. 음성 어노테이션 정보 구조 정의 및 구축 사례

No	항목	설명	타입	어노테이션 파일 구축 형태
	dataSet	데이터셋		
1	version	데이터셋 버전	String	
2	mediaUrl	녹취된 음원의 URL	String	
3	date	녹취된 날짜	String	
4	typeInfo	음원 데이터 상세 정보		
4-1	category	음원 카테고리 정보 : 강의, 회의, 고객응대 등	String	
4-2	subcategory	음원 서브카테고리	String	
4-3	place	음원 녹취 장소	String	
4-4	speakers	화자 목록	List	
4-3-1	type	화자 유형 : 강사, 상담사, 고객, 기타	String	
4-3-2	gender	인입 유형 : 유선, 모바일, 인터넷 등	String	
4-5	inputType	화자 성별 : 남성, 여성	String	
5	dialogs	전사 데이터 목록 : 화자가 변경될 때 생성	List	
5-1	speaker	화자 아이디 : speakers에 등록된 순번	String	
5-2	text	전사된 텍스트	String	
5-3	startTime	전사된 텍스트의 음원 재생 시작 위치	String	
5-4	endTime	전사된 텍스트의 음원 재생 끝 위치	String	
5-5	tags	전사된 텍스트 문장과 관련된 태그 리스트	String	

```
{
  "dataSet":
  {
    "version": "1.0",
   "mediaUrl": "http://..../23skskjdsfsks.wav",
    "date": "2020/05/20",
    "typeInfo":
    {
     "category": "conference"
     "speakers":
     [
      {
       "type": "representative", "gender":
 "female"
      },
      {
       "type": "customer", "gender":
 "female"
      }
     ],
     "inputType": "mobile"
    "dialogues":
    [
     {
      "speaker": 0,
      "text": "안녕하세요, 고객님. 무엇을
 도와드릴까요?",
      "startTime": "10.212",
      "endTime": "12.432",
      "tags": [ "Question", "Intro" ]
     },
     {
      "speaker": 1,
      "text": "환불할려고요",
      "startTime": "12.569",
      "endTime": "13.698",
      "tags": [ "Answer", "Refund" ]
     }
    ]
  }
}
```

4.4 데이터 라벨링 완료 후 관리 방법

- 데이터 관리 기본사항
  - 목적에 맞는 데이터 어노테이션 기준을 수립하고 데이터 사용 목적에 맞게 관리
  - 데이터의 사용 목적에 맞는 일관된 자료인지 확인한다.
  - 데이터들의 편향성을 확인 후 필요에 따라 데이터 추가한다.
  - 보존 일정 및 규정 준수 요구 사항에 따라 데이터 보관, 관리한다.

- 데이터 저장 관리
  - 원천데이터에 추가된 라벨링 정보를 저장하고 관리하는 기준을 수립한다. 파일을 체계적으로 분류하기 위해 데이터 종류 및 분류에 따른 라벨링데이터 파일 명명법과 파일 저장구조를 정의한다. 정의된 내용에 맞게 파일을 저장하도록 작업자에게 안내한다.
  - 작업자들이 원천데이터 및 라벨링 정보 저장 구조에 맞게 저장할 수 있도록 저장 절차를 정의하고, 작업자를 대상으로 배포한다.

- 데이터 백업 관리
  - 원천데이터 및 라벨링데이터의 훼손 및 멸실을 방지하기 위해 안전한 보관방법 및 백업방안(백업 시스템 및 프로세스 구축, 관리 절차 등)을 마련한다.

- 데이터 관리 조직 운영 방안
  - 데이터셋 제작 책임자는 품질관리 책임자로서 획득되는 데이터의 품질을 주기적으로 검사 및 관리한다.
  - 주기적인 실무협의체와의 미팅을 통해 데이터 품질에 대한 피드백을 공유하고 논의한다.
  - 데이터 품질 제고를 위해 데이터 라벨링 방안에 대하여 전문 컨설턴트 등 외부 기관의 조언을 받을 수 있다.

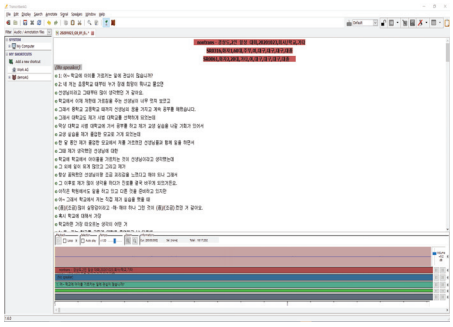
4.5 데이터 라벨링 방식에 적합한 도구 선정

- 라벨링 도구 선정
  - 데이터 구축 목적 달성을 위해 원천데이터 형태, 구축 목적에 부합하는 라벨링 도구를 선정한다.
  - 기존의 도구를 가지고 인공지능 학습용 데이터 구축 목표 달성이 어려울 경우, 기존 라벨링 도구의 기능을 추가하거나 완전히 새로 개발하는 방법을 고려한다.

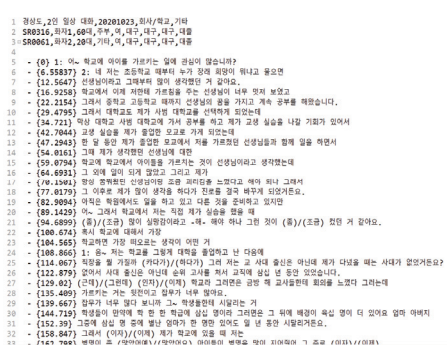
표 4-4. 국내외 주요 음성 데이터 라벨링 도구

No	도구명	요 금	설 명
1	SELECTSTAR	유료	• 크라우드소싱 기반의 데이터 수집 및 라벨링 도구 • 음성 클래스 분류 및 텍스트 전사 기능 제공 • 사이트 주소 : https://selectstar.ai/
2	크라우드웍스	유료	• 크라우드소싱 기반의 데이터 라벨링 도구 (PC, 모바일 앱 지원) • 음성→텍스트(STT) 전사 기능 제공 • 사이트 주소 : https://www.crowdworks.kr/main.do
3	AIWORKS	유료	• 크라우드소싱 기반의 데이터 라벨링 도구 • 음성→텍스트(STT) 전사 기능 제공 • 사이트 주소 : https://aiworks.co.kr/
4	소리자바	유료	• AI기반 음성 데이터 라벨링 도구 • 음성→텍스트(STT) 전사, 다국어 음성 인식 기능 제공 • 사이트 주소 : https://www.sorizava.co.kr/

- 라벨링 도구 활용 매뉴얼 작성
  - 작업자가 활용할 도구의 사용법에 대한 매뉴얼을 작성한다. 매뉴얼은 Step-by-Step 형태로 쉽게 따라할 수 있도록 작성하며, 이미지, 동영상 등을 활용하여 이해를 도울 수 있다.



[전사 저작도구 활용 예시]



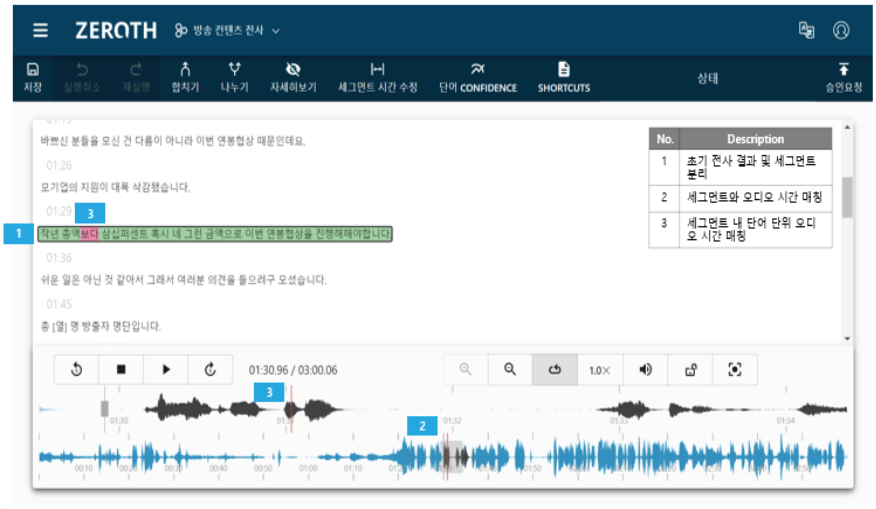
[전사 완료 후 txt 파일]

그림 4-1. 라벨링 도구 활용 매뉴얼 작성 예시

- 라벨링 도구 선정 및 개발 시 고려사항
  - 표준적인 어노테이션 및 라벨링 작업 가능 여부, 표준 파일 포맷 지원 여부 등을 고려한다.
  - 한국어 및 다국어를 입력할 때 깨짐현상이 없는 인코딩을 지원하는지 고려한다.
  - 특히 크라우드소싱 방식을 적용하는 경우, 다양한 작업환경(컴퓨터 성능, OS, 네트워크 등)에서의 실행 가능 여부를 확인한다.
  - 라벨링 결과를 효과적으로 관리하고 작업배분을 할 수 있는 관리 기능이 충실한 지 확인한다.
  - 작업자 또는 검사자가 어노테이션 결과를 눈으로 바로 확인할 수 있는 시각화 기능이 있는 지 고려한다.

【참고】 어노테이션 시각화 예시

- 라벨링된 값, 어노테이션 구간 등을 눈으로 확인할 수 있는 기능을 함께 제공한다.



5 처리 데이터 검사

5.1 검사 절차 정의

- 개요
  - 인공지능 학습용 데이터 구축을 위한 품질 검사 절차·방법은 데이터 유형, 도메인, 목표 서비스에 따라 달라질 수 있으며 사업 기간 및 예산 등 현실적인 여건을 고려하여 수립한다.
  - 데이터 검사 절차 및 규격은 데이터 구축 목적 정의 단계에서 수립한 데이터 활용 분야·목적을 달성할 수 있도록 정의한다.

- 검사 절차 정의
  - 다량의 데이터를 한정된 시간 내에 최적의 품질로 검사할 수 있도록 하는 검사 단계 및 절차를 수립한다.
  - 검사 프로세스는 학습용 데이터 구축 공정(획득, 정제, 라벨링) 각 단계별로 검사가 수행되는 형태를 기본으로 하며, 데이터 구축 공정 및 데이터 특성을 반영하여 적합한 절차를 수립할 수 있다.

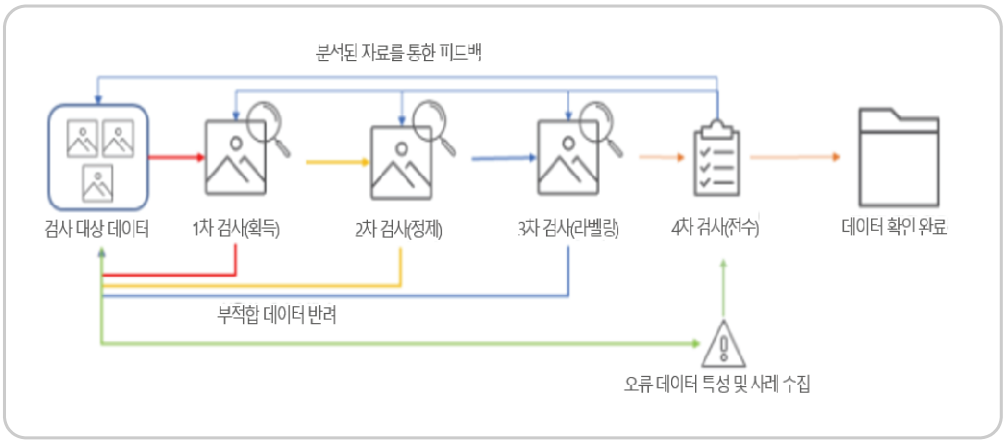
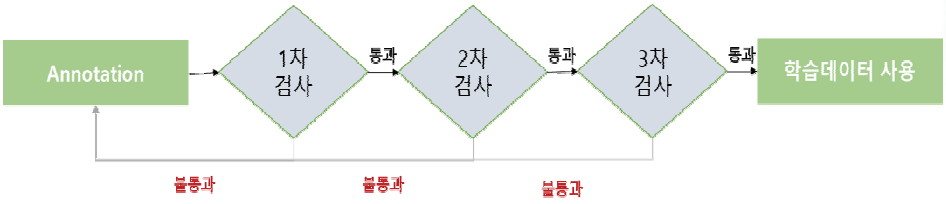


그림 5-1. 데이터 검사 절차 정의

- 검사 규모
  - 데이터 구축 설계 단계에서 구축될 데이터에 대한 품질 수준을 미리 정의하고, 품질 검사를 위한 검사 규모 및 방법*을 설정한다.
    * 전수 검사, 샘플링(00%), 다단계 샘플링 등
  - 전수 검사가 아닌 샘플링 방법으로 데이터를 검사할 경우, 검사 대상 데이터가 편향되지 않으면서 무작위로 추출될 수 있도록 한다.
    ※ 각 클래스별 동일한 비율로 추출되도록 함(층화추출)
    ※ 데이터가 구축된 순서 등이 특정 타이밍에 집중되지 않도록 함(파일명 정렬 후 무작위 추출 방법 등 적용)

【참고】 품질 검사 절차 정의 예시



- 데이터 검사는 양질의 데이터를 얻는 데 필요한 작업으로 정성적/정량적 평가를 통해 데이터의 유효함을 판별함
- 웹 저작도구를 사용하여 전사한 어노테이션 결과에 대해 1차, 2차, 3차 검사자를 걸쳐 음성품질 / 어노테이션 정확도 / 대화 주제 및 저작권확인 / 목표데이터 수집량 달성 여부를 확인함
- 1, 2차 검사자는 음성품질 및 어노테이션 정확도, 대화 주제 및 저작권 확인 위주로 검사를 진행하며, 교차 검증을 통해 데이터의 오류를 최소화함
- 1, 2차 검사는 전수검사로 모든 데이터에 대하여 진행함
- 3차 검사자는 목표데이터의 음성품질 및 도메인별 수집량 달성 여부 위주로 검사를 진행하며, 데이터의 다양성을 보장할 수 있도록 함
- 3차 검사는 검사도구를 통해 자동검사를 진행하며, 통과한 데이터에 대해서도 샘플링 검사를 진행하여 데이터 완결성을 보장함
- 각 차수의 검사자는 피드백을 통해 데이터가 한쪽에 치우치지 않도록 하며, 작업자가 숙달할 수 있도록 지원함

5.2 검사 방식

- 검사 항목 정의
  - 구축 공정(획득, 정제, 라벨링)별로 공통적으로 적용할 수 있는 검사 요구사항을 고려하여 검사 항목을 정의할 수 있다.
  - 검사 항목은 데이터 및 절차 측면에서 적합성·정확성·유효성, 준비성·완전성·유용성 지표를 측정할 수 있도록 한다.

※ 자세한 절차 및 내용은 'Ⅳ. 품질검사 방법' 내용을 참고하며 아래 표 내용 참고 가능

표 5-1. 구축 공정별 주요 검사 항목

검사 절차	검사 항목	요구사항
1차 검사 (획득)	법·제도 준수	원시데이터 획득 시 관련 법·제도적 규정 등을 반드시 준수해야 함
	사실적인 획득 환경 구성	원시데이터를 인위적인 환경과 조건 하에 획득해야 하는 경우 사실적인 획득 환경을 구성하여야함
	데이터 동기화	다중 데이터 소스 간 정교한 동기화를 위한 절차를 마련하여야 함
	편향성 방지	데이터 편향을 방지하기 위한 절차를 마련하여야함
2차 검사 (정제)	정제 기준의 명확성	데이터 사용 목적에 적합한 정제 기준 수립 여부
	중복성 방지	데이터 정제 후 정보 비교 후 중복도 여부
	정제 작업 매뉴얼	정제 작업을 위한 매뉴얼 작성 및 관리 여부
	정제 도구	정제 작업에 사용될 SW 도구를 확보 및 사용 방법을 숙지
	정제 작업 방식	데이터 특성 및 활용 목적에 맞는 적절한 정제 방식 선정 여부 및 선정 기준 타당성 여부
3차 검사 (라벨링)	라벨링 가이드	목적에 맞게 작성된 라벨링 가이드에 대한 타당성 여부를 검사 후 라벨링 작업자들에게 내용 가이드 전달
	어노테이션 항목	목적에 맞는 어노테이션 구성인지 여부를 검사 후 확인된 내용을 포함하도록 작업자들에게 전달
	라벨링 검사 도구	자동화 도구를 통해 검사 후 검사자가 육안으로 부적합 데이터 여부 2차 확인과 촬영된 영상(동적/정적) 이미지의 누락, 번짐 및 조건 오류를 전수 검사
4차 검사 (전수)	부적합 판정 데이터 분포 확인	데이터의 오류율, 특성 분포 확인을 통한 데이터 수집, 정제, 라벨링, 부문 최적화
	외부 검사자	외부 검사자(TTA 등), 도메인 전문가, 데이터 요청자

- 점검 기준 및 점검표 작성
  - 데이터를 일관된 기준으로 검사하기 위해, 데이터 정확성 및 구축 취지에 부합할 수 있는 참값(ground truth)을 정의하고 이 참값을 기준으로 검사 항목 및 채점 기준(통과 기준)을 정의한다.
  - 검사항목 및 채점 기준(통과 기준)을 검사자가 쉽게 확인하고 적용할 수 있도록 체크리스트 등의 형태로 작성하여 배포한다.

표 5-2. 검사 항목 점검표 정의 예시(한국어 상담 음성 데이터)

대항목	데이터 항목	어노테이션	판정 기준
음성품질	음성 파일 형식	N/A	검사도구를 통한 자동 검사 및 샘플링 검사
	명료성	N/A	작업자와 검사자에 의한 음성 명료함 판단
라벨링 정확도	텍스트	전사규칙에 따라 전사된 발화 내용	3인 이상의 검사자 판단하에 과반수 통과
	발화자	화자 고유번호, 화자 성별, 나이	3인 이상의 검사자 판단하에 과반수 통과, 검사도구를 통한 자동 검사, 샘플링 검사
	싱크	음성과 텍스트의 시작과 끝	검사도구를 통한 자동 검사 및 샘플링 검사
대화 주제 및 저작권 확인	대화주제	강의 전체 도메인에 대한 23종 태깅	3인 이상의 검사자 판단하에 과반수 통과
	저작권 및 개인정보	N/A	3인 이상의 검사자에 의한 저작권 및 개인정보 침해여부 판단
목표데이터 수집량 달성	도메인	강의 도메인별 음성 길이	검사도구를 통한 자동 검사 및 샘플링 검사
	화자	화자별 음성 길이	검사도구를 통한 자동 검사 및 샘플링 검사

- 검사 도구
  - 검사자가 데이터를 빠르게 확인하고 검사할 수 있는 도구를 마련한다. 검사 결과를 작업자에게 즉각적으로 피드백하려는 경우에는 데이터 라벨링 도구와 연계한 시스템으로 검사 도구를 구축할 수 있다.
  - 검사 내용이 단순·반복적인 경우 검사 항목 일부의 자동화 처리를 고려한다.

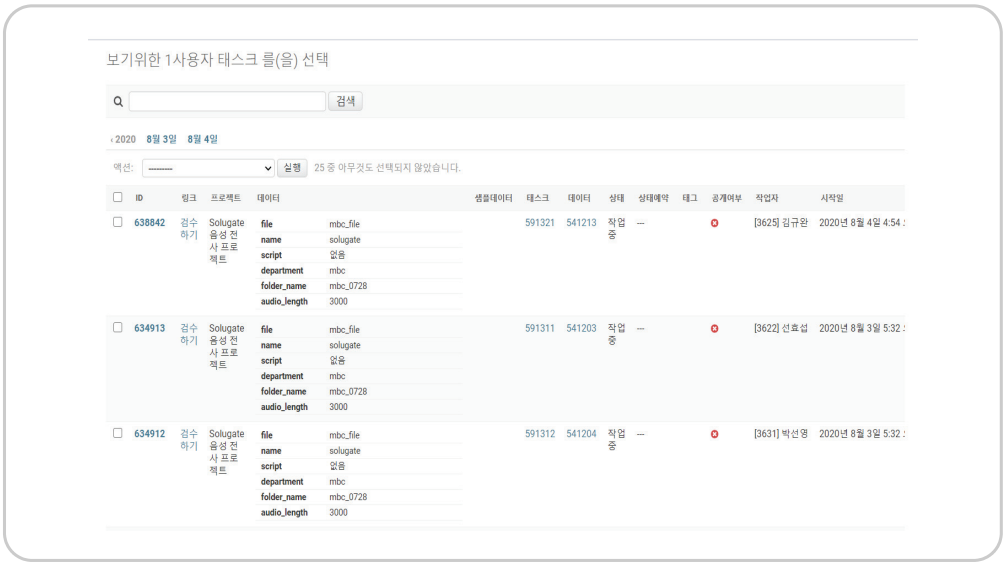


그림 5-2. 데이터 검사 플랫폼 구축 예시

5.3 검사 결과

- 결과 피드백
  - 각 검사 단계별로 검사자가 검사한 결과를 작업자 및 관리자, 고객에게 피드백하는 체계를 마련한다.
  - 검사결과 집계, 검사내용 확인, 의견사항 전달, 작업자·검사자 재배정, 이슈사항 공유, 피드백 시기 및 주기 등에 대한 절차 및 방법을 수립하여, 검사 현황 및 결과가 빠르게 공유되고 검사를 통과하지 못한 데이터에 대한 재 라벨링이 원활하게 될 수 있도록 한다.

【참고】 검사결과 피드백 프로세스 정의 예시

1. 검사 후의 피드백 프로세스는 다음과 같음
   - 검사 후 불합격 판정 → 검사 의견 제시 (라벨링 결과 문제점을 검사자가 기록) → 라벨링 작업자에게 재전달 → 2차 라벨링 → 2차 검사 → 불합격의 경우 재 라벨링 요청 (검사 의견 기록) → 재 라벨링 (3차 라벨링) → 3차 검사

2. 3차 검사 결과가 불합격인 경우 1) 문서가 매우 어려워 작업자가 라벨링할 수 없음, 2) 작업자의 역량이 해당 원천데이터를 라벨링할 능력이 없음, 3) 원천데이터의 형태가 라벨링이 불가능함 등의 사유로 라벨링 불능 데이터로 판정하고, 3차 검사 불합격 데이터를 별도 저장하여 향후 다른 작업자를 통해 재 라벨할 예정

3. 라벨링시 정확도를 높이기 위해 검사 합격된 문서에만 보상을 명문화하였음

4. 크라우드소싱의 특성상 작업자에 대한 교육이 어려운 관계로, 검사시 검사의견을 최대한 명확하고 구체적으로 전달

- 검사결과 처리
  - 미검사/검사 여부, 검사 통과/미통과 여부를 구분할 수 있도록 파일 관리 체계, 메타데이터 구조 등을 설계한다.
  - 목표 데이터, 원천데이터 대비 검사를 통과한 학습 데이터 현황을 모니터링하는 체계를 구성하여, 일정 내 학습 데이터를 확보할 수 있도록 한다.
  - 검사결과 및 피드백 사항을 데이터 제작 도구 및 시스템 상에서 어떻게 구현할 것인지 설계한다.

    ※ 데이터별 검사이력 보관방법, 데이터 전송, 작업자 대상 피드백 전달 방법 등

참 고 자 료

1. ㈜비플라이소프트·㈜위고·㈜테스트웍스·고려대학교·주식회사 에이아이닷엠, 문서 요약 텍스트 구축 가이드라인
2. 셀렉트스타 주식회사, 음성 수집 기준 정의서
3. ㈜소리자바, 음성(소리) DATA 수집 및 전사 가이드
4. ㈜솔루게이트, 한국인 대화음성 구축 가이드라인
5. ㈜솔트룩스, 인공지능 데이터 구축·활용 가이드라인-한국어 방언 AI 데이터
6. ㈜슈퍼브에이아이, 데이터 라벨링 사업운영 A to Z
7. ㈜아이스크림에듀, 인공지능 데이터 구축·활용 가이드라인-상담 음성 데이터
8. ㈜포티투마루, 관광분야 지식베이스 구축 절차서
9. 한국정보통신기술협회(TTA), AI 학습용 데이터 구축사업 공통기준
10. 한국정보통신기술협회(TTA), 2020년 인공지능 학습용 데이터 구축 사업(1차) 중간산출물 20종 검토
11. 한국지능정보사회진흥원(NIA), 2020년 인공지능 학습용 데이터 구축사업(2차) 가이드라인 참조

Ⅲ OCR (광학문자인식) 이미지 데이터

목 차

표 목 차

그림 목 차

제1장 | 개 요

1 작성 배경

- 인공지능 학습용 데이터 구축 사업 확대에 따라 다양한 역량의 수행 및 참여기관 참여로 사업 진척도 및 데이터 품질의 편차가 발생
- 인공지능 학습용 데이터 품질 향상 및 성공적인 사업 추진을 위해 수행 및 참여기관을 대상으로 데이터 구축 기준, 절차 등 노하우 공유 필요

2 작성 목적

- 인공지능 학습용 데이터 구축에 보편적으로 적용되는 데이터 유형별로 데이터 구축에 필요한 절차 및 구성요소를 제시하여 데이터 구축 과정에서의 시행착오를 줄이고 체계적인 계획 수립을 지원한다.
- 국내 인공지능 학습용 데이터 구축 시 활용된 다양한 가이드라인 사례를 제시하여, 향후 수행 및 참여기관에서 수립해야할 데이터 구축 가이드라인 작성에 참조할 수 있도록 한다.
- 향후 인공지능 학습용 데이터 구축 사업 추진 시 본 구축 안내서를 배포하여 다양한 수행 및 참여기관의 역량 향상 및 성공적인 사업 수행을 지원한다.

- 궁극적으로 양질의 인공지능 학습용 데이터 구축 및 개방을 통해 국내 인공지능 산업 활성화 및 발전에 기여한다.

3 작성 범위

- 인공지능 학습용 데이터 구축에 필수 공정단계인 데이터 획득·정제·라벨링·검사 단계를 본 구축 안내서의 작성 범위로 한다.
- (정적) 영상(이미지) 타입의 원시데이터를 텍스트 타입으로 라벨링하는 데이터 구축을 본 구축 안내서의 작성 범위로 하며, 특히 OCR 등 문자 이미지를 텍스트 타입으로 라벨링하는 데이터 구축을 적용 대상으로 한다.
- 기존 TTA 인공지능 학습용 데이터 구축 가이드라인, '20년 1차·추경(2차) 인공지능 학습용 데이터 구축사업 수행 및 참여기관의 데이터 구축 가이드라인, 국내 주요 인공지능 전문기업이 인공지능 학습용 데이터 구축을 위해 제작한 인공지능 데이터 구축 지침·가이드 등의 자료들을 검토하여 인공지능 학습용 데이터 구축 시 공통적으로 고려해야 할 사항들을 도출하여 구축 안내서에 반영한다.

4 용어 정의

- 데이터 획득 (Data Acquisition)
  - 인공지능의 기계학습에 필요한 데이터를 현실 세계에서 직접 수집 또는 생성하거나, 이미 보유하고 있는 조직이나 시스템 등으로부터 법률적 제약이 없도록 '원시데이터'를 확보하는 활동

- 데이터 정제 (Data Refinement)
  - 획득한 원시데이터를 기계학습에 필요한 형식으로 맞추거나 불필요한 중복을 제거하며, 개인정보를 비식별화하여 처리하는 등 일련의 전처리 과정을 통해 '원천데이터'를 확보하는 활동

- 데이터 라벨링 (Data Labeling)
  - 인공지능이 기계학습에 활용할 수 있도록 기능이나 목적에 부합하는 정보를 원천데이터에 부착하는 활동

- 라벨링데이터 (Labeled Data)
  - 원천데이터에 부여한 '참값', 파일형식이나 해상도 등의 속성, 그리고 설명이나 주석 등이 포함된 '어노테이션'의 집합

- 원시데이터 (Raw Data)
  - 기계학습을 목적으로 획득 단계에서 수집 또는 생성한 음성, 이미지, 영상, 텍스트 등의 데이터

- 원천데이터 (Source Data, Unlabeled Data)
  - 원시데이터를 라벨링 공정에 투입하기 위해 필요한 전처리 등 정제 작업을 수행한 데이터로 라벨링데이터가 부여되지 않은 상태의 데이터

- 인공지능 학습용 데이터 구축
  - 임무정의, 데이터 획득, 데이터 정제, 데이터 라벨링 등 인공지능 학습용 데이터를 구축하는 일련의 활동

- 참값 (Ground Truth)
  - 인공지능의 기계학습 목적에 따라 원천데이터에 라벨링된 정확한 값이나 사실의 의미적 표현

- 어노테이션 (Annotation)
  - 데이터 라벨링 시 원천데이터에 주석을 표시하는 작업을 의미하며, 추가 부착되는 설명정보 데이터는 기능 목적에 따라 다양한 형태로 표현될 수 있으며 이러한 설명정보 표현방식을 지칭

    ※ 용어사용 예 : 사물 바운딩박스 어노테이션, 클래스 라벨링 어노테이션 등

- 광학문자인식 (OCR, Optical Character Recognition)
  - 사람이 쓰거나 기계로 인쇄한 문자의 영상을 기계가 읽을 수 있는 문자로 변환하는 것

    ※ 자세한 용어 정의는 '인공지능 학습용 데이터 품질관리 가이드라인 Ⅴ.부록-1 용어정의'를 참조

제2장 | 구축 가이드라인 작성 방법

1 데이터 구축 목적 정의

- 데이터 구축 목적 정의
  - 인공지능 학습용 데이터 구축 목적은 단순한 데이터 수집, 모음이 아닌 구축된 데이터를 인공지능 학습 모델에 적용하여 의미있는 수준의 정확도를 확보하고 서비스 등에 유용하게 활용되는 것을 목표로 정의한다.
  - 목적 정의에는 데이터의 구축 배경 또는 필요성, 구축되는 데이터에 대한 명확한 정의, 구축 방향 및 활용(예상) 분야 등을 포함한다.
  - 구축될 학습용 데이터가 실제로 어떤 산업, 서비스, 연구분야에서 활용될 수 있는 지 정의하여 데이터 구축 방향에 대한 타당성을 재확인한다.
  - 데이터의 저장, 기록이나 해석에서 오류의 가능성이 없도록 명확한 단어, 어휘체계를 사용하여 정의한다.

【참고】 크데이터 구축 목적 정의 예시
(고서한자인식 OCR 데이터 구축)

- 데이터 구축 목적
  - 고품질의 인공지능 학습용 데이터 구축과 모델 개발 학습을 통해 인공지능 기반 한자 글자체 인식(OCR) 분야에서의 글로벌 주도권 선점
- 데이터 구축 필요성
  - 동아시아 한자문화권(한국, 중국, 일본, 베트남)에서 한자(고문헌) 글자체 인식(OCR) 기술 상용화 사례 없음
  - 민·관을 아울러 국가적으로 소장중인 방대한 분량의 세계적인 한자(고문헌) 기록유산을 효과적으로 활용하기 위한 필수요소인 디지털 텍스트를 수요자들에게 신속하게 제공할 수 있는 안정적인 기술기반 확보
  - 인공지능 기반 한자 글자체 인식(OCR) 학습데이터 구축과 인식 모델을 개발함으로써 비약적으로 발전하고 있는 최신 ICT 기술이 접목된 한자(고문헌) 텍스트 DB구축 솔루션의 개발 토대를 마련하여, 20년 전 기술에 의존하고 있는 관련 분야의 변화와 혁신을 주도
- 데이터 활용 분야
  - 최신 인공지능 기술을 통해 방대한 한자(고문헌) 자료의 획기적인 이용전기를 마련함으로써 주제와 내용의 다양성으로 무궁무진한 활용성을 지닌 콘텐츠의 보고 한자(고문헌) 기록유산의 현대적 활용과 연구에 기여

2 데이터 구축 시 고려사항

- 데이터 종류 및 규모
  - 획득해야할 데이터의 규모를 설정한다. 이때 대상으로 하는 산업분야 및 서비스에서 요구되는 수준과, 사업기간과 획득에 드는 시간과 비용을 종합적으로 고려하여 구축 규모를 선정한다.

- 어노테이션 타입
  - 데이터 활용 분야를 고려하여 구축되는 데이터의 어노테이션 타입을 정의한다.

    ※ 대부분의 OCR 이미지 데이터는 바운딩 박스, 폴리곤 타입을 활용

표 2-1. 이미지 데이터 어노테이션 타입 및 용도

어노테이션 타입	주요 활용 용도
• 클래스 라벨(단일, 다중)	• 이미지 분류(Image Classification)
• 바운딩 박스(사각형) • 폴리곤(다각형)	• 객체 인식(Object Recognition)
• 픽셀(점)*	• 영역 구분(Segmentation)
• 기타	• 그 밖의 용도

※ 영역 구분은 점 단위로 어노테이션되나, 라벨링 작업의 편의를 위해 라벨링 영역 지정은 바운딩 박스, 폴리곤, 브러시(자유형태) 등의 도구를 활용할 수 있음

- 데이터 구축 프로세스 정의
  - 데이터 구축 목적 정의, 데이터 획득, 데이터 정제, 데이터 라벨링, 데이터 검사에 이르는 일련의 데이터 구축 프로세스를 사전에 정의하고, 각 프로세스에 따르는 이슈 및 검토사항 등을 도출한다.
  - 데이터 구축 프로세스는 구축 단계별 주요 작업에 대해 서술하나, 순서도·표 등을 활용해 구조화하여 구축 관계자 및 작업자들이 쉽게 이해할 수 있도록 한다.

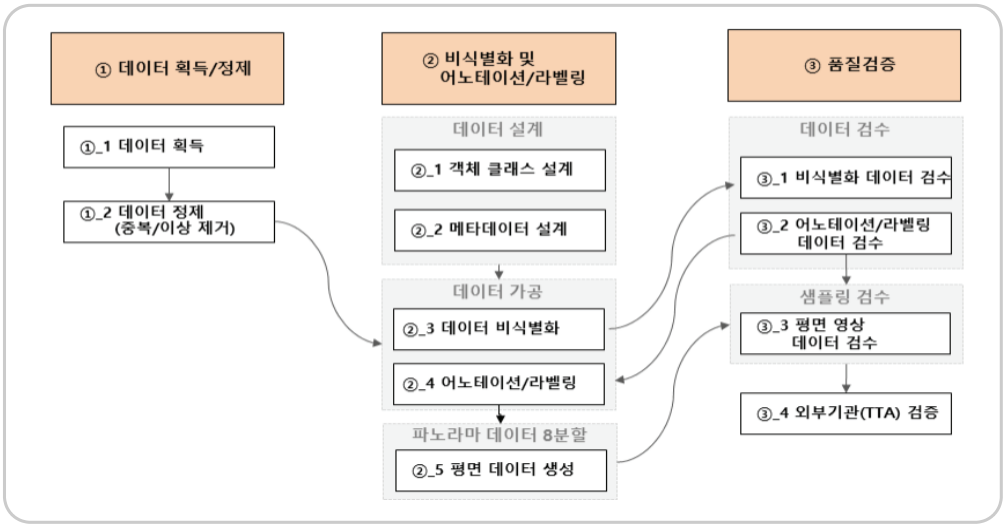


그림 2-1. 데이터 구축 프로세스(순서도 형식) 정의 예시(도로환경 파노라마 이미지)

표 2-2. 데이터 구축 프로세스(표 형식) 정의 예시(고서 한자 인식)

구축단계	세부공정	내용
획득	원시데이터 분석	• 원시데이터 : 고서(고도서) 원문 이미지 • 구축 대상 서체 여부 확인 : 해서체(95%), 행서체(5%) • 이미지 품질 확인 : 해상도, 기울기, 훼손 여부 등
	원시데이터 수집	• 개인정보 비식별화 불필요 : 일체의 개인정보 불포함 • 저작권, 특허권, 초상권 부존재 • 대상 자료 전량은 참여기관인 한국국학진흥원 소장 데이터
정제	세그먼테이션	• 원문이미지 상의 한자를 낱자별로 추출 • 데이터 저작도구 사용, 자동/수동 병행
라벨링	군집 (클러스터링)	• 한자 세그먼트들을 동일 한자끼리 하나의 그룹으로 클러스터링 • 데이터 저작도구 사용, 자동/수동 병행 • 크라우드 소싱
	입력	• 한자 전문인력들이 군집별 입력 • 데이터 저작도구 사용, 전량 수동 • 크라우드 소싱
검사	검사/교정	• 한자 전문인력들이 입력 완료된 한자를 원문이미지와 대조 검사/교정 • 데이터 저작도구 사용, 전량 수동 • 크라우드 소싱
	품질검사	• 한자 검사 전문인력들이 샘플링(5%) 검사 • 데이터 저작도구 사용, 전량 수동

- 데이터 품질 수준
  - 데이터 제작을 의뢰하는 고객이 있는 경우 고객이 요구하는 데이터 품질수준을 기본으로 하며, 세부적인 사항은 협의하여 결정한다.
  - 특정 고객 없이 범용적으로 활용할 수 있는 데이터를 제작하는 경우에도 해당 산업 및 서비스 분야에서 요구되는 품질 수준을 갖추기 위해, 해당 분야 산업 관계자 및 전문가 등의 검토를 통해 적절한 품질 수준을 설정한다.
  - 데이터 활용 목적에 맞는 데이터를 구축하는지 구축 데이터의 시간대, 주제, 비효율성, 인과관계 등을 검토한다.
  - 구축되는 데이터는 모집단 및 프로세스에 대한 충분한 정보를 얻을 수 있는지, 구축되는 도메인의 모집단 또는 프로세스를 대표할 수 있는지 검토한다.
  - 품질 수준 및 측정 방법은 논문, 연구, 유사 사업(사례) 등을 통해 객관적이고 명확하게 제시되어야 하며, 기존에 구축된 인공지능 학습용 데이터와 최소한 동일 품질 수준 이상을 갖추는 것을 목표로 해야 한다.

- 데이터 제작 도구
  - 구축 대상 데이터가 수행 및 참여기관에서 보유한 구축 도구(소프트웨어)로 목표로 하는 수준으로 제작할 수 있는지 검토한다.
  - 제작 도구는 자체 개발한 솔루션을 활용하거나, 타사의 상용 솔루션, 또는 오픈소스 도구를 활용할 수 있으며 이 중 적합한 방법을 선택한다.
  - 구축할 데이터의 특성에 맞게 구축 도구에 관한 환경설정을 진행한다.

- 가이드라인 수정 및 이력관리
  - 인공지능 학습용 데이터 구축 진행 중 발생되는 예외상황(edge case), 애매모호한 상황 등 데이터 구축 설계단계에서 제작한 가이드라인에서 변경이 필요한 사항이 발생 시, 가이드라인을 업데이트 하고 작업자에게 신속히 배포할 수 있는 방안을 마련한다.

- 구축 과정에서 큰 영향을 미치는 작업방법, 라벨링 세부조건 등에 관한 변경사항 발생 시 고객사 또는 각 산업 전문가 및 관계자 등의 검토·협의를 통해 적합한 방법을 도출할 수 있도록 변경 검토 절차를 마련한다.

  ※ 변경사항에 대한 검토 및 배포가 제대로 이루어지지 않을 때, 최악의 경우 라벨링을 다시해야 할 수 있음

- 작업 및 검사 인력 운영 방식
  - 대량의 데이터를 구축해야 하는 인공지능 학습용 데이터 특성 상 필요한 작업자 수와 수행 및 참여기관의 시설·노동 환경을 고려하여 내부조직, 아웃소싱, 크라우드소싱 또는 혼합 방식 등 적합한 작업자 운영 방식을 선정한다.
  - 작업자의 성과를 측정하는 기준을 마련하고, 특히 크라우드소싱 작업자 등 정확한 노동시간을 측정하기 어려운 작업자에 대한 임금 또는 보상 지급기준을 정의한다.

표 2-3. 작업자 운영 방식 특성 비교

구분	내부조직	아웃소싱	크라우드소싱 (crowd-sourcing)
특징	• 품질에 대한 상시 교육 및 피드백 가능 • 작업환경을 위한 운영비, 작업공간 및 인프라 필요(전기·통신시설, 컴퓨터 등)	• 높은 업무 전문성 및 경험 보유 • 요구사항 정의 및 기준 합의에 많은 시간 소요	• 높은 업무 접근성 (장소제약없음) • 품질 교육 및 피드백에 한계가 있음 • 크라우드소싱 대가 산정에 대한 명확한 기준 마련이 어려움
적합 용도	• 머신러닝 훈련에 대한 높은 수준의 이해가 필요한 작업 • 라벨링 결과에 대한 공정간(수집, 정제, 라벨링, 학습모델 등) 긴밀한 피드백을 요구하는 작업	• 데이터 구축에 전문적인 지식과 숙련도가 요구되는 작업	• 단기간에 대량의 데이터를 처리해야 하는 작업 • 작업 난이도가 비교적 낮고, 데이터 보안수준이 낮은 작업

【참고】 크라우드소싱 작업인력 운영 방식 수립 사례

- 기본 원칙
  - 입사 후 7일간 수습 신분으로 Half time (주20시간) 근무
  - 성실성 등 근무태도 평가 후 단기계약 가능
  - 계약 후 한달 근무 완료 시 아래의 세 가지 근무 타입 중 선택

[크라우드소싱 작업자 근무 타입별 운영방식 및 특성]

업무 장소	근무 유형	운영방식 및 고려사항
사무실 (인하우스)	하프타임	• 13시~18시 근무 • 단시간 집중도 높은 근무 • 전체 근무자 중 64.8% 비중
	풀 타임	• 09시~18시 근무 • 작업 이해도 및 작업 역량 단기간 내 확보 가능 • 매니저, 리더 포함 근무자 중 35.2% 비중
재택	자율근무	• 일정 퀄리티 이상의 작업 결과에 대해 건별로 보수 지급 • 작업 내용에 대한 피드백 지연 발생 • 인하우스 근로자 대비 퇴사율이 높아 계약 베이스로 운영하지 않을 경우 리소스 예측과 관리가 어려움

- 작업자 대상 매뉴얼 작성
  - 데이터 획득·정제·라벨링·검사 단계에 참여하는 작업자들이 인공지능 학습용 데이터셋 구축 취지에 부응하여 데이터 제작이 이루어질 수 있도록 작업자들이 직접 활용하는 매뉴얼을 제작한다.
  - 작업자 대상 매뉴얼에는 구축 목적·정의, 제작 절차, 제작 도구 활용방법과 작업 기준, 작업 결과 처리·저장 방법 등의 내용을 포함한다.
  - 작업자 관점에서 데이터 제작 과정에서 발생할 수 있는 다양한 유사사례 및 예시 등을 포함하여 매뉴얼을 제작한다.

3 데이터 획득 및 정제 방법

3.1 데이터 정의

- 원시데이터 분석
  - 인공지능 학습용 데이터 구축에 필요한 원시데이터 항목을 검토하고, 각 항목 별로 데이터 획득에 필요한 정보(데이터 획득정보, 획득방법, 획득 단계에서 필요한 요건 등)들을 검토하여 문서화한다.
  - 원시데이터 대상 및 획득 방법을 아래와 같이 육하원칙에 따라 정의할 수 있다.

표 3-1. 인공지능 학습용 데이터 구축 시 검토사항 및 예시

5W1H	항 목	예 시
What	• 측정대상 • 획득 시 포함되어야 할 변수들	• 일반인이 대상을 식별 할 수 있는 피사체 • 장비별, 객체별, 시간별, 종류별, 사람별, 지역별 검토 (필요시 도메인 전문가, 인공지능 전문가 협의 후 대상 객체를 명확히 함)
When	• 획득 기간 (From, To)	• 2주간(11.14 ~ 11.28), 아침 9:00, 점심 12:00 저녁 18:00 일 3회, 획득 시간 1시간
Where	• 획득장소 / 프로세스	• 충남 대전역 역사 내 대합실 외 동일 공간 3곳으로 이동과 고정을 병행하여 획득하고 동선에 겹치지 않는 장소를 선정(직접 지정함)하여 획득
Who	• 획득 담당자 / 획득하는 사람	• 00 주식회사 미디어센터 내 류XX • 그 외 크라우드 소싱 인력 20명
How	• 획득 방법, 측정주기, 샘플 크기, • 데이터 양식	• 직접 샘플링 모니터 후, 시간당 1회, 1일 3회, 회당 20개 별도 제작 체크시트 등 확인 후 개수 증감
Why	• 측정 목적 / 기대 결과	• 목적에 맞는 획득 데이터 이해와 프로세스 능력의 파악 / 추세분석

- 획득할 원시데이터 내역에 대한 정의 및 현황정보 등의 사항을 정리한다.

표 3-2. 원시데이터 현황정보 작성 기준 및 예시(관광분야 이미지)

항목	데이터량	획득기간	획득지역	획득유형	획득 방법
매장 전경	50만장	2020.06~2020.10.	서울시 주요 관광지	직접 촬영	• 내부직원 및 크라우드 소싱
표지판	100만장	2020.06~2020.10.			
매장 전경	2만장	2020.06~2020.10.	-	크롤링	• 한국관광공사 제공 POI 정보 - POI 키워드 속성이 정형화 되지 않은 축제/공연 행사, 여행코스는 제외
매장 내부	2만장	2020.06~2020.10.			
메뉴판	2만장	2020.06~2020.10.			
표지판	2만장	2020.06~2020.10.			

- 원시데이터 포맷
  - 원시데이터의 파일 형식은 특정 수집 장비 및 처리 도구에 종속되지 않으며, 보편적으로 통용되는 포맷을 활용한다.
    ※ jpg, png 등 (단 의료 등 전문분야의 경우 해당분야의 표준을 준수)

- 원시데이터 획득 규모
  - 원시데이터 획득 후 정제, 라벨링, 검사 과정에서 기준 미충족으로 버려지는 데이터 양을 고려하여 구축 목표치 이상의 데이터를 획득하도록 계획한다.
    ※ 구체적인 목표치 대비 획득량은 데이터 구축 공정 난이도 및 구축기간 등을 고려하여 설정

3.2 획득 데이터 특성 분석

- 원시데이터 획득 관련 이슈사항 도출
  - 획득할 원시데이터의 범위 및 방법을 명확히 하기 위해 데이터 규모·획득범위·수집처 등에 대한 세부 이슈사항을 도출하여 가이드라인에 기술한다.

【참고】 원시데이터 특성 분석 예시(관광분야 이미지)

- 직접 촬영 데이터
  - 매장 전경, 메뉴판, 매장 내부의 경우 TourAPI 데이터 분석 결과를 통해, 서울시 관광특구 소재 6개소를 중심으로 TourAPI에 포함된 음식점 588개 및 외국인들이 많이 찾아가는 음식점을 해외 사이트 랭킹 등을 참고한 자체 음식 랭킹으로 선정하여 원천데이터 후보군으로 선정
  - TourAPI 데이터를 분석하여, 서울시 관광특구 6개소(명동/남대문/북창, 이태원, 동대문 패션타운, 종로 청계, 잠실, 강남 지역)를 촬영 지역으로 선정
  - 메뉴판, 매장 내부의 경우 저작권 이슈가 있어 촬영 동의서 작성한 음식점만을 대상으로 촬영을 진행하려 했으나, 인테리어 도용 등의 영업비밀에 속하는 부분이라 대부분의 점주들이 촬영 동의서를 작성하지 않음
  - 이에, 매장 내부의 경우 크롤링 방식으로만 지원하고, 메뉴판의 경우 'R사'에서 이전에 구축한 데이터를 크롤링 데이터와 함께 대신 제공
  - 'R사'에서 제공하는 메뉴판의 경우 다양한 형태의 메뉴판 학습을 위해 각 POI별로 확보한 다양한 형태의 메뉴판을 제공하는 데 포커싱을 맞춤

* POI(Point of Interest) : 관심 지점

- 크롤링 데이터
  - 한국관광공사의 협조를 얻어 크롤링 작업 및 관광 메타데이터 구축에 사용할 TourAPI POI 정보를 크롤링 진행(총 21,000건, 전체의 89%)

항목	내 용
형태	• 정형 데이터
형식	• XML 형태 (API 데이터 수집)
구조	• 데이터 수집 후 Table 형태로 DB 저장(기존 필드 그대로 사용)
접근	• API 데이터를 활용한 크롤링
저작권	• 한국관광공사 제공

  - 수집 방법
    1. TourAPI를 수집하여 크롤링 대상 키워드 추출
    2. 키워드 기반으로 타겟 사이트에서 이미지 url 등의 정보 수집
    3. 수집된 이미지 URL 기반으로 실제 원천데이터 수집

- 원시데이터 적합성 검토
  - 원시데이터 항목별 데이터 획득 방법, 법적문제 발생가능여부 등을 검토하여 실제로 인공지능 학습용 데이터 구축에 활용할 수 있는 데이터를 선정한다.

- 원시데이터 선정
  - 데이터 품질, 획득 가능성(가능여부 및 획득량), 획득 비용, 수행 및 참여기관의 기술수준, 법적 요건 등을 검토하여 획득할 데이터를 최종 선정한다.
  - 선정된 원시데이터를 획득하기 위해 필요한 정보, 또는 원시데이터 획득현황을 파악하기 위한 데이터 명세서 또는 정의서를 작성하여 데이터 획득 기준으로 활용한다.

표 3-3. 원시데이터 명세서 작성 예시

데이터 명		매장 전경 데이터
데이터 포맷		jpg(이미지 파일)
데이터 요약		주요 관광지역 내 관광객들이 많이 찾는 매장들의 전경사진 이미지
데이터 출처		직접 촬영(내부 직원 및 크라우드소싱 활용)
데이터 이력	배포버전	Store_FrontImageDataSet_ver1.
	개정이력	신규
	작성자/배포자	작성자/배포자 : 최OO
데이터 통계	데이터 구축 규모	총 50만 건
	데이터 분포	서울시 주요 관광지구 : 1지구(명동)(25%), 2지구(이태원)(10%), 3지구(동대문)(10%), 4지구(종로)(20%), 5지구(잠실)(10%), 6지구(강남)(25%)
기타 정보	대표성	별도문서 참고
	독립성	
	유의사항	
	관련 연구	해당사항 없음

3.3 획득 절차 및 항목

- 데이터 획득·정제 절차 수립
  - 원시데이터 획득 및 정제 절차 수립 시 데이터 획득 방법별로 명확하게 획득·정제 절차가 정의될 수 있도록 한다.

    1) 원시데이터 직접 제작

    ※ 사진 촬영, 이미지 스캔 등

    2) 수행 및 참여기관 내·외부에 있는 데이터 수집

    ※ API, 크롤링, 직접수령 등

  - 데이터 관점 뿐만 아니라, 기관간 역할, 작업자 업무, 작업자-관리자 간 관계, 행정요소 등 사람 관점에서 실질적인 구축작업에 필요한 사항을 종합적으로 고려하여 절차를 수립한다.

표 3-4. 데이터 획득 방안 정의 예시

	데이터 획득 형태	수집장비	데이터 형식	수집처(장소)	담당 인원
1	야외현장 촬영	디지털카메라 (모델명 : 000)	RAW → JPG	000시 주요 관광지역 (00, 00, 00 등)	과제 인력 및 크라우드소싱 인력
2	문서 스캔	스캐너 (모델명 : 000)	JPG	크라우드소싱 수집사이트 (00.com)	크라우드소싱 인력
3	웹 크롤링	크롤링 서버	여러 이미지 포맷 → JPG	00사이트 (oo.com)	000사 크롤링 담당자

- 촬영을 통한 데이터 획득 절차 수립

  1) 계획 수립

  - 실제 현장에서 촬영을 통한 데이터 획득 시, 실제 환경을 반영한 다양한 조건의 인공지능 학습데이터 구축하기 위한 촬영전략을 수립한다. 이때 구축할 데이터의 범위, 목적, 획득기간 등을 고려하여 적절한 획득 범위(coverage)와 깊이(depth)를 설정한다.

    ※ 획득 범위 : 얼마나 많은 대상을 촬영할지 / 깊이 : 한 대상을 얼마나 다양하게 촬영할지

- 구축 대상이 되는 데이터를 획득할 수 있는 촬영 장소 및 일정을 수립한다.
- 직접 촬영을 통한 데이터 획득 절차를 마련한다. 이때 ①촬영 장비 제작 및 환경 세팅, ② 촬영 시나리오 교육/실습, ③실제 촬영 진행, ④촬영 결과 저장 순서로 진행할 수 있다.

2) 촬영방법 수립

- 이미지의 일관된 품질을 유지하고 신속한 촬영을 달성할 수 있는 촬영장비 및 환경을 세팅한다.
- 특히 획득 단계에서 크라우드소싱을 활용할 경우 전문적인 장비 없이도 획득할 수 있도록 스마트폰 등을 활용한 촬영방안을 마련한다.

※ 특히 스마트폰 촬영 시 품질 기준을 충족할 수 있도록 기준 해상도 등 환경설정 방법을 구체적으로 제시한다.

- 일관된 이미지 품질을 유지하기 위해 촬영 높이, 각도, 거리, 조명 등 세부 기준을 마련하고, 올바른 예시와 잘못된 예시를 마련하여 작업자가 쉽게 이해하고 올바르게 촬영할 수 있도록 한다.
- 동일한 대상을 시간 간격을 두고 촬영해야 하는 경우 제약조건을 명시한다.

※ 동일 대상은 동일한 각도, 위치에서 촬영 등

[올바른 촬영]

[잘못된 촬영]

그림 3-1. 올바른 촬영 / 잘못된 촬영 안내 예시

표 3-5. 촬영기준 수립 예시

	항목	촬영 기준
1	높이	• 5단계 (눈높이 기준 -50cm / -25cm / 0cm / +25cm / +50cm)
2	조명	• 2단계 (조명없음/ 기본 플래시(야간일 경우에만))
3	각도	• 3단계 (정면 / 좌측 다섯 보폭 / 우측 다섯 보폭)
4	거리	• 2단계 (근거리(3m) / 원거리(6m)
5	기타	• 매장의 주요 풍경과 간판이 하나의 이미지에 전부 들어오도록 함
6	시점	• 2단계 (주간/야간)

- 다양한 각도, 시점에서 동시에 촬영이 필요할 경우, 별도의 장비를 제작하여 신속하고 높은 품질의 데이터를 획득할 수 있다.

【참고】 촬영장비 제작 및 세팅 예시

- 트라이포드에 나무 패널을 붙이고, 패널에 스마트폰(갤럭시노트 4) 5대를 고정
  - 실제(Real) 사용자들이 스마트폰을 활용하여 촬영하기 때문에, 디지털카메라 등의 전문적인 기기가 아닌 스마트폰을 활용하여 원시데이터 촬영 진행
  - 스마트폰의 경우에도 사용자들이 많이 사용하는 삼성 갤럭시 시리즈를 선정하고, 최신기기가 아닌 어느 정도 보급된 보급형 기기를 선정
- 동시 촬영을 위해 카메라(스마트폰)에 블루투스 볼 앱 설치
- 나무패널 위 보조배터리를 설치하여 촬영 지속성 강화, 스마트폰에 보조배터리 연결
- 원시데이터 촬영의 경우 해상도 5,312*2,988(16:9) 기준으로 촬영
  * 해당 해상도는 갤럭시노트 4의 16:9 기준 촬영 최대 해상도(1장당 약 4.58MB)

- 문서·이미지 스캔을 통한 데이터 획득 절차 수립
  - 문자 스캔 이미지를 확보할 수 있는 자료 유형 및 확보 방법(장소)을 검토한다.

    ※ 예) 도서 : 국립중앙도서관 / 고문서 : 00박물관, 나라기록관 / 일반 필기글씨문서 : 크라우드소싱

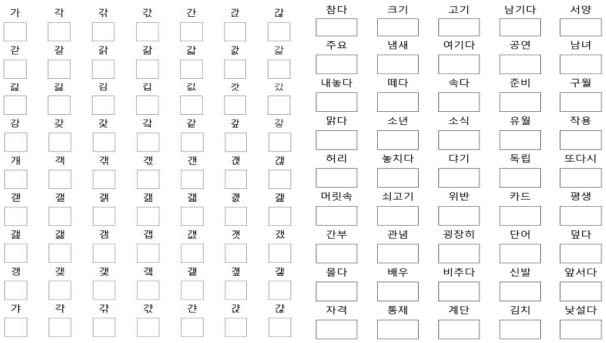
  - 데이터 획득에 필요한 자료 확보 및 스캔하기 위한 장소 및 일정을 수립하며, 자료 활용가능여부 확인 및 저작권 보호를 위해 자료 보유 기관(개인)과 필요한 사항을 협의한다.

- 확보 자료를 안전하고 스캔하기 위한 방법을 마련한다. 이 때 원본 자료 보존상태가 중요한 경우 자료를 훼손하지 않고 이미지를 확보하기 위한 방법을 마련한다.
- 크라우드소싱을 활용하여 자료를 확보할 경우 데이터 구축 목적 및 의도에 맞는 자료를 얻을 수 있도록 일정한 서식을 만들거나 작업 안내서 등을 작업자에게 배포할 수 있다.
- 일관된 이미지 품질을 유지하기 위해 변환 파일 포맷, 해상도, 올바른 스캔결과 기준 등을 정의한다.

【참고】 스캔 이미지 획득절차 수립 예시(한국어 손글씨 이미지)

- 데이터 수집자 모집
  - 다양한 손글씨를 확보하기 위해 성비, 나이대에 대한 기준을 수립하여 데이터 입력원 선발
  - 약 200여명의 표본 집단을 선정하여 채용

- 데이터 수집 : 데이터 수집자에게 데이터 획득용 워크시트 배포 후, 작성 요청
  - 아래 형태로, 안내 문구를 아래칸에 그대로 작성하는 형태로 진행
  - 데이터 수집에 활용되는 필기구는 실생활에서 가장 많이 사용되는 볼펜으로 제한



- 데이터 변환 : 수집된 워크시트를 대규모 네트워크 스캔 작업을 통해 디지털 데이터로 변환
  - 스캔 이미지는 jpg파일로 저장
  - 스캔 이미지는 글자(음절), 단어(어절), 문장 이미지 카테고리별로 분류하여 저장함

- 데이터 획득항목 정의
  - 원시데이터 함께 데이터 획득 시 확보해야할 정보를 정의한다.
    1) 이미지 메타데이터 : 촬영일시, 촬영위치(GPS 좌표), 노출도 등
       ※ 사진 촬영 시 저장되는 메타데이터 포맷(EXIF) 정보를 활용할 수 있음
    2) 도메인 정보 : 클래스 정보, 촬영대상의 이름, 상태정보 등
  - OCR 이미지 데이터 획득 시 수집 및 저장할 정보는 **'부록1. 인공지능 학습용 데이터셋 구축 공통참조기준'**을 준용하여 정의한다.
    1) 라벨링 메타정보 공통참조항목은 OCR 이미지 데이터 획득 시 공통적으로 기록해야할 메타정보이다.

표 3-6. OCR 이미지 데이터 획득 시 라벨링 메타정보 공통참조항목

No.	속성명	항목 설명	Type	필수여부	작성예시
1	Dataset.identifier	데이터셋 식별자	string	필수	IMG_OCR_01 (데이터유형_목적_순번)
2	Dataset.name	데이터셋 이름	string	필수	이미지 내 간판 텍스트 인식을 위한 학습용 데이터셋
3	Dataset.src_path	데이터셋 폴더 위치	string	필수	/dataSet/text/
4	Dataset.label_path	데이터셋 레이블 폴더 위치	string	필수	/dataSet/text/
5	Dataset.category	데이터셋 카테고리	number	필수	0: OCR, 1: 객체인식 등
6	Dataset.type	데이터셋 타임	number	필수	0: 텍스트, 1: 이미지, 2:영상, 3: 음성 등

2) 라벨링 이미지 파일 공통참조항목은 OCR 이미지 데이터 획득 시 공통적으로 기록해야할 이미지 파일 정보이다.

표 3-7. OCR 이미지 데이터 획득 시 라벨링 이미지 파일 공통참조항목

No.	속성명	항목 설명	Type	필수여부	작성예시
1	Images.identifier	이미지 식별자 (파일명)	string	필수	IMG_OCR_01_00001 (Dataset ID_순번)
2	Images.type	이미지 파일 확장자	string	필수	JPG, PNG 등
3	Images.width	이미지 가로 크기 (픽셀)	number	필수	1012
4	Images.height	이미지 세로 크기 (픽셀)	number	필수	768
5	data_captured	이미지 생성 일자	string	필수	yyyy.mm.dd HH:MM:SS

3) 라벨링 선택항목(저작권 정보)은 데이터 획득 시 저작권 정보가 필요한 경우 기록하는 정보이다.

표 3-8. OCR 이미지 데이터 획득 시 라벨링 선택항목(저작권 정보)

No.	속성명	항목 설명	Type	필수여부	작성예시
1	licenses.id	라이센스 고유 번호	string	선택	http://www.apache.org/licenses/LICENSE-1.0
2	licenses.name	라이센스 이름	string	선택	Apache License 1.0
3	licenses.url	문서 식별자	string	선택	NEWS_000001

- 획득 데이터 저장 및 관리
  - 획득 파일에 대한 저장, 전송, 백업 등 관리 절차 및 방안을 수립한다.

【참고】 획득 데이터 저장 방안 수립 예시

- 촬영 진행 후 사무실에 복귀하여 촬영 데이터를 노트북 3대에 순차적으로 옮긴 후, 노트북, 클라우드 및 외장하드 활용하여 3중 백업 진행
  - 총 3팀으로 나눠 원시데이터 촬영 진행하기 때문에, 팀마다 1대의 노트북을 활용하여 원시데이터 백업
  - 스마트폰 외장메모리를 활용하여 원시데이터 저장하고, 외장메모리에 있는 원시데이터를 파일 복사를 통해 원시데이터 전송 손실에 대비
  - 복사된 원시데이터를 노트북을 활용하여 노트북, 클라우드, 외장하드에 3중 백업
  - 외장하드 고장에 대비하기 위해 NAS* 등의 추가 장비를 활용하여 주기적으로 백업
- 촬영자 관리 및 촬영분 일일 관리를 위해 직접 촬영 데이터를 “촬영일자”_“촬영자” 기재한 폴더 형태로 구성하여 원시데이터 저장

* NAS(Network Attached Storage) : 네트워크 결합 스토리지

  - 획득한 파일을 체계적으로 분류하기 위해 데이터 종류 및 분류에 따른 라벨링데이터 파일 명명법과 파일 저장구조를 정의하고, 정의된 내용에 맞게 파일을 저장한다.

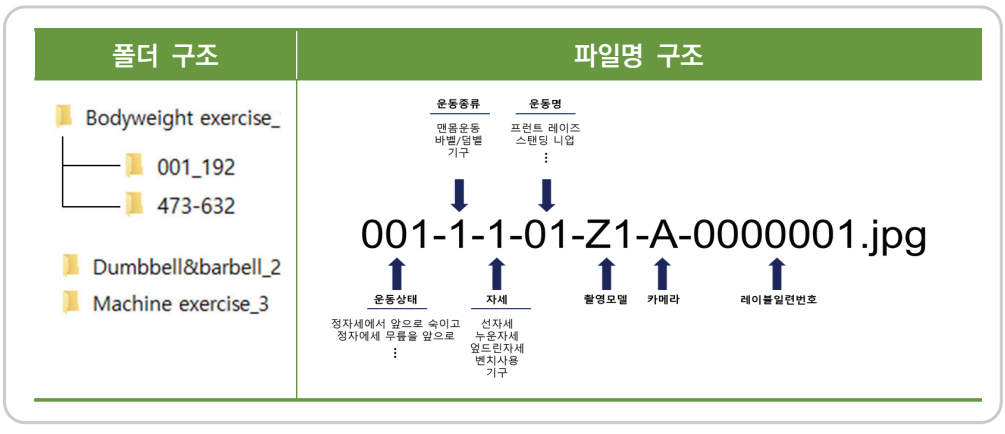


그림 3-2. 획득 데이터 폴더구조 및 파일명 코드화 예시

- 동일한 대상을 시간 간격을 두고 촬영해야 하는 경우, 파일 및 폴더 명명규칙, 메타데이터 저장 등에서 시계열 정보가 반영될 수 있는 규칙을 정의한다.

3.4 획득 데이터 정제 방식

- 정제 프로세스 수립
  - 어노테이션 단계에 들어가기 전에 학습용 데이터로 적합한 데이터를 선별하고 처리하는 정제 프로세스를 획득방법별로 수립한다.
  - 데이터 정제는 도구(소프트웨어)를 활용하여 정해진 규칙에 따라 제외 또는 변환하는 방법, 작업자가 직접 눈으로 확인하여 검사하는 방법 등을 적용할 수 있다.

- 정제 기준 수립
  - 데이터 구축 목적, 데이터 유형, 도메인 특성에 따른 데이터 정제 기준을 수립한다.
  - 촬영 이미지 데이터의 정제 기준으로는 촬영 장비, 이미지 크기, 비율, 화질 등의 요소가 있고, 스캔 이미지 데이터의 정제 기준으로 스캔 품질, 노이즈 여부 등이 있으며, 어떤 기준으로 정제하는지 정의한다.
  - 데이터 라벨링에 포함하지 않아야 할 개인정보 등을 필터링하는 정제 기준을 마련한다.

표 3-9. 촬영 이미지의 주요 정제기준

기준	고려사항
촬영수단	• 촬영 수단의 제한여부(스마트폰, 카메라, 캠코더, 그 밖의 특수장비 등)
객체의 크기비율	• 촬영 대상이 이미지 내에서 차지하는 적정크기 또는 크기 제한
촬영대상제한	• 이미지 내에서 촬영 대상 객체 외 다른 것들이 포함되어도 되는지 여부 • 한 이미지 내에 포함될 수 있는 객체의 개수 제한(최소, 최대 등)
이미지 비율	• 촬영 시 가로, 세로 방향 • 4:3, 16:9 등 가로세로 비율
화질 및 필터	• 해상도 제한 여부(최소 해상도, 최대 해상도 등) • 필터가 적용된 이미지 가능 여부
잘못된 촬영 허용수준	• 촬영 대상의 초점 안맞음 허용 여부 • 아웃포커싱 사진 허용 여부 • 이미지 흔들림 허용 여부 • 촬영 대상의 잘림 또는 가려짐 허용 여부 • 그 밖의 사유로 촬영 대상이 잘 보이지 않거나 흐릿함 허용 여부 • 기울어진 사진 허용 여부
개인정보처리	• 개인정보보호법 위배 여부
저작권	• 저작권 침해 가능성 여부

표 3-10. 스캔 이미지의 주요 정제기준

기준	고려사항
오탈자 여부	• 스캔 이미지 내 오탈자 및 발견 시 허용 여부
화질 및 필터	• 해상도 제한 여부(최소 해상도, 최대 해상도 등)
스캔 품질	• 이미지 흔들림 허용 기준 • 빛 노출 과다 허용 기준 • 스캔 대상의 잘림 또는 가려짐 허용 여부 • 기울어짐. 찌그러진 이미지 허용 여부 • 그 밖의 사유로 촬영 대상이 잘 보이지 않거나 흐릿함 허용 여부
개인정보처리	• 개인정보보호법 위배 여부
저작권	• 저작권 침해 가능성 여부

【참고】 이미지 정제 기준 처리방법 예시

- 해당 과정에서 관련 없는 항목이 찍힌 사진, 피사체와 맞지 않는 사진, 공통 촬영가이드에 어긋나는 사진, 첫 촬영시 매장정보 메모지가 포함된 사진 등을 1차 정제 진행
- 직접 촬영의 경우 서울시 관광특구 6개소에서 촬영 진행하여, 불가피하게 사람들의 얼굴들이 노출되는 경우가 있어 개인정보 보호 처리를 위해 해당 원시데이터에 노출된 얼굴 이미지를 blur 처리 진행
  - 딥러닝 모듈 중 Face Recognition(안면인식) 모듈을 사용하여 직접 촬영 데이터에서 사람의 얼굴을 별도 추출하고 blur 처리를 진행하는 기술을 적용
  - 해당 모델을 활용하여 사진에서 얼굴 영역을 인식하고, 인식한 얼굴 영역에 blur처리를 진행하는 2단계 과정으로 작업 진행하여 원시데이터에서 개인정보 관련 내용 정제
- 개인정보 처리된 원시데이터를 기준으로 하여, 각 사진별 해상도를 고/중/저 3단계로 분류하여 별도 생성/저장
  - 원시데이터의 경우 1장당 4.58MB 정도로 다량의 이미지 학습을 하기에는 용량 및 GPU 등의 컴퓨팅 파워가 다수 들어가는 문제점이 존재
  - 이를 해결하기 위해 2017년 얼굴 데이터 구축 조건과 같이, 원시데이터에서 고/중/저 해상도 3단계를 기준으로 이미지 전처리 수행하여 필요에 따라 해상도에 맞춰 사용자들이 골라서 사용할 수 있도록 함
  - 고/중/저 해상도의 경우 이미지 비율을 유지하면서, 사용자들이 실제로 많이 촬영하는 해상도 및 동영상 재생 조건에 맞춰 설정
    ※ 고해상도 : 2560*1440(16:9) 갤럭시 기본 사진 앱 기준,
    중해상도 : 1280*720(16:9), 저해상도 640*360(16:9)

* NAS(Network Attached Storage) : 네트워크 결합 스토리지

3.5 획득 도구 및 정제 도구

- 획득 및 정제도구
  - 데이터 획득 및 정제도구 개발 및 활용 계획을 작성한다.
  - 학습용 데이터 구축에 필요하지 않은 사람 얼굴 등 특정 이미지를 인식하여 처리해야 하는 사항은 안면인식모델 등 딥러닝 모델을 사용하여 처리할 수 있다.
  - 데이터 획득, 정제도구를 자체적으로 개발하기 어려운 경우, 시중의 제작 도구 또는 그와 유사한 역할을 할 수 있는 서비스·애플리케이션을 활용할 수 있다.

3.6 획득 시 고려사항

- 법·제도 준수
  - 데이터 획득 대상, 획득방법이 법·제도를 저촉하거나 또는 사회 윤리에 어긋나지 않도록 한다.
  - 개인정보 및 사생활 보호가 필요한 항목 획득 시, 개인정보보호법 등에 따라 적절한 법적, 기술적 절차를 거친 데이터를 활용하며, 그렇지 않은 데이터는 정제 과정에서 처리될 수 있도록 한다.

    ※ 법적 절차 : 개인정보 활용 동의, 초상 활용 동의, 명예훼손 가능성 여부 검토 등

    ※ 기술적 절차 : 데이터 유형별로 적용할 수 있는 익명처리 기법 적용

    ① 수치형 데이터 : 데이터 범주화 등

    ② 텍스트 데이터 : 이름, 민감정보 키워드 데이터 변환 등

    ③ 이미지·동영상 데이터 : 모자이크·블러처리, 크롭(자르기) 등

    ④ 음성 데이터 : 크롭(자르기) 등
  - 데이터가 3자 제공 및 대중에 개방에 문제가 없도록 법적요건 및 동의서 내용 등을 검토한다.

- 저작권 보호 대상인 데이터 획득 시 법에 저촉되지 않는 범위 내에서 획득할 수 있는 방안을 마련하며, 저작권 보호 대상 저작물 활용 필요 시 가급적 동의서, 계약서 등을 활용한 서면 자료 확보를 권장한다.

  ※ 예) 방송국 동영상을 활용할 경우 특정 방송국 이름(로고) 노출 가능 여부
  ※ 예) 이미지·동영상 내 특정 기업의 로고, 제품 형태 등 노출 가능 여부
  ※ 예) 뉴스, 출판물 등 저작권 보호 대상 저작물 활용 시 관련 당사자·기업·기관과 협의방안

- 개인정보활용동의서 및 저작물 활용 동의서 등 법적 요건을 준수하기 위한 관리방안을 마련한다.

  ※ '인공지능 학습용 데이터 품질관리 가이드라인'의 Ⅲ.품질관리기준, 3.1.5 체계준수성-보안준수의 【참고】개인정보보호 및 보안관련 법령·고시·권고 참조

【참고】 온라인 서명 플랫폼 활용 예시

- 원시데이터 제공자와 개인정보 수집 및 이용 동의 및 저작권 활용 계약을 전자서명 계약 기반으로 체결함으로써 민감한 정보 제공에 따른 위험 부담을 최소화함
- 저작권 이용 허락 계약 체결은 온라인 서명 플랫폼을 활용해 계약서를 업로드하여 법적으로 유효한 온라인 서명 기능을 구현

- 데이터 다양성 확보
  - 인공지능 학습모델이 현실을 잘 반영하고 본래의 구축 목적을 달성할 수 있도록, 획득하는 데이터가 일부 범주에만 치우치지 않고 가능한 다양한 시간·공간·집단·수준이 포함될 수 있도록 한다.

【참고】 데이터 다양성 미확보 예시

1. 랜드마크 건축물 학습 데이터 획득 시, 비슷한 구도의 이미지만 반복적으로 획득
2. 광화문, 서울N타워, 4.19학생혁명기념탑 등 서울시 일부 유명 랜드마크 위주로만 획득
3. 촬영 시간대와 날씨가 클래스별로 거의 고정됨

- 데이터 편향 방지 및 윤리 준수
  - 인공지능 학습모델이 인간의 비윤리 또는 편견을 학습하지 않고 사회적 윤리를 준수할 수 있도록 비윤리적 내용, 편견·편향된 데이터의 획득은 지양한다.

    ※ 딥페이크 분류, 가짜뉴스 분류, 비속어 필터링 등 비윤리·편향·왜곡된 정보 특성을 학습하는 것을 목적으로 구축하는 데이터는 예외로 할 수 있다.

- 사업계획서 및 데이터 구축 요건 일치
  - 사업계획 당시 정의한 데이터 구축 기준에 맞춰 데이터를 획득·정제하도록 구축 현황을 모니터링한다.

【참고】 사업계획서 및 데이터 구축 요건과 실제 데이터 획득 간 불일치 사례

1. 사업계획 시 JPG 포맷으로 획득하기로 했으나, 실제 구축되는 데이터에 JPG 외의 포맷(GIF, TIF 등)이 존재
2. 사업계획 시 FHD(1920px x 1080px)파일로 구축하기로 했으나, 실제 획득한 데이터에 더 낮은 해상도(HD 등)가 포함됨
3. 사업계획 시 '매장간판', '표지판', '메뉴판' 등 다양한 형태의 데이터를 구축하기로 하였으나, '간판' 분야의 이미지만 획득함

4 데이터 라벨링 작업

4.1 데이터 특성 식별 분류 체계 및 고려 사항

- 라벨링 작업 대상 및 범위 정의
  - 원천데이터 내에서 어떤 항목들을 라벨링 해야 하는지 대상과 범주를 정의한다.
  - 원천데이터 내에서 데이터 구축 목적에 부합하는 내용을 최대한 반영할 수 있는 정보를 라벨링할 수 있도록 라벨링 대상 범위를 정의하며, 데이터 품질 및 구축 목적과 무관한 내용을 불필요하게 라벨링하는 사항의 존재 여부 등을 검토한다.
  - 특히 이미지 전체에 대한 라벨링이 아닌, 하나의 이미지 안에 특정 영역, 객체 등을 라벨링해야 하는 경우 작업자들이 어떤 대상을 라벨링해야 하는지 판단할 수 있도록 세부적인 기준을 마련한다.
  - 원천데이터에 포함된 개인정보는 라벨링 대상에서 제외하거나, 익명처리 등 비식별화를 통해 개인정보를 알아볼 수 없게 라벨링한다.

    ※ 개인정보 활용 및 3자제공 동의를 받은 경우 동의 범위 내에서 개인정보를 라벨링 데이터로 활용할 수 있음

- 클래스 정의 및 관리
  - 원천데이터의 특성을 바탕으로 부여할 수 있는 클래스 리스트 또는 클래스의 범주를 정의한다.
  - 클래스를 정의할 때는 원천 데이터 내에 존재하는 다양한 값들을 모두 커버할 수 있도록 정의하고, 클래스 이름이 중복되거나 모호한 의미를 갖지 않도록 한다.
  - 클래스 이름은 의미를 바르고 명확하게 나타낼 수 있도록 적절한 어휘를 선택한다.

【참고】 잘못된 클래스 정의 예시 (텍스트 주제 및 감성 분류)

1. 클래스 간 의미 중복이 있어 구분이 애매함
   - (학업) vs (학교폭력) : 학업의 개념 안에 학교폭력을 포함
   - (학업 및 진로) vs (진로/취업/직장) : 전자와 후자가 서로 교집합이 있음
   - (상처) vs (슬픔) : 전자와 후자의 경계를 나누기 모호함
2. 라벨링 표기법에 일관성이 없음
   - (청소년) vs (청소년(10대)) : 전자와 후자를 같은 의미로 라벨링에 활용하였으나 표기법에 일관성이 없음
   - (여성) vs (FEMALE) : 전자와 후자를 같은 의미로 라벨링에 활용하였으나 언어·문자가 일관성이 없음
3. 범위와 주제에 맞지 않는 클래스 존재
   - 성별 구분에 (기타) 클래스 존재 : (남성), (여성) 외에 (기타) 성별을 정의할 수 있는지, 또는 의미가 무엇인지 불명확

- 라벨링 진행 중에 이전에 정의되지 못했거나 새롭게 정의가 필요한 클래스가 발견될 경우 클래스 항목 업데이트 방안을 마련한다.
- 클래스를 정해진 목록에서 선택하지 않는 경우에도, 작업자마다 일관된 기준 및 규칙에 따라 속성값을 부여할 수 있도록 하는 기준을 마련한다.

  ※ 주로 OCR 이미지, 텍스트, 음성 전사 등 문장, 텍스트로 값을 부여하는 데이터가 해당

- 동일한 의미라도 작업자마다 맞춤법 준수 여부, 외국어 표기 방식 등의 차이에 따라 다르게 표현될 수 있기 ,때문에, 표기법에 대한 세부적인 방식을 지정한다.

【참고】 라벨링 입력 범위 및 기준 마련 예시(다국어 입력)

- 매장 인식 관련 과업 수행에 적합하도록 매장 전경을 영역 지정 후, 해당 매장 정보를 어노테이션 다국어 데이터(영문, 일문, 중문)의 경우 매장 자체의 다국어 표기를 따르며, 다국어 표기가 없을 경우 영문은 로마자 고유명사 표기를 따름
  - 1순위 : 간판의 외래어 표기 기입(매장 POI* 정보)
  - 2순위 : 영문의 경우 한국어 매장 POI 정보를 로마자로 표기한 내용을, 중문 및 일어의 경우는 TourAPI 다국어 데이터에 기초하여 매칭되는 값이 있을 경우 입력 진행

* POI(Point of Interest) : 관심 지점

4.2 데이터 라벨링 방법 및 절차

- 개요
  - 획득→정제 과정을 통해 도출된 원천데이터를 라벨링하여 학습 데이터를 생성하기 위한 과정 및 고려사항을 작성한다.
  - 라벨링 지원 도구를 활용하며, 용어 및 분류체계를 준수하여 라벨링한다.

- 라벨링 작업 방식
  - 라벨링할 정보의 특성에 따라 자동, 반자동, 수동 방식을 결정한다. 원천데이터로부터 추출하는 방식이 정형화되어있고 자동화할 수 있는 사항인 경우 자동 방법을 고려할 수 있으며, 기계가 판단하기 어려운 사항은 반자동 또는 수동 방식이 적절하다. 반자동 방식은 자동으로 라벨링한 이후 사람이 다시 확인하여 수정하는 방식으로 작동된다.

- 작업 배분
  - 획득된 데이터를 라벨링에게 작업자에게 배분하고 라벨링 결과를 다시 저장하는 파일 저장체계 및 프로세스를 정의한다.

- 라벨링 작업 기준
  - 데이터별 어노테이션 기준, 라벨링 기준 등을 상세히 기술하며, 구체적인 예시를 들어 설명하여 작업자들이 혼동없이 명확한 기준을 갖고 빠르게 작업할 수 있도록 한다.

    ※ 레이블 범주, 레이블 부여기준(ground truth) 제시, 레이블 부여 예시, 애매한 내용이 나올 경우의 처리 기준, 자주 실수하는 예시, 검사 기준 등

표 4-1. 어노테이션 대상 고려사항 검토 사례(OCR 이미지)

항목	기준 및 고려사항
한글	• 옛 한글은 작업 대상으로 할 것인지
영문	• 대문자, 소문자 구분할지 • 로마자 기반으로 한 언어에서 사용되는 문자도 작업 대상인지(é, ð 등)
한자	• 한자는 작업 대상으로 할 것인지
아라비아 숫자	• 아라비아 숫자(0~9)는 작업 대상으로 할 것인지
기호	• 기호 문자는 작업 대상으로 할 것인지, 그리고 어떤 것을 대상으로 할 것인지 ex) 컴퓨터 키보드에서 입력 가능한 모든 기호, 문자+한자키로 표현할 수 있는 모든 기호
문장부호	• 문장 부호는 작업 대상으로 할 것인지, 그리고 어떤 것을 대상으로 할 것인지 • 한국어에서 사용하지 않는 문장부호를 포함할 것인지(⁉, ; 등)
기타 외국문자	• 한글, 로마자, 한자 외 외국어에서 사용되는 문자는 작업 대상으로 할 것인지

【참고】 라벨링 작업 안내 예시

1. 본 작업은 간판 내 문자를 대상으로 진행하는 작업입니다. 객체가 크기는 다양하며, 이미지에서 파악 가능한 객체는 모두 작업합니다.
2. 객체를 감싸는 박스의 크기는 객체보다 작거나 크지 않도록 해야하며, 객체의 그림자는 제외해야 합니다.
3. 100% 확대 배율에서 육안으로 확실히 문자를 식별할 수 있는 것만 라벨링하며, 불확실한 문자는 라벨링하지 않습니다.
4. 라벨링할 대상 객체가 없는 구획은 작업하지 않고 다른 구획으로 넘어갑니다.

※ 주의할 점

1. 식별가능한 객체는 모두 작업합니다. 크기 및 주변 환경을 꼭 확인해주세요.
2. 객체의 분류(Class)가 올바른지 확인합니다. (판단이 어려운 경우 필수로 리더에게 확인받은 후 진행해주세요.)
3. 객체의 분류(Class)마다 영역에 맞게끔 작업이 이루어졌는지 확인합니다. 분류에 따라 작업이 달라질 수 있습니다.

- OCR 이미지 데이터 라벨링
  - OCR 이미지 데이터 라벨링 방법 및 기준은 구축 목적, 도메인, 활용 분야를 고려하여 영역설정(바운딩 박스 등) 기준 및 텍스트 입력 기준을 수립한다.
  - 일반 텍스트가 아닌 타이포그래피 등 다양한 텍스트 형태 및 레이아웃의 OCR 이미지 데이터 구축 시 다양한 상황을 반영하여 기준을 수립한다.
  - 다국어 OCR 이미지 학습데이터 구축 시 라벨링 범위 및 입력 기준을 수립한다.

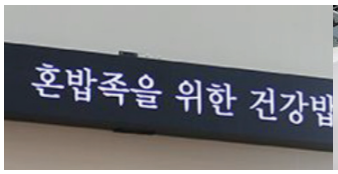
표 4-2. 촬영 OCR 이미지 데이터 라벨링 기준 안내 예시(매장 간판)

1. 글자의 영역을 최대한 여백없이 작성

[올바른 예]

[잘못된 예 : 여백이 있음, 글씨가 잘려있음]

2. 한글, 영문, 한자, 숫자 및 혼합 텍스트를 라벨링함



[한글]

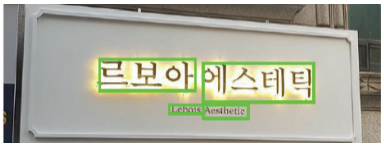
[영문]

[영문+숫자]

3. 기호는 라벨링에서 제외하고, 띄어쓰기는 분리하여 라벨링함

'화재' '지진' 으로 boxing
'화재' '지진' 으로 받아쓰기

[기호(특수문자)는 박싱하지 않음]
['화재' / '지진' 두 개를 나눠서 박싱]



[띄어쓰기가 있는 경우 띄어쓰기 단위로 분리하여 박싱]
['르보아' / '에스테틱' / 'Lebois' / 'Aesthetic'을 나눠서 박싱]

4. 특수 알파벳은 개별 박싱 후 기호로 처리

['PH'와 'o'를 분리하여 박싱, 'o'는 기호로 처리]

['é'. 'Å'는 개별 박싱 후 기호로 처리]

5. 영문 필기체는 박싱 후 일반 영문으로 입력

['Skechers'로 입력]

['Oatmeal'로 입력]

6. 글자로 인식할 수 있는 타이포그래피는 박싱 후 원래 글자로 입력

['매머드커피'로 입력]

['백곰' / '막걸리' / '크래프트비어' / 기호(%)로 입력]

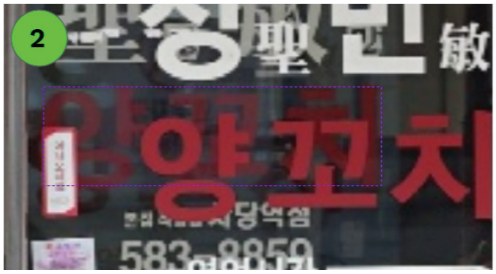
7. 텍스트가 1획 이상 가려져있거나 삭제되어 있을 경우 박싱하지 않음

['미'의 'ㅁ'에 세로 장애물이 글자의 획으로 보일 수 있기 때문에 제외]
['옥'의 'ㄱ'이 가려져 있기 때문에 제외]

['4'의 일부가 가려져 있어서 박싱하지 않음]

8. 글자가 겹쳐진 경우

[겹친 글자는 더 선명한 글자를 기준으로 박싱]



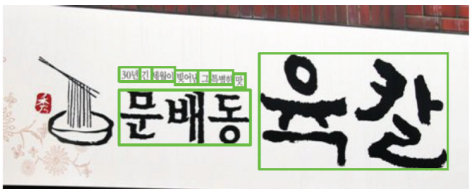
[비슷한 색상, 비슷한 투명도인 경우 박싱하지 않음]

9. 가로, 세로 레이아웃은 별도로 박싱함

[가로방향, '친구네' / 세로방향, '당구장'을 별도로 박싱]

['BAR', 'RA'를 별도로 박싱]

10. 글자 크기가 다른 경우 개별 단어로 박싱함



['문배동' / '육칼'을 별도로 박싱]

11. 원형 간판의 박싱 방법

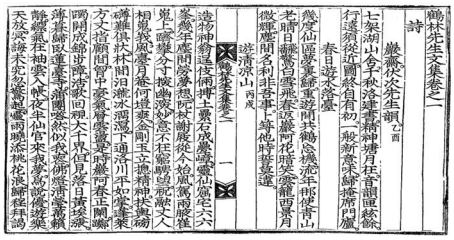
찾을 수 있는 항목

1. PROFESSIONAL
2. FOR
3. HAND
4. DRIP
5. COFFEE
6. DELPINO
7. GOOD
8. CAFE
9. ROASTERY
10. 핸드드립
11. 로스터리
12. 카페
13. 델피노

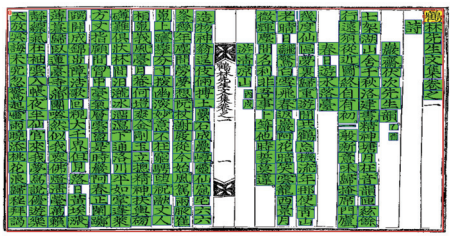
[원형으로 기울어진 텍스트는 적당한 구간으로 분리하여 박싱]

【참고】 스캔 OCR 이미지 데이터 라벨링 방법 예시(고서 한자 인식)

- 문자추출
  - 하나의 바운딩 박스 내에는 한 글자의 한자만 포함하도록 함
  - 1차로 데이터 저작도구를 활용해 자동으로 바운딩 박스를 생성함
  - 2차로 라벨링 작업자가 수동으로 잘못된 작업을 수정함
  - 원천 데이터 내 한자들의 자간이 좁아 한자의 획들이 정사격형의 문자 영역을 상호 침범하는 경우에는 불가피하게 라벨링 대상 한자의 획 일부가 잘리거나, 다른 한자의 획 일부가 라벨링 대상 한자의 바운딩박스에 포함되도록 할 수 있다. 단, 라벨링 대상 한자의 획 일부가 잘리거나 다른 한자의 획 일부가 바운딩 박스에 포함되었을 때 라벨링 대상 한자의 판독(인식)에 지장을 주거나 다른 한자로 오인되지 않도록 해야 함

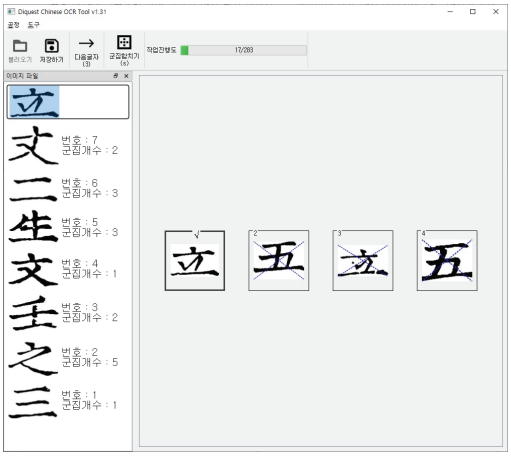


[문자추출 전 원문이미지]



[문자추출 후 원문이미지]

- 클러스터링
  - 같은 한자 세그먼트를 하나의 그룹(클러스터)으로 묶어주는 절차
  - 데이터 저작도구가 우선 자동으로 클러스터를 생성해주며, 사용자가 자동 생성된 클러스터를 수정하는 방법으로 작업을 진행함
  - 불러온 하나의 클러스터 안에 포함된 한자 낱자 이미지들을 확인하여 대표 한자 이미지와 다른 한자가 포함되어있을 경우 해당 클러스터에서 제거함
  - 불러온 하나의 클러스터 안에 포함된 한자 낱자 이미지들을 확인하여 대표 한자 이미지와 같은 한자인데도 클러스터에서 제외한다는 표시가 되어있는 한자가 있을 경우 해당 클러스터에 포함함
  - 대표한자가 동일한 복수의 클러스터가 존재할 수 있으며, 이 때는 가능한 한 하나의 클러스터로 묶음



[동일 한자에 대한 클러스터링]

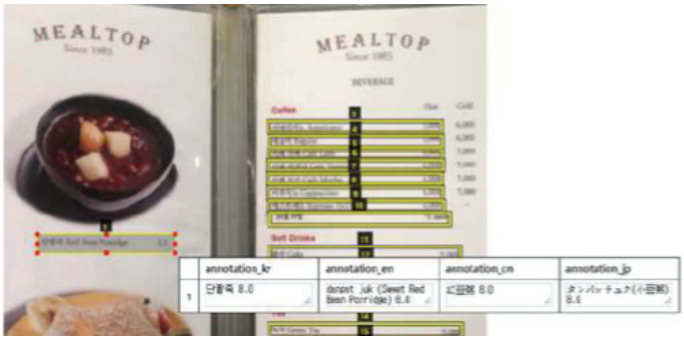
- 라벨링 값 입력
  - 클러스터링이 끝난 데이터를 열어 불러온 클러스터별로 한자 텍스트를 입력해주는 절차
  - 대표 한자의 이미지를 확인하여 해당 한자 텍스트를 입력함
  - 유니코드에 반영되어 있지 않아 입력이 불가능한 한자(신출한자, 고한자)나 문자로 인식은 되지만 원본의 훼손, 마멸 등으로 판독이 불가능하여 역시 입력이 불가능한 한자는 약속된 특수기호 '▼'로 입력함
  - 동형이음(의)자는 통일된 하나의 한자로 입력해야 함(동형이음(의)자 입력 기준 별도 수립)
  - 비슷한 형태의 글자가 많으므로 많으므로 한자 판독 및 입력 숙련자가 작업

[입력작업]

[글자별 라벨링 완료 후 모습]

【참고】 OCR 이미지 라벨링 사례 작성 예시(다국어 메뉴판 텍스트)

- 메뉴명+가격 중심으로 영역 설정 후 라벨링 진행(해당 부분에 필요없는 내용의 경우 되도록 영역설정에서 제외하는 것을 원칙으로 함)
- 사람이 육안으로 식별 불가능한 문자의 경우 라벨링에서 제외
- 다국어 기입의 경우 아래 순서에 따라 작성
  - 1순위 : 지정한 영역에 한국어 이외의 영문·중문·일문 표기가 있는 경우 해당 텍스트를 그대로 각 언어의 어노테이션 테이블에 작성. 영문·중문·일문 중 일부만 표기가 있는 경우는 해당 언어만 그대로 작성하고, 다른 언어는 2순위의 다국어 입력 방식을 따름
  - 2순위 : 지정한 영역에 다국어 표기가 존재하지 않을 시, 지정한 영역에 나와 있는 메뉴명을 서울시 외국어 표기사전에서 검색하여 검색한 결과를 각 언어 어노테이션 영역에 전체 기입



- OCR 이미지 데이터 어노테이션 기준 정의
  - 이미지 내 텍스트를 폴리곤, 바운딩 박스 등 닫힌 형태의 다각형 형태로 영역을 설정하는 기준을 마련한다.
  - 일반적으로 대상이 잘리지 않으면서 불필요한 공간이 남지 않도록 대상의 크기에 꼭 맞춘 사이즈로 박스를 치나, 인공지능모델 및 데이터 활용 목적에 따라 기준을 달리 정할 수 있다.
  - 이미지 내에 대상의 경계가 분명하지 않아 박스 영역을 설정하기 모호한 상황에서는 어떤 모양과 크기로 작업해야 하는 지에 대한 기준을 마련하여, 작업자가 헷갈리지 않고 명확한 기준을 갖고 작업할 수 있도록 한다. (표 참조)

표 4-3. 이미지 어노테이션 시 주요 기준 및 고려사항

항 목	기준 및 고려사항
초점	• 대상이 초점이 제대로 안맞아 경계선이 명확하지 않은 경우
흔들림/움직임	• 대상이 흔들리게 나와 경계선이 명확하지 않은 경우
밝기	• 대상이 너무 밝거나 어두워서 형체가 또렷하지 않은 경우
해상도(화질)	• 이미지 해상도가 낮아 대상의 경계가 또렷하게 보이지 않는 경우
잘림	• 대상이 이미지 경계에서 잘려서 전체가 드러나지 않는 경우
가려짐	• 대상이 다른 오브젝트에 가려서 일부분만 드러나는 경우

  - 이미지의 공통적인 어노테이션 주요 기준 이외에 작업 도메인에 따라 세부적인 어노테이션 기준을 마련한다.

표 4-4. OCR 이미지 어노테이션 시 주요 기준 및 고려사항 검토 사례

항 목	기준 및 고려사항
작업대상과 비대상의 섞임	• 한 어절 안에 작업 대상과 비대상이 함께 있는 경우 어노테이션 기준
기타 안보임	• 그림자, 빛으로 인한 가림으로 경계가 불분명한 텍스트의 어노테이션 기준 • 반사되어 표시된 텍스트에 대한 어노테이션 기준
회전 및 반전	• 회전되어 있거나 뒤집힌 형태의 텍스트에 대한 어노테이션 기준
왜곡	• 찌그러지거나 왜곡된 모양의 텍스트에 대한 어노테이션 기준
텍스트 겹침	• 두 개 이상 텍스트가 겹쳐진 경우 어노테이션 기준

4.3 데이터 어노테이션 포맷과 형식 정의 및 입력

- 어노테이션 포맷 및 저장 형식
  - OCR 이미지 데이터는 고정된 필드나 스키마가 존재하지 않는 비정형 데이터이기 때문에, 학습용 데이터로서 가치를 부여하는 어노테이션 정보를 저장할 수 있는 별도의 데이터 구조와 파일 포맷을 정의한다.
  - 어노테이션 파일 포맷은 특정 소프트웨어에 종속되지 않고 쉽게 열고 편집할 수 있는 포맷으로 선택하며, 구조화된 어노테이션 정보를 저장하기 적합한 포맷을 선택한다.

    ※ json, xml 등

- 어노테이션 정보 저장 구조
  - 어노테이션 정보(라벨링데이터)에 포함되어야 할 사항을 데이터 유형별(텍스트, 이미지, 동영상, 음성 등) 라벨링 참조 기준과 구축 목적에 따라 필요한 항목을 종합적으로 고려하여 정의한다.
  - OCR 이미지 데이터 어노테이션 정보 구조는 **'부록1. 인공지능 학습용 데이터셋 구축 공통참조기준'**을 준용하여 정의한다.

    1) 바운딩박스 어노테이션 구조
    - 이미지 내 텍스트 영역에 대한 바운딩박스 어노테이션 공통항목은 아래와 같다.

표 4-5. OCR 이미지 바운딩박스 어노테이션 공통항목

No.	속성명	항목 설명	Type	필수여부	작성예시
1	annotations[].bbox.id	바운딩박스 식별자	string	필수	BBX_0001(분류_순번)
2	annotations[].bbox.text	바운딩박스 내 텍스트	string	필수	교보문고
3	annotations[].bbox.x	바운딩박스 시작점 x 좌표	number	필수	100 (좌측상단 기준)
4	annotations[].bbox.y	바운딩박스 시작점 y 좌표	number	필수	120 (좌측상단 기준)
5	annotations[].bbox.width	바운딩박스 가로 길이(픽셀)	number	필수	273
6	annotations[].bbox.height	바운딩박스 세로 길이(픽셀)	number	필수	125

2) 장소정보 어노테이션 구조

- 이미지 내 텍스트와 매핑되는 실제 장소가 존재할 때, 매핑 관계를 표현하기 위한 선택적인 어노테이션 구조는 아래와 같다.

표 4-6. 장소정보 어노테이션 공통항목

No.	속성명	항목 설명	Type	필수여부	작성예시
1	annotations[].place.id	장소 식별자	string	선택	PLC_0001
2	annotations[].place.title	장소명	string	선택	교보문고 광화문점
3	annotations[].place.addr	장소 주소	string	선택	서울특별시 종로구 종로1가 종로 1
4	annotations[].place.longitude	위치정보(경도)	string	선택	126.977759
5	annotations[].place.latitude	위치정보(위도)	string	선택	37.570975

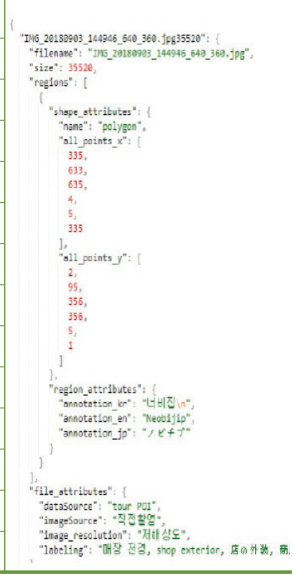
※ 관심지점(POI, Point of Interest)의 위치 정보는 도로명 주소와 WGS84 좌표 체계 혹은 국가지점번호 체계 활용

※ 공개 제한된 관심지점의 위치 정보는 표시하지 않음

- 어노테이션 정보(라벨링데이터)가 어떤 원천데이터와 매칭되는 지 확인할 수 있도록 어노테이션 구조 및 내용을 정의한다.

※ 학습용 데이터는 원천데이터 + 라벨링데이터로 구성됨을 고려

표 4-7. OCR 어노테이션 정보 구조 정의 및 구축 사례

항 목			설 명	json 포맷 구축 형태
ID			이미지 파일 고유 ID	
filename			이미지 파일명, 예) IMG_191022_434.jpg	
size			이미지 파일 크기	
regions	shape	name	어노테이션 형태, 예) polygon, rect 등	
		points	어노테이션 포인트 좌표(x,y)	
	region_attributes	annotation_kr	한글 기반의 어노테이션 데이터	
		annotation_en	영문 기반의 어노테이션 데이터	
		annotation_jp	일문 기반의 어노테이션 데이터	
file_attributes	datasource		수집 출처, 예) 직접 촬영, tourAPI	
	image_resolution		이미지 해상도	
	title		매장명	
	addr		매장 주소	
	longitude		위치정보(경도), 예) 126.983432	
	latitude		위치정보(위도), 예) 37.646221	

4.4 데이터 라벨링 완료 후 관리 방법

- 데이터 관리 기본사항
  - 목적에 맞는 데이터 어노테이션 기준을 수립하고 데이터 사용 목적에 맞게 관리
  - 데이터의 사용 목적에 맞는 일관된 자료인지 확인한다.
  - 데이터들의 편향성을 확인 후 필요에 따라 데이터 추가한다.
  - 보존 일정 및 규정 준수 요구 사항에 따라 데이터 보관, 관리한다.

- 데이터 저장 관리
  - 원천데이터에 추가된 라벨링 정보를 저장하고 관리하는 기준을 수립한다. 파일을 체계적으로 분류하기 위해 데이터 종류 및 분류에 따른 라벨링데이터 파일 명명법과 파일 저장구조를 정의한다. 정의된 내용에 맞게 파일을 저장하도록 작업자에게 안내한다.
  - 작업자들이 원천데이터 및 라벨링 정보 저장 구조에 맞게 저장할 수 있도록 저장 절차를 정의하고, 작업자를 대상으로 배포한다.

- 데이터 백업 관리
  - 원천데이터 및 라벨링데이터의 훼손 및 멸실을 방지하기 위해 안전한 보관방법 및 백업방안(백업 시스템 및 프로세스 구축, 관리 절차 등)을 마련한다.

- 데이터 관리 조직 운영 방안
  - 데이터셋 제작 책임자는 품질관리 책임자로서 획득되는 데이터의 품질을 주기적으로 검사 및 관리한다.
  - 주기적인 실무협의체와의 미팅을 통해 데이터 품질에 대한 피드백을 공유하고 논의한다.
  - 데이터 품질 제고를 위해 데이터 라벨링 방안에 대하여 전문 컨설턴트 등 외부 기관의 조언을 받을 수 있다.

4.5 데이터 라벨링 방식에 적합한 도구 선정

- 라벨링 도구 선정
  - 데이터 구축 목적 달성을 위해 원천데이터 형태, 구축 목적에 부합하는 라벨링 도구를 선정한다.
  - 기존의 도구를 가지고 인공지능 학습용 데이터 구축 목표 달성이 어려울 경우, 기존 라벨링 도구의 기능을 추가하거나 완전히 새로 개발하는 방법을 고려한다.

표 4-8. 국내외 주요 이미지 데이터 라벨링 도구

No	도구명	요금	설 명
1	labellimg	무료	• object detection 학습을 위해 영상에서 Bounding box를 지정하여 라벨링을 수행하고, 그 bounding box 정보들을 xml 형태로 저장 • 사이트 주소 : https://github.com/tzutalin/labelImg
2	LabelMe	무료	• Bounding box, Polygon, Polyline, Point 등 다양한 형태의 도형과 Classification, Segmentation 등 다양한 task의 라벨링 지원 • 사이트 주소 : https://github.com/wkentaro/labelme
3	Superb AI	유/무료	• 크라우드소싱 기반의 데이터 라벨링 도구 • 바운딩 박스, 폴리라인, 세그멘테이션, 키포인트, 클래스 분류 어노테이션 기능 • 이미지 작업·검사에 대한 오토 라벨링 기능 제공 • 사이트 주소 : https://www.superb-ai.com/
4	SELECTSTAR	유료	• 크라우드소싱 기반의 데이터 라벨링 도구 • 이미지 어노테이션 및 OCR, 문자인식 기능 제공 • 사이트 주소 : https://selectstar.ai/
5	BlackOlive	유료	• 크라우드소싱 기반의 데이터 라벨링 도구 • 바운딩박스, 폴리라인, 폴리곤, 키포인트, 큐보이드 어노테이션 기능 • 사이트 주소 : https://www.testworks.co.kr/contents/blackolive.html
6	크라우드웍스	유료	• 크라우드소싱 기반의 데이터 라벨링 도구(PC, 모바일 앱 지원) • 클래스 분류, 바운딩 박스, OCR 기능 • 사이트 주소 : https://www.crowdworks.kr/main.do
7	Amazon Textract	유료	• Amazon AWS 기반의 클라우드 서비스 • 인쇄 텍스트, 필기, 테이블 등 다양한 형태의 OCR 이미지에 대한 수동, 자동 텍스트 추출이 가능 • 사이트 주소 : https://aws.amazon.com/ko/textract/

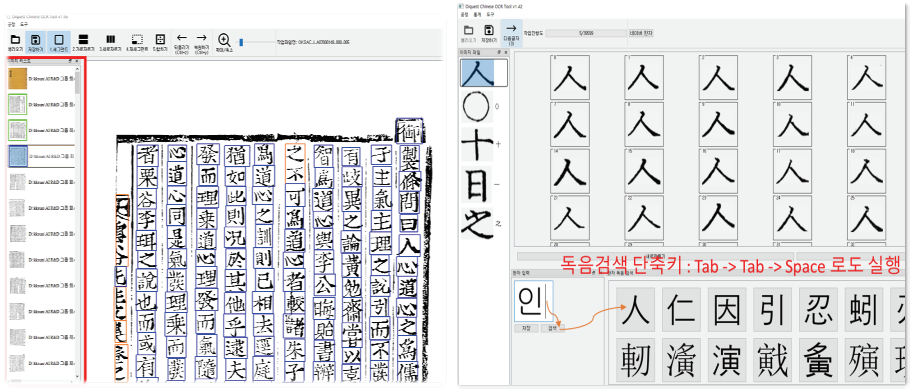
- 라벨링 도구 활용 매뉴얼 작성
  - 작업자가 활용할 도구의 사용법에 대한 매뉴얼을 작성한다. 매뉴얼은 Step-by-Step 형태로 쉽게 따라할 수 있도록 작성하며, 이미지, 동영상 등을 활용하여 이해를 도울 수 있다.

그림 4-1. 라벨링 도구 활용 매뉴얼 작성 예시

- 라벨링 도구 선정 및 개발 시 고려사항
  - 표준적인 어노테이션 및 라벨링 작업 가능 여부, 표준 파일 포맷 지원 여부 등을 고려한다.
  - 한국어 및 다국어를 입력할 때 깨짐현상이 없는 인코딩을 지원하는지 고려한다.
  - 특히 크라우드소싱 방식을 적용하는 경우, 다양한 작업환경(컴퓨터 성능, OS, 네트워크 등)에서의 실행 가능 여부를 확인한다.
  - 라벨링 결과를 효과적으로 관리하고 작업배분을 할 수 있는 관리 기능이 충실한 지 확인한다.
  - 작업자 또는 검사자가 어노테이션 결과를 눈으로 바로 확인할 수 있는 시각화 기능이 있는 지 고려한다.

【참고】 어노테이션 시각화 예시

- 라벨링된 값, 바운딩박스를 눈으로 확인할 수 있으며, 하나의 글자와 동일한 라벨링 값을 지니는 글자들의 모음, 동일한 독음에 해당되는 다른 글자들의 모음 등 여러 편의 기능을 제공한다.



5 처리 데이터 검사

5.1 검사 절차 정의

- 개요
  - 인공지능 학습용 데이터 구축을 위한 품질 검사 절차·방법은 데이터 유형, 도메인, 목표 서비스에 따라 달라질 수 있으며 사업 기간 및 예산 등 현실적인 여건을 고려하여 수립한다.
  - 데이터 검사 절차 및 규격은 데이터 구축 목적 정의 단계에서 수립한 데이터 활용 분야·목적을 달성할 수 있도록 정의한다.

- 검사 절차 정의
  - 다량의 데이터를 한정된 시간 내에 최적의 품질로 검사할 수 있도록 하는 검사 단계 및 절차를 수립한다.
  - 검사 프로세스는 학습용 데이터 구축 공정(획득, 정제, 라벨링) 각 단계별로 검사가 수행되는 형태를 기본으로 하며, 데이터 구축 공정 및 데이터 특성을 반영하여 적합한 절차를 수립할 수 있다.

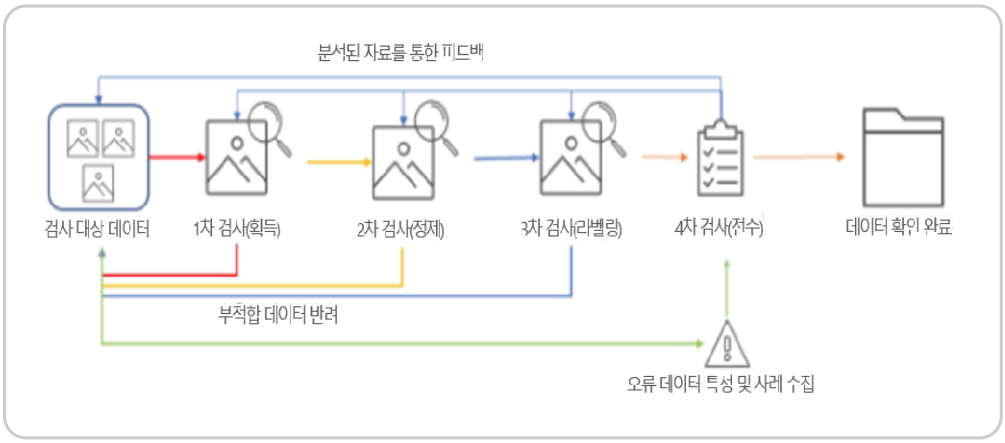


그림 5-1. 데이터 검사 절차 정의

- 검사 규모
  - 데이터 구축 설계 단계에서 구축될 데이터에 대한 품질 수준을 미리 정의하고, 품질 검사를 위한 검사 규모 및 방법*을 설정한다.
    ※ 전수 검사, 샘플링(00%), 다단계 샘플링 등
  - 전수 검사가 아닌 샘플링 방법으로 데이터를 검사할 경우, 검사 대상 데이터가 편향되지 않으면서 무작위로 추출될 수 있도록 한다.
    ※ 각 클래스별 동일한 비율로 추출되도록 함(층화추출)
    ※ 데이터가 구축된 순서 등이 특정 타이밍에 집중되지 않도록 함(파일명 정렬 후 무작위 추출 방법 등 적용)

5.2 검사 방식

- 검사 항목 정의
  - 구축 공정(획득, 정제, 라벨링)별로 공통적으로 적용할 수 있는 검사 요구사항을 고려하여 검사 항목을 정의할 수 있다.
  - 검사 항목은 데이터 및 절차 측면에서 적합성·정확성·유효성, 준비성·완전성·유용성 지표를 측정할 수 있도록 한다.
    ※ 자세한 절차 및 내용은 'Ⅳ. 품질검사 방법' 내용을 참고하며 아래 표 내용 참고 가능

표 5-1. 구축 공정별 주요 검사 항목

검사 절차	검사 항목	요구사항
1차 검사 (획득)	법·제도 준수	원시데이터 획득 시 관련 법·제도적 규정 등을 반드시 준수해야 함
	사실적인 획득 환경 구성	원시데이터를 인위적인 환경과 조건 하에 획득해야 하는 경우 사실적인 획득 환경을 구성하여야함
	데이터 동기화	다중 데이터 소스 간 정교한 동기화를 위한 절차를 마련하여야 함
	편향성 방지	데이터 편향을 방지하기 위한 절차를 마련하여야함
2차 검사 (정제)	정제 기준의 명확성	데이터 사용 목적에 적합한 정제 기준 수립 여부
	중복성 방지	데이터 정제 후 정보 비교 후 중복도 여부
	정제 작업 매뉴얼	정제 작업을 위한 매뉴얼 작성 및 관리 여부

검사 절차	검사 항목	요구사항
2차 검사 (정제)	정제 도구	정제 작업에 사용될 SW 도구를 확보 및 사용 방법을 숙지
	정제 작업 방식	데이터 특성 및 활용 목적에 맞는 적절한 정제 방식 선정 여부 및 선정 기준 타당성 여부
3차 검사 (라벨링)	라벨링 가이드	목적에 맞게 작성된 라벨링 가이드에 대한 타당성 여부를 검사 후 라벨링 작업자들에게 내용 가이드 전달
	어노테이션 항목	목적에 맞는 어노테이션 구성인지 여부를 검사 후 확인된 내용을 포함하도록 작업자들에게 전달
	라벨링 검사 도구	자동화 도구를 통해 검사 후 검사자가 육안으로 부적합 데이터 여부 2차 확인과 촬영된 영상(동적/정적) 이미지의 누락, 번짐 및 조건 오류를 전수 검사
4차 검사 (전수)	부적합 판정 데이터 분포 확인	데이터의 오류율, 특성 분포 확인을 통한 데이터 수집, 정제, 라벨링, 부문 최적화
	외부 검사자	외부 검사자(TTA 등), 도메인 전문가, 데이터 요청자

- 점검 기준 및 점검표 작성
  - 데이터를 일관된 기준으로 검사하기 위해, 데이터 정확성 및 구축 취지에 부합할 수 있는 참값(ground truth)을 정의하고 이 참값을 기준으로 검사 항목 및 채점 기준(통과 기준)을 정의한다.
  - 검사항목 및 채점 기준(통과 기준)을 검사자가 쉽게 확인하고 적용할 수 있도록 체크리스트 등의 형태로 작성하여 배포한다.

【참고】 검사절차 및 기준 수립 예시 (고서 한자 인식)

- 품질검사 수량 : 구축 및 공개 목표량인 1천만 건(자)의 5% 이상 ※:50만 건(자) 이상
- 품질검사 대상 선정 : 구축 완료 후 집계된 책별 글자수를 기준으로 글자수 합계가 50만자 이상이 되도록 책 단위로 무작위 샘플링 실시
- 품질검사 절차
  - 문자인식 정확도 검사 : 샘플링된 책들의 전체 세그먼트 내 오입력 한자 포함 여부를 육안 식별하여 집계

- 세그먼테이션 정확도 검사 : 샘플링된 책들의 전체 세그먼트 내 오류 세그먼트 포함 여부를 육안 식별하여 집계
- 세그먼테이션 유효성 검사 : 샘플링된 책들의 전체 세그먼트에 대한 기계적 검사
- 클러스터링 정확도 검사 : 샘플링된 전체 클러스터에 내 오류 세그먼트 포함 여부를 육안 식별하여 집계함

[절차별 검사방법 및 기준]

데이터	구분		측정지표	검사기준	자동화	검사율	샘플링 단위
문자 인식 데이터	정확도	구조 및 형식	• 문자입력 정확도 • 단위 : 이미지 상에서 세그먼테이션된 한자 낱자 1개와 입력된 디지털 텍스트 1자의 쌍 • 방법 : 자료이미지 상의 총 문자수 대비 정확하게 입력된 문자수의 비율 측정	정확도 99.9% 이상	수동 (육안)	5% 샘플링 (50만자 이상)	책
세그먼테이션 데이터	정확도	구조 및 형식	• 세그먼테이션 정확도 • 단위 : 이미지 상에서 세그먼테이션된 한자 낱자 1개 • 방법 : 자료이미지 상의 총 문자수 대비 정확하게 세그먼트된 문자수의 비율 측정	정확도 99% 이상	수동 (육안)	5% 샘플링 (50만자 이상)	책
	유효성	한자 객체 검출 정확도	• mAP	mAP 0.64 이상 (Precision 80%, Recall 80% 이상)	자동	5% 샘플링 (50만자 이상)	책
클러스터링 데이터	정확도	구조 및 형식	• 클러스터링 정확도 • 방법 : 자료이미지 상의 총 문자수 대비 정확하게 클러스터링 된 문자수의 비율 측정	정확도 99% 이상	수동 (육안)	5% 샘플링 (50만자 이상)	클러스터 생성 단위

- 검사 도구
  - 검사자가 데이터를 빠르게 확인하고 검사할 수 있는 도구를 마련한다. 검사 결과를 작업자에게 즉각적으로 피드백하려는 경우에는 데이터 라벨링 도구와 연계한 시스템으로 검사 도구를 구축할 수 있다.
  - 검사 내용이 단순·반복적인 경우 검사 항목 일부의 자동화 처리를 고려한다.

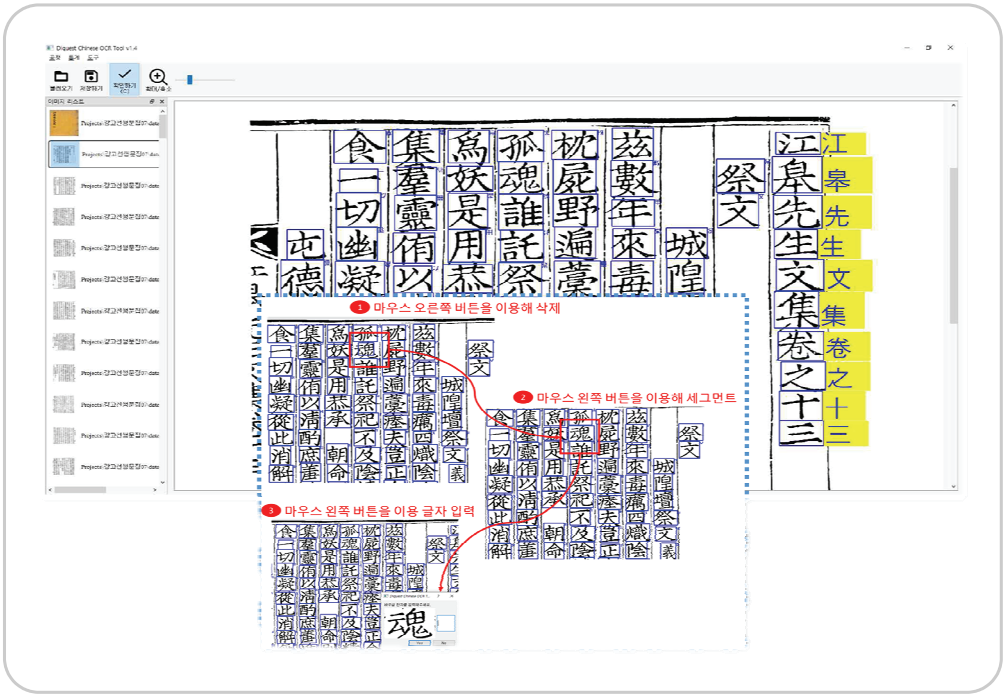


그림 5-2. 데이터 검사 플랫폼 구축 예시(고서 한자 인식)

5.3 검사 결과

- 결과 피드백
  - 각 검사 단계별로 검사자가 검사한 결과를 작업자 및 관리자, 고객에게 피드백하는 체계를 마련한다.
  - 검사결과 집계, 검사내용 확인, 의견사항 전달, 작업자·검사자 재배정, 이슈사항 공유, 피드백 시기 및 주기 등에 대한 절차 및 방법을 수립하여, 검사 현황 및 결과가 빠르게 공유되고 검사를 통과하지 못한 데이터에 대한 재 라벨링이 원활하게 될 수 있도록 한다.

【참고】 검사결과 피드백 프로세스 정의 예시

1. 검사 후의 피드백 프로세스는 다음과 같음
   - 검사 후 불합격 판정 → 검사 의견 제시 (라벨링 결과 문제점을 검사자가 기록) → 라벨링 작업자에게 재전달 → 2차 라벨링 → 2차 검사 → 불합격의 경우 재 라벨링 요청 (검사 의견 기록) → 재 라벨링 (3차 라벨링) → 3차 검사

2. 3차 검사 결과가 불합격인 경우 1) 문서가 매우 어려워 작업자가 라벨링할 수 없음, 2) 작업자의 역량이 해당 원천데이터를 라벨링할 능력이 없음, 3) 원천데이터의 형태가 라벨링이 불가능함 등의 사유로 라벨링 불능 데이터로 판정하고, 3차 검사 불합격 데이터를 별도 저장하여 향후 다른 작업자를 통해 재 라벨할 예정
3. 라벨링시 정확도를 높이기 위해 검사 합격된 문서에만 보상을 명문화하였음

4. 크라우드소싱의 특성상 작업자에 대한 교육이 어려운 관계로, 검사시 검사의견을 최대한 명확하고 구체적으로 전달

- 검사결과 처리
  - 미검사/검사 여부, 검사 통과/미통과 여부를 구분할 수 있도록 파일 관리 체계, 메타데이터 구조 등을 설계한다.

- 목표 데이터, 원천데이터 대비 검사를 통과한 학습 데이터 현황을 모니터링하는 체계를 구성하여, 일정 내 학습 데이터를 확보할 수 있도록 한다.
- 검사결과 및 피드백 사항을 데이터 제작 도구 및 시스템 상에서 어떻게 구현할 것인지 설계한다.

※ 데이터별 검사이력 보관방법, 데이터 전송, 작업자 대상 피드백 전달 방법 등

【참고】 데이터 검사결과 저장을 위한 메타데이터 구조 예시례

- Amazon Sagemaker Groundtruth의 바운딩 박스 어노테이션에 대한 검증(verification) 메타데이터 항목

```
"verify-bounding-box":"1",
  "verify-bounding-box-metadata":
  {
     "class-name": "bad",
     "confidence": 0.93,
     "type": "groundtruth/label-verification",
     "job-name": "verify-bounding-boxes",
     "human-annotated": "yes",
     "creation-date": "2018-11-20T22:18:13.527256",
     "worker-feedback": [
        {"comment": "The bounding box on the bird is too wide on the right side."},
        {"comment": "The bird on the upper right is not labeled."}
     ]
  }
```

참 고 자 료

1. ㈜누리아이디티, 인공지능 데이터 구축·활용 가이드라인 – 고서 한자 인식
2. ㈜데이터연구소, 간판 OCR 이미지 라벨링 가이드
3. ㈜데이터연구소, 위성영상 데이터 라벨링 가이드
4. ㈜비플라이소프트·㈜위고·㈜테스트웍스·고려대학교·주식회사 에이아이닷엠, 문서 요약 텍스트 구축 가이드라인
5. 셀렉트스타 주식회사, 바운딩박스 가공 기준 정의서
6. 셀렉트스타 주식회사, 이미지 수집 기준 정의서
7. 셀렉트스타 주식회사, 한국어 글자체 이미지 AI 데이터 구축 가이드라인
8. ㈜슈퍼브에이아이, 데이터 라벨링 사업운영 A to Z
9. ㈜슬릭코퍼레이션, 서울대학교, ㈜위힐드, 인공지능 데이터 구축·활용 가이드라인 – 피트니스 자세 이미지 데이터
10. ㈜지디에스컨설팅그룹, 인공지능 데이터 구축·활용 가이드라인 – 도로환경 파노라마 이미지 AI데이터
11. ㈜포티투마루, 관광분야 지식베이스 구축 절차서
12. 한국정보통신기술협회(TTA), AI 학습용 데이터 구축사업 공통기준
13. 한국정보통신기술협회(TTA), 2020년 인공지능 학습용 데이터 구축 사업(1차) 중간산출물 20종 검토
14. 한국지능정보사회진흥원(NIA), 2020년 인공지능 학습용 데이터 구축사업(2차) 가이드라인 참조
15. Amazon, AWS Sagemaker Groundtruth 도큐멘테이션 (https://docs.aws.amazon.com/sagemaker/latest/dg/sms.html)

Ⅳ 영상(동적/정적 이미지) 데이터

목 차

표 목 차

그림 목 차

제1장 | 개 요

1 작성 배경

- 인공지능 학습용 데이터에서 수집 · 생산 · 획득되는 다양한 영상(동적/정적) 이미지 데이터에 대한 구체적인 획득 및 수집 방법과 기준, 절차, 방향성 등을 포함한 사례와 방안을 작성하여 누구나 손쉽게 인공지능을 위한 데이터를 확보하는 과정의 참고서가 필요함

2 작성 목적

- 인공지능 학습용 데이터 구축사업에서 생산되는 학습데이터의 영상(동적/정적) 이미지를 바탕으로 목적에 맞는 원시데이터 획득 방법에 대해 인공지능 데이터로서 정확하게 적용이 가능하도록 기준을 마련함
- 2021년 인공지능 학습용 데이터 구축사업에서 데이터 영상(동적/정적) 이미지 획득 방법에 적용하는 것을 목적으로 신규 인공지능 학습용 데이터 구축사업 시 사업계획서에 사전 배포하여 사용하고자 함
- 인공지능 기술 분야에 대한 맞춤형 영상(동적/정적) 이미지 데이터 수집, 획득을 통해 신뢰성 있는 결과 확보

- 영상(동적/정적) 이미지 데이터에 대한 설명과 예시를 통해 데이터를 체계적으로 확보하고 활용도를 높일 수 있도록 필요한 절차와 요구사항 등을 정의하고 데이터의 획득 목적에 맞는 방법을 제시
- 인공지능 학습용 데이터에서 학습에 가장 중요한 요소인 고품질 학습데이터 획득을 통한 양질의 학습데이터를 다량으로 확보하는 방법을 제시하고 획득에 필요한 영상(동적/정적) 이미지 촬영 방식에 대한 구축방식과 방법을 제공하여 범용적으로 활용할 수 있는 방안 마련

3 작성 범위

- '20년 추경(2차) 인공지능 학습용 데이터 구축사업 8개 영역 48개 분야별 150종 인공지능 학습용 영상(동적/정적) 이미지 데이터 획득 공통 참조기준 가이드라인과 기존 TTA 인공지능 학습용 데이터 구축공정 가이드라인을 분석 검토하여 최소한의 기준으로 공통 참조 활용 가능한 기준을 수립하여 도출
- 해외 인공지능 주요 사례에서 확인 후 수집 방법에 대한 방법 참조
- 국내 학습데이터 관련 인공지능 기업 사례 조사 등 다양한 기업들의 선행 기법들에 대해 확인 후 가능 방법 적용

표 1-1. 획득 작성 범위

획득 자료	획득 영역	획득 분류
데이터 획득 분석	'20년 추경(2차) 인공지능 학습데이터 구축사업 8개 영역 및 48개 분야별 150종	헬스케어, 자율주행, 농축산, 국토환경, 미디어, 안전, 기타, 지역 자유
데이터 획득 방식	디지털카메라, 스마트폰, 드론, CCTV, 특수촬영, 특정 분야 및 영역 분류	150개 데이터 유형에서 영상(동적/정적), 이미지 부분을 추출하여 획득

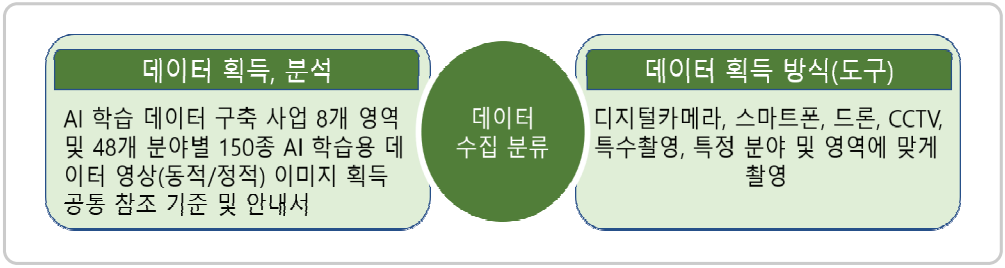


그림 1-1. 데이터 획득 분류

- TTA에서 제공한 가이드라인에 정의된 데이터 획득 → 정제 → 라벨링 → 저장 → 적용의 전주기 프로세스 단계를 인공지능 학습용 데이터셋 구축 목적을 확인하는 내용으로 활용

그림 1-2. 인공지능 학습용 영상(동적/정적) 이미지 데이터셋 구축 단계

표 1-2. 8개 분야 150종 영상 분야 획득 범위 확인

유형	자연어	헬스케어	자율주행	농축산업	국토환경	미디어	안전	기타	지역	자유	합계
오디오	17	-	-	-	-	-	-	-	-	1	18
텍스트	12	-	-	-	-	-	-	1	-	-	13
이미지	3	1	14	6	7	-	6	4	-	8	49
영상	-	14	-	6	-	11	10	2	1	8	52
멀티모달	-	3	-	-	-	-	-	-	1	3	7
정량수치	-	2	-	-	2	-	-	1	4	1	10

- 수집 데이터 대상 획득 범위와 해당 유형 분석 영상(동적/정적) 이미지를 통한 수집 방법 확인

- 전체 영상(동적/정적) 이미지의 내용을 담고 있는 종은 총 101종, 해당 사업계획서 및 가이드라인을 전체 확인한 후, 텍스트 영역의 이미지와 헬스케어 이미지, 자율주행 등 일부영역은 본 내용에서 목적과 내용에 대해 참고만 하고 본문에서 제외하고 작성

  ※ 특수 영상에 해당하는 헬스케어와 열화상 등은 이미지 자체의 해상도와 내용 등에 대해 공통 기준안을 포함하기 어려움이 존재하여 추후 추가 예정임

4 용어 정의

- 데이터 획득 (Data Acquisition)
  - 인공지능의 기계학습에 필요한 데이터를 현실 세계에서 직접 수집 또는 생성하거나, 이미 보유하고 있는 조직이나 시스템 등으로부터 법률적 제약이 없도록 '원시데이터'를 확보하는 활동

- 데이터 정제 (Data Refinement)
  - 획득한 원시데이터를 기계학습에 필요한 형식으로 맞추거나 불필요한 중복을 제거하며, 개인정보를 비식별화하여 처리하는 등 일련의 전처리 과정을 통해 '원천데이터'를 확보하는 활동

- 데이터 라벨링 (Data Labeling)
  - 인공지능이 기계학습에 활용할 수 있도록 기능이나 목적에 부합하는 정보를 원천데이터에 부착하는 활동

- 라벨링데이터 (Labeled Data)
  - 원천데이터에 부여한 '참값', 파일형식이나 해상도 등의 속성, 그리고 설명이나 주석 등이 포함된 '어노테이션'의 집합

- 원시데이터 (Raw Data)
  - 기계학습을 목적으로 획득 단계에서 수집 또는 생성한 음성, 이미지, 영상, 텍스트 등의 데이터

- 원천데이터 (Source Data, Unlabeled Data)
  - 원시데이터를 라벨링 공정에 투입하기 위해 필요한 전처리 등 정제 작업을 수행한 데이터로 라벨링데이터가 부여되지 않은 상태의 데이터

- 인공지능 학습용 데이터 구축
  - 임무정의, 데이터 획득, 데이터 정제, 데이터 라벨링 등 인공지능 학습용 데이터를 구축하는 일련의 활동

- 참값 (Ground Truth)
  - 인공지능의 기계학습 목적에 따라 원천데이터에 라벨링된 정확한 값이나 사실의 의미적 표현

- 어노테이션 (Annotation)
  - 데이터 라벨링 시 원천데이터에 주석을 표시하는 작업을 의미하며, 추가 부착되는 설명정보 데이터는 기능 목적에 따라 다양한 형태로 표현될 수 있으며 이러한 설명정보 표현방식을 지칭

    ※ 용어사용 예 : 사물 바운딩박스 어노테이션, 클래스 라벨링 어노테이션 등

- 광학문자인식 (OCR, Optical Character Recognition)
  - 사람이 쓰거나 기계로 인쇄한 문자의 영상을 기계가 읽을 수 있는 문자로 변환하는 것

    ※ 자세한 용어 정의는 '인공지능 학습용 데이터 품질관리 가이드라인 Ⅴ.부록-1 용어정의'를 참조

제2장 | 구축 가이드라인 작성 방법

1 데이터 구축 목적 정의

- 목적을 정의하고 진행하는 것은 인공지능 기술 분야에 대한 적합한 영상(동적/정적) 이미지 데이터 획득을 통해 신뢰성 있는 결과를 확보하는 목적을 정의함을 의미하므로 전체 공정에서 매우 중요한 요소임
- 구축되는 데이터의 기록을 쉽게 할 수 있어야 하고 이를 위해 명확한 지침 필요
- 데이터에 대한 명확한 운영 정의, 획득 방법, 기간, 일정, 규모(Sample Size)등을 포함
- 데이터의 저장, 기록이나 해석에서 오류의 가능성이 없도록 명확한 단어 및 어휘체계 사용
- 구축단계 절차상 향후 서비스 모델에 따른 분석, 참조 및 재확인을 위한 추가 정보를 포함
- 체크시트 혹은 데이터 확인요청서는 전문성과 객관성을 두고 작성해야 하며 누구나 쉽게 알아볼 수 있도록 일목요연하게 작성 필요

【참고】 농업 영상 데이터 구축 목적 정의 예시

※ 농업 영상 인공지능 데이터 구축 목적 정의(드론 촬영)

- 목적 : 드론을 통한 작물 재배 현황과 생산량등을 유추하여 농사 효율성을 제공하기 위한 데이터 구축 필요(인공지능에서 활용 가능한 필요 데이터 제시)
- 데이터 구축 프로세스 정의 : 데이터 획득 → 정제 → 라벨링 → 검사 단계 (각각의 단계 세부사항 서술 필요)
- 획득 방법 : 대상 영역 사전 답사 〉 대상지 선정 〉 촬영허가 〉 비행코스 설계, 기상 확인 〉 기체 및 센서 점검 〉 촬영 수행(각종 드론 장비 체크리스트 작성 및 기입 필요)
- 규모(Sample Size) : 고추, 콩, 배추, 무, 벼, 고구마, 기타 100만건 이상 구축, 데이터의 편향성을 막기 위해 강원도, 경상도, 충청도, 전라도에서 각각 25만건 이상의 데이터를 구축한다.
  (실제 수집될 데이터가 사전에 유의미한 데이터로서의 가치를 얻기 위해 도메인 전문가, 인공지능 개발자와 사전 협의 후 검토 이후에 진행)
- 제약 조건 : 사업 기간 및 제한(작물 수확시기 도래)을 최대한 회피하는 범위 내에서 촬영 시기 결정 필요, 간격을 두어 생육단계에 따른 영상 확보 필요, 촬영 시 그림자, 위치 정보, 개인정보 등 사전 제재 항목 검토 필요
- 데이터 검사 : 수집 도구 및 학습데이터로서의 필수요소와 충족 조건에 부합 여부 검사

2 데이터 구축 시 고려사항

- 데이터 종류 및 규모를 선정함
- 데이터 품질 수준을 정의하고, 검사 규모 및 방법에 대해 해당 산업 분야 전문가 검토를 통해 품질 수준을 설정함
- 구축할 데이터를 어떤 도구로 제작할 것인지 촬영 장비와 도구 형식, Log, Check sheet, Computer 등의 각종 촬영 도구를 검토하고 구축할 도구를 선정함
- 데이터 획득 종류, 군집객체, 단일객체, 영상기반, 활용목적에 맞게 획득
- 데이터 수집 전 목적에 맞는 샘플링 획득을 수행하여 샘플의 크기, 빈도, 샘플 선정 방법을 사전 정의, 분석 후 획득
- 데이터 획득 적절성으로 누가, 어디서, 언제, 어떠한 방법으로 획득할지에 대한 여부 확인
- 데이터 사용 목적에 맞는 시간별, 주제별, 비효율성, 인과관계 등을 파악하여 획득
- 샘플링은 작은 양의 데이터를 사용하면서도 좋은 정보를 얻을 수 있어야 함(특히, 샘플 채취 구간에 대해서는 인공지능 모델러와 협의 후 기준을 정하고 진행)
- 구축 진행 중에 발생되는 변경사항 등을 구축설계서에 반영하여 업데이트하고 작업자에게 배포할 수 있는 절차 및 방안을 마련함
- 연구 대상인 모집단 혹은 획득 단계별 프로세스를 대표하도록 샘플링 계획수립

- 모집단 및 획득 단계별 프로세스에 대한 충분한 정보를 얻을 수 있어야 함
- 현실적 문제 중요 항목(적절한 비용, 인력 자원 등)을 포함해서 작성 필요

표 2-1. 데이터 구축 시 고려사항 5W1H 원칙 예시

5W1H	항목	예시
What	• 측정대상 • 획득 시 포함되어야 할 변수들	• 일반인이 대상을 식별할 수 있는 피사체 • 장비별, 객체별, 시간별, 종류별, 사람별, 지역별 검토 (필요시 도메인 전문가, 인공지능 전문가 협의 후 대상 객체를 명확히 함)
When	• 획득 기간 (From, To)	• 2주간(11.14 ~ 11.28), 아침 9:00, 점심 12:00 저녁 18:00 일 3회, 획득 시간 1시간
Where	• 획득장소 / 프로세스	• 충남 대전역 역사 내 대합실 외 동일 공간 3곳으로, 이동과 고정을 병행하여 획득하고 동선에 겹치지 않는 장소를 선정(직접 지정함)하여 획득
Who	• 획득 담당자 / 획득하는 사람	• 00 주식회사 미디어센터 내 류XX • 그 외 크라우드 소싱 인력 20명
How	• 획득 방법, 측정주기, 샘플 크기, • 데이터 양식	• 직접 샘플링 모니터 후, 시간당 1회, 1일 3회, 회당 20개 별도 제작 체크시트 등 확인 후 개수 증감
Why	• 측정 목적 / 기대 결과	• 목적에 맞는 획득 데이터 이해와 프로세스 능력의 파악 / 추세분석

3 영상(동적/정적) 이미지 획득 및 정제 방법

3.1 영상(동적/정적) 이미지 데이터 정의

- 동적 영상(Dynamic image)
  - 역동적으로 움직인다는 뜻으로 연속적인 프레임을 연결하여 만든 스트리밍 이미지로서 일정 시간 동안 연결을 가진 동영상을 의미하며 일정 기간, 시간 동안 지속적으로 영상을 촬영하는 것을 의미. 기본적으로 동적 영상은 주로 동영상(영화), CCTV, 자율주행 등에 사용된 영상 결과물
  - 파일 포맷 : MP4, AVI 형식(압축율과 고해상도에 따른 손실율 확인 필요)

    ※ 용어사용 예 : 동영상, 영상 FPS(Frame per Second), CCTV 영상, 블랙박스 동영상 등

- 정적 이미지(Static image)
  - 시간의 흐름에 따라 지속적인 움직임을 담지 않고 피사체를 고정 촬영하는 것을 의미하며 일반적으로 스틸 이미지(still image), 사진이 이에 해당됨. 또한, 동적 영상에서 초당 생성되는 프레임의 일부를 추출하여 잘라낸 낱장 이미지도 정적 영상에 해당
  - 파일 포맷 : JPG, PNG, TIFF 형식(압축 형식을 사용하므로 손실율 확인 필요)

    ※ 용어사용 예 : 정지영상, 30FPS 중 1 Frame, 스마트폰 사진, 일반 사진, 스틸컷 등

3.2 영상(동적/정적) 이미지 획득 데이터 특성 분석

- 영상(동적/정적) 이미지 데이터 특성 분석
  - 획득 데이터는 목적에 필요한 영상 데이터 정보를 데이터 획득 정보, 획득 방법 등, 획득 단계에서 필요한 요건에 맞게 데이터 특성에 대한 분석 필요

- 특성 분석 시 도메인 전문가나 인공지능 전문가 등 다양한 전문가 참여 시 용이함
- 데이터 획득 방법, 법적 문제 발생 가능 여부 등을 검토하여, 실제로 인공지능 학습용 데이터 구축에 활용 가능 여부를 판단하여 데이터를 선정(사전에 샘플링 획득 후 가능 판단 여부를 각 전문가들과 검토하는 방법도 매우 중요)
- 영상 데이터 획득 시 필요한 메타데이터와 기본 속성 정보는 본 구축 안내서 내 제2장 3.3 - '영상(동적/정적) 이미지 획득 절차 및 항목' 내용을 참고

● 영상(동적/정적) 이미지 데이터 획득 방법

- 인공지능 학습데이터 영상(동적/정적) 이미지 획득은 크게 원천데이터 선정과 획득의 2단계로 나뉨

표 3-1. 데이터 획득 방법

세부 절차	작 업
1. 원시데이터 선정	• 저작권 확인, 국가 보안, 개인정보 등 법률적 확인 • 영상(동적/정적) 이미지 적절성 : 분야, 길이, 분량 • 기술 문제 검토 : 획득 작업 적절성, 난이도 등
2. 원시데이터 획득	• 획득 방법 결정 • 획득 기준 설정 : 영상(동적)길이, 해상도, 분야, 분량 • 메타데이터 결정 • 데이터베이스(DB)화

1) 원시데이터 선정

- 원시데이터를 선정할 때 가장 먼저 데이터 목적에 부합되도록 정함
- 외부에서 획득, 선정된 영상(동적/정적) 이미지인 경우, 데이터가 온라인 혹은 오프라인에서 저작권의 유무를 필히 확인함
- 저작권 계약이 필요한 경우, 원시데이터 소유자와 구축 목적, 범위에 맞게 계약함

- 저작권과 더불어 데이터 수요자와 협의한 영상(동적/정적) 이미지의 길이, 분량을 해당 원시데이터에서 획득할 수 있는지 확인하고 부족할 경우 추가 원시데이터를 선정(분량은 구축 목표량의 1.5배 혹은 필요에 따라 10배 이상 확보되어야 라벨링 단계에서 어려움이 적음)
- 선정 시 원시데이터에 대한 획득기술 및 방법도 검토

※ 예) 획득 촬영 장비와 해당 원시데이터의 제작환경을 통한 결과물일 경우 획득 방법과 구체적인 제작환경 서술 필요

2) 원시데이터 획득

- 원시데이터 선정한 후 데이터 수요자의 확인이 완료되면 이를 획득하는데 형태에 따라 획득 방법을 결정

※ 예) 파일인 경우 데이터 추출 방법으로 획득

- 데이터 수요자와 협의를 통해 획득에 필요한 기준 저장시간을 정의

※ 예) 영상(동적)의 길이(예: 30초), 해상도(예: 1920X1080), 분량(3시간) 등

- 데이터 구축 목적에 따라 분야별 메타데이터를 정의하고 이러한 기준에 맞춰 메타데이터와 함께 획득된 데이터는 정제 전 데이터베이스(DB)에 저장함

※ 필요시 디렉토리별로 관리해서 파일명과 기타 정보를 함께 이름에 명명하도록 사전 규칙이 필요함('그림 3-16. 파일 자료 모음과 보관 예시' 참조)

- 영상(동적/정적) 이미지 데이터 확보 방안
  - 원시데이터는 직접 촬영을 기본으로 하며, 환경조건 및 수집 불가능한 항목에 대한 보완책으로 크롤링, 유튜브, 기관별 보유 데이터, 기존 크라우드 플랫폼 수집을 통한 데이터 등 다양한 사전 수집된 데이터 저장소를 이용하거나 비용을 지불한 후 제공 받음
  - 촬영 데이터는 기본적으로 해상도 만족을 위해 Full HD 장축 1920(Pixel)픽셀 이상 확보

- 획득 방법 및 계획
  - 데이터의 획득 계획 이전 샘플링 획득 기준을 확대하여 더 자세하게 수립

- 데이터의 기록을 쉽게 할 수 있어야 하고 이를 위해 명확한 지침 필요
- 데이터에 대한 명확한 운영 정의, 획득 방법, 기간, 일정, 규모(Sample Size) 등을 포함
- 데이터의 저장, 기록이나 해석에서 오류의 가능성이 없도록 디자인
- 구축단계 절차상 향후 서비스 모델에 따른 분석, 참조 및 재확인을 위한 추가의 정보를 포함
- 체크시트 혹은 데이터 요청서는 전문성과 객관성을 두고 작성해야 하며 누구나 쉽게 알아볼 수 있도록 일목요연하게 작성

3.3 영상(동적/정적) 이미지 획득 절차 및 항목

- 영상(동적/정적) 이미지 데이터 획득 단계 절차

그림 3-1. 영상(동적/정적) 이미지 데이터셋 구축 절차 획득 단계

- 정제, 라벨링 단계에서의 요구사항 우선 획득 후 라벨링 목적에 맞는 영상(동적/정적) 이미지 데이터 획득 환경 구축 절차에 필요한 정보를 수집

- 데이터의 획득 방법에 따른 구조와 데이터 형식, 속성값을 데이터의 기준으로 활용
- 영상(동적/정적) 이미지 데이터가 인공지능이 처리해야 하는 목적에 사용될 수 있는 적합성을 기준으로 사용
- 언어, 사물, 시간, 장소, 언어 특성 등의 모든 특성 정보를 기준으로 사용
- 영상(동적/정적) 이미지 데이터의 신뢰성을 위한 출처를 기준으로 사용
- 영상(동적/정적) 이미지 데이터가 유의미한 결과를 낼 만한 양의 수량인지에 대한 수량을 기준으로 활용
- 농축산, 국토환경, 헬스케어, 의료 등 전문분야의 경우 데이터의 기준을 전문가를 통해 선정
- 이미지 데이터를 획득할 때 가로, 세로 픽셀 수에 대한 밀도 기준으로, 색 깊이는 픽셀 당 비트 수로 표시하는 것을 기준으로 활용
- 영상(동적/정적) 이미지 데이터를 획득할 때 획득한 이미지의 사실성을 기준으로 활용

※ 예) 강아지의 사진을 획득할 때 깨끗한 상태의 강아지 정보만을 습득하면 나중에 인공지능 학습 시 유기견이나 들개 등 관리되지 않은 강아지를 판별할 수 없을 것

표 3-2. 시나리오에 의한 촬영 시 원시데이터 획득 절차 예시

획득 절차	내 용	비 고
촬영 계획	• 촬영 방법과 시나리오별 촬영일정표 작성(직접 생성 작업) • 외부 데이터일 경우 활용할 수 있는 해당 데이터 수집 작업 • 촬영업체 섭외(촬영계획서 일정 검토 및 승인)	촬영 계획서 작성
촬영 장소	• 촬영 장소 선정(획득 환경 구축) • 출입 불가 구역 확인	촬영 장소 선정
촬영 장비	• 촬영준비 확인(카메라, 조명, 목적에 맞는 피사체 확인 등) • 시나리오에 따른 촬영 환경 세팅 (화각, 구도, 화질, 촬영 필요항목 등) • 촬영수행(데이터 획득)	촬영을 위한 장비 설치 및 촬영 시작
데이터 획득	• 촬영한 결과물을 지정된 포맷(4k mp4)에 맞게 파일로 생성 • 저장장치를 이용하여 검사팀에 제출	촬영결과물

표 3-3. 원시데이터 획득 시 확인사항 예시

목적별 필요 정보	획득 방법	획득 환경	촬영결과물
목적별 데이터로 활용하기 위한 최소 요건(데이터 규모, 해상도, 길이, 획득 단계) 등	직접 촬영(사람, 드론, 이동체 등), 외부 획득, 크라우드소싱 등	획득 기간 획득 지역	이미지: JPG, PNG 동영상: MP4, AVI

- 원시데이터 획득 방법

  1) 직접 촬영을 통한 데이터 획득

  - 촬영 전문가, 촬영 경험이 있는 인력을 투입하여 사진 촬영의 전문성을 확보
  - 직접 촬영을 통한 데이터 획득 방법으로 촬영 기법, 촬영 방법, 촬영 조건 방식을 촬영 목적에 적합한 촬영 데이터 확보 기준을 명시함

표 3-4. 직접 촬영을 통한 데이터 획득 방법

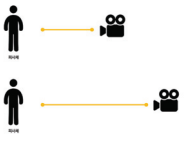
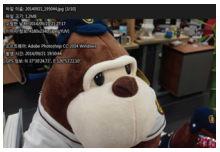
촬영 방법	설 명	비 고
촬영 기법	• 촬영 렌즈 : 마크로 렌즈(접사 촬영), 망원 렌즈(멀리 있는 피사체를 가깝게), 광각렌즈(넓은 앵글) 목적별 촬영 렌즈 제시 • 초점/ 초점 거리 : 피사체가 선명하게 보이도록 촬영 (예: 아웃포커싱) • 프레임 : 1초당 이미지가 몇 장인지 초당 프레임 수 (예: 1초당 30프레임 동영상 표시) • 화이트 밸런스 : 흰색에 대한 균형을 잡아주는 빛의 소스 값 • 해상도 : 가로, 세로 픽셀 수 (예: 1920*1080) 고해상도(FHD이상) 촬영 • 컬러심도 : 24비트 이미지는 RGB 기준으로 함 (예: 24비트 이미지가 제일 많이 쓰임) • ISO값/화질 : 기본 100만 화소 이상의 스틸컷 이미지 • 조리개/셔터속도 목적별 구분 제시 • 가로 촬영을 기준으로 함	촬영 세부사항

촬영 방법	설 명	비 고
촬영 방법	• 촬영 목적에 적합한 장비 사용 (예: 카메라, 스마트폰, 캠코더, 특수 카메라, 삼각대, 렌즈) • 촬영 대상을 명확하게 보여주도록 촬영 • 피사체 전체가 나오도록 정확한 화면 구도와 비율에 맞게 촬영 • 동일한 피사체를 촬영 각도, 촬영 거리 등 다르게 촬영 각도별 수량을 다르게 진행 (예: 360도, 180도, 90도, 45도, 15도로 촬영 각도를 구분하여 각도별 수량 범위를 정하여 수량 일관성 유지) • 일관된 이미지 품질 및 수량 범위를 위해 촬영 장비, 높이, 각도, 거리, 조명 등 기준 명시	촬영자 교육 및 분야별 촬영 진행
촬영 환경	• 실내/실외 촬영 : 목적에 맞는 데이터 획득 환경 구축 후 촬영 • 위치정보(GPS) 활성화 • 플래시를 사용한 촬영 금지 (카메라에서 플래시 없음 설정) 필요시 조명 설치 권고 • 주제에 맞춰 깨끗한 배경에서 촬영 • 촬영 시간대와 다양한 주변 환경에 맞는 촬영 일정 계획을 세워 작업 • 피사체에 따른 환경 세팅값에서 촬영 (예: 피복 지도 촬영 시 장비 드론촬영, 날씨, 계절, 높이, 시간, 각도 등)	촬영 환경 세팅

- 직접 촬영을 통한 영상(동적/정적) 이미지 확보 시 촬영과 함께 환경조건의 다양성 고려하되 공통참조기준을 참고해서 일관성 있는 데이터 획득이 가능하도록 함

가) 영상(동적/정적)촬영 방법

표 3-5. 영상(동적/정적) 이미지 촬영 방법 예시

직접 촬영	촬영 거리	촬영 각도	GPS 위치 정보 포함
			
주제나 소재에 초점을 맞춰 촬영	깨끗한 배경에서 촬영	FHD(1920px × 1080px) 이상의 해상도 지원	가로 촬영
			

나) 영상(동적/정적)이미지 촬영 각도

- 데이터 획득의 정확도를 높이기 위해 이미지 획득 시 촬영 각도 참조기준에 따라 촬영

* 촬영 환경에 따라 촬영 각도 외 특정 각도 가능

※ 촬영 각도

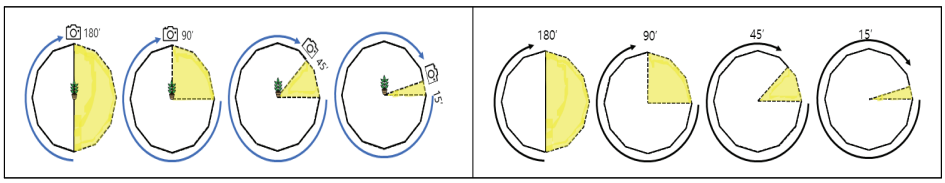


그림 3-2. 촬영 각도 예시

- 촬영 데이터 각 건당 최소 8장 이상 촬영 : 원거리 1장 필수, 정면, 위, 아래, 각 15도, 45도, 90도, 180도 등 각도를 달리한 이미지 데이터

다) 영상(동적/정적) 이미지 화면 구성

- 화면비율 : HD가 디지털 영상의 표준으로 자리 잡았기 때문에 4:3보다는 16:9 화면이 일반적임
- 화면을 좌, 우 또는 위, 아래로 나누었을 때 어느 쪽에 피사체를 위치시키는지에 따라서 균형이 유지되거나 깨지기 때문에 균형 있는 이미지 획득 필요
- 사진이 규정된 촬영 조건과 촬영 방식을 준수하였는지 확인

라) 황금비율 영상(동적/정적) 이미지 사진(예시)

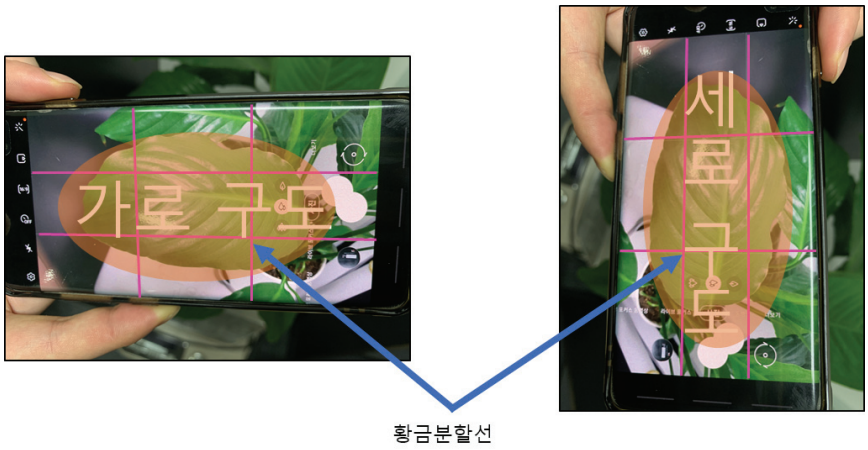


그림 3-3. 황금비율 촬영 기법 예시

2) 웹 크롤링과 웹 검색을 통한 데이터 획득

- 자체 보유 크롤링 Tool 활용과 웹 검색을 통하여 데이터 획득하는 경우
- 획득기술, 조건, 내용, 다운로드, 미리보기 기타 등 다양한 방식을 이용한 획득 가능
  여부 검토 필요하고, 지적재산권과 법률적 문제가 없도록 철저하게 확인 해야함.

가) 올바른 예시와 잘못된 예시를 구분하여 작업자에게 올바른 촬영을 제시

○ (올바른 촬영 예시)

× (잘못된 촬영 예시)

그림 3-4. 영상(동적/정적) 이미지 획득 데이터의 올바른 촬영/잘못된 촬영 예시 첨부

- 목적에 맞는 획득 데이터가 아닌 해당 과정에서 관련 없는 항목이 찍힌 사진, 공통참조기준 내용에서 어긋나는 사진 등 품질의 일관성 유지를 위해 촬영 매뉴얼을 제작하여 배포 필요

나) 웹 크롤링 도구(예시)

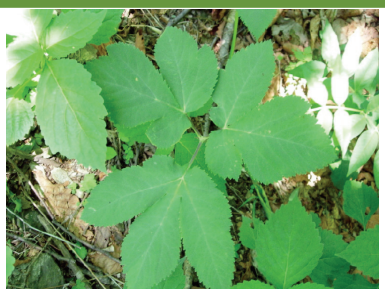
※ 웹 크롤링 타게팅 대상은 사용권이 해결된 한국문화정보원, 한국저작권위원회, 특허청, 국가기록원 등 공공포털내의 이미지로 설정하고 관련 기관과 협의하여 진행

그림 3-5. 웹 크롤링 이미지 예시

- 데이터 획득 시 식별 필요 사례

1) 사례 1 - 획득 단계부터 식별해서 약용/독초 이미지 획득 필요
- 해당 데이터는 획득 시부터 데이터의 특징과 선별 분류 작업이 필요한 과정임. 따라서 현재 일반적 획득 데이터와는 다른 획득 형식을 가지고 있음
- 올바른 획득 데이터 이미지 판별 기준으로 약초/독초 식별 기준 수립(잎, 뿌리, 열매, 줄기, 꽃, 씨)등 식별 가능한 데이터 유무 확인 필요
- 획득 계절 요인과 환경 요인 등 다양한 변수를 검토 해야함
- 독초의 경우 1%의 오류라도 복용 시 생명 위협이 존재하므로 육안으로 명확한 식별이 어려울 시 판단 유보 기능 등이 필요

표 3-6. 데이터 획득 시 식별 필요 사례(약용/독초 이미지)

① 당귀(참당귀)	② 지리강활
	
	
○ (잎·뿌리/약초)	× (잎·뿌리/독초)

※ 당귀(참당귀)로 그 뿌리는 보혈작용이 뛰어난 한약재로 사용

③ 당귀	④ 지리강활
	
○ 꽃잎 하얀색(뿌리/약초)	× 꽃잎 하얀색(뿌리/독초)

※ 왜당귀(신선초)와 지리강활이 꽃잎이 비슷하여 잘못 아는 사례 빈번

⑤ 당귀(참당귀)	⑥ 바디나물
	
○ 꽃잎 보라색(뿌리/약초)	○ 꽃잎 보라색(뿌리/약초)

※ 참당귀와 비슷하게 꽃이 보라색을 띠고 있으며 뿌리를 전호(前胡)라는 약재로 사용, 참당귀와는 전혀 다른 호흡기 질환을 치료하는 효능

- 독초와 약초의 판별 기준에 대한 명확한 기준을 도메인 전문가의 자문을 통해 사전에 설정 후 판별 기준에 따른 획득 방식으로 진행
- 당귀는 한의학에서 빈번하게 사용되는 약초이나 독성이 있는 지리강활과 매우 흡사하여 당귀로 잘못 알고 음용하여 중독 또는 사망 사례가 빈번하게 발생

- 약초와 독초 영상(동적/정적) 이미지 데이터 획득 목적은 약초와 독초의 오인으로 인한 피해를 예방하기 위해 인공지능 학습에 필요한 데이터 획득
- 획득 장비 : 디지털카메라/스마트폰 단말기/스마트폰 단말기(전용App설치)

2) 사례 2 - 획득 단계부터 식별해서 병충해(병명, 벌레 명 등) 이미지 획득 필요

표 3-7. 데이터 획득 시 식별 필요 사례(병충해 이미지)

① 토마토(정상작물 VS 해충작물)


※ 정상작물과 해충작물 피해 증상 데이터 획득

② 벼룩잎벌레(해충)


※ 생육기별 형태 (좌: 유충, 우: 성충) 영상(동적/정적) 이미지 획득

※ 촬영 시 Zoom-In, Zoom-Out에서 객체의 화면이 [그림 3-3]의 황금비율의 예시와 동일하게 해상도와 색상, 화소 등이 최대한 원판과 흡사하게 표현되어야 하며, 객체의 상이 틀어지거나 변경되지 않도록 확인 필요.(전체 화면의 70% 이상 식별이 용이하도록 채움)

표 3-8. 작물별 주요 발생 해충 데이터와 생육기별 질병해충 영상(동적/정적) 이미지 획득 데이터

• 주요 시설/노지 작물 데이터 • 질병 작물 데이터 • 해충 피해 데이터 • 해충 데이터 • 과수화상병 데이터	 정상작물 데이터 / 작물 충해 데이터 / 작물 질병 데이터 / 해충(알)데이터 / 화상병 데이터
• 촬영 도구 : 고정 / 이동 / 단말기를 사용한 데이터 획득 (정확한 식별을 위한 개수 표기 최소 기준) • 병충해 데이터는 각각 최소 기준(몇 장) 이상 • 과수화상병 데이터는 최소 기준(몇 장) 이상	

- 노지 작물 10종의 다발생 질병에 대한 단계별 표현형 영상(동적/정적) 이미지 데이터 획득/식별 가능한 데이터의 유무 확인 필요
- 정상작물을 시작으로 초기, 중기, 말기로 해충 피해 구분별 영상(동적/정적) 이미지 획득
- 작물 부위별, 감염 정도별, 시기별 다양한 피해 증상에 대한 데이터 확보
- 노지 작물별 발생하는 해충의 생육기별 형태 및 피해 증상 영상 데이터 획득
- 조사대상, 해충 종류, 계절의 환경적 요인, 생육단계, 획득지역, 획득 시간 등 정확한 식별 가능한 수준의 데이터 확보 필요
- 다양한 형태의 학습데이터를 획득하기 위해 거리, 각도, 해상도, 조도 등 활용
- 획득 장비 : 디지털카메라/스마트폰 단말기/스마트폰 단말기(전용App설치)/자동획득 장비 활용

※ 자동획득 장비를 통한 영상(동적/정적) 이미지 획득의 경우 하루 4회 촬영[(오전 2회, 오후 2회) 08시, 11시, 14시, 17시] 촬영 장비, 환경 데이터, 각도, 거리, 대상, 조도, 해상도 촬영 방법 선정 후 목적에 맞게 영상(동적/정적) 이미지 획득

1) 사례 3 - 사람 인체·자세 3D 이미지 식별을 통한 추론 데이터

- 2D 인체 영상을 3D 모델로 변환할 때 자세와 형태를 정확하게 인식하고 추론하기 위한 이미지 데이터

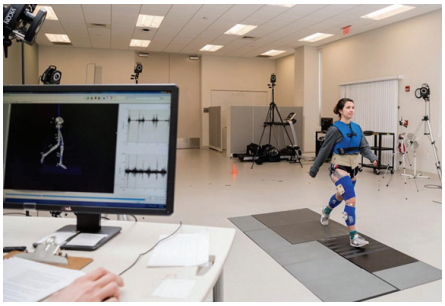
※ 자세 교정, 행동 인식, 이상 행동 감지, 증강현실 등 사람 자세 및 형태 추정 연구 목적

※ 2D 자세 추정을 활용한 산업 중 홈트레이닝, 게임, 가상피팅, 보안, 스마트홈 분야에 사용

※ 3D 형상 데이터를 활용한 산업 중 5G AR컨텐츠, 맞춤의류, 가상피팅 분야에 데이터 사용

- 이미지(monocular)에서 3차원 좌표를 추정하는 Single view 3D pose estimation방식과 다중 카메라 이미지(multi-view)에서 3차원 좌표를 추정하는 Multi-view 3D pose estimation 방식 촬영
- 액션캠을 이용하여 다수의 카메라를 이용해 다양한 앵글에서 대상인 사람을 촬영 데이터 획득
- 촬영된 동영상에서 Skeleton 데이터를 취득

① 2D 인체 영상 촬영



② 3D 모델로 변환

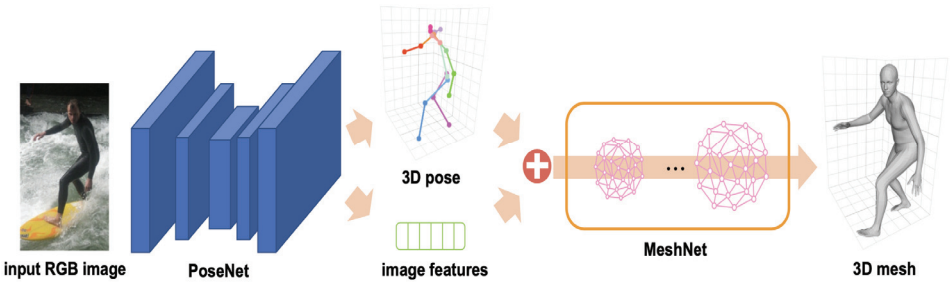
※ 출처 : 구글 이미지 참고

그림 3-6. 데이터 획득 시 식별 필요 사례(사람 인체,자세 3D 이미지)

표 3-9. 데이터 확보 절차

1단계	2단계	3단계	4단계	5단계
영상 데이터 획득	촬영 데이터 분석	영상 데이터에서 포인트 데이터생성	포인트 데이터를 3D 메쉬 형태로 구축	메쉬 데이터를 랩핑 기능을 사용 후 정리 데이터로 변환

※ 랩핑 기능: 3D 메쉬 기능에 표준 모델의 데이터로 대치하는 방식으로 데이터를 최적화하며 이를 위해서 마치 3차원 물체에 랩을 씌운 것처럼 작업하기 때문에 랩핑이라 함



※ 출처 : 사람 인체자세 3D 모델 이미지 참고

그림 3-7. Image features 추출을 통한 3D mesh 추정안 이미지 예시

- 획득한 촬영 데이터로부터 야외 활동 및 실내 일상생활을 하는 사람에 대한 자세와 형태의 입력 이미지 데이터를 추출
- 입력 이미지 : 입력 이미지는 1개+의 HxW(세로x가로) 해상도를 가지는 RGB이미지로, 야외 활동 및 실내 일상생활을 하는 사람 한 명을 담고 있음
- 3D 인간 자세 정보는 위 3D mesh에서 관절의 3D 위치를 따로 복원함으로써 얻을 수 있음

● 가공 및 라벨링 단계에서의 일반적인 데이터 식별

- 가공 및 라벨링데이터 획득의 가치를 정의하기 위해서는 목적에 맞게 데이터를 활용할 수 있도록 획득 계획을 세우고 데이터 활용과 영상(동적/정적) 이미지 획득 단계의 가공 및 라벨링 우선 식별데이터를 획득

① 가공 및 라벨링 단계에서 요구사항 우선 획득 후 라벨링 목적에 맞는 데이터 환경 구축에 필요한 정보 획득

② 영상 촬영 후 업로드한 영상물을 라벨링하고 데이터 라벨링 과정으로 영상 메타데이터 생성 후 목적에 맞는 이미지 추출을 통해 데이터 획득

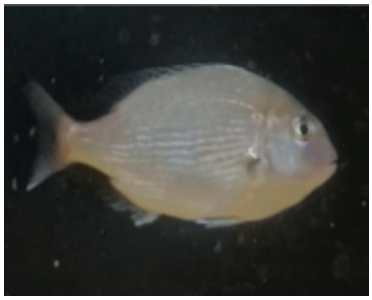
③ 추출된 프레임을 바탕으로 이미지 파일 제작

- 데이터의 획득 방법에 따른 구조와 데이터 형식, 속성값을 데이터의 기준으로 활용
- 학습용 데이터로 적합한 데이터를 선별하는 정제 프로세스를 획득 방법별로 수립
- 작업자가 원하는 데이터를 목적에 맞게 획득하고 도구(소프트웨어)를 활용하여 정해진 규칙에 따라 제외 또는 변환

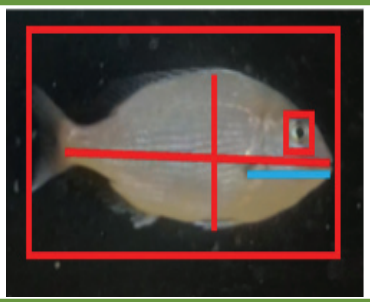
1) 사례 1 - 어류 행동 영상(동적/정적) 이미지 데이터

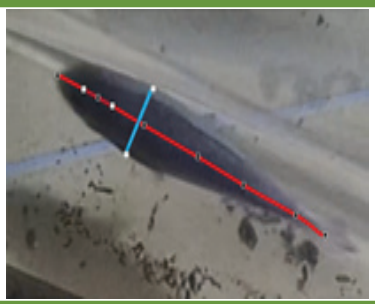
- 어류의 행동은 영상을 찍은 환경과 조건을 가지고 해당 도메인 전문가의 판단에 따라 행동 양식에 맞는 근거를 가지고 해당 영상을 라벨링 진행해야 함
- 제시된 이미지의 어종 원시데이터와 라벨링데이터는 단순 예시이며, 정확한 해법은 아니고 예시로서 다양한 방법과 접근이 필요한 상황임을 표현한 것임

표 3-10. 영상(동적/정적) 이미지 데이터셋 획득 및 구축 기준

① 어종(수평 VS 수직) 원시데이터 : 최대한 흔들림 없고 개체 분류가 가능한 이미지 선정	
	

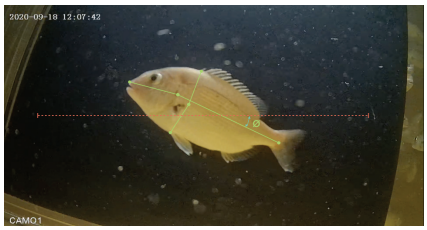
② 어종(수평 VS 수직) 데이터 가공 : 수평과 수직의 시점과 좌표의 시작점이 매우 중요함





※ 출처 : 어류 행동 이미지 참고

- 실제 양식장 내 수조를 활용, 수직, 수평과 수중 CCTV 등 다각도의 카메라 사용
- 각 카메라는 고정 틀에 설치 흔들림 없도록 고정, 동일한 거리에서 촬영된 데이터 획득
- 원시데이터 정제절차는 수직, 수평, 수중 영상(동적/정적) 이미지의 시작과 재생 시간을 동기화하여 한 쌍의 데이터로 묶어서 처리해야 함
- 수직, 수평, 수중 등 모든 동일 시점의 영상에서 동일 간격의 이미지와 상호 교집합에 해당하는 프레임 추출



[수평 어류 행동 이미지]

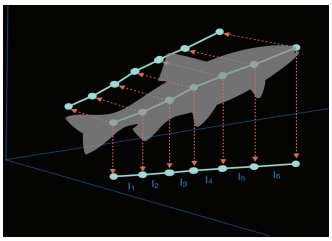
[수직 어류 행동 이미지]

※ 출처 : 어류 행동 이미지 참고

그림 3-8. 수평,수직 어류 행동 이미지 예시

- 위의 영상에서 KeyPoint를 기반으로 3D 매핑(길이 측정의 예시이며 성장 크기에 대한 추론 알고리즘을 포함해야 하므로 단순 그림으로 표현한 예시임)

- 각 포인트 간의 길이를 합하고, 수평과의 각도를 고려하여 길이를 산출



체장(추정) = $(l_1+l_2+l_3+l_4+l_5+l_6)/\cos(\Phi)$
체폭(추정) = W
체중(추정) = weight(*statistics)
***statistics: 어류의 크기, 무게 등에 대한 통계함수**

그림 3-9. 물고기의 KeyPoint를 3D로 매핑한 모습과 길이 추론 방법 제시 내용

2) 사례 2 - 감정인식 요약 영상(동적/정적) 이미지 데이터

- 감정인식 영상(동적/정적) 이미지는 인물(인물들)의 주어진 환경/조건에 따라 감정의 분류가 달라지기 때문에 해당 사항을 고려하기 위해서 주어진 환경/조건을 통해 데이터 라벨링을 우선 수행함
- 영상(동적/정적) 이미지 정보 카테고리에서 영상(동적/정적) 이미지 내 대상의 얼굴 크기가 10%~40% 이하인 FHD의 영상(동적/정적) 이미지 파일로 성별, 연령, 감정(분노, 슬픔, 불안, 상처, 중립, 당황, 기쁨)과 장소/배경(실내, 도시, 자연) 카테고리로 분류한 정제·라벨링 된 JPG 데이터셋

표 3-11. 가공 및 라벨링 우선식별 데이터 획득 기준

항목	상세 내용
데이터 획득 기준	• 구축 목적 인물의(감정인식)에 적절한 데이터를 선별하기 위한 명확한 기준 수립 • 기준 미달 또는 활용 불가능한 데이터를 효과적으로 제거할 수 있는 방법 수립 • 다중 원시데이터의 경우 개별 데이터 획득 후 동기화를 위한 절차 마련
데이터 추출 및 메타 생성	• 주어진 영상에서 초당 1프레임 수준 이미지 추출 및 메타 생성/샘플 수급

그림 3-10. 이미지 데이터 포맷

- 한국인의 얼굴 표정과 상황적 맥락이 담긴 다양한 카테고리의 이미지 파일과 해당 감정 정보가 표시된 감정인식 데이터셋 획득
- 효과적인 데이터 활용을 위하여, 원시데이터 획득 대상을 일반인과 전문인(연극배우 등)을 통해 감정별 안면 이미지 데이터를 획득함
- 인물별 감정 카테고리 이미지 획득의 '정확성' 확보를 위해 촬영 당시 본인의 감정, 장소/배경, 일시 정보를 포함하도록 해서 데이터 획득

● 영상(동적/정적) 이미지 획득 세부 항목 기준

- 필수항목(공통 참조 항목) 획득 기준으로 일부 참조한 형식은 EXIF (EXchangeable Image File Format)
- 메타데이터의 경우 카메라마다 표시 형식이 각기 다르므로 제시된 공통참조기준 항목의 내용을 보고 반영 필요

※ 교환이미지 파일 형식(EXIF, EXchangeable Image File format) 정보는 일본 전자산업진흥협회(JEIDA)가 개발한 이미지 파일의 메타데이터 포맷으로 카메라가 촬영한 사진의 노출값, 촬영 시간, 카메라 메타 정보를 저장하기 위해 개발된 이미지 파일 포맷이며 이미지 파일의 메타데이터 저장 포맷의 표준 지위를 가지고 있음

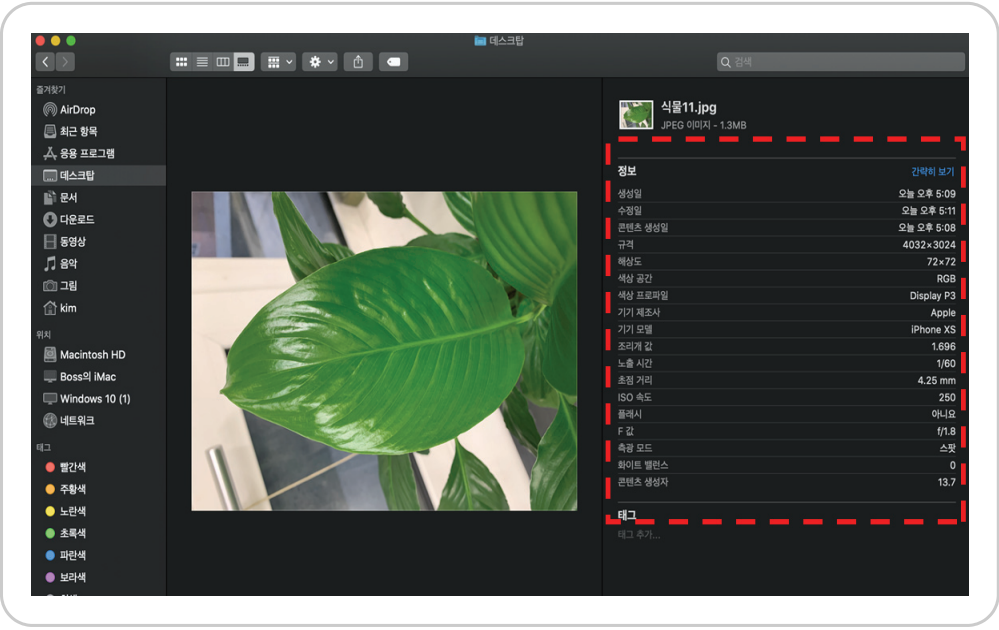


그림 3-11. EXIF 포맷 사진 예시

- 공통참조기준 항목 (디지털 카메라, 스마트폰)

표 3-12. 공통참조기준 항목(디지털 카메라, 스마트폰)

No.	속성명	항목 설명	작성예시
1	identifier/filename	파일명	세글자 이름(예: “DSC_0001.JPG”)
2	date	촬영날짜(년,월), 시간	2020.11.20. 17:08:15
3	file format	파일 형식(포맷)	TIFF/JPG/PNG/MP4/AVI 등
4	imsize	이미지 파일 크기	4800KB
5	images_Photographer	촬영자	촬영한 사람
6	device(camera, lidar)	장비정보	스마트폰, 디지털 카메라, drone, CCTV
7	region_name	촬영 지역명	서울시 종로구
8	images_location	촬영위치	강남구 영동대로 스타벅스
9	copyright	저작권 정보	저작권 정보 첨부 필드 체크

No.	속성명	항목 설명	작성예시
10	Video Clip	촬영시간	2~6분, 20~40분
11	length	영상길이	5분 영상에서 3분10초 부분 사용
12	FPS/Frame Rate	1초/프레임 재생 속도	30fps
13	width, height	이미지 사이즈	이미지 크기 4031*3024
14	Aspect ratio	비율(종횡비)	16:9(동영상)4:3(이미지)/가로세로
15	resolution	해상도	가로X세로 예) FHD(1920X1080)
16	bit	비트값	컬러색상/기본 24bit
17	Pixel	화소	사진정보/색상정보값(이미지픽셀)
18	depth	RGB 여부	색대표 : RGB, sRGB 등/ 비트값과 연관
19	ISO	ISO 감도	밝기에 따른 필름 감도
20	definition	선명도	일반 낮음 높음
21	white balance	화이트 밸런스	K(캘빈)단위 색온도/백열등, 형광등
22	exposure time	노출시간	조리개+셔터스피드 값
23	Exposure mode	노출 모드	자동 노출
24	Metering mode	측광 모드	스팟(중앙 중심) 접사에서 주로 사용
25	F-Stop	조리개 값	f2.8~f11 까지 이미지 밝기 조절
26	flash	플래시	자동 / 플래시 터지지 않음
27	filter	필터	필터 여부
28	focal length	초점 거리	mm초점거리/35mm~50mm(표준)
29	FOV(Field of View)	시야각(화각)	35mm→63도 예)50mm→46도
30	angle	촬영각도	촬영(360도 회전하며 8가지 이상)
31	GPS(Latitude,Longitude)	GPS 정보(위도, 경도)	GPS/GLONASS/37°30'24.7", 126°53'22.1"
32	weather	날씨정보	1)맑음 2)흐림 3)비 4)눈 중 선택

※ 국가 시설물과 개인정보 등 정보보안요소와 위치정보들은 비식별화를 통한 재보정 후 적용 필요

- 획득 선택항목(CCTV) - 공통참조기준 항목을 포함 후 선택항목 추가

표 3-13. 획득 시 선택항목(CCTV)

No.	속성명	항목 설명	작성예시
1	Visible distance	가시거리	최소 10m 이내

※ 국가 시설물과 개인정보 등 정보보안요소와 위치정보들은 비식별화를 통한 재보정 후 적용 필요

- 획득 선택 항목(드론/위성) - 공통참조기준 항목을 포함 후 선택항목 추가

표 3-14. 획득 시 선택항목(드론/위성)

No.	속성명	항목 설명	작성예시
1	Mounted sensor	탑재 센서	1/2.3"유효픽셀수:12M
2	Flight time	비행 시간	23분
3	frequency	송신기 주파수	2.4GHz ISM
4	temperature	온도	- ~ 220℃
5	humidity	습도	0~100%,정확도
6	coordinates	영상좌상단, 후하단좌표	촬영 고도에 따른 지상기준점 설정 값
7	INS	카메라 회전각 정보	X,Y,Z 3축 짐벌
8	speed(hoboring,1m/s,2m/s, 4m/s, 8m/s)	비행속도	X(Pitch),Y(Roll),Z(Yaw) 방향의 전량
9	Range(m/s)	촬영범위	30~300cm
10	altitude	촬영고도	150m~
11	overlap	중복도	GSD(Ground Sampling Distance) 확보
12	Ascent, Descent speed	최대 상승, 하강 속도	5m/s, 3m/s
13	mission	촬영지 분류	산림지, 관광지, 도심지
14	Working temperature	작동 온도	0℃ ~ 40℃

※ 국가 시설물과 개인정보 등 정보보안요소와 위치정보들은 비식별화를 통한 재보정 후 적용 필요

- 특수 카메라 - 본 구축 안내서에서는 제외함

표 3-15. 특수 카메라별 사용 분야

No.	항목명	사용분야
1	CT, MRI, X-ray 등	헬스케어
2	열화상 카메라	피복, 체온, 상하수도, 시설물 등 주변 온도 변화 (지하 하수구 누수)
3	LiDAR (센서)	자율 주행, 3D 객체 인지, 지형 탐색
4	수중 카메라(CCTV)	수중 영상
5	짐벌, 액티브 캠, 초분광, 적외선	특수 촬영 목적으로 활용

3.4 영상(동적/정적) 이미지 획득 데이터 정제 방식

- 데이터 정제 단계 절차

그림 3-12. 영상(동적/정적) 이미지 데이터셋 구축 절차 정제 단계

- 데이터 획득→ 정제 → 라벨링 → 저장 → 적용의 전주기 프로세스에서 정제 단계에서 데이터 라벨링에 필요한 객체 설정 후 인공지능 학습용 데이터에 필요한 기준으로 활용
- 데이터 라벨링에 필요한 객체 설정하고 데이터 라벨링에 필요한 필수정보 확인
- 원시데이터 내의 필수정보 포함 여부 확인 후 라벨링 단계에서 정제 데이터의 품질 검사 요청 후 피드백 수행함
- 정제 작업의 각 절차에 대한 예시를 반드시 포함하여 사용자가 정제절차 전후의 데이터 모습을 확인할 수 있도록 함
- 원시데이터 획득을 통한 정제 오류를 처리하는 절차 등 데이터 획득 및 정제 과정에서 발생할 수 있는 절차를 모두 기술함

● 데이터 정제 방법

1) 유사한 획득 이미지의 중복성 제거를 통한 이미지 확보
- 분류/구별 : 동일한 사진부터 유사한 사진까지 분류하여 중복이미지 제거

○ (유사 이미지 예시) × (중복 이미지 예시)

그림 3-13. 영상(동적/정적) 이미지 정제 데이터의 유사·중복 예시

- 유사 이미지 간 중복성 제거 프로세스

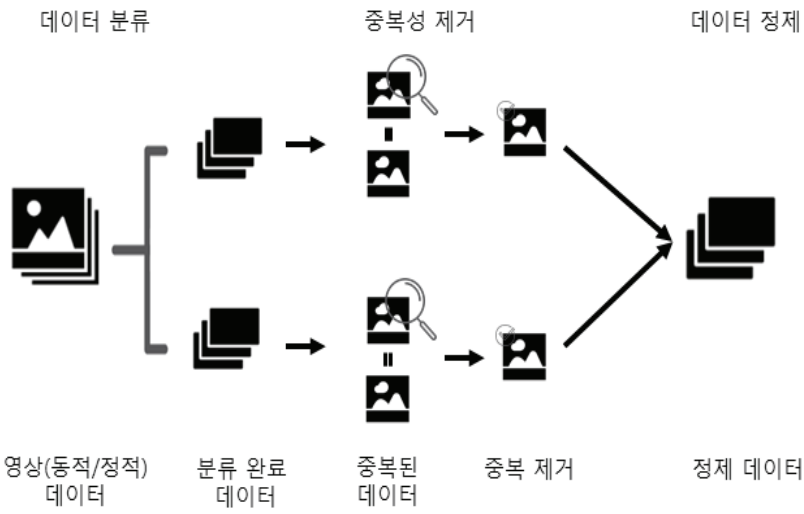


그림 3-14. 유사 이미지 간 중복성 제거 프로세스 예시

2) 개인정보 비식별화

- 비식별화 범위 : 인물 얼굴, 자동차 번호판
- 개인정보에 해당되는 얼굴 부분, 자동차 번호판 흐림 효과(blur)를 통해 비식별화

[인물얼굴 비식별화 이미지 예시] [자동차 번호판 비식별화 이미지 예시]

※ 출처 : 구글 이미지 참고

그림 3-15. 영상(동적/정적) 이미지 정제 데이터 개인정보 비식별화 예시

- 데이터 정제 기준
  - 획득 단계의 각 절차(방법)를 수행할 때, 따라야 할 기준을 제시
    ※ 예) 해상도 설정, 거리, 화각, 구도, 시간, 날짜, 저장 형태 등 작성
  - 원시데이터 정제 시 분류(class) 간의 균형, 분류 내 instance 간의 균형 확보 작성
    ※ 예) 자율 자동차 – 오전 시간대 위주 금지 / 사물 이미지 – 특정 상품을 각도 등을 달리하여 반복 촬영 금지/자연어 – 특정 단말 위주 획득 금지

표 3-16. 데이터 정제 기준항목 예시

기준항목	특 징	비 고
획득 배경	(데이터 획득 배경에 대한 설명)	
데이터 획득	• 목적에 따른 획득 방법 또는 제작 방법 수립 • 촬영전략 및 계획 수립(데이터 범위, 목적, 획득 기간 등) – 획득범위: 목적에 맞는 획득 데이터를 얼마나 많은 대상을 다양하게 촬영할지 – 획득방법 : 야외현장 촬영 , 웹 크롤링 – 획득장비 : 데이터 획득을 위해 필요 적합한 장비(디지털카메라, 크롤링 서버) – 데이터 형식 : 최초 획득 시 데이터 형식 및 정제된 데이터의 형식(파일 확장자 등) RAW→PNG, JPG – 데이터 처리 : 데이터 획득.정제에 사용되는 소프트웨어(라벨링 도구) – 담당 인원 : 획득.정제에 활용할 인원(크라우드소싱 인력 및 크롤링 담당자)	장비 모델명 (000)
데이터 정제	• 데이터 정제를 위한 작업 프로세스 – 불필요한 데이터, 무의미한 데이터 정의 및 제거 방법 – 데이터 정제를 위한 SW 구조 및 GUI 형태 – 변수의 유형 관리 방법(연속형, 정수형, 범주형 등)	
데이터 활용분야	• 연구 분야 – 예) 약초와 독초를 판별하는 가이드라인 구축 • 산업 분야 – 안전 복지: 일반 사용자 기준에서 필요한 정보를 정확하게 구분할 수 있으며 누구나 정확하고 안전하게 정보를 이용할 수 있도록 사고 상황 검출	
데이터 형태	• 동영상 비디오 클립 mp4 포맷 또는 이미지 파일 PNG , JPG • 목적에 맞는 획득 데이터 식별 기준에 따른(독초/약초 구분) 정보 포함	

- 목적에 맞게 일반 사용자, 데이터 사용자의 요구에 부합될 수 있도록 촬영 계획 및 기준을 적용하여 획득

표 3-17. 데이터 정제 방법 예시

항 목	상세 내용
데이터 획득·정제	• 획득 목적에 맞는 영상만 선별 • 공통참조기준 촬영 항목에서 어긋나는 사진 배제 • 어류 행동을 명확히 확인 가능하거나, 추적 대상 객체가 존재하는 경우만 선정 • 화질이 불량하거나(흔들림, 초점 불일지 등), 렌즈 이물질 등으로 잘 보이지 않는 영상의 경우 배제
데이터 추출 및 메타 생성	• 주어진 영상에서 초당 1프레임 수준 이미지 추출 및 메타 생성
환경요소	• 획득 영상(동적/정적) 이미지 시나리오에 맞는 계절, 날씨, 시간대인지 확인
해상도 기준	• 고/중/저의 해상도의 경우 이미지 비율을 유지, 촬영하는 해상도 및 동영상 재생 조건에 맞춰 설정 • FHD(1920px *1080Ppx)파일로 구축 또는 4k를 원칙으로 획득 • 60fps 이상을 원칙으로 획득.
화각, 구도	• 육상 양식장 수조를 활용, 수직, 수평 카메라는 수중 CCTV 영상의 화각 및 구도
저장 형태	• MP4, AVI 포맷 기본 외 (기본적으로 H.264, H.265, MPEG4 코덱 등 일반적인 코덱) 추출

3.5 영상(동적/정적) 이미지 획득 도구 및 정제 도구

- 종별 획득 도구 유형
  - '20년 추경(2차) 인공지능 학습용 데이터 구축사업의 150종을 분석한 결과, 종별 획득 도구의 다양성 확인

    ※ 수집 방식 변경 시 종과 관계없이 가변적임

표 3-18. 데이터 종별 획득 도구 유형

획득도구	자연어	헬스케어	자율주행	미디어	농축수산	국토환경	안전	기타	합계
디지털 카메라/스마트폰			○(18)	○(8)	○(6)	○(2)	○(2)	○(2)	38
위성/드론(초분광영상)			○(2)		○(2)	○(4)	○(1)		9
CCTV							○(6)		6
라이다 센서			○(2)					○(1)	3
수중 카메라					○(2)	○(1)			3
액션캠/고프로							○(3)	○(3)	6
CT, MRI		○(21)							21
열화상 카메라						○(2)			2
합 계		21	22	8	10	9	12	6	88

- 원시데이터 획득 시 촬영 목적에 맞게 용도별 장비를 사용하며 목적에 맞는 장비를 선택하여 촬영하는 것을 원칙으로 함
  - 모든 촬영 장비와 수집 방식은 수집 전 인공지능 개발자, 도메인 전문가, 수집 전문가들과충분한 의견 수렴 및 수집 데이터의 목적성에 대한 실험 결과를 반영한 수집 지시가 있어야 함

  1) 촬영 장비(예시로 제조사 스펙 포함)
  - 획득 장비 : 디지털 카메라 / 스마트폰 단말기 / 스마트폰 단말기(전용App 설치) / 자동획득 장비

  ① 디지털카메라 : Canon 5D Mark3 (추가구성 : 매크로 렌즈 EF100mm f2.8L MACRO IS USM)

  ※ 작은 미소 해충의 명확한 이미지 촬영에 이용

  ② 스마트폰 단말기(전용App설치 포함) : 100만 화소 이상 영상(동적/정적) 이미지 촬영 가능 단말기

  ※ 휴대 편리한 장비로 불특정 환경에서 촬영 가능 / 획득 앱을 통한 효율적인 업로드 가능

③ 자동획득 장비: 블랙박스

※ 문제 해결을 위한 이미지 촬영에 이용

④ CCTV : 주연전자 8MP 3.6mm UHD 820만화소 실외 카메라

※ 교통문제 해결을 위한 CCTV 영상(동적/정적) 이미지 촬영에 이용

※ 도시철도 역사 내 CCTV에서 관측될 수 있는 이상행동 영상 데이터 및 사회적 약자 보호, 범죄예방을 위한 객체 추적 영상 데이터 촬영에 이용

※ 어류 영상 등 특수 상황인 경우 실험체에 대한 실험 이후 선택 권고

⑤ 위성/드론 : ESA Sentinel-2 위성영상(국토지리정보원 제공 항공영상)/ DJI팬텀3 등

※ 국토환경 피복지도 및 산림수종 영상(동적/정적) 이미지 촬영에 이용

※ 자율주행 드론 이동체 인지영상, 제주 월동 작물 자동 탐지 드론

⑥ 열화상카메라 : 하수관로 CCTV 특수 카메라(광각렌즈(wide-angle lens)나 어안렌즈 (fish-eye lens) 촬영장비(90만 화소 이상)/ 특수 정보(영상: 파노라마, 레이저 프로 파일, 3D 스캔, 열화상 등)를 저장

※ 하수관로 내부 이미지 촬영에 이용

※ 한국가스안전공사 가스안전연구원, 국립소방연구원

⑦ 라이다(LiDAR) 센서 : 자율주행 자동차 센서 카메라(눈)으로 인식하고 주변을 파악하는 장비/ Velodyne사의 PUCK 라이다 등

※ 자율주행 도로상태 및 자율버스 주행 중 건물 영상(동적/정적) 이미지 촬영에 이용

2) 촬영 시 주의사항

- 안전 규칙 준수와 저작권 및 초상권 준수

① 교통사고, 낙상사고, 낙석 등 안전사고 주의

② 쓰레기 매립지, 재활용 처리업체, 소각장 등 촬영 시 안전 규칙을 철저히 준수

③ 촬영 시 배경에 사람, 차량 등을 최대한 배제

- 영상(동적/정적) 이미지 데이터 획득 환경 정보
  - 획득/촬영 당시의 주변 환경 정보(온/습도, 조도, 날씨 등)를 같이 기록
  - 날씨 단위는 맑음, 구름, 흐림, 비, 눈 5종으로 한정
  - 획득/촬영 당시의 위치, 지역 정보(예: 다른 피사체와의 비교 영상(동적/정적) 이미지)를 같이 기록
  - 영상(동적/정적) 이미지에 있는 색깔 정보로 판단 또는 판별하는 경우에는 분광 정보(채널별 신호 세기)를 획득할 수 있도록 주의 필요

- 영상(동적/정적) 이미지 형태

  ※ 영상(동적/정적) 이미지 획득 이미지 형태(예시)

표 3-19. 영상(동적/정적) 이미지 획득 형태

농업영상 (잎, 꽃, 가지, 줄기, 과실)	가축(소, 돼지, 닭)	생활폐기물(형태, 재질)
		
[시간, 계절, 기후 등]	[목적에 맞는 이미지 획득 형태]	[명확한 이미지 데이터 형태]

  - 시간, 계절, 기후, 장소, 꽃(성격), 음식, 인물 : 성격에 맞는 촬영 방법으로 영상(동적/정적) 이미지 획득
  - 목적에 맞는 데이터인지 영상(동적/정적) 이미지 확인과 개체별 촬영과 군집 촬영 목적에 근거를 두고 영상(동적/정적) 이미지 형태 획득
  - 형태나 재질의 구분이 명확한 영상(동적/정적) 이미지 데이터 획득
  - 배경 화면이 문제가 될 경우 배경 삭제 및 정제 필요

- 원시데이터 정제 도구
  - 영상(동적/정적) 이미지 유사도 분류 도구 활용
  - 유사 영상(동적/정적) 이미지 간 중복성 제거 프로그램(예:Dup Detector) 실행
  - 개인정보 비식별화 모자이크 처리 프로그램 : 정제 도구 실행(흐림 정도 픽셀값 및 사람 얼굴 인식 여부 확인)

- 데이터 제출 방법

  1) 폴더 정리
  - 폴더명 정확하게 기입
  - 데이터를 경로에 따라 정확하게 업로드
  - 최종 폴더명을 규칙에 맞게 기입(규칙 및 명명은 사전에 인공지능 개발자와 협의 후 진행)
  - 단일 객체 촬영 원칙, 혼합 객체의 경우 대체 객체만 표기

  2) 업로드 방식 및 순서
  - 저장소(웹하드&FTP) 서버로 로그인 후 업로드 → 촬영일과 상관없이 업로드할 때, 사전에 정의하고 구성된 목적 폴더에 업로드
  - 개인 컴퓨터에 아래 그림과 같이 자료 정리 → 내 컴퓨터 내 사진 정리 폴더는 촬영일 기준
  - 저장 → 적용의 전주기 프로세스에서 수집 →획득 단계부터 데이터 획득 기준이 적용
  - 획득 단계부터 타겟 객체의 판단 근거가 독극물인지 약용인지 특징을 명확한 목적에 맞게 작업

3) 개인 컴퓨터에 정리된 영상(동적/정적) 이미지 데이터 파일 자료 모음 예시

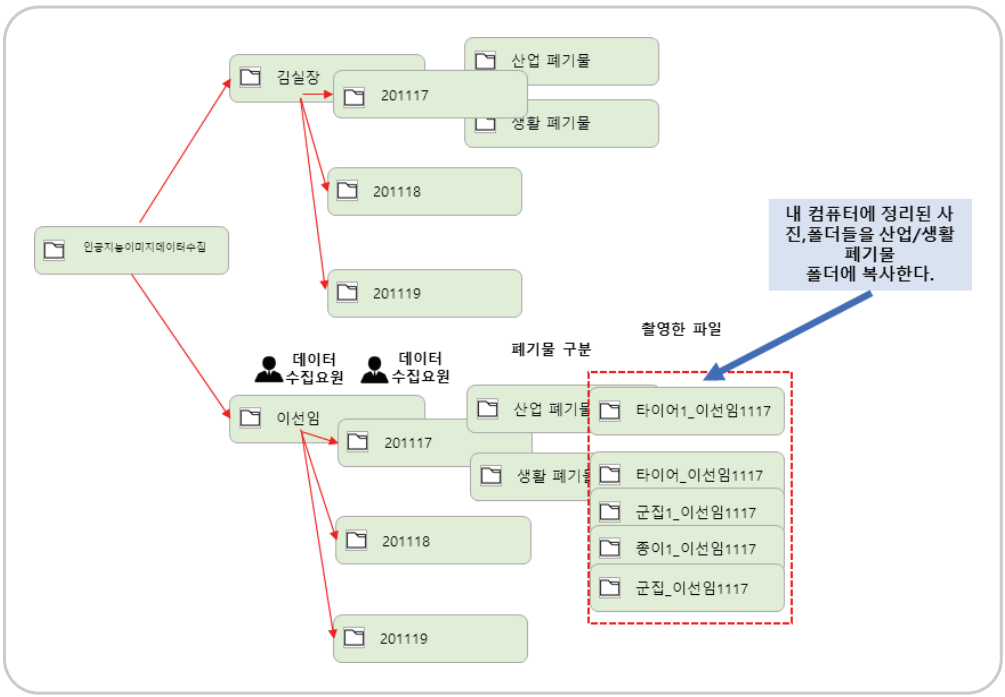


그림 3-16. 파일 자료 모음과 보관 예시

3.6 획득 시 고려사항

- 획득 데이터를 위한 필수사항 고려

표 3-20. 획득 시 고려사항 예시

획득 시 고려사항	설명	비고
획득 가능성	• 획득이 불가능하거나 통제 불가능한 주기를 가지고 있다면 원시 데이터의 정책에 의존하게 되므로 바람직하지 않음 • 획득이 용이하더라도 서비스 활용 측면에서 데이터를 활용하기 위해 가공 처리에 많은 비용이 드는 데이터는 선정하기 어려움 • 직접 산출이 어려운 경우 획득난이도 측면에서 트래픽량과 저장처리 장치의 용량등이 고려 대상, 획득 대상의 대안 필요	가능성 검토
데이터 정확성	• 서비스의 활용목적에 세부 항목이 정확히 존재하는가 검토 필요 • 획득 목적에 맞는 데이터를 획득하기 위해서는 사전, 사후 처리 방안 필요 • 사업계획서 당시 정의한 구축요건에 맞춰 데이터를 획득, 정제함 • 그 밖의 사례를 참고하여 데이터 획득, 정제에 필요한 사항을 가이드에 반영함	정확성 검토
보안사항 개인정보 및 저작권	• 데이터 저작권에 대한 라이센스 확보와 자체 데이터 제작 권장 • 가공된 데이터의 라이선스에 대한 확약 및 협약서 • 획득 데이터에 개인정보 획득 등 보안 사항이 없는지 검토	공개 데이터를 위한 필수사항
데이터 균형	• 개체의 다양성 • 목적 및 상황의 다양성 • 시간별, 종류별, 사람별, 지역별 다양성	활용 분야별 검토
신뢰성	• 데이터 획득대상, 획득방법이 법,제도를 저촉하거나 사회윤리에 어긋나지 않도록 해야함 • 이미지 데이터 획득 시 획득된 데이터의 종류에 따라 검증을 진행하여 데이터에 대한 신뢰성 부여 • 획득된 이미지 데이터의 중복 여부를 검사하고 신뢰할 수 있는 데이터인지를 확인	신뢰성 검토

4 데이터 라벨링 작업

4.1 데이터 특성 식별 분류 체계 및 고려사항

- 데이터 라벨링 시 식별 분류는 인공지능 학습의 품질을 결정하는 핵심 업무로서 특성을 명확하게 규정해 주는 것이 지도 학습에서의 결과물로서의 가치를 높이는 기준이 됨. 이를 위한 고려사항으로 획득 가능성과 활용데이터의 보안 문제, 데이터 정확성 등을 확인하고 고민해서 라벨링 단계를 진행해야 함

- 데이터 특성 식별 분류체계
  - 목적에 필요한 데이터 특성 식별 분류 후 정제 단계에 전달함
    ※ 예) 동의보감 독초와 약초
  - 식별된 특성을 통해 라벨링에 필요한 라벨링 기준 및 어노테이션 속성항목 작성
  - 전처리 진행 중에 추가사항 지속적 업데이트
  - 피쳐링이라고도 하는 특이점 표시로서, 데이터의 특이점이나 구별 혹은 식별이 가능한 영역을 입력하여, 인공지능이 학습해야 하는 개체의 가로X세로 좌표를 픽셀 단위로 읽을 수 있도록 하는 행위를 식별 분류라고 표현함

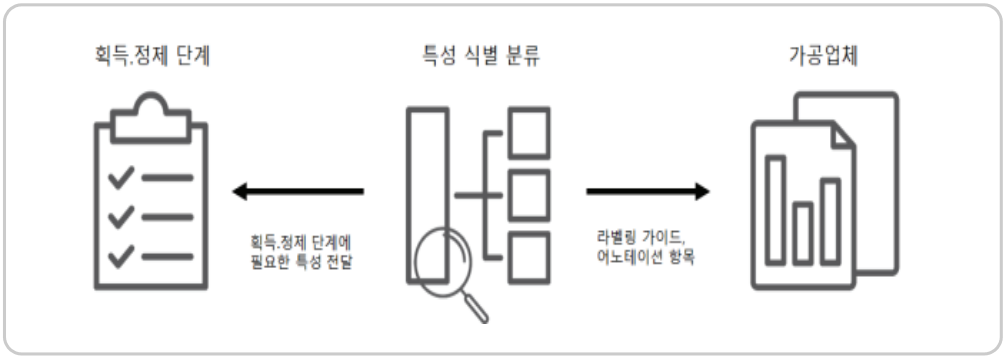


그림 4-1. 데이터 특성 식별 분류 체계 예시

- 데이터 식별 고려사항

1) 획득 가능성
- 데이터 선정 이후 라벨링 전 단계에서 가장 우선적으로 고려해야 할 사항으로, 서비스 활용에 좋은 데이터라 하더라도 획득이 불가능하거나 인공지능으로서 식별이 불가능한 형식을 갖고 있다면 가능성을 배제해야 함
- 획득이 아무리 용이하더라도 서비스 활용 측면에서 데이터를 활용하기 위한 가공 후 처리에 비용이 과다하게 되면 좋은 데이터 선정이라 할 수 없음

2) 보안 문제
- 라벨링 시 획득한 데이터의 개인정보, 저작권 등의 법, 제도적 부분 확인 필요
- 라벨링된 데이터에 대한 개인정보 문제나 저작권에 대한 문제 발생 시 서비스 활용에 대해 심각한 문제가 발생

3) 활용처
- 데이터의 활용 분야, 활용처에 따라 같은 데이터라도 라벨링의 기준이 달라질 수 있음

※ 예: 사람을 감지할 때 인체만 할 것인지 착용된 아이템까지 할 것인지 확인 필요

4) 데이터의 정확성
- 라벨링한 데이터의 정확성은 서비스의 활용목적에 세부 항목이 정확히 존재하는가에 대해 검토해야 함. 획득 목적에 맞는 데이터를 획득하기 위해서는 선 라벨링 과정이 필요하고 획득한 데이터의 사후처리 방안도 마련돼야 함

5) 라벨링 비용
- 라벨링 비용은 데이터를 라벨링하기 위해 직접적으로 들어가는 비용으로, 정량적 기준으로 적용된 라벨링기술에 사용 비용이 과다 발생할 경우 라벨링기술에 대한 검토가 필요

- 영상(동적/정적) 이미지 데이터가 인공지능 데이터로서 적합 여부 확인
  - 영상 데이터가 인공지능에 사용될 수 있는 정보를 포함하고 있는지에 대한 적합성이 매우 중요함
  - 데이터 파일 형식은 특정 획득 장비 및 처리 도구에 종속되지 않으며, 파일 포맷은 보편적으로 통용되는 대표적인 파일 포맷을 활용
  - 구체적인 데이터 포맷 관련 정보는 '부록1. 인공지능 학습용 데이터셋 구축 공통참조기준' 참조

표 4-1. 원천데이터 적합성 예시

원천데이터 구분	데이터 유형	파일 포맷	해상도	Frame Rate	Color bit	촬영장비	비고
획득하는 원천데이터 기준	동영상[1]	mp4[2]	FHD[3]	30 FPS[4]	24bit[5]	4K 지원 카메라	(원천데이터 관련 추가 정보 기술)

1) 데이터 유형은 이미지, 동영상, 오디오, 텍스트, 정량수치, 로그로 구분
2) 파일 포맷은 대중적으로 널리 사용하는 대표적인 파일 포맷 사용
3) 동영상의 경우 가로x세로 픽셀수로 표시하거나 SD(720x576 또는 720x480 또는, 640x480), HD(1280x720), FHD(1920x1080), UHD(3840x2160), 4K(4096x2160), 8K(7680x4320)로 표시
4) 동영상 내 1초당 몇 장의 이미지인지 초당 프레임 수로 표시
5) 픽셀당 색 깊이(depth)를 비트 수로 표시

4.2 데이터 라벨링 방법 및 절차

- 데이터 라벨링 단계 절차

그림 4-2. 영상(동적/정적) 이미지 데이터셋 구축 라벨링 단계

- 데이터 획득 → 정제 → 라벨링 → 저장 → 적용의 전주기 프로세스에서 라벨링 단계를 학습에 맞는 라벨링 작업 및 어노테이션을 설정하고 인공지능 학습용 데이터에 대한 공통참조기준으로 활용
- 데이터 라벨링은 학습에 필요한 어노테이션 정보 설정 및 정제 단계에서 전달함
- 학습데이터는 예측하고자 하는 데이터와 가장 유사 해야함. 또한, 공간상 유사 배경이나 수집된 데이터가 모두 흐리다면 학습 데이터도 흐리게 학습을 해야만 성능이 좋아짐
- 일반적인 상식으로 수집된 데이터에서 약 1000장 정도의 이미지를 가공하여 모델링에 접목시킨 후 모델과 적합성을 확인함
- 가장 정상적인 데이터와 비정상적인 데이터를 라벨링해서 구분 시 정상데이터가 100배에서 120배 정도의 배수를 가지는 것이 학습에 효율적임
- 인공지능 개발에서 일반적으로 사용되는 라벨링 용어 및 분류체계를 준수함

- 라벨링 지원 도구 사용 및 작업 방식에 따른 품질 자체 검사
- 정제 데이터의 품질 검사 및 피드백
- 라벨링 후 데이터 시험과 성능 등 정확성과 적합성 검증을 위한 데이터셋의 비율이 필요하며, 인공지능 개발자와의 협의를 통해 지정
  ※ 예) 학습데이터(80%), 검증데이터(10%), 시험데이터(10%)
- 모든 산출 데이터는 표준화된 포맷을 지원할 수 있도록 CSV, JSON 등으로 묶어서 제공 필요

● 데이터 라벨링 기준
- 절차별 규정과 기준을 제시
  ※ 예) 어노테이션 데이터 포맷 표준화 사용, 보편적으로 활용 가능한 포맷 형태 우선 채택
- 획득→정제가 끝난 중간데이터에서 최종 데이터 형태가 나오기까지의 모든 과정 작성
  ※ 예) 레이블링을 하는 경우 선정한 레이블의 구성, 각 레이블을 부여하는 기준(ground truth)과 방법, 레이블을 부여하는 모습 예시, 애매한 내용이 나올 경우의 처리 기준, 자주 하는 실수의 예시, 라벨링 완료 기준, 라벨링 작업의 검사 데이터 처리방법 제시
- 어노테이션의 구성은 데이터 확장을 고려하여 항목명, 타입, 필수 구분, 항목 설명을 반드시 포함하여 표기함
- 반드시 인공지능 개발자와 협의 후 관리 방식 공유 필요

● 라벨링데이터셋(예시) - 필요시 자체 작성 후 공유 필요

표 4-2. 라벨링데이터셋 속성 정의 예시

No.	속성명	하위 속성 및 내용
1	Dataset.name	데이터셋 명
2	Dataset.date_created	데이터셋 생성일자
3	Dataset.img_path	데이터셋 영상(동적/정적) 이미지 폴더 경로
4	Dataset.label_path	데이터셋 레이블 폴더 경로
5	Dataset.category	데이터셋 카테고리
6	Dataset.type	데이터셋 타입

- 라벨링데이터 식별 방식 유형

표 4-3. 라벨링데이터 식별 방식 유형

데이터 유형	라벨링 기능	어노테이션 방식
정적 영상	• 이미지 분류(Image Classification)	• 클래스 라벨(단일, 다중)
	• 객체 인식(Object Recognition)	• 바운딩 박스(사각형)
	• 영역 구분(Segmentation)	• 바운딩 박스(사각형) • 타원 바운딩 박스 • 폴리곤(다각형) • 브러시(자유형태) • 키포인트
동적 영상	• 동영상 분류(Video Classification)	• 클래스 라벨(단일, 다중)
	• 객체 인식(Object Recognition)	• 바운딩 박스(사각형) • 키 포인트(정점) • 폴리곤(다각형) • 폴리라인(선)
	• 객체 추적(Object Tracking)	

- 2D 어노테이션
  - 2D 영상기반의 객체 주석화 (예: 자동차, 보행자, 표지판 등)

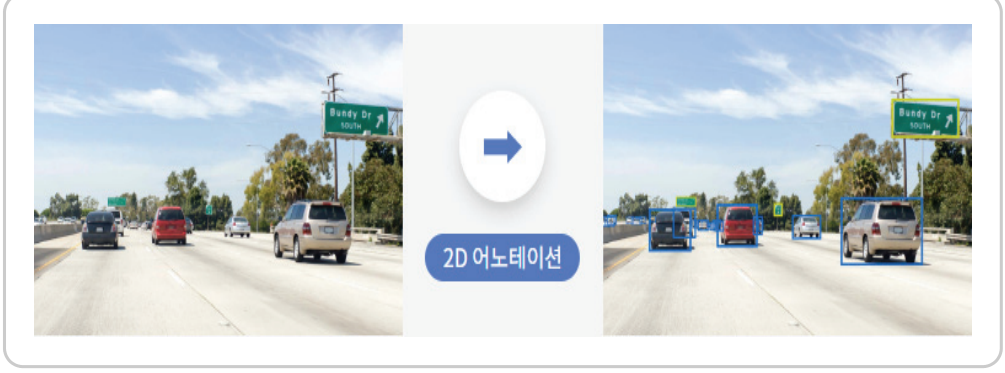


그림 4-3. 2D 영상기반의 객체 주석화 예시

- 3D 어노테이션
  - 3D 정보(LiDAR) 기반의 객체 주석화 (예: 자동차, 보행자 등)

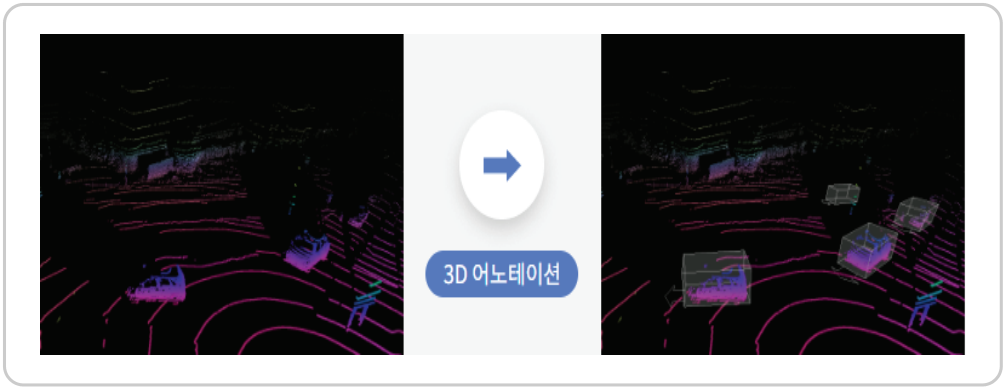


그림 4-4. 3D 영상기반의 객체 주석화 예시

- Segmentation
  - 정밀한 객체 구분을 위한 픽셀 단위 주석화

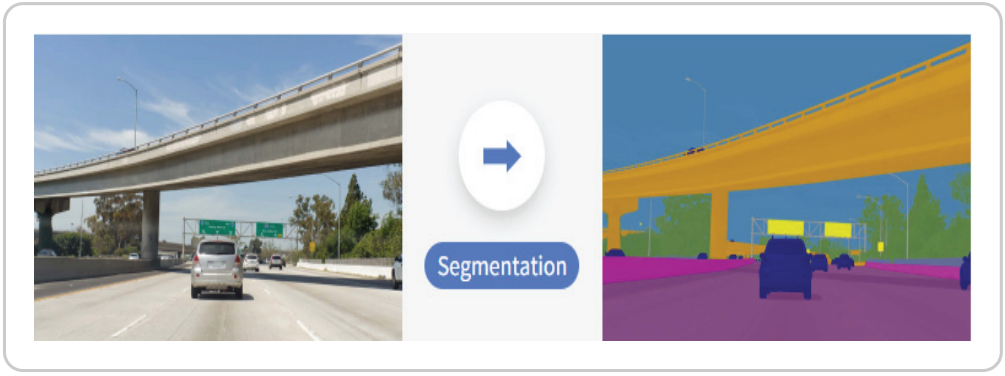
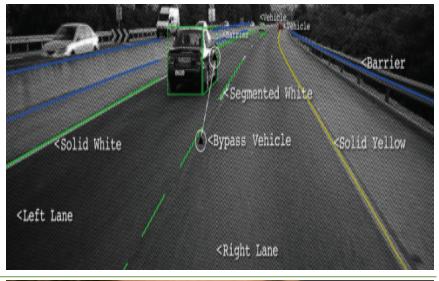
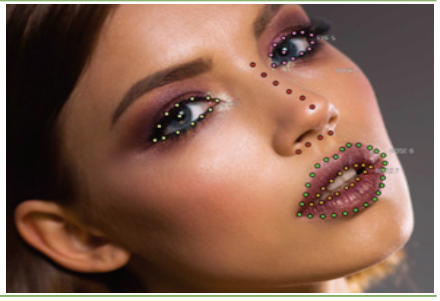
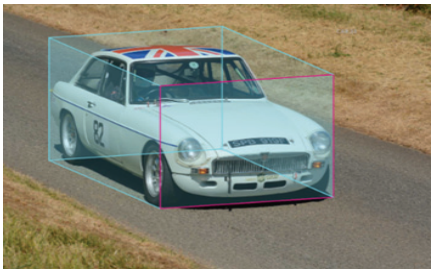
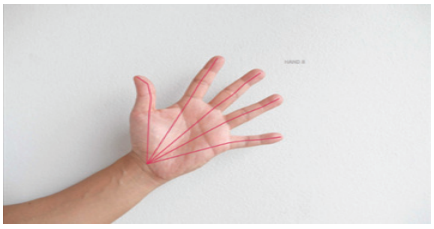


그림 4-5. 2D 영상기반의 픽셀 단위 주석화 예시

- 라벨링 작업 방식

표 4-4. 라벨링 도구 예시

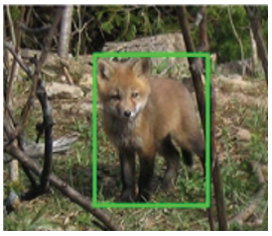
라벨링 방식	설 명	이미지 예시
Bounding Box	• 객체를 직사각형 모양의 박스 안에 포함되도록 그리는 라벨링 방법으로 데이터 라벨링 작업에서 가장 일반적으로 사용 – Bounding Box는 객체를 전체가 커버되도록 하며, 박스 안에 객체 이외의 여백을 최소화하도록 지정	
Polygon	• 다각형 모양으로 객체의 가시 영역 외곽선을 따라 점을 찍어 그리는 라벨링 방법, 개체 이외의 포함된 빈공간으로 인해 발생하는 오류에 대응할 수 있는 기능 – 물체를 정확하게 인식하기 위해서 사용하며, 사물의 테두리를 따라 그리는 것을 통해 여백 없이 정확히 물체만을 인식하기 위해 사용	
Polyline	• 여러 개의 점을 가진 선을 활용하여 특정 영역을 라벨링 함으로써 인도, 차선 등을 구분하기 위해 사용 – 인도와 차선, 경계라고 할 수 있는 모든 것들을 인식시키기 위해서 사용	
Keypoint	• 특정 지점을 라벨링 하는 작업으로 안면 인식을 통한 감정 분석과 같이 정밀하고 섬세한 작업을 요구하는 기술 – 객체의 중요 특징점을 지정하여 물체를 추적하고 인식할 수 있음	

라벨링 방식	설 명	이미지 예시
Cuboid	• 2D로 작업할 수 없는 3D 객체들을 정육면체로 생성하는 라벨링 방식으로 자동차, 건물 등 입체적인 객체들을 2D 형식으로 라벨링 작업하는 것에는 한계가 있는 부분을 해결하기 위한 기술 - 2D 사진에서의 객체에 대한 정보를 3D로 표현할 수 있도록 하는 방법	
Body	• 전체적인 모션 캡쳐나 이상 행동 등 사람의 움직임을 검출할 필요가 있는 경우 몸에 객체를 생성하는 방식 - 사람의 관절이나 기계의 움직임을 표현 할 수 있는 부분에 대한 정보를 얻기 위해 사용하는 방법	
Face	• 얼굴의 특징점을 검출할 필요가 있는 경우 얼굴에 객체를 생성하는 방식 - 얼굴 그 차제를 인식하기 위한 방식을 눈 코 입을 포함한 얼굴선에 대한 라벨링을 진행	
Hands	• 수어 등의 손의 움직임을 파악하기 위해 손의 마디를 검출할 필요가 있는 경우 사용	

- 데이터 라벨링 / 어노테이션 방법 예시
  - 영상(동적/정적) 이미지 데이터 라벨링이란 객체에 대한 Detection 과 annotation이므로 Bounding box와 Polygon을 예시로 함
  - Polygon을 사용해서 목적에 맞는 객체의 외곽선에 맞추어 라벨링을 실시
  - Bounding box는 목적에 맞는 객체를 모두 포함하면서 최대한 타이트하게 box처리 **(최대 약 2 Pixel)**
  - 목적에 맞는 객체와 다른 객체가 겹쳐 있을 경우에는 Polygon을 사용하여 정확도를 높임
  - Bounding box로 라벨링 할 경우 단일, 다중 객체가 정확히 인식되도록 해야함

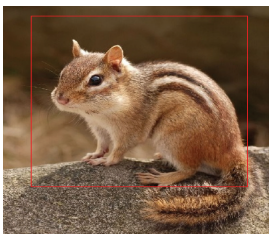
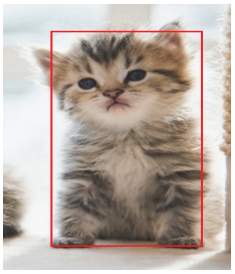
- 올바른 라벨링 예시 (Bounding box, Polygon)

표 4-5. 올바른 라벨링 작업 방법 예시

	• Boundin box가 다람쥐라는 객체를 모두 포함하며 box의 크기가 외각선에서 최대한 타이트하게 라벨링 • annotation : 다람쥐
	• 사진에서 여우의 다리가 잘 보이지 않는 것을 감안해 라벨링을 보이는 외각선에 최대한 타이트하게 라벨링 • annotation : OO산, 여우
	• 여러마리의 고양이가 겹쳐있을 경우 데이터의 정확성을 위해 polygon방식을 사용하여 라벨링을 실시 • 최대한 자세하게 외각선을 그려야함 • annotation : 고양이

- 잘못된 방법 예시

표 4-6. 잘못된 라벨링 작업 방법 예시

	• 사진에서 특정할 수 있는 여우의 외각선에 비해 Boundin box가 매우 크게 그려져 있어 잘못 라벨링 된 데이터로 분류
	• 사진에서 다람쥐라고 특정할 수 있는 꼬리 부분이 Boundin box안에 존재하지 않고 외각선에서 떨어져 있는 상태이므로 잘못 라벨링된 데이터로 분류
	• 사진에서 고양이의 모습을 특정할 수 있는 외각선 부분보다 Boundin box가 작게 라벨링 되어있어 라벨링된 데이터로 적합하지 않음

4.3 데이터 어노테이션 포맷과 형식 정의 및 입력

- 어노테이션 형식 및 정의

※ 라벨링 작업 시 예로 바운딩 박스의 시작 좌표와 이어지는 좌표, 끝점 좌표가 매우 중요함

표 4-7. 어노테이션 형식 및 정의 예시

No.	어노테이션 형태	항목 설명
1	annotations[].id	어노테이션 식별자
2	annotations[].image_id	연관 영상(동적/정적) 이미지 식별자
3	annotations[].classes	어노테이션 클래스
4	annotations[].segmentation	객체 영역 정보
5	annotations[].bbox	어노테이션 바운딩박스 정보
6	annotations[].polygon	어노테이션 폴리곤 정보
7	annotations[].polyline	어노테이션 폴리라인 정보
8	annotations[].cuboid	어노테이션 큐보이드 정보
9	annotations[].points	어노테이션 포인트 정보

- 영상(동적/정적) 이미지 데이터 라벨링 정보

- 공통 참조 필수항목(디지털 카메라, 스마트폰, CCTV, 드론/위성)
  - 인공지능 학습용 데이터셋 구축 공통참조기준을 바탕으로 어노테이션 항목으로 변환 (예시로서 활용 가능)

※ 국가 시설물과 개인정보 등 정보보안요소와 위치정보들은 비식별화를 통한 재보정 후 적용 필요

표 4-8. 영상(동적/정적) 이미지 데이터 라벨링 정보 - 공통참조 필수항목

No.	속성명	항목 설명	Type	필수여부	작성예시
1	videos[].filename	파일 이름	string	필수	DSC_0001 (분류_순번)
2	videos[].id	ID	string	필수	DSC_0001_개체번호 (분류_순번)

No.	속성명	항목 설명	Type	필수여부	작성예시
3	videos[].date_created	촬영일자	string	필수	2020.11.20. 17:08:15
4	videos[].type	데이터 형식	string	필수	mp4, PNG, JPG
5	videos[].format	포맷	string	필수	h.264/mpeg-4
6	videos[].filesize	크기	number	필수	4800KB
7	videos[].photographer	촬영자	string	필수	홍길동
8	videos[].device	촬영 장비	string	필수	디지털 카메라
9	videos[].location	촬영 지역명	string	필수	서울시 종로구 (동까지만 표기)
10	videos[].license	라이선스	string	필수	-
11	videos[].length	영상길이	string	필수	10M
12	videos[].FPS	프레임 재생속도	string	필수	30
13	videos[].frames	총 프레임 수 (FPS)	number	필수	60
14	videos[].aspect_ratio	종횡비	string	필수	4:3
15	videos[].width	너비	number	필수	4031
16	videos[].height	높이	number	필수	3024
17	videos[].resolution	해상도	string	필수	FHD
18	videos[].bit	비트값	string	필수	24bit
19	videos[].pixel	화소	string	필수	4K
20	videos[].color_depth	색심도	string	필수	sRGB
21	videos[].ISO	ISO 감도	string	필수	3200
22	videos[].whith balance	화이트 밸런스	string	필수	5500K
23	videos[].exposure_time	노출시간	string	필수	f2.8 1/80
24	videos[].F-stop	조리개값	string	필수	f2.8
25	videos[].flash	플래시	string	필수	자동
26	videos[].focal_length	초점거리	string	필수	50mm
27	videos[].angle_view	화각	string	필수	46
28	videos[].angle	촬영각도	string	필수	120도
29	videos[].weather	날씨정보	string	필수	맑음

- 장비별 영상(동적/정적) 이미지 참조 항목 라벨링 정보
  - cctv 영상(동적/정적) 이미지 - 공통참조기준을 바탕으로 어노테이션 항목으로 변환 (예시로서 활용 가능)

※ 국가 시설물과 개인정보 등 정보보안요소와 위치정보들은 비식별화를 통한 재보정 후 적용 필요

표 4-9. CCTV 영상(동적/정적) 이미지 데이터 라벨링 정보

No.	속성명	항목 설명	Type	필수여부	단위 (작성예시)
1	videos[].visible_distance	가시거리	number	필수	50M
2	videos[].temperature	온도	number	선택	3℃
3	videos[].humidity	습도	number	선택	32%
4	videos[].coordinates	좌표	number	선택	위경도 정보
5	videos[].cctv_name	CCTV 명	string	필수	광화문4거리3번 CCTV
6	videos[].range	촬영범위(360도 회전 혹은 고정)	string	필수	50
7	videos[].mission	촬영지 분류	string	필수	도심지
8	videos[].event_id	이벤트 분류	string	선택	ABA_0001 (분류_순번)
9	videos[].event_name	이벤트명	string	선택	특이 상황판별
10	videos[].event_name.start_time	이벤트 시작시간	string	선택	2020.11.20 17:08:15
11	videos[].event_name.end_time	이벤트 종료시간	string	선택	2020.11.20 17:08:20

- 드론/위성 영상(동적/정적) 이미지

※ 국가 시설물과 개인정보 등 정보보안요소와 위치정보들은 비식별화를 통한 재보정 후 적용 필요

표 4-10. 드론/위성 영상(동적/정적) 이미지 데이터 라벨링 정보

No.	속성명	항목 설명	Type	필수여부	단위 (작성예시)
1	videos[].drone_name	드론명	string	필수	DRN_0001 (분류_순번)
2	videos[].sensor	센서	string	선택	가속도
3	videos[].max_flight	최대비행시간	string	선택	1H
4	videos[].temperature	온도	number	선택	3℃
5	videos[].humidity	습도	number	선택	30%
6	videos[].coordinates	좌표	number	필수	위경도좌표
7	videos[].GPS	GPS 정보(위도, 경도)	string	필수	37.3024, 126.53221
8	videos[].INS	카메라 회전각 정보	string	선택	120, 60, 0
9	videos[].speed	비행속도	number	선택	8km/h
10	videos[].range	촬영범위 (360도 혹은 고정, 가시거리), 정사영상 등	string	필수	50M
11	videos[].altitude	촬영고도	number	필수	150M
12	videos[].mission	촬영지 분류	string	필수	산림지
13	videos[].overlap	중복도	string	선택	3M, 3M
14	videos[].GCP	지상기준점 (GPS와는 별개로 위치 보정 역할 지정)	string	선택	위경도좌표

- 라벨링 완료 후 데이터 저장 포맷
  - 사용자의 목적에 맞는 저장 포맷을 선정하고 해당 데이터를 라벨링 완료된 데이터의 항목에 따라 분류하여 저장

1) COCO dataset 예시 (JSON)
   - 아래 검은색 그림의 예시 어노테이션은 COCO dataset의 어노테이션으로 해당 정보에서는 라벨링 된 객체가 어떤 것인지 분류하는 category_id를 통해 데이터를 분류하여 저장

표 4-11. COCO instances JSON 예시

항 목		설 명
info - 기본 정보		
	description	데이터셋 이름
	url	데이터셋 제작자 url
	version	제작 버전
	year	제작 년도
	contributor	데이터셋 제공자
	date_created	데이터셋 제작 시간
licenses - 저작권 정보		
	url	문서 아이디
	id	라이센스 고유 번호
	name	라이센스 이름
images - 이미지 데이터 정보		
	license	라이센스 번호
	file_name	원천데이터 이름
	coco_url	coco 다운로드 url
	height	세로
	width	가로
	date_captured	데이터 제작 시간
	flickr_url	플리커 등록 url
	id	이미지 고유번호
annotations - 어노테이션 정보		
	segmentation	segmentation mask 정보
	area	어노테이션 정확도의 평균
	iscrowd	객체의 단일 여부 (단일 : 0, 다중 : 1)
	image_id	이미지 고유번호
	bbox	bounding box 정보
	category_id	카테고리 고유 번호
	id	어노테이션 고유 번호
categories - 카테고리 리스트		
	supercategory	큰 틀의 카테고리
	id	카테고리 고유 번호
	name	이름

json 포맷 구축 형태


```
{
  "info": {
    "description": "COCO 2017 Dataset",
    "url": "http://cocodataset.org",
    "version": "1.0",
    "year": 2017,
    "contributor": "COCO Consortium",
    "date_created": "2017/09/01"
  },
  "licenses": [
    {
      "url": "http://creativecommons.org/licenses/by-nc-sa/2.0/",
      "id": 1,
      "name": "Attribution-NonCommercial-ShareAlike License"
    },
    ...
  ],
  "images": [
    ...
    {
      "license": 1,
      "file_name": "000000324158.jpg",
      "coco_url": "http://images.cocodataset.org/val2017/000000324158.jpg",
      "height": 334,
      "width": 500,
      "date_captured": "2013-11-19 23:54:06",
      "flickr_url": "http://farm1.staticflickr.com/169/417836491_5bf8762150
      "id": 324158
    },
    ...
  ],
  "annotations": [
    ...
    {
      "segmentation": [
        [
          216.7,
          211.89,
          216.16,
          217.81,
          215.89,
          220.77,

          ...
          212.16
        ]
      ],
      "area": 759.3375500000002,
      "iscrowd": 0,
      "image_id": 324158,
      "bbox": [
        196.51,
        183.36,
        23.95,
        53.02
      ],
      "category_id": 18,
      "id": 10673
    },
    ...
    {

  "categories": [
    {
      "supercategory": "person",
      "id": 1,
      "name": "person"
    },
    ...
  ]
}
```

2) PASCAL VOC dataset 예시 (XML)

- PASCAL VOC의 dataset의 구조

표 4-12. PASCAL VOC dataset 예시

폴더명	세부 내용
Annotations	JPEGImages 폴더 속 원본 이미지와 같은 이름들의 xml파일들이 존재
ImageSets	특정 클래스가 어떤 이미지에 있는지 등에 대한 정보들을 포함하고 있는 폴더
JPEGImages	*.jpg 확장자를 가진 이미지 파일들이 모여있는 폴더
SegmentationClass	Semantic segmentation을 학습하기 위한 label 이미지
SegmentationObject	Instance segmentation을 학습하기 위한 label 이미지

- PASCAL VOC는 object detection를 구현하기 위한 학습 데이터로 많은 양의 이미지 데이터를 포함
- 아래의 어노테이션은 PASCAL VOC의 어노테이션으로 XML로 구현되어 있고, 해당 어노테이션에서는 object를 통한 분류를 통해 데이터를 분류하여 저장

```
<annotation>
    <folder>VOC2007</folder>
    <filename>000001.jpg</filename>
    <source>
        <database>The VOC2007 Database</database>
        <annotation>PASCAL VOC2007</annotation>
        <image>flickr</image>
        <flickrid>341012865</flickrid>
    </source>
    <owner>
        <flickrid>Fried Camels</flickrid>
        <name>Jinky the Fruit Bat</name>
    </owner>
    <size>
        <width>353</width>
        <height>500</height>
        <depth>3</depth>
    </size>
    <segmented>0</segmented>
    <object>
        <name>dog</name>
        <pose>Left</pose>
        <truncated>1</truncated>
        <difficult>0</difficult>
        <bndbox>
            <xmin>48</xmin>
            <ymin>240</ymin>
            <xmax>195</xmax>
            <ymax>371</ymax>
        </bndbox>
    </object>
    <object>
        <name>person</name>
        <pose>Left</pose>
        <truncated>1</truncated>
        <difficult>0</difficult>
        <bndbox>
            <xmin>8</xmin>
            <ymin>12</ymin>
            <xmax>352</xmax>
            <ymax>498</ymax>
        </bndbox>
    </object>
</annotation>
```

그림 4-6. PASCAL VOC 어노테이션 XML 예시

4.4 데이터 라벨링 완료 후 관리 방법

- 데이터 라벨링 완료 후 관리 기본 사항
  - 목적에 맞는 데이터 어노테이션 기준을 수립하고 데이터 사용 목적에 맞게 관리
  - 데이터의 어노테이션 항목은 기존의 데이터 사용 목적의 변화로 인해 수정이 불가피한 경우를 제외하고는 쉽게 바뀌지 않음
  - 데이터의 어노테이션 정보는 쉽게 이해할 수 있어야 하며, 의미가 불분명하여 발생하는 혼란을 최소화
  - 데이터의 사용 목적에 맞는 일관된 자료인지 수시 확인
  - 데이터들의 편향성을 확인 후 필요에 따라 데이터 지속적 추가
  - 보존 일정 및 규정 준수 요구사항에 따라 데이터 보관, 저장 방식을 이용하여 관리

- 지속적인 교육을 통한 관리 방법 (보관방식은 '그림 3-16. 파일 자료 모음과 보관 예시' 참조)
  - 촬영 영상 확인을 위한 가이드 제작 및 교육
  - 라벨링 작업을 위한 가이드 제작 및 교육
  - 검사를 위한 가이드 제작 및 교육
  - 해당 데이터를 사용한 응용 서비스 개발을 통해 성능시험 평가 내역 공개

- 데이터 관리 조직 운영 방안
  - 데이터셋 제작 책임자는 품질관리 책임자로서 획득되는 데이터의 품질을 주기적으로 검사 및 관리
  - 수행 및 참여기관 간 주기적인 미팅을 통해 데이터 품질에 대한 피드백을 공유하고 논의

- 외부 검증기관 통한 품질관리
  - 외부 인공지능 지원 기관이나 기업을 통한 관리
  - 인공지능 데이터 구축·활용 가이드라인 작성을 통한 사전, 사후 관리
  - 인공지능 데이터 정확도·유효성 검증 계획서 작성을 통한 사전 및 사후 관리

4.5 데이터 라벨링 방식에 적합한 도구 선정 및 사용설명

데이터 라벨링 방식에 적합한 도구를 사용 목적에 맞춰 선택. 폴리곤이나 세그멘테이션이 필요한 경우, 키포인트인 경우 등 다양한 사용 목적이 가능한 도구를 사전에 선택하거나 오픈소스를 통해 직접 제작하여 사용 가능

- labelImg
  - object detection 학습을 위해 영상에서 Bounding box를 지정하여 라벨링을 수행하고, 그 bounding box 정보들을 xml 형태로 저장
  - 사이트 주소 : https://github.com/tzutalin/labelImg
- CVAT (Computer Vision Annotation Tool)
  - 컴퓨터 비전 알고리즘의 데이터 레이블을 지정하는데 사용되는 웹 기반 이미지 및 비디오 주석 도구
  - 주로 object detection, image segmentation, image classification을 하는데 쓰임
  - 사이트 주소 : https://github.com/openvinotoolkit/cvat
- LabelMe
  - Bounding box, Polygon, Polyline, Point 등 다양한 형태의 도형과 Classification, Segmentation 등 다양한 task의 라벨링 지원
  - 사이트 주소 : https://github.com/wkentaro/labelme
- LabelBox
  - 라벨링 경과를 csv, json 형식뿐만 아니라 일반적으로 많이 사용하는 데이터셋의 포맷(COCO, VOC, TFRecord 등)으로 export 할 수 있음
  - 사이트 주소 : https://labelbox.com/
- YOLO Mark
  - 영상(동적/정적) 이미지 안의 객체를 bounding box로 표시한 레이어 안의 데이터를 레이블링하는 도구
  - 사이트 주소 : https://github.com/AlexeyAB/Yolo_mark

5 처리 데이터 검사

5.1 검사 절차 정의

- 검사 절차는 검사 유형, 기준, 방법에 따라 달라질 수 있으며 검사 절차를 정의할 때는 데이터 사용 목적을 중점으로 진행. 데이터 활용목적에 맞는 검사 규격을 명확히 하여 전체적인 절차를 마련

1) 검사 절차

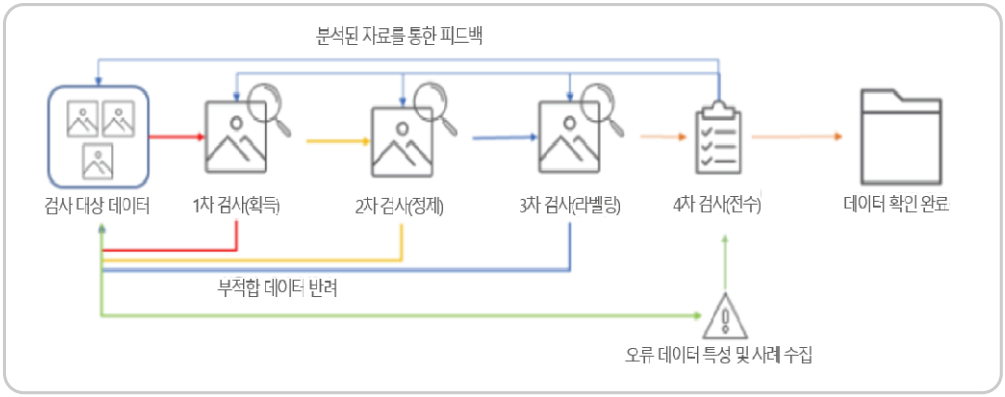


그림 5-1. 데이터 검사 절차

- 작업물의 상태로는 작업 중, 작업 완료, 검사 완료, 부적합(이유)으로 구성

2) 검사 기준

표 5-1. 검사 절차 기준 예시

검사 절차	영상(동적/정적) 이미지 공통 항목	요구사항
1차 검사 (획득)	법·제도 준수	원시데이터 획득 시 관련 법·제도적 규정 등을 반드시 준수하여야 함
	사실적인 획득 환경 구성	원시데이터를 인위적인 환경과 조건 하에 획득해야 하는 경우 사실적인 획득 환경을 구성하여야함

검사 절차	영상(동적/정적) 이미지 공통 항목	요구사항
	데이터 동기화	다중 데이터 소스 간 정교한 동기화를 위한 절차를 마련하여야 함
	편향성 방지	데이터 편향을 방지하기 위한 절차를 마련하여야 함
2차 검사 (정제)	정제 기준의 명확성	데이터 사용 목적에 적합한 정제 기준 수립 여부
	중복성 방지	데이터 정제 후 정보 비교 후 중복도 여부
	정제 작업 매뉴얼	정제 작업을 위한 매뉴얼 작성 및 관리 여부
	정제 도구	정제 작업에 사용될 SW 도구를 확보 및 사용 방법을 숙지
	정제 작업 방식	데이터 특성 및 활용 목적에 맞는 적절한 정제 방식 선정 여부 및 선정 기준 타당성 여부
3차 검사 (라벨링)	라벨링 가이드	목적에 맞게 작성된 라벨링 가이드에 대한 타당성 여부를 검사 후 라벨링 작업자들에게 내용 가이드 전달
	어노테이션 항목	목적에 맞는 어노테이션 구성인지 여부를 검사 후 확인된 내용을 포함하도록 작업자들에게 전달
	라벨링 검사 도구	자동화 도구를 통해 검사 후 검사자가 육안으로 부적합 데이터 여부 2차 확인과 촬영된 영상(동적/정적) 이미지의 누락, 번짐 및 조건 오류를 전수 검사
4차 검사 (전수)	부적합 판정 데이터 분포 확인	데이터의 오류율, 특성 분포 확인을 통한 데이터 수집, 정제, 라벨링, 부문 최적화
	외부 검사자	외부 검사자(TTA 등), 도메인 전문가, 데이터 요청자

3) 검사 조직

- 데이터 획득 검사팀, 데이터 정제 검사팀, 데이터 라벨링 검사팀, 데이터 전수 검사팀, 데이터 요청자 등으로 구성
- 해당 분야의 도메인 전문가, 데이터 요청자, 외부 업체를 통한 검사 필요(TTA 등)

4) 검사 도구

- 자동화된 도구를 사용하여 데이터 라벨링 여부 확인 및 어노테이션 항목 조건 충족 확인
- 수동으로 검사자가 자동화에서 검사 완료 상태의 데이터를 육안으로 검사

5.2 검사 방식

- 디지털 카메라, 스마트폰 이용한 수집 및 정제

표 5-2. 디지털 카메라, 스마트폰 검사 절차 방식 예시

검사 절차	영상(동적/정적) 이미지 공통 항목	요구사항
1차 검사 (획득)	법·제도 준수	촬영자, 저작권 정보 등 법적인 요소
	사실적인 획득 환경 구성	촬영 지역명, GPS 정보, 촬영 위치, 날씨 정보 등
	데이터 동기화	촬영 목적에 맞는 공통항목 및 추가 항목 검사
	편향성 방지	촬영 환경, 데이터 획득 항목별 편향 여부를 판단
2차 검사 (정제)	정제 기준의 명확성	데이터 사용 목적에 적합한 정제 기준 수립 여부
	중복성 방지	데이터 정제 후 정보 비교 후 중복도 여부
	정제 작업 매뉴얼	정제 작업을 위한 매뉴얼 작성 및 관리 여부
	정제 도구	정제 작업에 사용될 SW도구를 확보 및 사용 방법을 숙지
	정제 작업 방식	데이터 특성 및 활용 목적에 맞는 적절한 정제 방식 선정 여부 및 선정 기준 타당성 여부
3차 검사 (라벨링)	라벨링 가이드	목적에 맞게 작성된 라벨링 가이드에 대한 타당성 여부를 검사 후 라벨링 작업자들에게 내용 가이드 전달
	어노테이션 항목	목적에 맞는 어노테이션 구성인지 여부를 검사 후 확인된 내용을 포함하도록 작업자들에게 전달
	라벨링 검사 도구	자동화 도구를 통해 검사 후 검사자가 육안으로 부적합 데이터 여부 2차 확인과 촬영된 영상(동적/정적) 이미지의 누락, 번짐 및 조건 오류를 점검하기 위한 도구를 활용하여 전수 검사를 통해 중복 없이 빠른 검사를 진행
4차 검사 (전수)	부적합 판정 데이터 분포 확인	데이터의 오류율, 특성 분포 확인을 통한 데이터 수집, 정제, 라벨링, 부문 최적화
	외부 검사자	외부 검사자(TTA 등), 도메인 전문가, 데이터 요청자

- 위성/드론

표 5-3. 위성/드론 검사 절차 방식 예시

검사 절차	영상(동적/정적) 이미지 공통 항목	요구사항
1차 검사 (획득)	법·제도 준수	송신기 주파수, 촬영지 분류, GPS 정보, 개인정보, 보안 시설의 촬영 여부 등을 확인
	사실적인 획득 환경 구성	촬영지 분류, 촬영고도, 촬영범위, 온도, 습도, GPS 정보
	데이터 동기화	촬영 목적에 맞는 공통항목 및 추가 항목 검사
	편향성 방지	획득된 데이터의 항목을 비교하여 편향 여부를 판단
2차 검사 (정제)	정제 기준의 명확성	데이터 사용 목적에 적합한 정제 기준 수립 여부
	중복성 방지	데이터 정제 후 정보 비교 후 중복도 여부
	정제 작업 매뉴얼	정제 작업을 위한 매뉴얼 작성 및 관리 여부
	정제 도구	정제 작업에 사용될 SW 도구를 확보 및 사용 방법을 숙지
	정제 작업 방식	데이터 특성 및 활용 목적에 맞는 적절한 정제 방식 선정 여부 및 선정 기준 타당성 여부
3차 검사 (라벨링)	라벨링 가이드	목적에 맞게 작성된 라벨링 가이드에 대한 타당성 여부를 검사 후 라벨링 작업자들에게 내용 가이드 전달
	어노테이션 항목	목적에 맞는 어노테이션 구성인지 여부를 검사 후 확인된 내용을 포함하도록 작업자들에게 전달
	라벨링 검사 도구	자동화 도구를 통해 검사 후 검사자가 육안으로 부적합 데이터 여부 2차 확인과 촬영된 영상(동적/정적) 이미지의 누락, 번짐 및 조건 오류를 점검하기 위한 도구를 활용하여 전수 검사를 통해 중복 없이 빠른 검사를 진행
4차 검사 (전수)	부적합 판정 데이터 분포 확인	데이터의 오류율, 특성 분포 확인을 통한 데이터 수집, 정제, 라벨링, 부문 최적화
	외부 검사자	외부 검사자(TTA 등), 도메인 전문가, 데이터 요청자

- 특수 카메라
  - CT, MRI, X-ray, 열화상 카메라, LiDAR (센서), 수중 카메라 (CCTV), 짐벌, 액티브 캠, 초분광, 적외선, CCTV

표 5-4. 특수카메라 검사 절차 방식 예시

검사 절차	영상(동적/정적) 이미지 공통 항목	요구사항
1차 검사 (획득)	법·제도 준수	개인정보, 법·제도 관련 검사
	사실적인 획득 환경 구성	목적에 맞는 환경 구성 여부 검사
	데이터 동기화	영상(동적/정적) 이미지 추가 데이터 획득 항목 검사
	편향성 방지	획득된 데이터의 항목을 비교하여 편향 여부를 판단
2차 검사 (정제)	정제 기준의 명확성	데이터 사용 목적에 적합한 정제 기준 수립 여부
	중복성 방지	데이터 정제 후 정보 비교 후 중복도 여부
	정제 작업 매뉴얼	정제 작업을 위한 매뉴얼 작성 및 관리 여부
	정제 도구	정제 작업에 사용될 SW 도구를 확보 및 사용 방법을 숙지
	정제 작업 방식	데이터 특성 및 활용 목적에 맞는 적절한 정제 방식 선정 여부 및 선정 기준 타당성 여부
3차 검사 (라벨링)	라벨링 가이드	목적에 맞게 작성된 라벨링 가이드에 대한 타당성 여부를 검사 후 라벨링 작업자들에게 내용 가이드 전달
	어노테이션 항목	목적에 맞는 어노테이션 구성인지 여부를 검사 후 확인된 내용을 포함하도록 작업자들에게 전달
	라벨링 검사 도구	자동화 도구를 통해 검사 후 검사자가 육안으로 부적합 데이터 여부 2차 확인과 촬영된 영상(동적/정적) 이미지의 누락, 번짐 및 조건 오류를 점검하기 위한 도구를 활용하여 전수 검사를 통해 중복 없이 빠른 검사를 진행
4차 검사 (전수)	부적합 판정 데이터 분포 확인	데이터의 오류율, 특성 분포 확인을 통한 데이터 수집, 정제, 라벨링, 부문 최적화
	외부 검사자	외부 검사자(TTA 등), 도메인 전문가, 데이터 요청자

5.3 검사 결과

- 검사 결과에서 위 절차와 방법에 따라 수행하여 문제가 발생할 경우, 모든 데이터는 그 시점에서 획득→정제→라벨링→저장→적용의 모든 단계를 역추하거나 분석을 통해 문제점을 찾아내야 함
- 검사에서는 단계별 결과를 제공하지는 않지만 최종적으로 모델을 통한 검증이 진행되므로 해당 수치화 결과값이 제공될 뿐 기본적인 데이터의 문제점까지는 제공하지 않기 때문에, 사전에 샘플링 방법을 통해 검사를 수행하여 대비해야 함

참 고 자 료

1. 한국정보통신기술협회(TTA), AI 학습용 데이터 구축사업 공통기준
2. 한국정보통신기술협회(TTA), 2020년 인공지능 학습용 데이터 구축 사업(1차) 중간산출물 20종 검토
3. 한국지능정보사회진흥원(NIA), 2020년 인공지능 학습용 데이터 구축사업(2차) 가이드라인 참조

V 부 록

1. 인공지능 학습용 데이터셋 구축 공통참조기준
2. 인공지능 학습용 데이터셋 구축 계획서

부록 1. 인공지능 학습용 데이터셋 구축 공통참조기준

목 차

1 개 요

1.1 작성 배경

- 인공지능 학습용 데이터의 구축 수행 및 참여기관에 따라 라벨링 데이터의 유형(xml, json 등) 및 구조 등이 상이하여 기 구축된 인공지능 학습용 데이터를 다른 목적으로 활용하기 어렵고, 구축된 데이터 품질 관리 문제가 발생
- 인공지능 학습용 데이터 라벨링 공통 기준을 통해 원시데이터 유형과 목적별 라벨링 구조 및 포맷 등의 기준이 필요함

1.2 작성 목적

- 인공지능 학습용 데이터 구축사업의 수행 및 참여기업에서 인공지능 학습데이터의 수집 · 정제 · 라벨링 절차에 따라 구축 목적과 원천데이터 유형에 맞는 구축사업에서 참조할 수 있는 공통사항들을 위한 기준 마련
- 인공지능 학습용 데이터 구축사업에 참여하는 신규 수행 및 참여기관의 라벨링 구축 방식을 안내하는 것을 목적으로 인공지능 학습용 구축사업 시 사업계획서에 사전 반영 및 배포하여 사용하고자 함
- 인공지능 학습용 구축사업의 인공지능 기술 분야에 대한 라벨링 관련 공통참조기준 제공을 통해 고품질의 인공지능 학습용 데이터 셋을 확보하고자 함

1.3 작성 범위

- '20년 인공지능 학습용 데이터 구축사업 8개 영역 48개 분야별 150종 인공지능 학습용 데이터 획득 공통 참조기준 가이드라인과 기존 TTA 인공지능 학습용 데이터 구축공정 가이드라인을 분석 검토하여 최소한의 기준으로 공통 참조 가능한 기준을 수립하여 도출
- 해외 인공지능 주요 사례에서 확인 후 라벨링 방식 별 데이터 구조 참조
- 국내 학습데이터 관련 인공지능 기업 사례 조사 등 다양한 기업들의 선행 사례들을 확인 후 가능 방법 적용
- 인공지능 학습용 데이터 구축사업 중 텍스트와 광학문자인식 이미지, 자율주행, 영상(동적/정적)이미지 총 4종 영역을 분석하여 최소한의 기준으로 공통 활용 가능한 기준을 수립하여 도출

1.3.1 공통참조기준 도출 방법

- '20년 인공지능 학습용 데이터 구축사업 1차 구축 영역 중 관련 결과물 분석을 통한 라벨링 구조 도출
- 한국어 분석에 가장 많이 사용되는 한국전자통신연구원(ETRI)의 KorBERT 언어모델을 중심으로 SKT의 KoBERT 언어모델과 HanBERT 언어모델을 분석하여 텍스트 유형 공통참조기준 도출
- Google Vision API와 Tesseract OCR을 참조하여 이미지 유형 중에서 이미지 내 텍스트 추출(OCR)에 대한 공통참조기준 도출
- 자율주행 관련 '20년 인공지능 학습용 데이터 1차 구축사업 영역 결과물 분석 및 관련 기업의 라벨링 자료 획득 및 분석을 통한 자율주행 유형 공통참조기준 도출

- 영상(동적/정적)이미지의 경우 ’20년 인공지능 학습용 구축사업 1차 20개와 2차 150개를 확인 후 종별 101개 영역을 분석 후 획득 방식에 따른 분류 작업 진행

- 영상(동적/정적) 이미지 가운데 획득 방법에 대한 장비별 분류를 통하여 렌즈와 해상도가 동일할 수 있는 장비와 묶어서 분석 검토

- ’20년 추경 150종 영상(동적/정적) 이미지 내용 분석(디지털 일안반사식 카메라(Digital Single Lens Reflex Camera 이하 DSLR) 또는 디지털 카메라, 스마트폰, 드론/위성, 폐쇄회로 TV(Closed-circuit television 이하 CCTV), 특수촬영 장비 사용)
  - 광학 문자 인식(Optical Character Recognition 이하 OCR), 헬스케어(X-Ray, 자기공명영상(Magnetic Resonance Imaging 이하 MRI), 컴퓨터단층촬영(Computed Tomography 이하 CT), 특수렌즈(열화상, 초분광) 사용 카메라 등은 검토 및 확인하고 공통 참조 항목은 추가 사업을 통해 진행 예정

- 각각의 영상(동적/정적) 이미지 획득 방법과 공통 사용 항목 도출 방식
  - 개개별 과제 내 영상 속성 정보 확인
  - 공통 참조분모 활용 기준 도출
    · 수행기관이 제출한 구축공정 활용 가이드라인에서 수집 대상과 방법 검토 확인
    · 국내외 영상(동적/정적) 이미지 획득 사례 참조

1.4 용어 정의

- 데이터 획득(Data Acquisition)
  - 인공지능의 기계학습에 필요한 데이터를 현실 세계에서 직접 수집 또는 생성하거나, 이미 보유하고 있는 조직이나 시스템 등으로부터 법률적 제약이 없도록 ‘원시데이터’를 확보하는 활동

- 데이터 정제(Data Refinement)
  - 획득한 원시데이터를 기계학습에 필요한 형식으로 맞추거나 불필요한 중복을 제거하며, 개인정보를 비식별화하여 처리하는 등 일련의 전처리 과정을 통해 '원천데이터'를 확보하는 활동

- 데이터 라벨링(Data Labeling)
  - 인공지능이 기계학습에 활용할 수 있도록 기능이나 목적에 부합하는 정보를 원천데이터에 부착하는 활동

- 라벨링데이터(Labeled Data)
  - 원천데이터에 부여한 '참값', 파일형식이나 해상도 등의 속성, 그리고 설명이나 주석 등이 포함된 '어노테이션'의 집합

- 원시데이터(Raw Data)
  - 기계학습을 목적으로 획득 단계에서 수집 또는 생성한 음성, 이미지, 영상, 텍스트 등의 데이터

- 원천데이터(Source Data, Unlabeled Data)
  - 원시데이터를 라벨링 공정에 투입하기 위해 필요한 전처리 등 정제 작업을 수행한 데이터로 라벨링데이터가 부여되지 않은 상태의 데이터

- 인공지능 학습용 데이터 구축
  - 임무정의, 데이터 획득, 데이터 정제, 데이터 라벨링 등 인공지능 학습용 데이터를 구축하는 일련의 활동

- 참값(Ground Truth)
  - 인공지능의 기계학습 목적에 따라 원시데이터에 라벨링된 정확한 값이나 사실의 의미적 표현

- 어노테이션(Annotation)
  - 데이터 라벨링 시 원시데이터에 주석을 표시하는 작업을 의미하며,

추가 부착되는 설명정보 데이터는 기능 목적에 따라 다양한 형태로 표현될 수 있으며 이러한 설명정보 표현방식을 지칭

※ 용어사용 예 : 사물 바운딩박스 어노테이션, 클래스 라벨링 어노테이션 등

- 광학문자인식(OCR, Optical Character Recognition)
  - 사람이 쓰거나 기계로 인쇄한 문자의 영상을 기계가 읽을 수 있는 문자로 변환하는 것

※ 자세한 용어 정의는 본 가이드라인 Ⅴ.부록-1 용어정의를 참조

2 인공지능 학습용 데이터 셋 구축 공통참조기준

2.1 원시데이터 유형별 라벨링 기능 및 어노테이션 방식

No	데이터 유형	라벨링 기능	어노테이션 방식
1	텍스트	- 텍스트 분류(Text Classification)	- 클래스 라벨 (단일, 다중)
		- 개체명 인식(Named Entity Recognition)	- 단어(구문) 라벨
		- 관계-의존성 정의(Relation-Dependencies)	- 단어(구문) 라벨링 및 두 단어 사이의 관계
2	이미지	- 이미지 분류(Image Classification)	- 클래스 라벨 (단일, 다중)
		- 객체 인식(Object Recognition)	- 바운딩 박스(사각형) - 폴리곤(다각형)
		- 영역 구분(Segmentation)	- 픽셀(점)
3	동영상	- 동영상 분류(Video Classification)	- 클래스 라벨 (단일, 다중)
		- 객체 인식(Object Recognition)	- 바운딩 박스(사각형) - 키 포인트(정점) - 폴리곤(다각형) - 폴리라인(선)
		- 객체 추적(Object Tracking)	
4	오디오	- 오디오 분류(Audio Classification)	- 클래스 라벨
		- 오디오 세그멘테이션(Audio Segmentation)	
		- 음성인식(음성→텍스트 변환) (Speech to Text)	- 텍스트 전사
5	기타	- 시계열 세그멘테이션 (Time-Series Segmentation) - HTML 문서 분류(HTML Classification)	- 클래스 라벨

2.2 텍스트

2.2.1 라벨링 공통 항목

텍스트 유형의 인공지능 학습용 데이터는 '문서요약', '질의응답', '기계번역', '대화' 등 다양한 목적으로 구축되며, 가장 기본이 되는 언어 분석을 위한 라벨링 구조를 공통기준으로 도출

- 라벨링 메타정보 공통참조항목

No.	속성명	항목 설명	Type	필수여부	작성예시
1	Dataset.identifier	데이터셋 식별자	string	필수	TEXT_QnA_LAW_01 (데이터유형_목적_분야_순번)
2	Dataset.name	데이터셋 이름	string	필수	법률 관련 인공지능 질의응답 학습용 데이터 셋
3	Dataset.src_path	데이터셋 폴더 위치	string	필수	/dataSet/text/
4	Dataset.label_path	데이터셋 레이블 폴더 위치	string	필수	/dataSet/text/
5	Dataset.category	데이터셋 카테고리	number	필수	0: 텍스트 분류, 1: 문서요약, 2:질의응답, 3: 기계번역 등
6	Dataset.type	데이터셋 타입	number	필수	0: 텍스트, 1: 이미지, 2:영상, 3: 음성 등

● 라벨링 메타정보 선택항목

- '문서요약' 및 '문서분류' 등의 목적으로 원시데이터의 출처 정보가 중요한 경우 선택적으로 작성

No.	속성명	항목 설명	Type	필수여부	작성예시
1	info.filename	원시데이터 파일명	string	선택	NEWS_000001 (매체유형_순번)
2	info.title	원시데이터 제목	string	선택	이스라엘 75세 남성 화이자 백신 접종 후 사망… "백신 연관성 없는 듯"
3	info.mediatype	매체유형	string	선택	뉴스, 블로그, SNS 등
4	info.medianame	매체명	string	선택	중앙일보
5	info.category	원시데이터 카테고리	string	선택	정치, 경제, 연예, 스포츠 등
6	info.size	원시데이터 크기 (글자수)	string	선택	270
7	info.date	발행일자	string	선택	2020.12.29 12:40:23 (yyyy.MM.dd HH:mm:ss)

- 저작권 정보가 존재할 때, 선택적으로 작성

No.	속성명	항목 설명	Type	필수여부	작성예시
1	licenses.id	라이센스 고유 번호	string	선택	http://www.apache.org/licenses/LICENSE-1.0
2	licenses.name	라이센스 이름	string	선택	Apache License 1.0
3	licenses.url	문서 식별자	string	선택	NEWS_000001

2.2.2 어노테이션 공통 항목

- 자연어 처리 어노테이션 항목
  - 자연어의 의미 분석을 위해 문장 단위로 형태소 분석을 위한 어노테이션 항목

No.	속성명	항목 설명	Type	필수여부	작성예시
1	annotations[].morp.id	형태소 식별자 (출현 순서)	string	선택	NNG_00001 (형태소태그_순번)
2	annotations[].morp.lemma	형태소	string	선택	일반명사
3	annotations[].morp.type	형태소 태그	string	선택	NNG, NNP, NNB 등
4	annotations[].morp.position	문장 내 위치	number	선택	231 (형태소 위치)
5	annotations[].morp.weight	형태소 분석 결과 신뢰도	number	선택	0.92 (0 ~ 1)

  - 자연어의 의미 분석을 위해 단어 혹은 구문 단위 분석을 위한 어노테이션 항목

No.	속성명	항목 설명	Type	필수여부	작성예시
1	annotations[].text_id	원문 텍스트 식별자	string	선택	TXT_0001 (분류_순번)
2	annotations[].word[].string	단어	string	선택	네이버
3	annotations[].word[].label	단어 레이블	string	선택	기업
4	annotations[].word[].start	원문 내 단어 시작 지점	number	선택	100
5	annotations[].word[].end	원문 내 단어 종료 지점	number	선택	102

- 자연어의 의미 분석을 위해 문장 단위로 어휘의미 분석을 위한 어노테이션 항목

No.	속성명	항목 설명	Type	필수여부	작성예시
1	annotations[].WSD.id	어휘의미 식별자 (출현 순서)	string	선택	WSD_0001 (분류_순번)
2	annotations[].WSD.text	어휘 텍스트	string	선택	배
3	annotations[].WSD.weight	어휘의미 분석 결과 신뢰도	number	선택	0.92 (0 ~ 1)
4	annotations[].WSD.position	문장 내 위치	number	선택	2310
5	annotations[].WSD.begin	어휘의 첫 형태소 식별자	string	선택	LC_0012
6	annotations[].WSD.end	어휘의 끝 형태소 식별자	string	선택	LC_0017

- 텍스트 데이터 라벨링 기능 및 어노테이션 방식

No	라벨링 기능	어노테이션 방식
1	- 텍스트 분류(Text Classification)	- 클래스 라벨(단일, 다중)
2	- 개체명 인식(Named Entity Recognition)	- 단어(구문) 라벨
3	- 관계-의존성 정의(Relation-Dependencies)	- 단어(구문) 라벨링 및 두 단어 사이의 관계

- 클래스 어노테이션 예시

No.	속성명	항목 설명	Type	필수여부	작성예시
1	annotations[].id	어노테이션 식별자	string	선택	CL_0001 (분류_순번)
2	annotations[].class	클래스 분류 (클래스 정의 필요)	number	선택	0: 정치, 1: 사회, 2: 연예, 등

- 텍스트 내 개체명 인식을 위한 어노테이션항목

No.	속성명	항목 설명	Type	필수여부	작성예시
1	annotations[].NE.id	개체명 식별자	string	선택	LC_0012 (개체명분류_순번)
2	annotations[].NE.text	개체명 텍스트	string	선택	광화문
3	annotations[].NE.type	개체명 타입	string	선택	관광명소 (LC_TOUR)
4	annotations[].NE.begin	객체명 구성 첫 형태소 식별자	string	선택	LC_0009
5	annotations[].NE.end	객체명 구성 끝 형태소 식별자	string	선택	LC_0015
6	annotations[].NE.weight	개체명 인식 결과 신뢰도	number	선택	0.92 (0 ~ 1)

- 관계-의존성 어노테이션 라벨

No.	속성명	항목 설명	Type	필수여부	작성예시
1	annotations[].dependency[].id	어절 ID (출현 순서)	string	선택	DEF_0021 (분류_순번)
2	annotations[].dependency[].text	의존구문 텍스트	string	선택	안녕하세요. 좋은 아침입니다.
3	annotations[].dependency[].head	부모 어절의 ID	string	선택	DEF_0020
4	annotations[].dependency[].label	의존관계 레이블	string	선택	-
5	annotations[].dependency[].mod[]	자식 어절들의 ID	string	선택	-
6	annotations[].dependency[].weight	의존구문 분석 결과 신뢰도	number	선택	0~1

2.3 OCR 이미지

2.3.1 라벨링 공통 항목

- 라벨링 메타정보 공통참조항목

No.	속성명	항목 설명	Type	필수여부	작성예시
1	Dataset.identifier	데이터셋 식별자	string	필수	IMG_OCR_01 (데이터유형_목적_순번)
2	Dataset.name	데이터셋 이름	string	필수	이미지 내 간판 텍스트 인식을 위한 학습용 데이터셋
3	Dataset.src_path	데이터셋 폴더 위치	string	필수	/dataSet/text/
4	Dataset.label_path	데이터셋 레이블 폴더 위치	string	필수	/dataSet/text/
5	Dataset.category	데이터셋 카테고리	number	필수	0: OCR, 1: 객체인식 등
6	Dataset.type	데이터셋 타입	number	필수	0: 텍스트, 1: 이미지, 2:영상, 3: 음성 등

- 라벨링 이미지 파일 공통참조항목

No.	속성명	항목 설명	Type	필수여부	작성예시
1	Images.identifier	이미지 식별자 (파일명)	string	필수	IMG_OCR_01_00001 (Dataset ID_순번)
2	Images.type	이미지 파일 확장자	string	필수	JPG, PNG 등
3	Images.width	이미지 가로 크기 (픽셀)	number	필수	1012
4	Images.height	이미지 세로 크기 (픽셀)	number	필수	768
5	data_captured	이미지 생성 일자	string	필수	yyyy.mm.dd HH:MM:SS

※ 데이터 전처리를 통해 이미지 내 텍스트 영역(바운딩박스)만 따로 추출한 경우에는 바운딩 박스 정보(x, y, width, height)는 생략 가능

- 저작권 정보가 존재할 때, 선택적으로 작성

No.	속성명	항목 설명	Type	필수여부	작성예시
1	licenses.id	라이센스 고유 번호	string	선택	http://www.apache.org/licenses/LICENSE-1.0
2	licenses.name	라이센스 이름	string	선택	Apache License 1.0
3	licenses.url	문서 식별자	string	선택	NEWS_000001

2.3.2 어노테이션 공통 항목

- 이미지 데이터 유형의 라벨링 기능 및 어노테이션 방식

No	라벨링 기능	어노테이션 방식
1	- 이미지 분류(Image Classification)	- 클래스 라벨(단일, 다중)
2	- 객체 인식(Object Recognition)	- 바운딩 박스(사각형) - 폴리곤(다각형)
3	- 영역 구분(Segmentation)	- 픽셀(점)

- 바운딩박스(Bounding Box) 어노테이션 구조

  - 이미지 내 텍스트 영역에 대한 사각형 박스 형태의 어노테이션 구조

No.	속성명	항목 설명	Type	필수여부	작성예시
1	annotations[].bbox.id	바운딩박스 식별자	string	필수	BBX_0001 (분류_순번)
2	annotations[].bbox.text	바운딩박스 내 텍스트	string	필수	교보문고
3	annotations[].bbox.x	바운딩박스 시작점 x 좌표	number	필수	100 (좌측상단 기준)
4	annotations[].bbox.y	바운딩박스 시작점 y 좌표	number	필수	120 (좌측상단 기준)
5	annotations[].bbox.width	바운딩박스 가로 길이(픽셀)	number	필수	273
6	annotations[].bbox.height	바운딩박스 세로 길이(픽셀)	number	필수	125

- 이미지 내 텍스트와 매핑되는 실제 장소가 존재할 때, 매핑 관계를 표현하기 위한 선택적 항목

No.	속성명	항목 설명	Type	필수여부	작성예시
1	annotations[].place.id	장소 식별자	string	선택	PLC_0001
2	annotations[].place.title	장소명	string	선택	교보문고 광화문점
3	annotations[].place.addr	장소 주소	string	선택	서울특별시 종로구 종로1가 종로 1
4	annotations[].place.longitude	위치정보(경도)	string	선택	126.977759
5	annotations[].place.latitude	위치정보(위도)	string	선택	37.570975

※ 관심지점(POI, Point of Interest)의 위치 정보는 도로명 주소와 WGS84 좌표 체계 혹은 국가지점번호 체계 활용
※ 공개 제한된 관심지점의 위치 정보는 표시하지 않음

2.4 자율주행

2.4.1 라벨링 공통 항목

- 라벨링 메타정보 공통참조항목

No.	속성명	항목 설명	Type	필수여부	작성예시
1	Dataset.identifier	데이터셋 식별자	string	필수	AUD_01 (데이터유형_순번)
2	Dataset.name	데이터셋 이름	string	필수	자율 주차를 위한 학습용 데이터셋
3	Dataset.src_path	데이터셋 폴더 위치	string	필수	/dataSet/text/
4	Dataset.label_path	데이터셋 레이블 폴더 위치	string	필수	/dataSet/text/
5	Dataset.category	데이터셋 카테고리	string	필수	0: 운행중데이터, 1: 정지객체인식 등
6	Dataset.type	데이터셋 타입	string	필수	(복수개의 데이터가 존재하므로 빈값으로 설정)

- 라벨링 이미지 파일 공통참조항목

No.	속성명	항목 설명	Type	필수여부	작성예시
1	Images.identifier	이미지 식별자 (파일명)	string	필수	AUD_01_IMG_00001 (Dataset ID_유형_순번)
2	Images.type	이미지 파일 확장자	string	필수	JPG, PNG 등
3	Images.width	이미지 가로 크기 (픽셀)	number	필수	1012
4	Images.height	이미지 세로 크기 (픽셀)	number	필수	768
5	Images.data_captured	이미지 생성 일자	string	필수	2020-12-29 12:40:23 (yyyy-MM-dd HH:mm:ss)
6	Images.frame_num	영상 내 이미지 프레임 순서	number	필수	573

- 라벨링 영상 파일 공통참조항목

No.	속성명	항목 설명	Type	필수여부	작성예시
1	Images.identifier	영상 식별자 (파일명)	string	필수	AUD_01_MOV_00001 (Dataset ID_유형_순번)
2	Images.type	영상 파일 확장자	string	필수	JPG, PNG 등
3	Images.width	영상 가로 크기 (픽셀)	number	필수	1012
4	Images.height	영상 세로 크기 (픽셀)	number	필수	768
5	Images.data_captured	영상 생성 일자	string	필수	2020-12-29 12:40:23 (yyyy-MM-dd HH:mm:ss)
6	Images.play_time	영상 길이	string	필수	00:19:21 (HH:mm:ss)

- 저작권 정보가 존재할 때, 선택적으로 작성

No.	속성명	항목 설명	Type	필수여부	작성예시
1	licenses.id	라이센스 고유 번호	string	선택	http://www.apache.org/licenses/LICENSE-1.0
2	licenses.name	라이센스 이름	string	선택	Apache License 1.0
3	licenses.url	문서 식별자	string	선택	NEWS_000001

2.4.2 어노테이션 공통 항목

- 바운딩박스(Bounding Box) 어노테이션 구조
  - 이미지 내 객체 영역에 대한 사각형 박스 형태의 어노테이션 구조

No.	속성명	항목 설명	Type	필수여부	작성예시
1	annotations[].bbox.id	바운딩박스 식별자	string	선택	BBX_0001 (분류_순번)
2	annotations[].bbox.name	바운딩박스 내 객체명	string	선택	승용차
3	annotations[].bbox.category	바운딩박스 내 객체 유형	string	선택	Car
4	annotations[].bbox.x	바운딩박스 시작점 x 좌표	number	선택	100 (좌측상단 기준)
5	annotations[].bbox.y	바운딩박스 시작점 y 좌표	number	선택	120 (좌측상단 기준)
6	annotations[].bbox.width	바운딩박스 가로 길이(픽셀)	number	선택	273
7	annotations[].bbox.height	바운딩박스 세조 길이(픽셀)	number	선택	125
8	annotations[].bbox.longitude	객체 위치(경도)	string	선택	126.977759
9	annotations[].bbox.latitude	객체 위치(위도)	string	선택	37.570975

- 3D 바운딩박스(3D Bounding Box, Cuboid) 어노테이션 구조
  - 이미지 내 객체 영역에 대한 3차원 육면체 박스 형태의 어노테이션 구조

No.	속성명	항목 설명	Type	필수여부	작성예시
1	annotations[].3dbbox.id	3D 바운딩박스 식별자	string	선택	CBD_0001 (분류_순번)
2	annotations[].3dbbox.name	3D 바운딩박스 내 객체명	string	선택	승용차
3	annotations[].3dbbox.category	3D 바운딩박스 내 객체 유형	string	선택	Car
4	annotations[].bbox.vertices[]	3D 바운딩박스 꼭지점 좌표	number	선택	[(10, 10, -10), ...] [(x, y, z), ...]
5	annotations[].bbox.edges[]	3D 바운딩박스 꼭지점 2개를 연결하는 변 좌표	number	선택	[(10, 10, 30, 10),...] [(x_1, y_1, x_2, y_2), ...]
6	annotations[].bbox.longitude	객체 위치(경도)	string	선택	126.977759
7	annotations[].bbox.latitude	객체 위치(위도)	string	선택	37.570975

- 폴리곤(Polygon) 어노테이션 구조
  - 이미지 내 객체 영역에 대한 다각형 형태의 어노테이션 구조

No.	속성명	항목 설명	Type	필수여부	작성예시
1	annotations[].polygon.id	폴리곤 식별자	string	선택	PLG_0001 (분류_순번)
2	annotations[].polygon.name	폴리곤 내 객체명	string	선택	승용차
3	annotations[].polygon.category	폴리곤 내 객체 유형	string	선택	Car
4	annotations[].polygon.points[]	폴리곤 내 점(x, y)의 집합	number	선택	[(100, 105), ...,(160,104)]

- 폴리라인(Polyline) 어노테이션 구조
  - 이미지 내 영역에 대한 선 형태의 어노테이션 구조

No.	속성명	항목 설명	Type	필수여부	작성예시
1	annotations[].polyline.id	폴리라인 식별자	string	선택	PLL_0001 (분류_순번)
2	annotations[].polyline.name	폴리라인 내 객체명	string	선택	중앙선
3	annotations[].polyline.category	폴리라인 내 객체 유형	string	선택	CenterLine
4	annotations[].polyline.points[]	폴리라인 내 점(x, y)의 집합	number	선택	[(100, 105), ...,(160,104)]

- 세그멘테이션(Segmentation) 어노테이션 구조
  - 이미지 내 객체 영역에 대한 다각형 형태의 어노테이션 구조

No.	속성명	항목 설명	Type	필수여부	작성예시
1	annotations[].segm.id	세그멘테이션 식별자	string	선택	SEG_0001 (분류_순번)
2	annotations[].segm.name	세그멘테이션 내 객체명	string	선택	승용차
3	annotations[].segm.category	세그멘테이션 내 객체 유형	string	선택	Car
4	annotations[].segm.points[]	세그멘테이션 내 점(x, y)의 집합	number	선택	[(100, 105), ...,(160,104)]

- 이벤트(Event) 어노테이션 구조
  - 영상 내 발생 이벤트 어노테이션 구조

No.	속성명	항목 설명	Type	필수여부	작성예시
1	annotations[].event.id	이벤트 식별자	string	선택	EVT_0001 (분류_순번)
2	annotations[].event.name	이벤트명	string	선택	자동차추돌사고
3	annotations[].event.category	이벤트 유형	string	선택	Accident
4	annotations[].event.start	영상 내 이벤트 시작 시점	number	선택	00:19:21 (HH:mm:ss)
5	annotations[].event.end	영상 내 이벤트 종료 시점	number	선택	00:19:32 (HH:mm:ss)
6	annotations[].event.entities[]	이벤트 관련 엔티티 목록	string	선택	[BBX_0001, BBX_0002] 엔티티 ID 목록

2.5 영상(동적/정적) 이미지

2.5.1 라벨링 공통 항목

- 영상(동적/정적) 이미지 획득 시 라벨링을 수행하기 위한 필수 반영 요소
- 획득 공통 참조 항목 (디지털 카메라, 스마트폰)

  ※ 국가 시설물과 개인정보 등 정보보안요소와 위치정보들은 비식별화를 통한 재보정 후 적용 필요

No.	속성명	항목 설명	작성예시
1	identifier/filename	파일명	세글자 이름(예: “DSC_0001.JPG”)
2	date	촬영날짜(년,월), 시간	2020.11.20. 17:08:15
3	file format	파일 형식(포맷)	TIFF/JPG/PNG/MP4/AVI 등
4	imsize	이미지 파일 크기	4800KB
5	images_Photographer	촬영자	촬영한 사람
6	device(camera, lidar)	장비정보	스마트폰, 디지털 카메라, drone, CCTV
7	region_name	촬영 지역명	서울시 종로구
8	images_location	촬영위치	강남구 영동대로 스타벅스
9	copyright	저작권 정보	저직권 정보 첨부 필드 체크
10	Video Clip	촬영시간	2~6분, 20~40분
11	length	영상길이	5분 영상에서 3분10초 부분 사용
12	FPS/Frame Rate	1초/프레임 재생 속도	30fps
13	width, height	이미지 사이즈	이미지 크기 4031*3024
14	Aspect ratio	비율(종횡비)	16:9(동영상)4:3(이미지)/가로세로
15	resolution	해상도	가로X세로 예) FHD(1920X1080)
16	bit	비트값	컬러색상/기본 24bit
17	Pixel	화소	사진정보/색상정보값(이미지픽셀)

No.	속성명	항목 설명	작성예시
18	depth	RGB 여부	색대표 : RGB, sRGB 등/비트값과 연관
19	ISO	ISO 감도	밝기에 따른 필름 감도
20	definition	선명도	일반 낮음 높음
21	white balance	화이트 밸런스	K(캘빈)단위 색온도/백열등, 형광등
22	exposure time	노출시간	조리개+셔터스피드 값
23	Exposure mode	노출 모드	자동 노출
24	Metering mode	측광 모드	스팟(중앙 중심) 접사에서 주로 사용
25	F-Stop	조리개 값	f2.8~f11 까지 이미지 밝기 조절
26	flash	플래시	자동 / 플래시 터지지 않음
27	filter	필터	필터 여부
28	focal length	초점 거리	mm초점거리/35mm~50mm(표준)
29	FOV(Field of View)	시야각(화각)	35mm→63도 예)50mm→46도
30	angle	촬영각도	촬영(360도 회전하며 8가지 이상)
31	GPS(Latitude,Longitude)	GPS 정보(위도, 경도)	GPS/GLONASS/37°30'24.7", 126°53'22.1"
32	weather	날씨정보	1)맑음 2)흐림 3)비 4)눈 중 선택

- 획득 선택항목(CCTV)

No.	속성명	항목 설명	작성예시
1	Visible distance	가시거리	최소 10m 이내

- 획득 선택 항목(드론/위성)

※ 국가 시설물과 개인정보 등 정보보안요소와 위치정보들은 비식별화를 통한 재보정 후 적용 필요

No.	속성명	항목 설명	작성예시
1	Mounted sensor	탑재 센서	1/2.3"유효픽셀수:12M
2	Flight time	비행 시간	23분
3	frequency	송신기 주파수	2.4GHz ISM
4	temperature	온도	-30℃ ~ 220℃
5	humidity	습도	0~100%,정확도
6	coordinates	영상좌상단, 후하단좌표	촬영 고도에 따른 지상기준 점 설정 값
7	INS	카메라 회전각 정보	X,Y,Z 3축 짐벌
8	speed(hoboring,1m/s,2m/s, 4m/s, 8m/s)	비행속도	X(Pitch),Y(Roll),Z(Yaw)방향의 전량
9	Range(m/s)	촬영범위	30~300cm
10	altitude	촬영고도	150m~
11	overlap	중복도	GSD(Ground Sampling Distance) 확보
12	Ascent, Descent speed	최대 상승, 하강 속도	5m/s, 3m/s
13	mission	촬영지 분류	산림지, 관광지, 도심지
14	Working temperature	작동 온도	0℃ ~ 40℃

- 특수 카메라

No.	항목명	특수 촬영 목적
1	CT, MRI, X-ray 등	헬스케어
2	열화상 카메라	피복, 체온, 상하수도, 시설물 등 주변 온도 변화 (지하 하수구 누수)
3	LiDAR (센서)	자율 주행, 3D 객체 인지, 지형 탐색
4	수중 카메라(CCTV)	수중 영상
5	짐벌, 액티브 캠, 초분광, 적외선	특수 촬영 목적으로 활용

2.5.2 어노테이션 공통 항목

- 어노테이션 형식 및 정의

※ 라벨링 작업 시 예로 바운딩 박스의 시작 좌표와 이어지는 좌표, 끝점 좌표가 매우 중요함

No.	어노테이션 형태	항목 설명
1	annotations[].id	어노테이션 식별자
2	annotations[].image_id	연관 영상(동적/정적) 이미지 식별자
3	annotations[].classes	어노테이션 클래스
4	annotations[].segmentation	객체 영역 정보
5	annotations[].bbox	어노테이션 바운딩박스 정보
6	annotations[].polygon	어노테이션 폴리곤 정보
7	annotations[].polyline	어노테이션 폴리라인 정보
8	annotations[].cuboid	어노테이션 큐보이드 정보
9	annotations[].points	어노테이션 포인트 정보

- 영상(동적/정적) 이미지 데이터 라벨링 정보
  - 공통 참조 필수항목(디지털 카메라, 스마트폰, CCTV, 드론/위성) : 공통 참조기준을 바탕으로 어노테이션 항목으로 변환 (예시로서 활용 가능)

※ 국가 시설물과 개인정보 등 정보보안요소와 위치정보들은 비식별화를 통한 재보정 후 적용 필요

No.	속성명	항목 설명	Type	필수여부	작성예시
1	videos[].filename	파일 이름	string	필수	DSC_0001 (분류_순번)
2	videos[].id	ID	string	필수	DSC_0001_개체번호 (분류_순번)
3	videos[].date_created	촬영일자	string	필수	2020.11.20. 17:08:15
4	videos[].type	데이터 형식	string	필수	mp4, PNG, JPG
5	videos[].format	포맷	string	필수	h.264/mpeg-4

No.	속성명	항목 설명	Type	필수여부	작성예시
6	videos[].filesize	크기	number	필수	4800KB
7	videos[].photographer	촬영자	string	필수	홍길동
8	videos[].device	촬영 장비	string	필수	디지털 카메라
9	videos[].location	촬영 지역명	string	필수	서울시 종로구 (동까지만 표기)
10	videos[].license	라이선스	string	필수	-
11	videos[].length	영상길이	string	필수	10M
12	videos[].FPS	프레임 재생속도	string	필수	30
13	videos[].frames	총 프레임 수(FPS)	number	필수	60
14	videos[].aspect_ratio	종횡비	string	필수	4:3
15	videos[].width	너비	number	필수	4031
16	videos[].height	높이	number	필수	3024
17	videos[].resolution	해상도	string	필수	FHD
18	videos[].bit	비트값	string	필수	24bit
19	videos[].pixel	화소	string	필수	4K
20	videos[].color_depth	색심도	string	필수	sRGB
21	videos[].ISO	ISO 감도	string	필수	3200
22	videos[].whith balance	화이트 밸런스	string	필수	5500K
23	videos[].exposure_time	노출시간	string	필수	f2.8 1/80
24	videos[].F-stop	조리개값	string	필수	f2.8
25	videos[].flash	플래시	string	필수	자동
26	videos[].focal_length	초점거리	string	필수	50mm
27	videos[].angle_view	화각	string	필수	46
28	videos[].angle	촬영각도	string	필수	120도
29	videos[].weather	날씨정보	string	필수	맑음

- 장비별 영상(동적/정적) 이미지 참조 항목 라벨링 정보
  - cctv 영상(동적/정적) 이미지 - 공통 참조기준을 바탕으로 어노테이션 항목으로 변환 (예시로서 활용 가능)

※ 국가 시설물과 개인정보 등 정보보안요소와 위치정보들은 비식별화를 통한 재보정 후 적용 필요

No.	속성명	항목 설명	Type	필수여부	단위 (작성예시)
1	videos[].visible_distance	가시거리	number	필수	50M
2	videos[].temperature	온도	number	선택	3℃
3	videos[].humidity	습도	number	선택	32%
4	videos[].coordinates	좌표	number	선택	위경도 정보
5	videos[].cctv_name	CCTV 명	string	필수	광화문4거리 3번CCTV
6	videos[].range	촬영범위(360도 회전 혹은 고정)	string	필수	50
7	videos[].mission	촬영지 분류	string	필수	도심지
8	videos[].event_id	이벤트 분류	string	선택	ABA_0001 (분류_순번)
9	videos[].event_name	이벤트명	string	선택	특이 상황판별
10	videos[].event_name.start_time	이벤트 시작시간	string	선택	2020.11.20. 17:08:15
11	videos[].event_name.end_time	이벤트 종료시간	string	선택	2020.11.20. 17:08:20

- 드론/위성 영상(동적/정적) 이미지

※ 국가 시설물과 개인정보 등 정보보안요소와 위치정보들은 비식별화를 통한 재보정 후 적용 필요

No.	속성명	항목 설명	Type	필수여부	단위 (작성예시)
1	videos[].drone_name	드론명	string	필수	DRN_0001 (분류_순번)
2	videos[].sensor	센서	string	선택	가속도
3	videos[].max_flight	최대비행시간	string	선택	1H
4	videos[].temperature	온도	number	선택	3℃
5	videos[].humidity	습도	number	선택	30%
6	videos[].coordinates	좌표	number	필수	위경도 정보
7	videos[].GPS	GPS 정보(위도, 경도)	string	필수	37.3024, 126.53221
8	videos[].INS	카메라 회전각 정보	string	선택	120, 60, 0
9	videos[].speed	비행속도	number	선택	8km/h
10	videos[].range	촬영범위(360도 혹은 고정, 가시거리), 정사영상 등	string	필수	50M
11	videos[].altitude	촬영고도	number	필수	150M
12	videos[].mission	촬영지 분류	string	필수	산림지
13	videos[].overlap	중복도	string	선택	3M, 3M
14	videos[].GCP	지상기준점(GPS와는 별개로 위치 보정 역할 지정)	string	선택	위경도 좌표

부록 2. 인공지능 학습용 데이터셋 구축 계획서

[서식1] 인공지능 학습용 데이터셋 구축 계획서

인공지능 학습용 데이터셋 구축 계획서

※ 세부데이터별 인공지능 학습용 데이터셋 구축 계획서를 작성함.

세부데이터명	○○○데이터

2021. XX

○○○컨소시엄

목 차

〈일 러 두 기〉

- 이 문서는 구축하는 인공지능 학습용 데이터셋의 세부 구축 방법을 기술하는 문서로, 사업수행계획서 작성 시 '붙임' 문서로 작성하여 제출하셔야 합니다.
  - 목차별로 제공된 '작성요령'과 '인공지능 학습용 데이터셋 구축 안내서'를 참조하여 구체적으로 작성해 주시기 바랍니다.

  ※ 수행기관 및 참여기관은 사업 착수 초기에 전문기관(NIA)의 전문기술 지원 컨설팅 결과를 반영해서 세부 내용을 보완해야 합니다.

- 이 문서는 다음의 목적으로 활용됩니다.
  - '인공지능 학습용 데이터 품질관리 계획' 작성을 위한 기준 문서로 활용됩니다.
  - 인공지능 학습용 데이터 품질 검증 시 참고문서 등으로 활용됩니다.
  - 구축 완료 후 AI Hub를 통해 민간에 데이터 개방 시 함께 제공되는 '인공지능 학습용 데이터 구축 · 활용 가이드라인'을 작성하는 참고자료로 활용됩니다.

제1장 구축내용

1. 구축 배경

〈 작성요령 〉

- 인공지능 학습용 데이터 구축 배경, 필요성, 구축에 따른 활용 방안 등을 제시한다.
- 인공지능 학습용 데이터 구축을 위한 전체적인 설명을 요약하여 작성한다.

【참고】 개요 작성 예시

〈예시 – 문서요약 텍스트 구축〉

- 데이터 구축 필요성
  - 인공지능이 텍스트를 이해하고 핵심 내용을 요약적으로 전달하기 위해서는 인공지능 소프트웨어가(SW)가 해당 텍스트의 주요 내용이 무엇인지를 이해할 수 있는 형태로 가공된, 다양한 유형의 대규모 요약 텍스트 데이터 구축이 필요
  - 국내 인공지능 기반 요약 기술 개발과 관련된 다수의 연구들에서는 해당 텍스트의 제목을 본문의 요약문으로 가정하거나 뉴스 기사의 제목 혹은 첫 문장을 전체 기사의 요약문으로 가정하여 학습데이터로 활용함에 따라 본문 전체의 핵심 내용이나 의미 전달을 온전히 포함하지 못하는 한계점을 내포함
  - 특정 채널에 편향되지 않는 요약기술 개발을 위해서는 채널별로 균형 있는 데이터 원문 수집과 함께, 텍스트 성격에 따라 핵심내용에 영향을 미치지 않는 부분들에 대한 정제 작업이 필수로 요구됨 (중략)

2. 구축 목적

〈 작성요령 〉

- ‘인공지능 학습용 데이터셋 구축 안내서’(이하 ’구축 안내서’라 한다.)의 관련 목차를 참조하여 ‘구축 목적’을 구체적이고, 명확하게 정의한다.
  (**인공지능 학습용 데이터셋 구축 안내서** 제2장 - 1. 데이터 구축 목적 정의 참조)
  - 인공지능 학습용 데이터 구축은 단순히 데이터를 수집하고 생성하는 것이 아니며, 구축한 데이터는 인공지능 모델의 학습에 필요한 데이터로 사용된다는 점을 고려해서 구축 목적을 작성한다.
  - 데이터의 종류, 학습에 필요한 데이터 구축 규모, 데이터 어노테이션 유형, 데이터 품질 수준, 데이터 제작 도구, 작업 및 검사인력 운영 방식 등을 고려해서 작성한다.

제2장 구축 데이터 정의

1. 분석 방법

〈 작성요령 〉

- 제안하는 사업의 목적성에 맞는 인공지능 학습용 데이터 구축을 위해서 다양한 형태의 리서치 및 분석 필요하다. 따라서, 이에 대한 기초적인 사업 유사성과 관련 연구 및 데이터의 정확성과 적합성, 다양성을 위한 분석 내용을 제시한다.

2. 필요 요소 도출

〈 작성요령 〉

- 제안하는 인공지능 학습용 데이터를 구축하는데 필요한 사항(사람, 개체, 물건, 장비, 장소, 전문가 등)을 제시한다.

3. 적용방안(예시)

〈 작성요령 〉

- 본 사업을 통해 구축한 데이터와 인공지능 학습 모델을 활용한 다양한 서비스 모델이나 활용 방안 등을 도식화해서 제시한다.

4. 데이터 명세

4.1. 데이터 포맷 정의

〈 작성요령 〉

- 데이터 구축 목적 정의에 따라 획득하고, 정제하고, 라벨링하는 구축 과정에서의 데이터가 무엇인지를 식별하고, 식별된 데이터의 포맷을 제시한다.
  - 예를 들어, 라벨링 데이터는 JSON, XML 형식으로 제공하며, 획득단계의 원시데이터는 영상데이터로 데이터 포맷은 MP4로 구축 등

4.2. 데이터 속성 정의

〈 작성요령 〉

- **'인공지능 학습용 데이터셋 구축 공통참조기준'**의 공통항목 등을 참조해서 해당 데이터에 맞는 데이터 속성을 도출하고, '속성명', '속성 설명', '데이터 타입', '필수 여부', '예시' 등을 포함해서 작성한다.

4.3. 특성 분류 정의

〈 작성요령 〉

- **'인공지능 학습용 데이터셋 구축 안내서'**의 관련 목차를 참조하여 라벨링 작업 대상 및 범위, 클래스 분류기준 등을 제시한다.

4.4. 라벨링 및 어노테이션 구조 정의

〈 작성요령 〉

- **'인공지능 학습용 데이터셋 구축 안내서'**의 관련 목차를 참조하여 '라벨링 및 어노테이션 구조'을 구체적으로 제시한다.

4.5. 저장 구조 정의

〈 작성요령 〉

- **'인공지능 학습용 데이터셋 구축 안내서'**의 관련 목차를 참조하여 '저장 구조'를 구체적으로 제시한다.
  - 파일을 체계적으로 분류하고 저장하기 위해 데이터 종류 및 분류에 따른 라벨링 데이터 파일 명명법, 파일의 저장 구조, 폴더 구조 등을 작성한다.

제3장 구축 방법

1. 구축환경

1.1. 구축을 위한 작업환경 구성방안

〈 작성요령 〉

- 데이터 획득, 정제, 라벨링을 위한 작업환경, 작업도구(PC 등) 및 시설(전기·통신시설 등)에 대한 세부적인 구성방안을 제시한다.
- 작업환경, 작업도구(PC 등) 및 시설(전기·통신시설 등) 구성 시 접근통제 등 보안을 위한 구체적인 내용을 작성한다.

1.2. 구축 인프라 구성방안

〈 작성요령 〉

- 대량의 데이터를 구축해야 하는 인공지능 학습용 데이터 특성을 고려한 인프라 구성방안을 상세하게 제시한다.
- 인프라 구성 시 개인정보보호 및 보안 등에 대한 사항을 반영하여 구성되도록 해야 한다.

2. 획득방법 (인공지능 학습용 데이터셋 구축 공통참조기준 중심으로)

2.1. 획득 데이터 정의

〈 작성요령 〉

• 원시 데이터 정의
- 인공지능 학습용 데이터 구축에 필요한 원시 데이터 항목을 검토하고, 각 항목 별로 데이터 획득에 필요한 정보(데이터 획득정보, 획득방법, 획득 단계에서 필요한 요건 등)들을 검토하여 제시한다.
- 원시 데이터 대상 및 획득 방법은 육하원칙(5W1H)에 따라 구체적으로 작성한다.
(**인공지능 학습용 데이터셋 구축 안내서** 제2장 - 3.1. 데이터 정의 참조)
- 획득할 원시데이터 내역(원시데이터 종류)에 대한 정의 및 현황정보 등의 사항을 구체적으로 작성한다.
(**인공지능 학습용 데이터셋 구축 안내서** 제2장 - 3.1. 데이터 정의 참조)
- 원시데이터 포맷과 획득 규모는 아래 참고사항을 고려하여 정의한다.

【참고】 크라우드소싱 작업인력 운영 방식 수립 사례

• 원시 데이터 포맷
- 원시 데이터의 파일 형식은 특정 수집 장비 및 처리 도구에 종속되지 않으며, 보편적으로 통용되는 포맷을 활용하도록 한다.
※ hwp, docx 등 특정 워드프로세서에서만 호환되거나 pdf 등 기계 가독이 어려운 포맷은 배제
※ 텍스트 인코딩은 특정 OS, 특정 프로그램이 아닌 보편적으로 활용되는 UTF-8 인코딩을 준수

• 원시 데이터 획득 규모
- 원시 데이터 획득 후 정제, 라벨링, 검사 과정에서 검사기준 미충족으로 제외되는 데이터 수량을 고려하여 구축 목표치 이상의 데이터를 획득하도록 계획에 반영해야 한다.
※ 구체적인 목표치 대비 획득량은 데이터 구축 공정 난이도 및 구축기간 등을 고려하여 설정

2.2. 획득 데이터 특성 분석

〈 작성요령 〉

- 원시 데이터 획득 관련 이슈사항 도출
  - 획득할 원시 데이터의 범위 및 방법을 정의하기 위해 데이터 규모·획득범위·수집처 등과 관련된 세부 이슈사항을 도출하여 구체적으로 제시한다.
- 원시 데이터 획득 시 적합성 검토 및 원시 데이터 선정 시 아래 참고 사항을 반영하여 내용을 작성하도록 한다.

【참고】 원시데이터 적합성 검토 및 선정

- 원시 데이터 적합성 검토
  - 원시 데이터 항목별 데이터 획득 방법, 법적문제 발생가능여부 등을 검토하여 실제로 인공지능 학습용 데이터 구축에 활용할 수 있는 데이터를 선정한다
- 원시 데이터 선정
  - 데이터 품질, 획득 가능성(가능여부 및 획득량), 획득 비용 및 기술수준, 법적 요건 등을 검토하여 획득할 데이터를 최종 선정한다.
  - 선정된 원시 데이터를 획득하기 위해 필요한 정보 또는 원시 데이터 획득현황을 파악하기 위한 데이터 명세서 또는 정의서를 작성하여 데이터 획득 기준으로 활용한다.

[원시 데이터 명세서 작성 예시(문서요약 텍스트)]

데이터 명	문서요약 텍스트 AI 데이터
데이터 포맷	txt(텍스트 파일)
활용 분야	뉴스기사 요약, 법률분서 요약, 사업보고서 요약 등 핵심내용을 신속하고 정확하게 파악할 수 있는 인공지능 요약기술 개발에 활용
데이터 요약	다양한 한국어 원문데이터로부터 정제된 추출 및 생성 요약문을 도출하고 검증한 한국어 문서요약 인공지능 데이터셋

데이터 출처		전국종합일간, 지역종합일간, 경제일간, 스포츠일간, 전문일간, 전문주간 등 60개 언론매체로부터 신문(30만 건), 기고문 (3만 건), 잡지 (1만 건), 법률 (3만 건), 논문 (3만 건)의 원시데이터 확보
데이터 이력	배포버전	TextSummaryAIDataSet_ver1.
	개정이력	신규
	작성자/배포자	수행기관(000)
데이터 통계	데이터 구축 규모	원문 총 40만 건, 요약문 총 80만 건 (추출요약 40만 건/생성요약 40만 건)
	데이터 분포	매체별 분포 : 전국종합일간(45%), 지역종합일간(12.5%), 경제일간(12.5%), 전문일간(12.5%), 잡지(2.5%), 판결 해설문(7.5%), 논문(7.5%) 주제별 분포 : 신문-종합(22.5%), 신문-정치(9.3%), 신문-경제(9.3%)...(중략)... , 기고문(7.5%), 잡지-시사(1.25%), 문화예술(0.75%)...(중략)...
기타 정보	대표성	별도문서 참고
	독립성	
	유의사항	
	관련 연구	해당사항 없음

2.3. 획득 절차 및 항목

〈 작성요령 〉

- 데이터 획득·정제 절차, 데이터 획득 항목, 획득 데이터 저장 및 관리에 대한 사항을 구체적으로 제시한다.
- 데이터 획득·정제 절차 수립 작성
  - 원시 데이터 획득 및 정제 절차 수립 시 데이터 획득 방법별로 명확하게 획득·정제 절차를 정의한다.
  - 또한, 기관간 역할과 책임, 작업자와 관리자 역할과 책임, 행정요소 등 작업을 수행하는 인력관점에서 실질적인 구축작업에 필요한 사항을 종합적으로 고려하여 절차를 수립 한다.
- 데이터 획득항목 정의 작성
  - 획득단계에서 확보해야할 정보를 메타 데이터, 도메인에 대한 항목을 상세하게 정의한다.
    ※ 참고 사항은 텍스트에 대한 예시이며, 텍스트 데이터 획득 시 수집 및 저장할 정보는 '부록1. 인공지능 학습용 데이터셋 구축 공통참조기준(텍스트)'을 준용하여 정의한다.
- 획득 데이터 저장 및 관리 사항 작성
  - 획득 파일에 대한 저장, 전송, 백업 등 관리 절차 및 방안을 수립한다.
  - 획득한 파일을 체계적으로 분류하기 위해 데이터 종류 및 분류에 따른 라벨링 데이터 파일 명명법과 파일 저장구조를 정의하고, 정의된 내용에 맞게 파일을 저장한다.

  (세부적인 사항은 **인공지능 학습용 데이터셋 구축 안내서**를 참조)

【참고】 획득·정제 절차 방안 및 절차 예시 – 이미지

[데이터 획득·정제방안 예시]

	데이터 획득 형태	수집장비	데이터 형식	데이터 처리	담당 인원
1	야외현장 촬영	디지털카메라 (모델명 : 000)	RAW → PNG	데이터 수집 및 라벨링 툴	과제 인력 및 크라우드소싱 인력
2	웹 크롤링	크롤링 서버	여러 이미지 포맷 → JPG	데이터 수집 및 라벨링 툴	000사 크롤링 담당자

【참고】 획득 데이터 저장 방안 수립 예시

- 촬영 진행 후 명동에 위치한 사무실에 복귀하여 촬영 데이터를 노트북 3대에 순차적으로 옮긴 후, 노트북, 클라우드 및 외장하드 활용하여 3중 백업 진행
  - 총 3팀으로 나눠 원시데이터 촬영 진행하기 때문에, 팀마다 1대의 노트북을 활용하여 원시 데이터 백업
  - 스마트폰 외장메모리를 활용하여 원시 데이터 저장하고, 외장메모리에 있는 원시데이터를 파일 복사를 통해 원시 데이터 전송 손실에 대비
  - 복사된 원시데이터를 노트북을 활용하여 노트북, 클라우드, 외장하드에 3중 백업
  - 외장하드 고장에 대비하기 위해 NAS* 등의 추가 장비를 활용하여 주기적으로 백업
- 촬영자 관리 및 촬영분 일일 관리를 위해 직접 촬영 데이터을 “촬영일자”_“촬영자” 기재한 폴더 형태로 구성하여 원시데이터 저장

* NAS(Network Attached Storage) : 네트워크 결합 스토리지

3. 정제 방법

3.1. 원시 데이터 정제 방식

〈 작성요령 〉

- 원시 데이터 정제 프로세스와 정제 기준에 대한 사항을 구체적으로 제시한다.
- 정제 프로세스 수립 작성사항
  - 어노테이션 단계에 들어가기 전에 학습용 데이터로 적합한 데이터를 선별하고, 처리하는 정제 프로세스를 획득방법별로 정의한다.
  - 데이터 정제는 도구(소프트웨어)를 활용하여 정해진 규칙에 따라 제외 또는 변환하는 방법, 작업자가 직접 눈으로 확인하는 검사하는 방법 등을 적용할 수 있다.
- 정제 기준 수립 작성 사항
  - 데이터 구축 목적, 데이터 유형, 도메인 특성에 따른 데이터 정제 기준을 수립한다.
  - 라벨링 단계에서 작업자가 쉽게 라벨링 할 수 있도록, 작업효율 향상을 위해 데이터를 정제할 수 있다.
  - 데이터 획득 담당자가 정제기준에 맞게 데이터를 획득할 수 있도록, 획득 활동 이전에 정제기준을 미리 안내한다.

3.2. 획득 도구 및 정제 도구

〈 작성요령 〉

- 원시 데이터 획득·정제 정제를 위한 도구에 대한 사항을 구체적으로 제시한다.
(**인공지능 학습용 데이터셋 구축 안내서** 2장 - 3.5. 획득 도구 및 정제 도구 참조)

3.3. 획득 / 정제 시 고려사항

〈 작성요령 〉

- 원시 데이터 획득·정제 시 개인정보보호 및 보안, 저작권, 초상권 등 관련 법·제도에 대한 사항, 데이터 다양성 확보, 데이터 편향 방지 및 윤리 준수, 데이터 획득 시 품질 고려사항 등에 대한 세부적인 내용을 구체적으로 제시한다.
(**인공지능 학습용 데이터셋 구축 안내서** 2장 3.6. 획득 시 고려사항 참조)

4. 라벨링 방법

4.1. 데이터 라벨링 방법 및 절차

〈 작성요령 〉

- 획득→정제 과정을 통해 도출된 원천 데이터를 라벨링하여 학습 데이터를 생성하기 위한 과정 및 고려사항을 작성한다. 이 때 인공지능 학습 데이터 구축 목적, 도메인, 활용 분야를 고려하여 방법 및 절차(기준)를 수립한다.
(**인공지능 학습용 데이터셋 구축 안내서** 2장 4절 데이터 라벨링 작업 참조)

4.2. 데이터 어노테이션 포맷과 형식 정의

〈 작성요령 〉

- 어노테이션 포맷 및 저장 형식, 저장구조에 대한 내용을 구체적으로 제시한다.
(**인공지능 학습용 데이터셋 구축 안내서** 2장 - 4.3 데이터 어노테이션 포맷과 형식 정의 및 입력 참조)

4.3. 데이터 라벨링 도구 선정

〈 작성요령 〉

- 원시 데이터 획득·정제 시 개인정보보호 및 보안, 저작권, 초상권 등 관련 법·제도에 대한 사항, 데이터 다양성 확보, 데이터 편향 방지 및 윤리 준수, 데이터 획득 시 품질 고려사항 등에 대한 세부적인 내용을 구체적으로 제시한다.
(**인공지능 학습용 데이터셋 구축 안내서** 2장 3.6. 획득 시 고려사항 참조)

5. 저장 방법

〈 작성요령 〉

- 데이터관리를 위한 관리사항, 조직, 저장 및 백업에 대한 세부적인 내용을 제시한다.
(**인공지능 학습용 데이터셋 구축 안내서** 2장 - 4.4 데이터 라벨링 완료 후 관리 방법 참조)

6. 모델 적용 방안 (최소 2개 이상 비교 분석)

6.1. 모델 선정방안

〈 작성요령 〉

- 문제해결을 위한 인공지능 학습모델 선정 기준과 고려사항 등을 제시한다.
- 인공지능 학습 모델 선정 절차에 대한 세부적인 내용을 제시한다.

6.2. 모델 적합성 검토

〈 작성요령 〉

- 문제해결을 위한 인공지능 학습 모델 후보군을 제시한다. 이 때 후보군은 최소 4개 이상을 제시해야 한다
- 인공지능 학습 모델 선정기준에 따라 후보군에 대한 적합성 등을 검토한다.

6.3. 모델 선정 및 적용방안

〈 작성요령 〉

- 적합성 검토결과 인공지능 학습 모델을 선정한다. 이때 인공지능 학습 모델은 최소 2개 이상이 선정되어야 한다.
- 인공지능 학습 모델 적용을 위한 세부적인 내용을 상세하게 제시한다.

제4장 검사 방법

1. 검사기준요소 정의

1.1. 검사 절차 정의

〈 작성요령 〉

- 검사 절차는 검사 유형, 기준, 방법에 따라 달라질 수 있으며, 검사 절차를 정의할 때는 데이터 사용 목적을 중점으로 진행. 데이터 활용목적에 맞는 검사 규격을 명확히 하여 전체적인 절차를 마련하고 세부적인 내용을 제시한다.
(**인공지능 학습용 데이터셋 구축 안내서** 2장 - 5.1 검사 절차 정의 참조)

1.2. 검사 기준

〈 작성요령 〉

- 인공지능 학습용 데이터에 대한 검사를 위한 절차별 기준 정의하고 내용을 구체적으로 제시한다.
(**인공지능 학습용 데이터셋 구축 안내서** 2장 - 5.2 검사 방식 참조)

2. 단계별 검사 방법

2.1. 수행단계별 검사 항목 : 인공지능 학습데이터 품질관리계획 참조

〈 작성요령 〉

- 인공지능 학습용 구축안내서를 참조하여 수행단계별 검사 기준 및 항목에 대해 구체적으로 제시한다.
(**인공지능 학습용 데이터셋 구축 안내서** 2장 - 5.2 검사 방식 참조)

2.2. 1cycle 중심 검사 방법 (필수)
- AI 학습 모델 선정 근거 제시 요망

〈 작성요령 〉

- 인공지능 학습용 데이터 구축에 대한 절차 및 데이터에 대한 품질확보를 위해 검사에 필요한 적정 데이터 수량을 결정하여 획득·정제·라벨링·AI모델 적용까지 1회 수행하고, 그 결과를 사업수행계획서와 품질관리계획서에 반영하도록 세부적인 계획을 수립하여 제시한다. AI 학습 모델 선정근거 자료를 제시한다.

인공지능 학습용 데이터셋 구축 안내서

2021년 2월 발행

발행처 : 한국지능정보사회진흥원

〈 구축 안내서 개발 참여진 〉	
한국지능정보사회진흥원 고윤석 본부장	한국지능정보사회진흥원 박정은 단장
한국지능정보사회진흥원 오현목 팀장	한국지능정보사회진흥원 유호진 팀장
한국지능정보사회진흥원 박현우 수석	주식회사 티지 서경석 대표
주식회사 케이앤컨설팅 김학철 대표	비트레스 주식회사 류동주 대표
세종대학교 산학협력단 구영현 교수	

〈 자문 위원 〉	
서울시립대학교 이재호 교수	광주과학기술원 이용구 교수
국민대학교 윤상민 교수	이화여자대학교 민동보 교수
국립암센터 황보율 팀장	NHN다이퀘스트 주식회사 김경선 박사
주식회사 리노스 정휘웅 소장	주식회사 딥네츄럴 박상원 대표
셀렉트스타 주식회사 신호욱 대표	주식회사 소리자바 김창환 이사
슈퍼브에이아이 주식회사 이현동 이사	주식회사 스위트케이 이준호 소장
주식회사 에이모 이승택 부사장	주식회사 크라우드웍스 김대영 이사
주식회사 테스트웍스 금호영 이사	주식회사 포티투마루 손아림 이사

■ 본 구축 안내서는 지능정보산업 인프라 조성을 위한 인공지능 학습용 데이터 구축 사업 중 'AI 학습용 데이터 품질관리체계 및 공통기준 가이드라인' 용역 사업의 결과 산출물입니다.

▶ 한국지능정보사회진흥원 지능데이터본부
전화번호 (02)6747-2179 / phw@nia.or.kr (박현우 수석)