

21년도 인공지능 학습용 데이터 구축 가이드라인

< 내용량 손글씨 OCR >

인공지능 데이터 구축	사업 총괄	 쇼우테크
	데이터 설계	 쇼우테크
	데이터 수집 및 정제	 쇼우테크
	데이터 가공 (라벨링, 어노테이션)	 쇼우테크
	데이터 검수 (자체 검수)	 쇼우테크
	클라우드 소싱	 쇼우테크
	저작도구 개발	 쇼우테크
	AI모델 개발	nhn
가이드라인 작성	 쇼우테크	황강연
가이드라인 버전	ver 1.0 ('22. 1. 30)	

목 차

1. 데이터 명세 정보 1

1.1 데이터 정보 요약	1
1.2 데이터 포맷	1
1.3 어노테이션 포맷	2
1.4 데이터 구성	3
1.5 데이터 통계	4
1.6 원시데이터 특성	5
1.7 기타 정보	5

2. 데이터 구축 가이드 6

2.1 데이터 구축 개요	6
2.2 문제정의	6
2.3 수집·정제	6
2.4 어노테이션/라벨링	7
2.5 검수	10
2.6 활용	10

1. 데이터 명세 정보

1.1 데이터 정보 요약

데이터 이름	대용량 손글씨 OCR	
활용 분야	기존 OCR의 한계를 극복한 손글씨 OCR 인식 서비스 개발 (병원 의무기록 및 의료비처리, 회계 감사와 증빙자료 OCR 서비스)	
데이터 요약	일상생활에서 많이 사용하는 손글씨 데이터를 종이와 태블릿으로 구분하여 수집 및 가공을 거쳐 인공지능 학습용 데이터 구축	
데이터 출처	일반인이 많이 사용하는 민원서식과 자주 쓰는 단어위주의 자유필사 데이터를 성별, 연령대별로 구성된 클라우드워커에게 제공하여 손글씨 필사본 회수	
데이터 이력	배포버전	ver 1.0
	개정이력	ver 1.0 (주)쇼우테크 황강연 신규 ('22.01.10)
	작성자/ 배포자	(주)쇼우테크 황강연

1.2 데이터 포맷

이미지 데이터(PNG)

JSON 형식

```

{
  "Annotation": {
    "object_recognition": 0,
    "text_language": 0
  },
  "Dataset": {
    "category": 0,
    "identifier": "HW_OCR_53",
    "label_path": "HW_OCR/4.Validation/Paper(R,Free)",
    "name": "대용량 손글씨 데이터셋",
    "src_path": "HW_OCR/4.Validation/Paper(R,Free)",
    "type": 1
  },
  "Images": [
    {
      "acquisition_location": "자판",
      "application_field": "기타",
      "background": 0,
      "data_captured": "2021.10.05",
      "height": 3503,
      "identifier": "HW_OCR_53_4P8_17040",
      "media_color": "black",
      "media_type": 0,
      "pen_color": "black",
      "pen_type": 0,
      "type": "img",
      "width": 2475,
      "writer_age": 36,
      "writer_sex": 0,
      "written_content": 1
    },
    {
      "data": "017040",
      "id": 1,
      "x": [1882, 1882, 2179, 2179],
      "y": [139, 220, 139, 220]
    },
    {
      "data": "대구광역시",
      "id": 3,
      "x": [299, 299, 659, 659],
      "y": [1020, 1141, 1020, 1141]
    }
  ]
}

```

1.3 어노테이션 포맷

1.3.1 라벨링 규격

No.	속성명
1	Dataset
1-1	Dataset.identifier
1-2	Dataset.name
1-3	Dataset.src_path
1-4	Dataset.label_path
1-5	Dataset.category
1-6	Dataset.type
2	Images
2-1	Images.identifier
2-2	Images.type
2-3	Images.width
2-4	Images.height
2-5	Images.background
2-6	Images.pen_type
2-7	Images.pen_color
2-8	Images.distortion
2-9	Images.clearness
2-10	Images.noise
2-11	Images.acquisition_location
2-12	Images.media_type
2-13	Images.application_field
2-14	Images.writer_age
2-15	Images.writer_sex
2-16	Images.data_captured
2-17	images.written_content
3	Annotation
3-1	Annotation.object_recognition
3-2	Annotation.text_language
4	BBox
4-1	BBox[.].id
4-2	BBox[.].text
4-3	BBox[.].x[.]
4-4	BBox[.].y[.]

1.4 데이터 구성

1.4.1 저장구조 정의

1레벨	설명	2레벨	설명	3레벨	설명
1.Raw	원시 데이터	P.Paper	종이	R.Free	자유필사형
				O.Form	정보제공형
		T.Tablet	태블릿	R.Free	자유필사형
				O.Form	정보제공형
2.Source	원천 데이터	P.Paper	종이	R.Free	자유필사형
				O.Form	정보제공형
		T.Tablet	태블릿	R.Free	자유필사형
				O.Form	정보제공형
3.Annotation	라벨링 데이터	P.Paper	종이	R.Free	자유필사형
				O.Form	정보제공형
		T.Tablet	태블릿	R.Free	자유필사형
				O.Form	정보제공형
4.Validation	검수 데이터	P.Paper	종이	R.Free	자유필사형
				O.Form	정보제공형
		T.Tablet	태블릿	R.Free	자유필사형
				O.Form	정보제공형
5.Failure	반려 데이터	P.Paper	종이	R.Free	자유필사형
				O.Form	정보제공형
		T.Tablet	태블릿	R.Free	자유필사형
				O.Form	정보제공형
6.Compensate	수정 데이터	P.Paper	종이	R.Free	자유필사형
				O.Form	정보제공형
		T.Tablet	태블릿	R.Free	자유필사형
				O.Form	정보제공형

1.4.2 파일 명명규칙 정의



1.5 데이터 통계

1.5.1 데이터 구축 규모

항목	데이터량
정보제공	61,695장
자유필사	44,824장
합계	106,519장

- 정보제공형 : 공공기관 및 산업에서 활용하는 서식
- 자유필사형 : 자유형태의 손글씨 데이터 및 목민심서 필사본, 시험답안지, 노트필기류 등

1.5.2 데이터 분포

1.5.2.1 수집 매체별 분포

항목	
종이	
태블릿	
합계	

1.5.2.2 내용별 분포

항목	
정보제공	
자유필사	
합계	

1.5.2.3 작성자의 성별 분포

항목	
여성	
남성	
불명	
합계	

1.5.2.4 작성자의 연령 분포

항목	데이터량	비율
10대 미만(~9)	1,275장	1.20%
10대(10~19)	10,100장	9.48%
20대(20~29)	30,472장	28.61%
30대(30~39)	22,363장	20.99%
40대(40~49)	29,863장	28.04%
50대 이상(50~)	12,446장	11.68%
합계	106,519장	100%

1.5.3 기타 활용 통계

해당없음

1.6 원시데이터 특성

1.6.1 대상분류

실제 데이터

일상생활에서 많이 사용하는 단어를 선정하여 손글씨로 제작

1.6.2 제약조건

제약있음

손글씨 데이터의 편향성 방지를 위하여 일상생활에서 많이 사용하는 단어를 선별하여 클라우드 워커에게 제공하여 작성

1.6.3 속성

1.6.3.1 종이부문 원시데이터

회수된 종이 문서를 300dpi이상, 24bit 컬러 이미지로 취득

1.6.3.2 태블릿부문 원시데이터

태블릿에서 작성한 원시데이터를 이미지로 Output(300dpi이상)

1.7 기타정보

1.7.1 포괄성

//데이터가 특정 모집단의 실제 특성을 어느 정도를 표현하고 있는지 작성 ex) 전체 지역 범위 중 몇 개의 도시 대상

1.7.2 독립성

//데이터가 원시데이터에 의존하고 있는 사항이 있는 확인하여 표기 ex) 법률 개정 ex) 민감 정보, 법적 문제 등

1.7.3 유의사항

//데이터 배포 시 파급효과, 데이터 활용 시 유의사항 요약 설명

1.7.4 관련 연구

- 해당없음

2. 데이터 구축 가이드

2.1 데이터 구축 개요

구분	데이터유형
데이터획득	종이
	태블릿
데이터정제	종이
	태블릿
데이터가공	종이
	태블릿
데이터검수	종이
	태블릿

2.2 문제정의

2.2.1 임무 정의

2.2.1.1 추진 배경

- 현대의 많은 정보는 만들어질 때부터 디지털 데이터로 만들어지지만, 아직도 일선 관공서 나 민간의 많은 부분에서는 서로 다른 수많은 양식에 일일이 수기 작성하여 제출하는 자료가 많이 있음
- 이들 자료를 디지털로 변화기 위해서는 사람이 별도의 작업이 필요로 함
- 현재 OCR 기술은 영문이나 한글 인쇄체에 대해서는 높은 인식률을 보이고 있으나 손글씨에 대해서는 인식률이 높지 않은 수준인데 이를 극복하기 위한 대용량 손글씨 학습데이터 구축으로 인공지능 기술에 기반한 OCR 기술 개발이 필요함

2.2.1.2 추진 목적

- 공공 및 민간의 다양한 문서 양식의 OCR 인식을 위한 손글씨 데이터의 확보와 다양한 기관 및 환경에서 생산되는 자유필사 손글씨 데이터의 확보로 산업 및 일상생활 전반에 적용될 수 있는 OCR 개발을 위한 각종 손글씨 OCR 데이터 구축의 필요
- 손글씨는 성별, 연령별 뿐만 아니라 개개인마다 글자체의 개성이 두드러져 필적학이라는 심리학적 필체 연구도 행해지고 있으므로 대량의 손글씨 학습데이터를 다양하게 확보하여 인공지능 기술을 접목하여 OCR 인식성능의 고도화
- 산업 및 실생활 전반에서 사용하는 손글씨 데이터를 확보하여 인공지능 기반 한글 손글씨 인식(OCR) 학습 데이터 구축과 인식모델을 개발함으로써 비약적으로 발전하고 있는 최신

ICT 기술이 적용된 DB 구축 솔루션의 개발 토대를 마련하여 관련 분야의 변화와 혁신을 주도

2.2.2 데이터 구축 유의사항

2.2.2.1 민감정보의 비식별화

- 수집데이터에 포함된 민감정보(개인정보 등)에 대한 비식별화 처리를 통하여 민감정보의 배포가능성의 원천 차단
- 학습데이터의 수집단계부터 개인정보가 포함되지 않은 익명정보의 사용으로 개인정보 수집을 사전에 차단하고, 데이터 제공협약을 통하여 수집되는 원시데이터에 포함된 개인정보는 비식별화 처리하여 민감정보의 유출 방지

2.2.2.2 법·제도적인 검토 방안

- 개인정보보호법 및 저작권, 초상권 관련 관계 법령 준수 및 법률자문을 통하여 위배사항이 발생되지 않도록 데이터 수집 및 정제

2.3 수집·정제

2.3.1 원시데이터 선정

2.3.1.1 원시데이터 특성

구분	내용
구조	종이, 태블릿 / 정보제공형, 자유필사형
형태	이미지 데이터
포맷	png
출처	<ul style="list-style-type: none"> 정보제공형 : 민원24에서 많이 사용하는 민원서식 자유필사형 : 데이터 제공협약을 체결한 공공기관, 대학교 및 자체제작

2.3.2 수집·정제 절차

2.3.2.1 데이터 선정

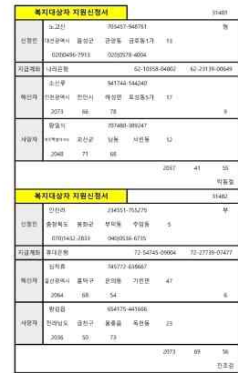
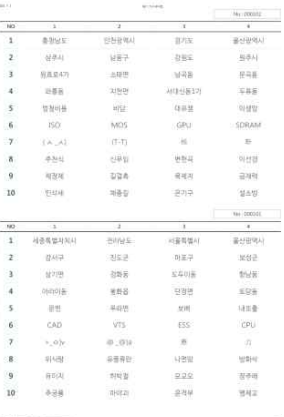
구분	내용
정보제공형	<ul style="list-style-type: none"> 민원24에서 많이 사용하는 민원서식 선정
자유필사형	<ul style="list-style-type: none"> 데이터제공협약 기관(김해시청, 창원시청, 부산경상대학교)에서 제공하는 자유필사형 데이터 수집 <ul style="list-style-type: none"> 공무원 목민심서 필사본 논술형 시험답안지 노트필기 자체 제작 데이터시트 <ul style="list-style-type: none"> 손글씨로 많이 사용되는 인명, 주소, 연락처 등의 단어 조합 데이터 생성

2.3.2.2 데이터 시트 제작 및 배포

- 성별, 연령대별로 클라우드워커 모집





구분	내용
성별	<ul style="list-style-type: none"> 남성, 여성 비율 50:50
연령대	<ul style="list-style-type: none"> 10대:20대:30대:40대:50대 비율 10:30:20:30:10

- 데이터 유형별 작성 가이드 및 데이터 시트 배포

구분	내용
정보제공형 데이터시트	
자유필사형 데이터시트	



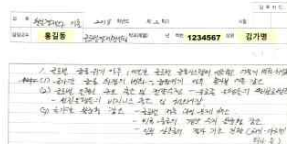
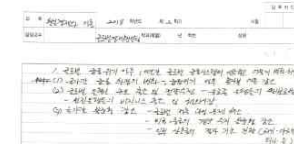
2.3.2.3 데이터 시트 회수 및 스캐닝

- 클라우드워커가 작성한 손글씨 데이터시트 우편 회수
- 회수된 데이터시트의 디지털화(스캔)

우편회수	디지털화(스캔)
 	 

수집·정제기준	예시
이미지 품질 불량	글자 식별불가, 접혀서 일부만 표시 등
이미지 중복	동일한 데이터시트의 중복
민감정보 포함	개인정보가 포함되어 있는 경우
정량적 목표	유형별 목표 수량에 맞게 이미지 확보 (종이/테블릿, 정보제공/자유필사)

2.3.2.4 수집데이터 정제

구분	정제전	정제후
품질검사		
민감정보 비식별화		

• 원시데이터의 내용이 잘리거나 알아볼수 없는 경우 재스캔
 • 수집 데이터의 포맷(png, 24bit 컬러) 기준에 미달시 재스캔
 • 원본상태의 불량(잘림 등)인 경우 수집대상에서 제외 처리

• 원시데이터에 포함된 민감정보(개인정보 등)에 대한 비식별화 처리
 • 비식별화 대상이 대량으로 포함된 경우 수집 대상에서 제외 처리(삭제 및 원시데이터 폐기)

2.3.3 수집·정제 기준



2.3.4 수집·정제 조직(필요 시 작성)

//수집 조직은 원시데이터를 수집하는 조직의 구성과 각 구성원의 역할별 책임과 권한을 구분하여 작성합니다.

//조직도를 삽입하여 표현하면 작성이 용이합니다.

//데이터를 절차에 맞게 수집하기 위해 실시하는 교육 및 훈련 계획을 작성합니다.


2.3.5 수집·정제 도구(필요 시 작성)




구분	내용
데이터 수집도구 (워크크리에이터)	
데이터 정제도구 (Panasonic ICP*)	

(* Panasonic ICP : 파나소닉 스캐너 번들 소프트웨어)

2.4 어노테이션/라벨링


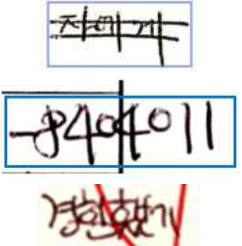
2.4.1 어노테이션/라벨링 절차

순서	구분	내용
1	작업대상조회	

2	어노테이션 (바운딩박스)		<ul style="list-style-type: none"> 작업대상 데이터를 선택하여 손글씨 영역의 바운딩 박스
3	어노테이션 (라벨링 텍스트)		<ul style="list-style-type: none"> 바운딩박스영역에 해당하는 라벨링 정보 등록
4	정보저장		<ul style="list-style-type: none"> 해당 이미지의 어노테이션이 완료된 경우 작업 정보 저장

2.4.2 어노테이션/라벨링 기준

2.4.2.1 어노테이션 기준

구분	예시	상세내용
바운딩박스		<ul style="list-style-type: none"> 손글씨로 작성된 영역에 바운딩박스 후 바운딩박스의 내용을 라벨링 텍스트 등록
라벨링 텍스트	<input type="text" value="제주특별자치도"/>	
제외대상		<ul style="list-style-type: none"> 손글씨 작성 영역중 다른 글자, 양식과 겹쳐져 육안으로 판독이 어려워 오독의 위험성이 높은 경우 어노테이션 제외

2.4.2.2 어노테이션 예시

구분	예시	상세내용
----	----	------

올바른 경우	
오류로 인한 반려 대상	

2.4.3 어노테이션/라벨링 조직 (필요 시 작성)

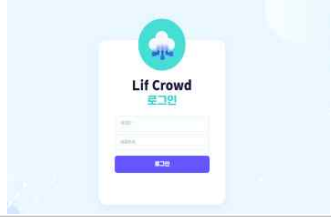

//데이터를 가공하는 조직의 구성과 각 구성원의 역할별 책임과 권한을 구분하여 작성합니다.

//조직도를 삽입하여 표현하면 작성이 용이합니다.

//데이터를 절차에 맞게 가공하기 위해 실시하는 교육 및 훈련 계획을 작성합니다.

2.4.4 어노테이션/라벨링 도구 (필요 시 작성)

2.4.4.1 종이부문 라벨링 도구

순서	구분	화면구성
1	클라우드플랫폼 로그인	
2	업무선택	
3	데이터라벨링	

2.5.4 검수 도구 *(필요 시 작성)*

//검수과정에서 사용하는 자동화 검수도구들을 설명 합니다. 어떤 도구를 사용하는지, 해당 도구의 어떤 기능을 활용하여 어떤 결과를 얻는지 등을 작성합니다. 검수를 통해 걸러지는 오류의 예시를 제시하여 사용자의 이해를 도울 수 있습니다.

2.5.5 기타 품질관리 활동 *(필요 시 작성)*

//전문성을 요하는 데이터(번역, 의료 등)의 외부 전문가의 별도검수 내용, 데이터의 구축 공정 외에 따로 관리하는 품질 지표, 데이터 구축 중 품질 향상을 위해 진행한 활동 등을 여기에 작성합니다.

2.6 활용

2.6.1 활용 모델

2.6.1.1 모델 학습

//데이터를 학습시키기 위한 기본 학습모델(알고리즘)을 제시하고 그 구조를 기술합니다.

//위에서 제시한 학습 모델에 구축한 데이터를 학습시키기에 적당한 권장 학습 분배량(학습 : 검증 : 평가) 및 학습 적용 방법을 안내합니다. 사용자가 안내를 따라 공개 모델을 학습시킬 수 있도록 학습 환경 및 파라미터 설정 등을 자세히 설명합니다.

2.6.1.2 서비스 활용 시나리오

//모델학습을 완료한 알고리즘/ 네트워크를 활용한 응용서비스 및 서비스 활용 시나리오를 설명합니다.

2.6.2 데이터 제공

//인공지능 데이터를 다운받기 위해 어떤 절차를 거쳐 제공되는지, 사용자가 다운받기 위한 자격 및 경로를 안내합니다.

//특히 보안을 요하는 데이터의 경우, 합법적인 데이터 배포방안과 그 접근수단을 확보하여 안내합니다.

2.6.3 데이터 유지보수 *(필요 시 작성)*

//인공지능 데이터가 지속성은 있는지, 데이터에 대한 피드백은 어떻게 처리하는지 등을 작성합니다.

//사용자가 발견할 수 있는 오류를 수정할 것인지, 수정한 데이터를 재배포할 계획이 있는지, 규모를 확장하여 다음버전을 제공하는지 등의 유지보수 계획을 작성합니다.