●○ OCR 과제

공공행정문서 OCR

●○ 개요: 공공행정문서OCR 데이터셋

기존 한글의 광학글자인식(OCR)의 성능은 활자체(인쇄체)로 구성된 문자 단위의 기계적인 구조 학습으로 글자의 판독 결과를 제공하는 분야에서 벗어나 공공행정문서에 포함된 다양한 형태의 한글 글자체(손글씨, 인쇄체 등)의 학습과 공공행정문서에 포함된 공공행정 용어의 학습으로 행정용어의 문자 인지 서비스를 제공함으로서 다양한 정부 공공 행정문서의 OCR 인식률을 높이고자 공공행정문서에 특화된 문자 인식 AI 모델 개발로 공공행정문서의 포함된 문자 인식 모델을 개발을 목표로 하고 이를 통하여 공공행정에 특화된 용어의 인식, 공공기록물의 판독 서비스 등에 활용할 수 있다.

그림1 | 공공행정문서의 판독 서비스

●○ 데이터셋의 구성

본 데이터셋은 다양한 형태가 포함된 공공행정문서 900,000면과 식별가능한 단어 1,500만단어 이상의 원시 데이터를 수집하여 공공행정문서에 특화된 데이터셋을 제작을 위하여 데이터 제공기관에서 제공하는 원본 기록물, 디지털 이미지를 수집하여 원시데이터를 구성하며,

원시데이터에 포함된 다양한 행정용어의 어노테이션 정보와 공공행정문서의 추가 메타정보(생산기관, 출처, 생산연도 등)를 포함한 JSON 파일을 제작하여 AI학습용 데이터셋을 구성한다.

●○ 데이터셋의 설계 기준과 분포

공공행정문서의 특성상 민감정보(개인정보 등)를 포함한 정보가 다수 포함되어 공개될 경우 개인정보를 침해하는 사례가 발생될 수 있어 원시데이터의 획득 단계부터 데이터 제공기관과의 긴밀한 협의를 통하여 공개 가능한 공공행정문서의 획득을 통하여 비공개 대상 정보가 공개되지 않도록 원시 데이터를 선별한다. 또한, 데이터셋의 선별, 정제, 검수 단계별로 민감정보(개인정보 등)에 대한 비식별화 공정을 추가함으로써 개인정보 보호를 위한 활동을 추가하여 데이터셋을 제작한다.

본 데이터셋은 다양한 형태가 포함된 공공행정문서 900,000면을 대상으로 원시 데이터를 수집하여 데이터셋을 구성하며, 공공행정문서에 포함된 행정용어, 포함된 문자(단어)의 유형, 글자체의 유형에 따라 아래와 같은 기준으로 공공행정문서 원시 데이터를 수집한다.

이를 위하여 공공기관(지방자치단체)의 고유 사무를 중심으로 국민 생활과 밀접한 연관관계가 높은 주요 행정업무를 선정하고, 학습 데이터의 다양성을 위하여 생산연대별로 구분하여 데이터셋을 구성한다.

- 공공행정문서의 유형별 선별 기준

데이터 유형	포함 내용
공공행정업무 유형	일반행정, 인허가, 상하수도, 도시계획 등 지방자치단체의 주요사무 10여종 선정
문자형태	수기체, 타자체, 인쇄체를 포함하는 공공행정문서 (시대별 흐름에 따른 다양한 생산연도에 생산된 기록물)

공공행정문서 보유 현황 분석을 통한 시대별 문서 분포도

구분	1980년대	1990년대	2000년대	2010년대	계
지방자치단체 A	6.3%	30.4%	56.5%	5.8%	100%
문자형태	수기체 타자체	타자체 수기체	인쇄체 위주	인쇄체 위주	

행정업무 유형에 따른 원시 데이터 수집 계획(약 10개 이상의 지방자치단체 고유 사무)

- 일반행정문서 : 10%
- 인허가문서 : 30%
- 주민자치문서 : 20%
- 지방도시계획문서 : 10%
- 지역개발문서 : 20%
- 기타 : 10%

●○ 데이터 구조

데이터셋에 포함되는 원시 데이터는 공공행정문서 디지털 이미지(TIF, JPG)와 어노테이션 정보가 수록된 JSON파일 셋트로 구성한다.

JSON파일에 포함된 데이터의 포맷 기준을 아래 테이블과 같이 구성한다.

구분	항목		길이	타입	필수여부	비고
	한글명	영문명				
데이터셋	데이터셋명	info.name			Y	
	데이터셋상세설명	info.description	1,000	string		
	데이터셋 생성일자	info.date_created	100	string	Y	
이미지정보	이미지 생산기관	image.make_code	7	number	Y	
	이미지 생산연도	image.make_year	4	number	Y	
	이미지 카테고리	image.cartegory	100	string	Y	
	이미지 파일명	image.file_name	100	string	Y	
	이미지 높이	image.height	4	number	Y	
	이미지 너비	image.width	4	number	Y	
	이미지 해상도	image.dpi	3	number	Y	
	이미지 생성일자	image.date_created	100	string		
	이미지라이선스	image.licence	100	string		
어노테이션정보	어노테이션 타입	annotation.type	100	list	Y	
	어노테이션 바운딩박스	annotation.bbox	100	list	Y	
	어노테이션 텍스트	annotation.text	1,000	string	Y	
라이선스정보	라이선스명	licence.name	100	string		
	라이선스상세설명	licence.description	1000	string		
	라이선스url	licence.url	200	string		

●○ 데이터 예시

본 데이터는 이미지의 추가 메타정보와 어노테이션의 바운딩박스 정보, 라벨링 정보를 포함한 데이터를 저장하며 아래의 예시와 같은 데이터 구조를 가진다.

```
{
    "images": [
        {
            "image.make.code": “1234567”,
"image.make.year": 1997,
“image.category”: ”일반행정”
“image.width": 2492,
            "image.height": 3500,
            "image.file.name": "1234567-1997-0001-0001.jpg",
            "image.create.time": "2020-09-08 11:01:38"
        }
    ],
    "annotations": [
        {
            "id": 1,
            "annotation.type": rectangle,
            "annotation.text": "건축허가",
            "annotation.bbox": [
                1088,
                355,
                261,
                98
            ]
        },
        {
            "id": 2,
            " annotation.type": rectangle,
"annotation.text": "도시건축과장",
            "annotation.bbox": [
                410,
                776,
                241,
                74
            ]
        },}, ...]}
```

●○ 데이터 구축 과정

구축단계	세부절차		세부설명
데이터 획득	수집 대상 선정	필수	• 공공행정문서의 생산연대별, 유형별 원시데이터 수집 대상 선별 • 공공행정문서 - 면수 기준 : 900,000면 - 단어 기준 : 15,000,000단어 (1면당 15개 이상의 단어)
	학습대상 객체 정의	필수	• 공공행정문서에 포함된 주요 행정정보 객체 정의 (발신기관명, 문서번호, 제목 등) • 항목별 입력 내용의 유형 정의 • 예외사항에 대한 처리 절차 정의
	어노테이션 툴 개발	필수	• 학습 대상 정보에 대한 정의서를 기준으로 툴 제작
	매뉴얼 작성	필수	• 공공행정문서에 대한 유형별 이해할 수 있는 매뉴얼 작성 • 어노테이션/라벨링 작업을 위한 매뉴얼 작성
데이터 정제	데이터 확보	필수	• 공공기관 행정문서의 디지털 이미지 확보 • 원문 상태의 행정문서의 디지털화(국가기록원 이미지 디지털화 기준 준수)
	데이터 선별 및 분류	필수	• AI 학습에 적합한 디지털 이미지 선별 • 행정용어, 자연어 기준 1면당 15단어 이상 • 생산연대별 행정 절차의 파악이 용이한 공공 행정문서의 선별 • 수기체, 타자체, 활자체 등의 유형별 분류 및 선별 • 민감정보(개인정보 등)에 대한 비식별화
어노테이션/ 라벨링	boundig box	필수	• 유형별 수집 항목에 대한 bounding box
	라벨링	필수	• bounding box 항목에 대한 라벨링
데이터 검수	1차 전수검사	필수	• 어노테이션/라벨링이 완료된 데이터에 대한 전수검사
	2차 샘플링 검사	필수	• 대상문서의 약 10% 데이터에 대한 2차 샘플링 검사 • 동일 데이터에 대하여 2개 조직으로 구성하여 동일 데이터 이중 검사
	3차 샘플링 검사	필수	• 기록물관리전문요원에 의한 샘플링 검사

●○ 검수와 품질 확보

구축 완료되는 공공행정문서OCR 학습 데이터셋 900,000장의 고품질을 유지하기 위하여 3단계에 걸친 품질검사 단계를 확보하여 품질검사를 수행한다.

- 1단계 품질검사는
  - 현장 일반 검수자가 제작된 원시 데이터셋 전량 전수검사를 실시하여 데이터셋의 오류 사항을 점검하고

- 2단계 품질검사는
  - 약 10%의 샘플링 발췌하여 동일 이미지에 대하여 2개의 2차 검수자 조직을 구성하여 동일 데이터의 이중검사를 실시하여 일치 여부를 확인한다.
- 3단계 품질검사는
  - 전체 데이터셋의 약 5%를 선별하여 전문 품질검사요원(기록물관리전문요원 등)을 투입하여 데이터의 품질검사를 수행한다.

각 단계별 품질검사는 어노테이션 정보의 품질관리를 포함하여 공공행정문서에 포함된 비공개정보(개인정보를 포함한 각종 민감정보)의 비식별화 대상의 마스킹(Masking) 상태를 점검하여 민감정보의 유출을 사전에 차단하는 품질관리 활동을 수행하여 최종 산출물인 데이터셋의 품질을 보장한다.

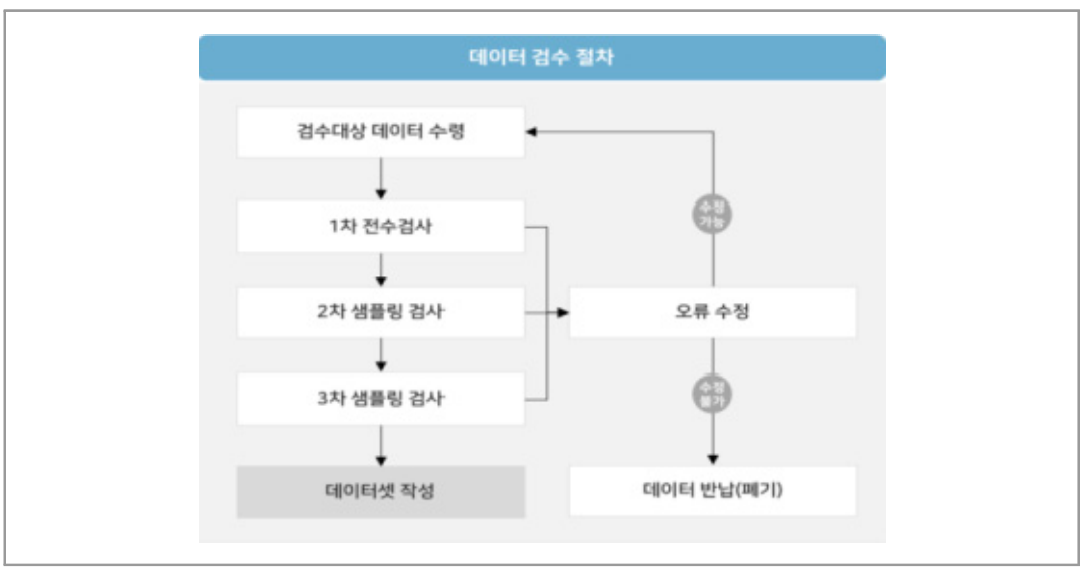

그림2 | 품질확보를 위한 3단계 품질 검사