<table>
<tr><td rowspan="8">메타테이블 정보 (다중기입가능)</td><td>분야</td><td>데이터 유형</td><td>구축 데이터량</td><td>원천데이터 형식</td><td>라벨링 형식</td><td>라벨링 유형</td></tr>
<tr><td>NIA 기입</td><td>텍스트</td><td>102,300,332 어절</td><td>txt</td><td>json</td><td>텍스트</td></tr>
<tr><td>데이터 출처</td><td>데이터 구축년도</td><td>구축기관 (총괄)</td><td>가공기관</td><td>검수기관</td><td></td></tr>
<tr><td>도서, 특허, 판결문 자체 수집</td><td>2022년</td><td>고려대학교 산학협력단</td><td>고려대학교 산학협력단</td><td>㈜나라지식정보</td><td></td></tr>
<tr><td rowspan="2">데이터 문의처</td><td>기관명</td><td>문의담당자명</td><td>전화번호 (유선전화번호기입)</td><td>메일주소</td><td></td></tr>
<tr><td>고려대학교 산학협력단</td><td>김종호</td><td>02-919-8565</td><td>mecey@naver.com</td><td></td></tr>
<tr><td>데이터 소개</td><td colspan="5">저작재산권 이용허락을 확보한 의료 및 법률 분야 핵심 전문 도서와 분야 전문성과 최신성을 갖춘 특허(의료), 판결문(법률)을 기반으로 1억 어절의 데이터셋을 구축</td></tr>
<tr><td>주요키워드</td><td colspan="5">의료, 법률, 말뭉치 데이터, 자연어처리, 인공지능</td></tr>
<tr><td colspan="2">카테고리 정의서</td><td colspan="5">별첨</td></tr>
</table>

<table>
<tr><td rowspan="2">데이터셋 명</td><td>국문</td><td>의료, 법률 전문 서적 말뭉치</td></tr>
<tr><td>영문</td><td>Medical and Legal Specialized Book Corpus</td></tr>
<tr><td colspan="2">구축목적</td><td>전문 분야 한국어 말뭉치 표본을 제공하고 자연어 처리 산업 활성화와 의료 및 법률 분야 인공지능 솔루션 개발 확대에 기여하고자 함</td></tr>
<tr><td colspan="2">활용서비스</td><td>ㅇ 의료 및 법률 분야에 대해 정보검색엔진 성능 개선
ㅇ 의료 및 법률 분야에 대한 질의응답 및 상담 서비스 개발
ㅇ 의료 분야에 대한 진료과 추천 서비스 시스템 개발
ㅇ 법률 분야에 대한 판례 요약 서비스 플랫폼 개발
ㅇ 법률 분야에 대한 기업 불공정 계약 탐지 시스템 개발
ㅇ 의료 및 법률 인공지능 연구개발 분야에 대해 한국어 인공지능 성능 평가용 벤치마크 개발</td></tr>
<tr><td colspan="2">소개</td><td>ㅇ 의료 및 법률 분야의 다양한 주제를 포괄하고, 전문적인 언어 사용 양상이 반영할 수 있는 핵심적인 한국어 전문 도서에 대한 인공지능 학습용 데이터 활용 라이선스 확보함.
ㅇ 의료 및 법률 분야의 핵심적인 최신 전문 도서를 대상으로 저작물의 디지털 파일을 수집하고 말뭉치 원시 자료를 정제 및 가공함. 또한, 의료 전문가, 법률 전문가, 언어 전문가의 전문 지식을 통해 체계화된 말뭉치 구축 형식을 확보함.
ㅇ 인공지능 학습을 위해 구축된 말뭉치를 효율적으로 활용할 수 있도록 자세한 메타 정보를 선정하여 정밀하게 라벨링 된 인공지능 학습용 데이터세트 구축함. 데이터 세트 구축은 크라우드 소싱을 활용하여 대규모 학습 데이터를 구축하고, 크라우드 소싱을 원활하게 진행하기 위한 인력관리 플랫폼을 자체적으로 개발, 도입함.
ㅇ 구축된 의료 및 법률 분야 인공지능 학습용 전문 도서 말뭉치 데이터세트를 기반으로 다양한 하위 작업(downstream task)용 인공지능 개발을 위한 프로토타입 모델을 개발함.</td></tr>
<tr><td colspan="2">데이터셋 통계 (구축 규모 및 분포)</td><td>1. 데이터 구축 규모
- 원천데이터 : 저작재산권 이용허락을 확보한 의료 및 법률 전문 도서와 특허(의료), 판결문(법률)을 기반으로 126,512건 원천데이터 구축
- 라벨링데이터 : 확보된 원천데이터로 총 102,300,332어절 구축</td></tr>
</table>

언어	구분	원천데이터	라벨링데이터
한국어	의료	46,883건	49,217,568어절
	법률	79,630건	53,082,764어절
	합	**126,513건**	**102,300,332어절**

1.1 개방된 파일 및 데이터 건수 대비

구분	Training	Validation	Test	합	Sample
원천 데이터	101,211개	12,651개	12,651개	126,513개	100개
	81,888,182어절	10,174,608어절	10,237,542어절	102,300,332어절	104,321 어절
라벨링 데이터	2개	2개	2개	6개	2개
	81,888,182어절	10,174,608어절	10,237,542어절	102,300,332어절	104,321 어절

2. 데이터 분포

2.1 문장 어절 수 분포

어절 구간	원천데이터 수	비율
10 어절 미만	1	0.00%
10 어절 ~ 20 어절	10	0.01%
21 어절 ~ 30 어절	27	0.02%
31 어절 ~ 40 어절	33	0.03%
41 어절 ~ 50 어절	61	0.05%
51 어절 ~	126,381	99.90%
합계	126,513	100%

2.2 도서 출판 기간 분포

출판 기간	원천데이터 수	비율
2010년 미만	7,710	6.09%
2010년 ~ 2011년	6,017	4.76%
2012년 ~ 2013년	9,059	7.16%
2014년 ~ 2015년	13,202	10.44%
2016년 ~ 2017년	15,050	11.90%
2018년 ~ 2019년	23,637	18.68%
2020년 ~	51,838	40.97%
합계	126,513	100%

2.3 전문 용어 분포

대분류	중분류	전문 용어	건수	비율
의료	STUDY_FIELD	FD_MEDICINE	44,230	0.65%
	THEORY	TR_MED_PROCEDURE	398,906	5.86%
		TR_MED_MEASURE	169,416	2.49%
		TR_MED_OBSERVATION	12,193	0.18%
	ARTIFACTS	AFW_MED_DEVICE	180,709	2.66%
	ORGANIZATION	OGG_MEDICINE	11,397	0.17%
	ANIMAL	AM_PART	268,171	3.94%
	TERM	TMM_DISEASE	685,799	10.08%
		TMM_SYMPTOM	305,514	4.49%
		TMM_DRUG	238,190	3.50%
		TM_CELL_TISSUE_ORGAN	1,018,956	14.98%
법률	ORGANIZATION	OGG_LAW	374,884	5.51%
	CIVILIZATION	CV_LAW	1,352,497	19.88%
		CV_TAX	32,807	0.48%
	EVENT	EV_OTHERS	12,405	0.18%
	TERM	TML_PARTY	470,796	6.92%
		TML_PROCEDURAL_ACTS	208,849	3.07%
		TML_LITIGATION	165,950	2.44%
		TML_PROVISION	508,108	7.47%
		TML_CASE	187,276	2.75%
		TML_JUDGMENT	155,177	2.28%
합계			6,802,230	100%

2.4 분류(독해 난이도) 분포

난이도 수준	원천데이터 수	비율
하(일반인이 쉽게 이해)	18,911	14.95%
중(일반인이 어느 정도 이해)	55,643	43.98%
상(일반인이 이해하기 어려움)	51,959	41.07%
합계	126,513	100%

2.5 문서 표준 분류 분포

2.5.1 분야 구성비

분야	원천데이터 수	어절 수	비율(어절 수 기준)
의료	46,883	49,217,568	48.11%
법률	79,630	53,082,764	51.89%
합계	126,513	102,300,332	100%

2.5.2 분야별 카테고리 구성비

분야	카테고리	원천데이터 수	비율
의료	가정의학	431	0.34%
	간호학	7	0.01%
	내과학	6,963	5.50%
	마취과학	434	0.34%
	면역학/병리학/임상병리학	2,461	1.95%
	미생물학/기생충학	1,370	1.08%
	방사선과학	3,864	3.05%
	비뇨기과학	1,110	0.88%
	산부인과학	1,630	1.29%
	생리학	391	0.31%
	생화학	679	0.54%
	성형외과학	640	0.51%
	소아과학	936	0.74%
	신경과학	763	0.60%
	신경외과학	1,658	1.31%
	안과학/임상안광학	821	0.65%
	약학/약리학	3,817	3.02%
	예방의학/직업환경의학	1,453	1.15%
	응급의학	2,299	1.82%
	이비인후과학	2,100	1.66%
	일반외과학	1,573	1.24%
	재활의학/물리치료학/작업치료학	5,767	4.56%
	정신과학	496	0.39%
	정형외과학	1,873	1.48%
	치의학	2,003	1.58%
	피부과학	298	0.24%
	해부학	620	0.49%
	흉부외과학	426	0.34%

법률	경제법	1,547	1.08%
	교육법	150	0.12%
	국제법(공법)	305	0.24%
	국제법(사법)	1,252	0.99%
	노동법등	704	0.56%
	민법일반	12,941	10.23%
	민사소송법	4,548	3.59%
	법학일반	4,208	3.33%
	상사법등	11,102	8.78%
	소비자/보호법	555	0.44%
	의료/보건법	1,231	0.97%
	인권/국제인권법	10	0.01%
	조세/세법	1,006	0.80%
	중재법	6	0.00%
	지방자치법	395	0.31%
	지적재산권법	3,223	2.55%
	친족/상속법	695	0.55%
	토지/부동산/주택/농지법	1,328	1.05%
	해사/해상/해상운송/해상보험법	690	0.55%
	행정법	9,473	7.49%
	헌법	12,196	9.64%
	형법/형사소송법등	11,676	9.23%
	환경/교통법	389	0.31%
합계		126,513	100%

데이터셋 구성

1. 원시데이터(도서) 메타정보

No.	속성명	필수여부	속성 및 내용
1	BOOK ID	필수	도서 ID
2	Book.name	필수	도서명
3	Authors	필수	저자명
4	Translators	선택	번역자명
5	Year	필수	도서 출판일
6	Page	필수	총 페이지수
7	Publisher	필수	출판사명
8	Category	필수	도서 카테고리

2. 어노테이션 포맷

No.		속성명	필수여부	Type	항목 설명	예시
1		totalcount	Y	number	카운트	
2		data	Y	object	데이터셋	
	2-1	data[].book_id	Y	string	도서식별 ID	의료 : MTB000001 법률 : LTB000001

	2-2		data[].category	Y	string	카테고리	외과, 헌법 등 도서의 카테고리
	2-3		data[].popularity	Y	number	라벨링 난이도	1(하), 2(중), 3(상)
	2-4		data[].keyword	N	array	키워드	[“포도당”, “포도당신합성”, “단백질 이화작용”]
	2-5		data[].text	Y	string	텍스트 본문	한국의 동물 보호법은 1991년 단 12개의..
	2-6		data[].word_segment	Y	number	텍스트 어절 수	
	2-7		data[].publication_ymd	Y	string	도서 출판일 특허 출원일 판결문 선고일	예시) 20201020
	2-8		data[].NE	N	object	라벨링 오브젝트	
		2-8-1	data[].NE[].id	N	number	라벨링 순번	
		2-8-2	data[].NE[].entity	N	string	라벨링 대상 텍스트	동물 보호법
		2-8-3	data[].NE[].type	N	string	카테고리 타입	CV_LAW,AM_PART..
		2-8-4	data[].NE[].begin	N	number	Target 시작 위치	
		2-8-5	data[].NE[].end	N	number	Target 끝 위치	

3. 라벨링데이터 예시

[의료]


```
{
        “book_id“: “MTB001487“,
        “category“: “소아과학“,
        “popularity“: 3,
        “keyword“: [
            “신중론“,
            “손상“,
            “물리치료“,
            “뇌성마비“,
            “근거중심의 진단과 예후“
        ],
        “text“: “이를 위해 해리스(Harris, 1996)와 골든(Golden, 1980)은 임상의사 결정에 따른 중재의 과학적 근거를 분석할 때...(중략)...평가의 필요성을 설명하면서 신중론을 얘기했다.“,
        “word_segment“: 571,
        “publication_ymd“: “20220302“,
        “NE“: [
            {
                “id“: 1,
                “entity“: “물리치료“,
                “type“: “TR_MED_PROCEDURE“,
                “begin“: 847,
                “end“: 850
            },
            {
                “id“: 2,
                “entity“: “뇌성마비“,
                “type“: “TMM_DISEASE“,
                “begin“: 1216,
                “end“: 1219
```

```
        },
        {
            "id": 3,
            "entity": "소아물리치료",
            "type": "TR_MED_PROCEDURE",
            "begin": 1365,
            "end": 1370
        },
        {
            "id": 4,
            "entity": "의학",
            "type": "FD_MEDICINE",
            "begin": 1390,
            "end": 1391
        },
        ...(중략)
```


[법률]


```
    {
        "book_id": "LTB063045",
        "category": "상사법등",
        "popularity": 2,
        "keyword": [
            "파산선고",
            "보험금청구권",
            "고지의무",
            "보험금청구",
            "상법"
        ],
        "text": "하지만 보험계약자의 권리가 일정한 범위 이내에서 제한을 받는다. ...(중략)...비로소 보험계약을 해지 또는 해제할 수 있다(제650조 제3항).",
        "word_segment": 567,
        "publication_ymd": "20210720",
        "NE": [
            {
                "id": 1,
                "entity": "계약해지권",
                "type": "CV_LAW",
                "begin": 49,
                "end": 53
            },
            {
                "id": 2,
                "entity": "보험금청구권",
                "type": "CV_LAW",
                "begin": 167,
                "end": 172
            },
            {
                "id": 3,
                "entity": "제639조 제2항",
                "type": "TML_PROVISION",
                "begin": 348,
                "end": 356
            },
            {
                "id": 4,
                "entity": "보험금청구권",
```

<table>
<tr><td></td><td>
<pre>
"type": "CV_LAW",
"begin": 423,
"end": 428
},
...(중략)
</pre>
</td></tr>
<tr><td rowspan="6">데이터셋
구축
수행기관
담당자</td><td rowspan="2">주관기관</td><td>기관명</td><td>책임자명</td><td>전화번호
(유선전화번호기입)</td><td>메일주소</td><td>담당업무</td></tr>
<tr><td>고려대학교
산학협력단</td><td>주형준</td><td>02-919-8565</td><td>drjoohj@gmail.com</td><td>데이터
정제/가공 및
레이블</td></tr>
<tr><td rowspan="4">참여기관</td><td>기관명</td><td>담당업무</td><td>기관명</td><td>담당업무</td><td rowspan="4"></td></tr>
<tr><td>군자출판사(주)</td><td>의료 말뭉치
원천 데이터
수집</td><td>㈜나라지식정보</td><td>데이터 품질
검수</td></tr>
<tr><td>㈜박영사</td><td>법률 말뭉치
원천 데이터
수집</td><td>㈜범문
에듀케이션</td><td>의료 말뭉치
원천 데이터
수집</td></tr>
<tr><td>법문사</td><td>법률 말뭉치
원천 데이터
수집</td><td>한양대학교
산학협력단</td><td>모델 개발 및
검증</td></tr>
</table>