




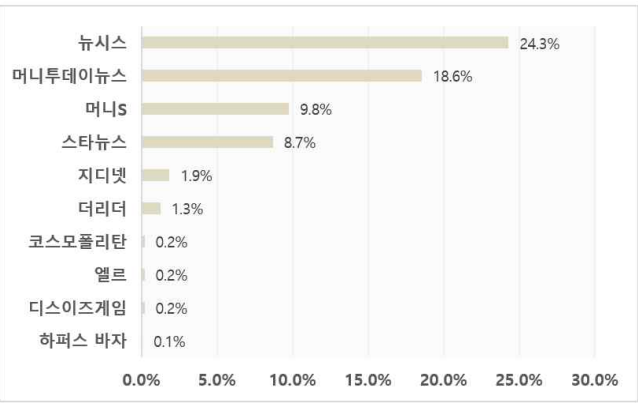
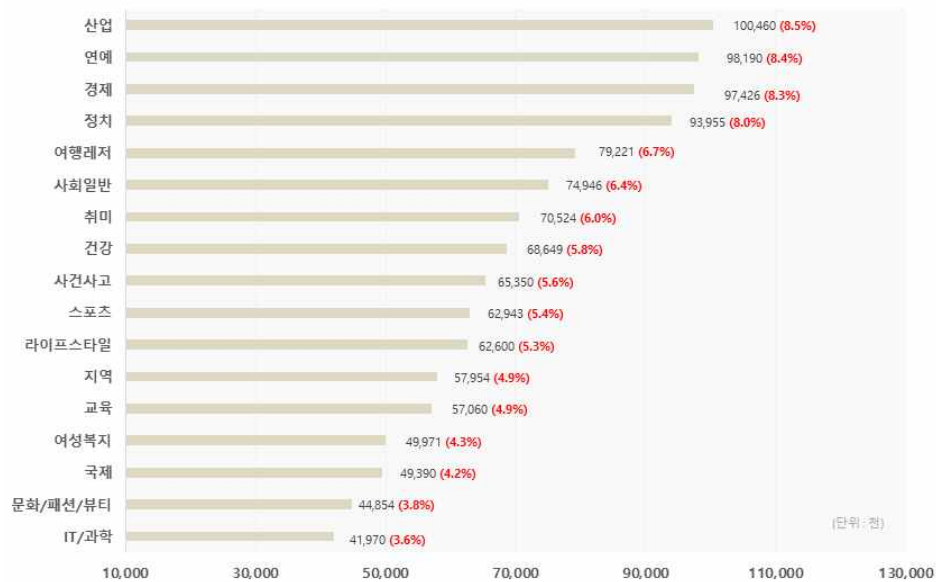


메타데이터 정보 (다중기입가능)	분야	데이터 유형 ¹⁾	구축 데이터량	원천데이터 형식 ²⁾	라벨링 형식 ³⁾	라벨링 유형 ⁴⁾
	한국어	텍스트	10억 어절	TXT	JSON	텍스트
	데이터 출처 ⁵⁾	데이터 구축년도	구축기관(총 괄)	가공기관	검수기관	
	언론기사	2021년	솔트룩스	비플라이소프트, 이노그루, 오픈큐비트	소리자바, 알체라, 솔트룩스	
	데이터 문의처	기관명	문의담당자명	전화번호	메일주소	
		솔트룩스	방재준	02-2193-1682	jjang@saltlux.co m	
	데이터 소개	웹사이트 기반 대용량의 텍스트 데이터를 수집 후, 전사도구를 활용하여 구축된 타이틀, 단락 제목, 본문 텍스트가 구조화된 10억 어절의 AI 학습 데이터 셋				
	주요키워드	챗봇, 스마트 스피커, AI , 인공지능 대규모 한국어 학습용 데이터, 데이터 생태계, 클라우드소싱, 문자인식, 웹기반 한국어 데이터				
카테고리 정의서		첨부의 카테고리 정의서 엑셀파일에 데이터카테고리 작성하여 제출(예시참고)				

1) 텍스트, 오디오, 이미지, 비디오,
2) txt, jpg,.....
3) json, csv,.....
4) 내용요약(텍스트), 번역(자연어), 질의응답(자연어), 바운딩박스(이미지/동영상), 키포인트(이미지/동영상), 세그멘테이션(이미지/동영상), 전자(음성)
5) 4대 언론기사, 자체 수집,.....

데이터셋명	국문영문	대규모 웹데이터 기반 한국어 말뭉치 데이터 셋																										
		Large web data-based Korean Corpus Dataset																										
구축목적		AI 학습용 데이터로서의 대용량 텍스트 데이터가 필요하며, 현재 웹 기반의 다양한 한국어 말뭉치 데이터가 부족하여 10억 어절 이상의 웹 데이터 기반 한국어 말뭉치 데이터를 구축																										
활용서비스		대규모 한국어 서비스, 한국어 문법 교정기 고도화, 한국어 문서 핵심 요약 서비스, 신조어 이해 사전 서비스 고도화 등에 활용																										
소개		<div>1. 데이터의 메타 정보와 단락 제목, 본문 텍스트의 구조화된 데이터를 수집하고, 수집된 범용 용어 및 고유명사, 우리말샘 용어 데이터 등으로 활용할 수 있도록 라벨링된 대용량 AI 학습데이터로, 대규모 데이터를 17개 카테고리로 분류된 데이터를 구축</div> <div>2. 데이터 라벨링 대상은 개체명과 신조어를 라벨링 항목으로 정의<ul style="list-style-type: none">- 개체명 인식을 위한 라벨링을 위한 대상 및 범주는 원시데이터 내에서 제목, 소제목, 본문 항목으로 함- 제목, 소제목, 본문에서 개체명 및 신조어를, AI 학습용 데이터로 활용할 목적으로 라벨링 수행- 기존 개체명 인식 모델이 사용하는 클래스를 사용하여 한정된 범위 내에서 신조어 라벨링을 수행</div> <div><div><div><div>원천데이터 수집</div><div>웹 기반 구조화된 텍스트 데이터</div></div><div><div>데이터 정제</div><div>비속어 사전을 통한 텍스트 클리닝 개인정보 비식별화</div></div><div><div>데이터 라벨링</div><div>저작도구를 활용하여 크라우드워커가 신조어 + 개체명 라벨링</div></div><div><div>검수</div><div>학습데이터 품질고도화 최종 개인정보 비식별화</div></div><div><div>AI학습데이터 공개</div></div></div></div>																										
데이터셋 통계 (구축 규모 및 분포)		<div>1. 데이터 구축 규모<ul style="list-style-type: none">- 웹 데이터 중 정제되지 않은 데이터를 수집하여 정제/라벨링이 수행된 대규모 텍스트 데이터 10억 어절</div> <div>2. 데이터 분포</div> <div>1) 출처별 분포<ul style="list-style-type: none">- 11개 웹 뉴스 채널에서 기사 분포<div><table><tr><th>항목명</th><th>비율</th></tr><tr><td>뉴스1</td><td>34.8%</td></tr><tr><td>뉴스스</td><td>24.3%</td></tr><tr><td>머니투데이뉴스</td><td>18.6%</td></tr><tr><td>머니S</td><td>9.8%</td></tr><tr><td>스타뉴스</td><td>8.7%</td></tr><tr><td>지디넷</td><td>1.9%</td></tr><tr><td>더리더</td><td>1.3%</td></tr><tr><td>코스모폴리탄</td><td>0.2%</td></tr><tr><td>엘르</td><td>0.2%</td></tr><tr><td>디스이즈게임</td><td>0.2%</td></tr><tr><td>하퍼스 바자</td><td>0.1%</td></tr><tr><td>합계</td><td>100%</td></tr></table><div></div></div></div>	항목명	비율	뉴스1	34.8%	뉴스스	24.3%	머니투데이뉴스	18.6%	머니S	9.8%	스타뉴스	8.7%	지디넷	1.9%	더리더	1.3%	코스모폴리탄	0.2%	엘르	0.2%	디스이즈게임	0.2%	하퍼스 바자	0.1%	합계	100%
항목명	비율																											
뉴스1	34.8%																											
뉴스스	24.3%																											
머니투데이뉴스	18.6%																											
머니S	9.8%																											
스타뉴스	8.7%																											
지디넷	1.9%																											
더리더	1.3%																											
코스모폴리탄	0.2%																											
엘르	0.2%																											
디스이즈게임	0.2%																											
하퍼스 바자	0.1%																											
합계	100%																											
		<div>2) 카테고리별 분포<ul style="list-style-type: none">- 17개 카테고리에서 10억 어절의 인공지능 학습 데이터 구축</div>																										

분류코드	분류명	어절 수	비율
SC	IT/과학	41,970,255	3.6%
CU	문화/패션/뷰티	44,853,911	3.8%
IN	국제	49,390,134	4.2%
WO	여성복지	49,970,850	4.3%
ED	교육	57,060,422	4.9%
LC	지역	57,954,012	4.9%
LI	라이프스타일	62,600,032	5.3%
SP	스포츠	62,942,643	5.4%
AC	사건사고	65,350,073	5.6%
HE	건강	68,648,804	5.8%
HB	취미	70,524,485	6.0%
SG	사회일반	74,946,093	6.4%
TL	여행레저	79,220,922	6.7%
PO	정치	93,955,104	8.0%
EC	경제	97,426,330	8.3%
EN	연예	98,190,180	8.4%
ID	산업	100,459,770	8.5%
		1,175,464,020	100.0%



1. 라벨링 파일 - json 파일 실제 예시

데이터셋 구성

```
{
  "header": {
    "identifier": "텍스트_웹데이터 말뭉치 구축_0007000000",
    "name": "컨테이너 텍스트 인식을 위한 학습용 데이터셋",
    "category": "0",
    "type": "0",
    "source_file": "BWCO210007000000",
    "source": "0",
    "subject": "CO"
  },
  "named_entity": [
    {
      "title": [
        {
          "sentence": "[세계속으로]중국인들의 언어습관에 어떤 변화가",
          "labels": [
            {
              "id": 0,
              "text": "중국인",
              "tag": "CV"
            }
          ]
        }
      ]
    }
  ]
}
```

```

    }
    },
    "subtitle": [],
    "content": [
      {
        "sentence": "5억6000만 네티즌을 보유한 중국은 몇 년 전부터 연말만 되면 언론매체마다 선별한 올해의 유행어로 떠들썩하다.",
        "labels": []
      },
      {
        "sentence": "마치 전국민이 참여하는 거대한 언어유희처럼 이들 유행어는 중국인들에게 한 해를 되돌아보게 하면서 회심의 미소와 많은 생각을 선사한다.",
        "labels": [
          {
            "id": 0,
            "text": "중국인",
            "tag": "CV"
          }
        ]
      },
      {
        "sentence": "지난해 말 한자 관련 중국의 저명한 월간지인 야오원자오즈(咬文嚼字)가 2013년 10대 유행어를 선정하면서 중국인들의 언어습관에 생긴 중요한 변화를 총결산했다.",
        "labels": [
          {
            "id": 0,
            "text": "야오원자오즈(咬文嚼字)",
            "tag": "OG"
          },
          {
            "id": 1,
            "text": "중국인",
            "tag": "CV"
          }
        ]
      },
      {
        "sentence": "1.",
        "labels": []
      },
      {
        "sentence": "긍정적 에너지를 얻을 수 있는 표현들10대 유행어 중 중국몽(中國夢 Chinese Dream)은 단연 1위를 기록했다.",
        "labels": []
      },
      {
        "sentence": "이 말은 중국공산당 제18차 전국대표대회가 거행된 후 시진핑 총서기가 국가박물관에서 주최한 부흥의 길이란 전시회를 참관하면서 중화민족의 위대한 부흥의 꿈을 실현하고자 제기한 데서 비롯됐다.",
        "labels": [
          {
            "id": 0,
            "text": "중국공산당",
            "tag": "OG"
          },
          {
            "id": 1,
            "text": "총서기",
            "tag": "CV"
          },
          {
            "id": 2,
            "text": "국가박물관",
            "tag": "OG"
          },
          {
            "id": 3,
            "text": "중화민족",
            "tag": "CV"
          }
        ]
      },
      {
        "sentence": "중국몽이란 표현은 탄생과 동시에 언론매체에 의해 순식간에 전국적으로 퍼져 나갔다.",
        "labels": []
      },
      {
        "sentence": "전문가들은 중국몽의 기본적 뜻을 국가 부강 민족 진흥 국민 행복으로 설명하고 있다.",
        "labels": []
      }
    ]
  },
  {
    "sentence": "중국몽이란 표현은 탄생과 동시에 언론매체에 의해 순식간에 전국적으로 퍼져 나갔다.",
    "labels": []
  },
  {
    "sentence": "전문가들은 중국몽의 기본적 뜻을 국가 부강 민족 진흥 국민 행복으로 설명하고 있다.",
    "labels": []
  }

```

```

    },
    {
      "sentence": "이 말 자체가 지닌 색다른 이념과 대중친화력으로 말미암아 일반 대중들에  
게 널리 인정받게 되고 긍정적인 에너지를 줄 수 있는 대표적인 유행어가 되었다.",
      "labels": []
    },
    {
      "sentence": "아메리칸 드림이 있듯이 중국인들도 이제 마음껏 국가 부흥의 꿈을 꿀 수  
있는 시대가 도래했다는 중국인들의 자부심을 엿볼 수 있는 대목이라 하겠다.",
      "labels": [
        {
          "id": 0,
          "text": "중국인",
          "tag": "CV"
        }
      ]
    }
  ],
},

```

2. 라벨링 데이터 구성 및 어노테이션 포맷

단계	수준1	수준2	수준3	수준4	데이터 타입	필수값 여부	설명	유효값/허용범위/예시
라벨링	header				Object	Y	데이터셋	
		identifier			string	Y	데이터셋 식별자	ex) 데이터유형_목적_순번
		name			string	Y	데이터셋 이름	
		category			number	Y	데이터셋 카테고리	0: 텍스트
		type			number	Y	데이터셋 타입	0: 텍스트
		source_file			string	Y	소스 파일명	1.1. 파일명 부여 방식 참조
		source			string	Y	데이터 출처	0: 뉴스 1: 게시판 2: 위키
		subject			string	Y	데이터 주제	source_file 항목의 3~4번째 문자
	named _entity	title			List	Y	개체명 / 신조어 목록	
			sentence		string	Y	제목 내용	
			labels		List	Y	제목 분석 목록	
				id	number	Y	제목 분석 일련 번호	
				text	string	Y	제목 분석 항목	
				tag	string	Y	제목 분석 태그	개체명/신조어 태그
		subtitle			List		소제목	
			sentence		string		소제목 내용	
			labels		List		소제목 분석 목 록	
				id	number		소제목 분석 일 련번호	
				text	string		소제목 분석 항 목	
				tag	string		소제목 분석 태 그	개체명/신조어 태그
		content			List		본문	
			sentence		string	Y	본문 내용	
			labels		List	Y	본문 분석 목록	
				id	number	Y	본문 분석 일련 번호	
				text	string	Y	본문 분석 항목	

				tag	string	Y	본문 분석 태그	개체명/신조어 태그
		board			string		게시판	태그 작업 비대상
		writer			string		작성자	태그 작업 비대상
		write_date			string	Y	작성일시	태그 작업 비대상, yyyy-MM-dd
		url			string		출처 URL	태그 작업 비대상
		source_site			string		출처 사이트	태그 작업 비대상
데이터셋 구축 수행기관 담당자	주관 기관	기관명	책임자 명	전화번호 (유선전화번호 가입)	메일주소	담당업무		
		솔트룩스	방재준	02-2193-1682	jjbang@saltlux.com	<ul style="list-style-type: none"> - 컨소시엄 총괄 및 관리 - 데이터셋 설계 - 데이터 정제(비식별화) 및 검수 - 저작도구 개발 및 학습모델 구현 - 경진대회 개최 		
	참여 기관	기관명	담당업무		기관명	담당업무		
		알토비전	데이터 수집 및 정제		알체라	데이터 교차 검수		
		비플라이소프트	데이터 검수		소리자바	데이터 교차 검수		
		이노그루	데이터 가공					
		오픈큐비트	데이터 가공					