문화체육관광부
국립국어원

국립국어원 신문 말뭉치 2022

(버전 1.0)

· **자료명:** 국립국어원 신문 말뭉치 2022

· **공개일**

· (버전 1.0) 2022. 12. 30.

· **자료 유형:** 텍스트

· **관련 사업:** 2022년 신문 기사 원문 수집 및 정제

· **자료 설명**

· 내용 및 분량

- 2021년 생산된 신문 기사 978,342건

· 포함 신문 매체 (총 34개)

매체 종류	매체 이름
전국 종합 일간 (6개)	국민일보, 내일신문, 서울신문, 조선일보, 한겨레, 한국일보
지역 종합 일간 (17개)	강원일보, 경기일보, 경북일보, 경인일보, 기호일보, 남도일보, 대구신문, 대전일보, 동양일보, 매일신문, 부산일보, 전남일보, 전북도민일보, 중도일보, 충북일보, 충청일보, 충청투데이,
경제 일간 (7개)	머니투데이, 서울경제, 아시아경제, 아주경제, 이데일리, 이투데이, 헤럴드경제
스포츠 일간 (1개)	스포츠서울
종합 전문 주간 (1개)	미디어오늘
인터넷 신문 (2개)	노컷뉴스, 뉴스핌

· 주제(TOPIC)별 기사 건수

사회	경제	생활	정치	IT/과학	미용/건강	스포츠	문화	연예
340,220	169,640	150,109	103,927	66,772	48,806	45,317	32,530	21,021

· 파일 형식: JSON(UTF-8 인코딩)

· 파일 수 및 크기: 파일 34개(zip 압축파일 1개, 898MB)

· 인용:

(국문) 국립국어원(2022). 2022년도 구축 국립국어원 신문 말뭉치(버전 1.0). URL: https://kli.korean.go.kr/corpus

(영문) The National Institute of Korean Language(2022). NIKL Newspaper Corpus 2022 (v1.0) URL: https://kli.korean.go.kr/corpus

· 파일 명명 규칙

자리	1	2	3 4	5 6	7 8 9 10 11 12 13 14
속성	매체	장르	주석 단계	구축 연도	일련번호(8자리)
정의값	N: 신문	W: 전국 종합지 L: 지역 종합지 P: 전문지 I: 인터넷 기반 신문 Z: 기타	RW: 원시 말뭉치	22: 2022년	00000001 ~ 99999999 (여덟 자리 일련번호)
※ 예시: NWRW2200000001.json 2022년도에 구축한 '신문 전국 종합지 매체의 기사 원시 말뭉치' 첫 번째 파일					

· 예시


```
{
  "id": "NLRW2200000010.3",
  "metadata": {
    "title": "매일신문 2021년 기사",
    "author": "장성혁",
    "publisher": "매일신문",
    "date": "20210101",
    "topic": "미용/건강",
    "original_topic": "경제,경제일반||"
  },
  "paragraph": [
    {
      "id": "NLRW2200000010.3.1",
      "form": "'변이 바이러스' 잡는 모더나 백신 2000만명 올 2분기 한국 온다"
    },
    {
      "id": "NLRW2200000010.3.2",
      "form": "미국 제약사 모더나 백신이 올해 2분기 한국에 들어올 예정이다."
    },
    {
      "id": "NLRW2200000010.3.3",
      "form": "정부는 모더나와 코로나19 감염증 백신 2000만명분 구매 계약을 체결했다고 밝혔다."
    },
```

```
        {
          "id": "NLRW2200000010.3.4",
          "form": "이날 모더나와의 계약으로 한국이 확보한 코로나19 백신은 5600만명분으로 늘어나 집단 면역이 충분히 가능한 수량을 확보하게 됐다."
        },
        {
          "id": "NLRW2200000010.3.5",
          "form": "정은경 질병관리청장은 “모더나와 코로나19 백신 2000만명 분인 4000만 회 구매 계약을 체결했으며, 2분기부터 국내에 공급될 예정이다”고 밝혔다."
        },
        {
          "id": "NLRW2200000010.3.6",
          "form": "그는 “본 계약은 문 대통령과 반셀 CEO와의 영상 통화에서 2000만 명분의 코로나19 백신 공급 합의 이후, 후속 협상을 통해 체결된 것”이라며 “백신 구매물량이 당초 계약 협상 추진하던 1000만 명분 보다 두 배로 늘어났으며, 공급 시작 시기는 올해 3분기에서 2분기로 앞당겨졌다”고 말했다."
        },
        {
          "id": "NLRW2200000010.3.7",
          "form": "또 정 청장은 “정부가 구매한 백신은 총 5600만 명분(1억600만회분)으로 우리나라 전체 인구의 100%를 초과한다. 통상적인 집단 면역을 확보하는 데에는 충분한 물량이다”라며 “선구매한 백신의 공급 시작 시기는 아스트라제네카 2021년 1분기, 얀센과 모더나 2분기, 화이자 3분기로 단계적으로 도입될 예정이다”고 설명했다."
        },
        {
          "id": "NLRW2200000010.3.8",
          "form": "특히 2000만명분을 확보한 모더나 백신은 영국에서 유행 중인 코로나19 변이 바이러스에도 예방 효과가 있는 것으로 알려졌다. 모더나는 “동물과 사람의 혈청을 통해 시험한 결과 코로나19 유행 초기부터 나타난 몇 종류의 변이 바이러스에도 똑같이 효과가 있다는 것을 확인했다”고 밝혔다. 모더나는 변이 바이러스에 대한 예방 효과를 입증하기 위해 추가 시험을 진행할 계획이다."
        },
        {
          "id": "NLRW2200000010.3.9",
          "form": "한편 코로나19 백신에 대한 접종 계획은 질병관리청이 1월 중 발표할 계획이다."
        }
```


※ “original_topic”: 신문 매체의 자체 분류 주제
“topic”: 매체 통합 분류 주제(정치, 경제, 사회, 생활, IT/과학, 연예, 스포츠 문화, 미용/건강)

· **자료 내용 문의:** 02-2669-9636