문화체육관광부
국립국어원

국립국어원 신문 말뭉치

(버전 2.0)

· **자료명:** 국립국어원 신문 말뭉치

· **공개일**
  · (버전 1.0) 2020. 8. 25.
  · (버전 2.0) 2021. 3. 30. - 기사 2.136건 추가

· **자료 유형:** 텍스트

· **관련 사업:** 2018년 국어 말뭉치 연구 및 구축(2018), 신문 기사 원문 자료 수집 및 정제(2019)

· **자료 설명**
  · 내용
    - 2009년부터 2018년까지 10년 동안 생산된 신문 기사 연 1억여 어절

  · 분량
    - 기사 3,536,491건

  · 포함 신문 매체(총 42개 매체)

매체 종류	매체 이름
중앙 종합지 (5개)	경향신문, 내일신문, 동아일보, 조선일보, 한겨레신문
지방지 (26개)	강원도민일보, 강원일보, 경기일보, 경남도민일보, 경남일보, 경상일보, 경인일보, 광주매일신문, 광주일보, 국제신문, 대구일보, 대전일보, 매일신문, 무등일보, 부산일보, 영남일보, 울산매일신문, 전남일보, 전북도민일보, 전북일보, 제민일보, 중부매일, 중부일보, 충청일보, 충청투데이, 한라일보
전문지 (7개)	매일경제신문, 스포츠경향, 스포츠동아, 전자신문, 한국경제신문, 환경일보, EBN산업뉴스
인터넷 매체 (4개)	노컷뉴스, 미디어오늘, 비엔티뉴스, 오마이뉴스

  · 파일 형식: JSON(UTF-8 인코딩)

  · 파일 수 및 크기: 파일 363개, 총 15.6GB

· **인용:**

(국문) 국립국어원(2021). 국립국어원 신문 말뭉치(버전 2.0). URL: https://kli.korean.go.kr/corpus

(영문) National Institute of Korean Language (2021). NIKL Newspaper Corpus (v.2.0). 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: 종합 전국지 L: 지역 종합지 P: 전국지 I: 인터넷 기반 신문 Z: 기타	RW: 원시 말뭉치	18: 2018년 19: 2019년	00000001 ~ 99999999 (여덟 자리 일련번호)
※ 예시: NWRW1900000001.json 2019년에 구축한 신문 전국 종합지 매체의 기사 원시 말뭉치 1번째 파일					

· **예시**


```
{
    "id": "NIRW1900000001",
    "metadata": {
        "title": "국립국어원 신문 말뭉치 NIRW1900000001",
        "creator": "국립국어원",
        "distributor": "국립국어원",
        "year": "2019",
        "category": "신문 > 인터넷 기반 신문",
        "annotation_level": [
            "원시"
        ],
        "sampling": "부분 추출 - 임의 추출"
    },
    "document": [
        {
            "id": "NIRW1900000001.1",
            "metadata": {
                "title": "오마이뉴스 2009년 기사",
                "author": "선대식",
                "publisher": "오마이뉴스",
                "date": "20090101",
                "topic": "사회",
                "original_topic": "경제"
            },
            "paragraph": [
                {
                    "id": "NIRW1900000001.1.1",
                    "form": "\"대통령, 시장 방문만 하지 말고 실천해달라\""
                },
```

```
                {
                    "id": "NIRW1900000001.1.2",
                    "form": "2008년의 마지막 새벽, 언론의 카메라는 서울 여의도를 향했다. 방송법 등 주
요쟁점 법안이 상정될 국회 본회의장을 두고 여야 의원들의 전쟁을 기다리고 있었던 것."
                },
                {
                    "id": "NIRW1900000001.1.3",
                    "form": "같은 시각, 국회 밖 세상에서 서민들은 경제 위기와 강추위 속에서 삶의 고단
함과 정치에 대한 절망감에 맞선 채 팍팍한 삶을 이어가고 있었다. 이들의 목소리를 듣기 위해 이날
새벽 3시 대한민국의 아침을 여는 서울 가락동 농수산물종합도매시장으로 향했다."
                },
                {
                    "id": "NIRW1900000001.1.4",
                    "form": "가락시장으로 가는 택시 안에서 2008년의 마지막 새벽을 맞는 서민들의 얘기
를 엿들을 수 있었다. 택시기사 서인철(가명·63)씨는 말한다."
                },
```


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

· **자료 내용 문의:** 02-2669-9636