국립국어원 2020-01-28

발간등록번호
11-1371028-000839-01

2020년 신문 기사 원문 자료 수집 및 정제

사업책임자

안 준 환

국립국어원

제 출 문

국립국어원장 귀하

국립국어원과 체결한 용역 계약에 따라 ‘2020년 신문 기사 원문 자료 수집 및 정제’에 관한 용역 보고서를 작성하여 제출합니다.

■ 사업기간: 2020년 06월 ~ 2020년 12월

2020 년 12 월 24 일

사업 책임자: 안 준 환 (주식회사 마인즈랩)

사업 수행 기관 주식회사 마인즈랩
사업 책임자 안준환
사업 참여자 임성모, 서상원, 송혜원, 이원문,
박지원, 정소라, 이미수,
손영효, 권영현

<사업 수행자> 주식회사 마인즈랩

사업 책임자	안준환(주식회사 마인즈랩 전무)
사업 참여자	임성모(주식회사 마인즈랩 이사)
	서상원(주식회사 마인즈랩 팀장)
	송혜원(주식회사 마인즈랩 매니저)
	이원문(주식회사 마인즈랩 매니저)
	박지원(주식회사 마인즈랩 매니저)
	정소라(주식회사 마인즈랩 매니저)
	이미수(주식회사 마인즈랩 매니저)
	손영효(주식회사 마인즈랩 매니저)
	권영현(주식회사 마인즈랩 매니저)

<국문 초록>

요 약 문

“21 세기 세종 계획” 사업으로 구축된 세종 말뭉치는 당시에는 세계 최대 규모였지만 지속적으로 구축되지 않아 현재는 미국, 중국, 일본 등 주요 국가의 말뭉치(코퍼스) 구축량에 비해 현저하게 뒤처지고 있는 실정이다. 이에 4 차 산업혁명 시대의 인공지능 서비스 개발 및 기술 혁신을 위한 공공재로 활용할 수 있는 한국어 말뭉치 구축 사업이 재개되었다.

본 사업은 2018 년부터 인공지능 산업계와 관련 연구 기관 등에서 공공재로 활용할 수 있는 국립국어원의 대규모 한국어 학습 자료 구축 사업의 일환으로, 2019 년 진행했던 신문 기사 원문 자료 수집 및 정제 사업의 연장선이다. 2020 년 신문 기사 원문 자료 수집 및 정제 사업은 2019 년 1 년 동안 발행된 신문 기사 원문을 월별 1,000 만 어절 이상 수집하여 공공으로 사용 가능한 최신 말뭉치로 구축하였다. 사업을 통해 구축된 신문 기사 말뭉치는 인공지능 산업 등 첨단 산업을 비롯하여 산업계 및 학계에서 각종 기술 개발과 연구 발전에 이바지할 수 있다.

본 사업의 수행범위는 신문 기사 원문 자료 수집, 매체 구성 및 2 차 저작권 확보, 정제 및 정규화 작업, 메타 데이터 태깅의 네 부분으로 나눌 수 있다. 또한 구축 준비 및 매체 선정, 원문 자료 수집 및 디지털화, 중복 기사 제거 및 정제, 메타 정보 부착 및 목록 작성 4 단계의 절차로 수행하였다.

신문 기사 원문 자료를 수집할 대상 매체는 발행 부수, 매체의 종류 및 사업 요구 사항 등을 종합적으로 고려하여 전국 종합지 3 개, 인터넷 매체는 2 개(전체 매체 수 대비 10% 이하)를 포함하여 최종적으로 35 개 매체를 선정하였다. 이후 진행된 저작권 이용 허락에 관련한 협상을 거쳐 국립국어원과 매체 간 및 사업 수행사 간 저작권 이용허락 계약 및 부속합의서를 체결하였다.

선정된 매체로부터 총 1,839,277 건 및 328,431,587 어절의 기사 데이터를 수집하였고, 이를 정제하는 작업자를 위한 도구를 개발하였다. 정제 도구는 다수의 작업자가 동시에 작업을 할 수 있는 시스템으로 구축하였다. 웹사이트에 로그인한 작업자는 배포한 매뉴얼을 바탕으로 적게는 3,000~4,000 건, 많게는 15,000 건의 기사로 묶인 프로젝트 단위로 작업할 수 있었다.

한 편 이와는 별도로 작업자들의 수작업 정제 작업 이전에 1 차로 자동 정제 작업을 실시하였다. 기사 길이에 따라 지나치게 짧은 기사, 기사 내용이 일정 수준 이상 중복된 기사 및 이번 사업 저작권 이용 허락 계약을 맺지 않은 매체의 기사를 배제하는 작업이 주를 이루었다. 이렇게 1 차 정제한 결과 기사 수로는 1,213,575 건 및 251,579,628 어절의 결과가 도출되었다.

2 차 수작업 정제 시에는 국립국어원과 협의한 기준에 따라 작업을 실시하였는데 그 기준은 이미지, 표, 그래프 등의 캡션 정보, 해당 기사의 저작권 관련 정보와 기사 정보, 기사 내용(맥락)과 관련 없는 정보, 저작권 문제의 가능성이 있는 타 매체의 기사, 외부 기고가가 작성한 기사, 일반적인 신문기사로 보기 어려운 기사 및 전체가 구어체로 된 기사이다. 작업자들은 온라인을 통해 상세 작업 기준을 공유하면서 작업하였고, 그 결과 기사 수로는 630,095 건 및 150,669,174 어절이 도출되었다.

1, 2 차 정제 이후 신문 말뭉치 구축 지침에 따라 최종 말뭉치 명명 규칙 및 인코딩 방식을 적용하고, 말뭉치 파일 내 포함해야 하는 메타 데이터의 범주를 결정하였다. 메타 데이터는 제목, 기사 작성자, 신문사, 기사 작성일, 주제, 키워드, 기사 요약 내용으로 구성되며, 기사의 주제 분류의 경우 매체 자체 분류와 통합 분류의 2 가지를 포함하였다.

최종 말뭉치 파일은 JSON 형식으로 제작하였으며, JSON 의 기본 구조는 id, metadata, document 부분으로 구성되어 있다.

2019 년 1 년간의 신문 기사 원문 수집과 이용권 확보를 통해 구축한 신문 원시 말뭉치는 실제 언어생활을 반영하는 언어 자원으로서 국어 문화, 국어 정책 수립과 같은 다양한 국어 연구와 AI 스피커, 챗봇과 같은 인공지능 한국어 처리 응용 시스템들의 성능 개선 및 평가 등 여러 산업 분야에서 활용할 수 있을 것으로 기대된다.

주요어: 신문 말뭉치, 신문 기사, 현대 한국어, 기사 말뭉치

차 례

제1장 서론

제2장 사업 수행 내용

제3장 사업 수행 결과

표 차례

그림 차례

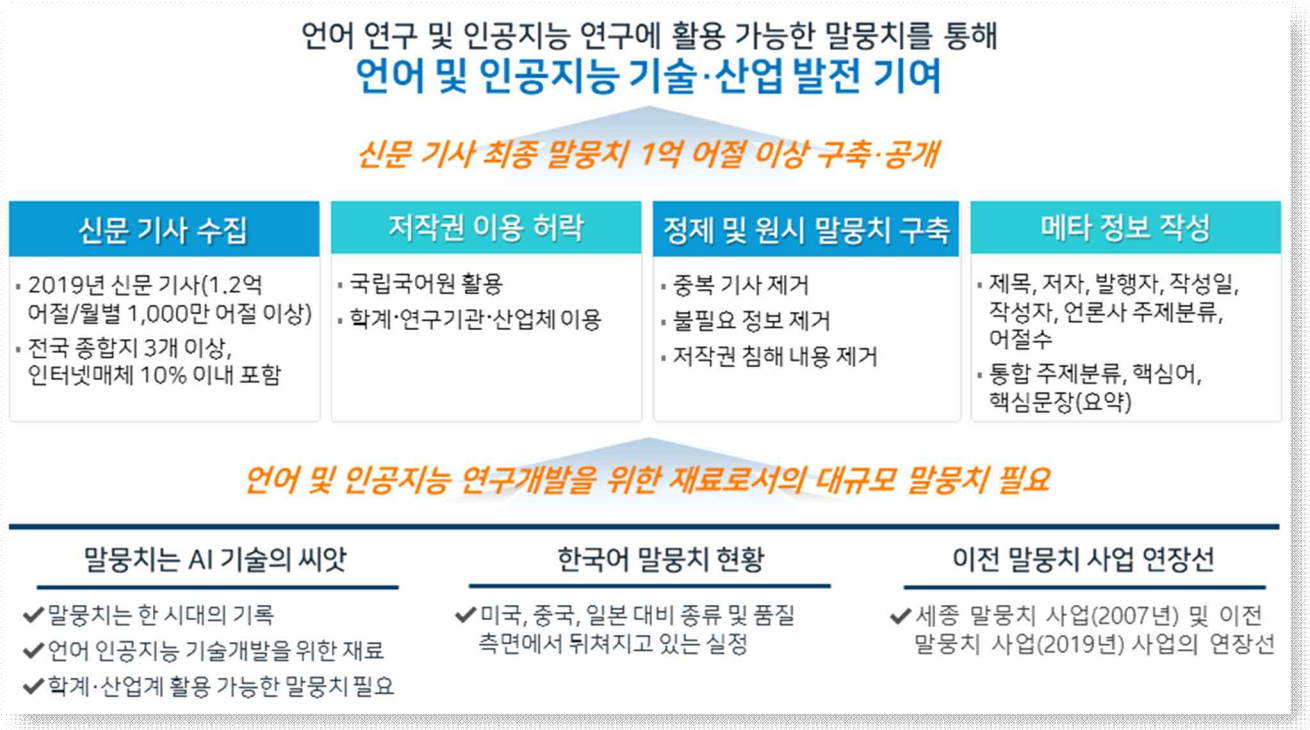
제 1 장

서 론

1. 사업 목적

1998 년부터 2007 년까지 진행된 “21 세기 세종 계획” 사업으로 구축된 2 억여 어절의 세종 말뭉치는 당시에는 세계 최대 규모였지만 지속적으로 구축되지 않아 현재는 미국, 중국, 일본 등 주요 국가의 말뭉치(코퍼스) 구축량에 비해 현저하게 뒤처지고 있는 실정이다. 이에 국립국어원에서는 4 차 산업혁명 시대의 인공지능 서비스 개발 및 기술 혁신을 위한 공공재로 활용할 수 있는 한국어 말뭉치를 구축하는 사업을 재개하였다.

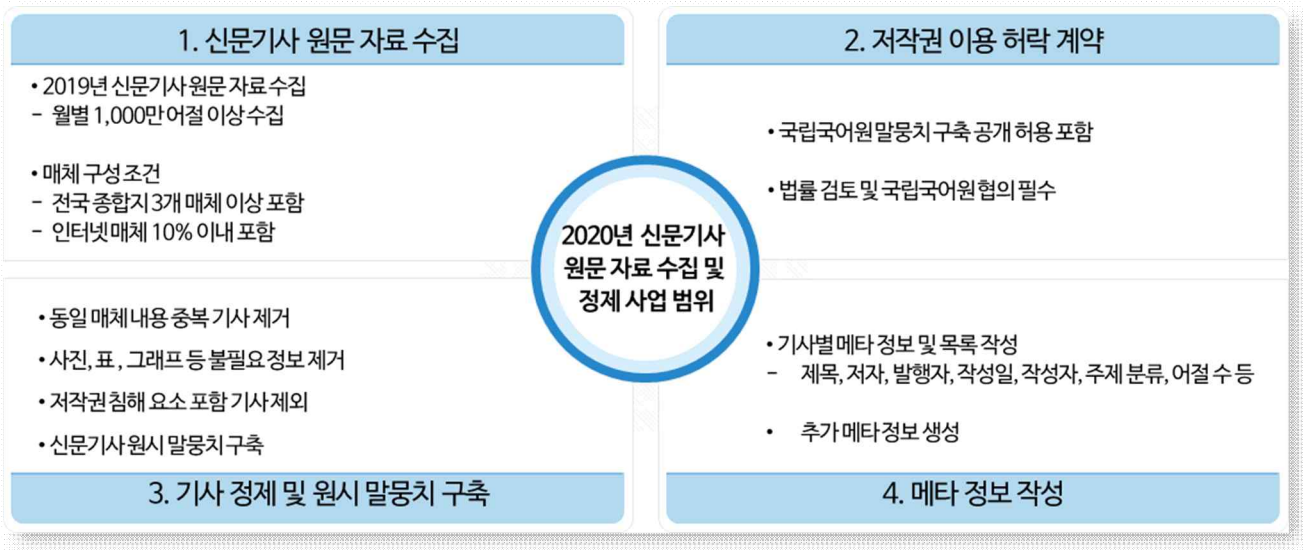
본 사업은 2018 년부터 시작된 인공지능 산업계와 관련 연구 기관 등에서 공공재로 활용할 수 있는 대규모 한국어 학습 자료 구축 사업의 일환으로, 2019 년 10 년치 신문 말뭉치 기사 수집 및 정제 사업의 연장선이다. 2020 년 신문 말뭉치 수집 및 정제 사업은 2019 년 1 월부터 12 월까지 발행된 다양한 분야의 신문 기사 원문을 월별 1,000 만 어절씩 총 1 억 어절 이상 수집하여 공공으로 사용 가능한 최신 말뭉치로 구축하는 사업이다. 본 사업을 통해 구축된 신문 기사 말뭉치는 인공지능 산업 등 첨단 산업을 비롯하여 산업계 및 학계에서 각종 기술 개발과 연구 발전에 이바지할 수 있을 것이다.



<그림 1 사업의 목적 및 배경>

2. 사업 수행 범위

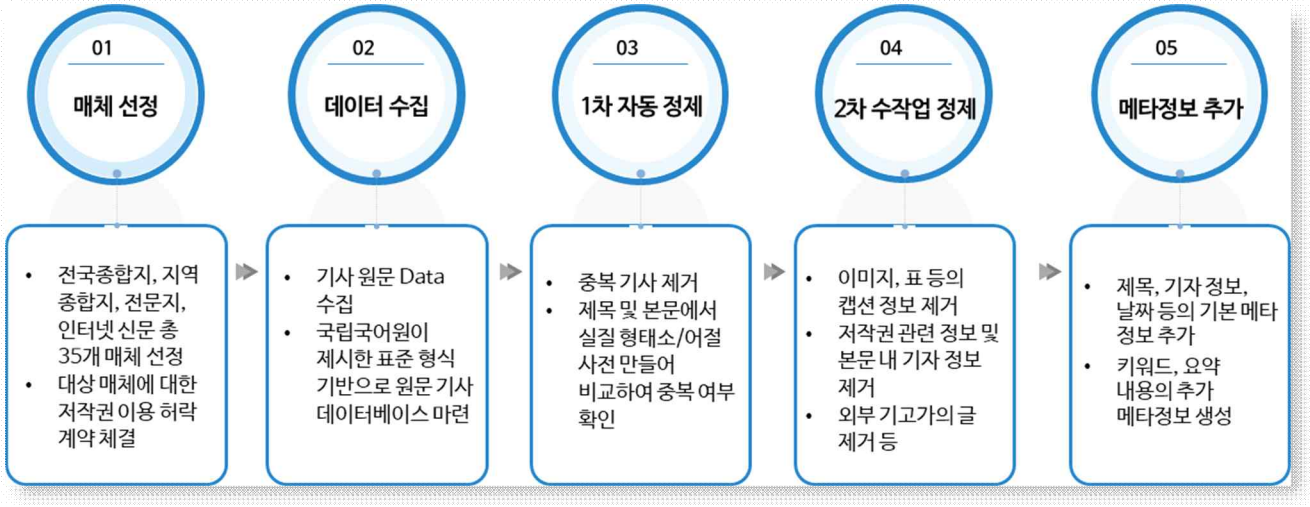
본 사업의 수행 범위는 크게 네 부분으로 나눌 수 있다. 첫째, 신문 기사 원문 자료 수집에서는 말뭉치 구축에 필요한 신문 기사 원문 자료를 수집하게 되는데 2019 년 1 년간 생산된 신문기사 월별 1,000 만 어절 이상 수집을 목표로 한다. 둘째, 매체 구성 및 저작권 확보에서는 전국 종합지 3 개 매체 이상 포함, 인터넷 기반 매체는 수집 전체 매체 수의 10% 이내로 하되 영구적·준영구적인 저작물 이용권을 확보하는 것을 목표로 한다. 셋째, 정제 및 정규화 작업으로는 중복 기사를 제거하고 사진, 도표, 그림 등 불필요한 텍스트를 제거하는 것이다. 넷째, 신문 기사에 대해 신문사, 기사 작성일, 주제 분류, 제목, 어절 수, 핵심어, 요약 등의 메타 데이터를 태깅하는 것이다.



<그림 2 사업의 범위>

3. 사업 수행 절차

본 사업은 매체 선정 및 저작권 확보, 매체별 데이터 수집, 공통/매체별 자동정제, 수작업 정제 및 검수, 메타정보 작성의 총 5 단계를 통하여 진행되었다.



<그림 3 사업 수행 절차>

첫 번째 매체 선정 단계에서는 전국 종합지, 지역 종합지, 전문지, 인터넷 신문을 대상으로 이전 사업의 일관성, 분야/지역별 형평성을 고려하여 총 35 개의 매체를 선정하고, 대상 매체의 원문 자료를 말뭉치 구축과 활용하는 데에 필요한 저작권 이용 허락 계약을 체결하였다.

두 번째로는 저작권 이용 허락 계약을 맺은 조선일보, 한국언론진흥재단으로부터 신문 기사 데이터를 입수하고, 국립국어원이 제시한 표준 형식을 기반으로 원문 기사의 데이터베이스를 마련하였다.

세 번째로는 매체별 자동 정제를 진행하였다. 자동 정제란 중복 기사 및 짧은 기사를 제거하는 단계이다. 제목 및 본문에서 실질 형태소와 어절 사전을 만들어 비교하여 이를 바탕으로 중복 여부를 확인하여 기준에 따라 제거하는 것과, 100 어절 이하의 너무 짧은 기사를 제거한다. 또한 해당 매체가 아닌 타 매체의 기사가 포함된 경우 제거한다.

네 번째로는 자동 정제된 텍스트를 수작업으로 정제한다. 2 차 수작업 정제 단계에서는 작업자가 이미지, 표 등의 캡션 정보, 저작권 문제가 있는 내용 및 본문 내 기자 정보, 외부 기고가의 글을 제거한다. 수작업 정제 결과는 검수자가 한 번 더

검토하여 이상 여부를 확인한다. 다시 말해, 수작업 정제 및 검수 과정을 거쳐 정제 작업이 완료된다.

마지막으로 메타 정보를 추가하였다. 기본 메타 정보 항목에 신문사, 기사 작성일, 주제 분류, 제목, 기자명, 본문 등의 내용을 넣고, 필수 메타 정보 항목에 핵심어와 요약 문장을 추가하였다.

4. 사업 추진 경과

본 사업의 추진 경과는 다음과 같다.

단계	작업 내용	7월	8월	9월	10월	11월	12월
준비	착수 보고						
수집	매체 선정						
	매체 계약						
	데이터 확보						
정제	자동 정제						
	수작업 정제						
정제 도구	수작업 정제 도구 설정 및 테스트						
	통계 추출						
메타데이터 생성	설계, 생성						
	검수						
납품 및 종료	샘플 데이터 납품						
	종료 보고						

<표 1 사업 추진 경과>

제 2 장

사업 수행 내용

1. 대상 매체 선정 및 저작권 이용 허락 계약

본 사업에서 신문 기사 원문 자료를 수집할 대상 매체는 매체의 종류 및 사업 요구사항 등을 종합적으로 고려하여 최종적으로 35 개 매체를 선정하였다. 그중 전국종합지는 3 개를 선정하였으며, 인터넷 매체는 2 개로 전체 매체 수 대비 10% 이하로 선정하였다.

매체 종류	매체 이름
전국종합일간 (3개)	세계일보, 조선일보, 한겨레
지역종합일간 (25개)	강원도민일보, 강원일보, 경기일보, 경남도민일보, 경상일보, 경인일보, 광주매일신문, 광주일보, 국제신문, 대구일보, 대전일보, 매일신문, 무등일보, 부산일보, 영남일보, 울산매일신문, 전남일보, 전북도민일보, 전북일보, 제민일보, 중부매일, 중부일보, 충청일보, 충청투데이, 한라일보
경제일간 (1개)	한국경제
스포츠일간 (1개)	스포츠서울
전문일간 (3개)	전자신문, 환경일보, 미디어오늘
인터넷신문 (2개)	노컷뉴스, EBN산업뉴스

<표 2> 선정 매체 목록

본 사업은 작년 사업과 동일하게 국립국어원과 매체 간의 2 자간 저작권 이용 허락 계약과 국립국어원, 매체, 사업수행사 3 자 간의 부속합의서 계약으로 진행하였으며, 계약 기간 및 금액을 제외하고 동일한 내용으로 계약을 진행하였기 때문에 큰 쟁점이나 이슈 없이 매체와의 저작권 이용 허락 계약을 진행할 수 있었다.

이용 허락 계약서에는 본 사업에서 수집하는 저작물(신문 기사)가 산업계 및 학계의 기술 개발 및 연구에 활용하기 위한 말뭉치 구축 및 활용의 목적이며, 별도의 콘텐츠 재판매 등의 수익 사업을 위한 것이 아님을 명시하였다.

대상저작물 및 복제·변형물의 이용허락 최소 기간은 10 년, 2031 년 12 월 31 일까지이며 최소 기간 만료 후 언론사가 이용허락 중지 의사를 밝혔을 경우 의사 내용에 따라 이용허락이 중지되나, 그렇지 않을 경우 이용허락이 1 년 단위로 자동 갱신되는 것으로 하였다.

또한 매체는 사업수행사에게 1 년치의 신문 기사 데이터를 인도하고, 사업 수행사는 이에 대한 검수 후 저작권 이용료를 지급할 수 있도록 하였다.

2. 원문 자료 수집

선정된 매체로부터 수집된 기사는 총 1,839,277 건 및 328,431,587 어절로 기사 당 평균 178 어절인 것으로 집계되었다. 각 매체별로 기사 건수와 어절 수가 가장 많은 5 개 매체와 가장 적은 5 개 매체를 정리하면 다음과 같다.

구분	상위 5	하위 5
기사 건수	한국경제 (약 19만건)	미디어오늘 (5,570건)
	스포츠서울 (약 15만건)	제민일보 (10,291건)
	노컷뉴스 (약 12만건)	대구일보 (17,776건)
	세계일보 (약 11만건)	경상일보 (19,281건)
	부산일보 (약9만 5천건)	무등일보 (20,521건)
어절 수	한국경제 (약 3,501만 어절)	제민일보 (약 147만 어절)
	세계일보 (약 2,661만 어절)	미디어오늘 (약 255만 어절)
	노컷뉴스 (약 2,453만 어절)	대구일보 (약 310만 어절)
	스포츠서울 (약 2,349만 어절)	울산매일 (약 345만 어절)
	전자신문 (약 1,957만 어절)	경상일보 (약 348만 어절)

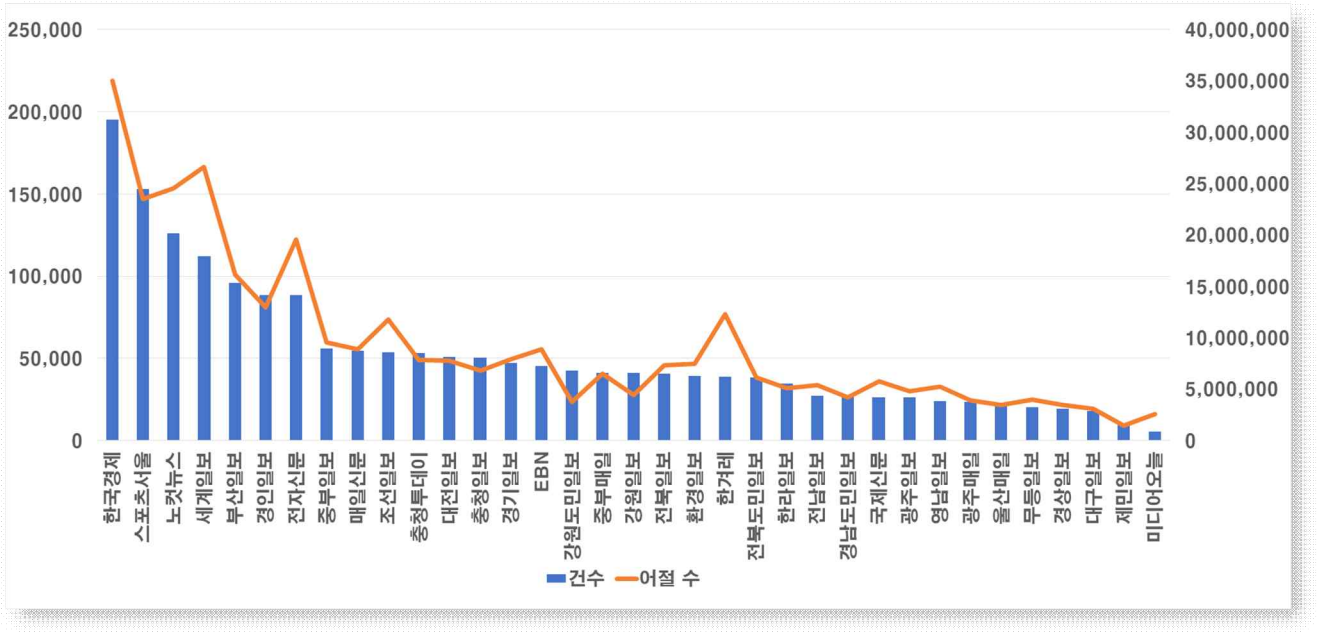
<표 3> 수집 매체 중 상위 5개, 하위 5개의 기사 수 및 어절 수

총 35 개 매체에서 수집된 전체 기사 수와 어절 수는 다음과 같다.

매체	기사 수	어절 수	매체	기사 수	어절 수
세계일보	112,143	26,615,961	중부일보	56,062	9,565,937
조선일보	53,556	11,781,086	충청일보	50,540	6,821,009
한겨레	38,790	12,296,688	충청투데이	53,292	7,868,388
강원도민일보	42,570	3,776,463	한라일보	34,709	5,081,180
강원일보	41,338	4,463,660	한국경제	195,065	35,016,761
경기일보	47,022	7,929,477	스포츠서울	152,881	23,490,375
경남도민일보	26,881	4,209,684	전자신문	88,610	19,577,357
경상일보	19,281	3,482,975	환경일보	39,394	7,450,097
경인일보	88,710	12,947,987	미디어오늘	5,570	2,559,425
광주매일	23,768	3,906,084	노컷뉴스	126,277	24,536,623
광주일보	26,312	4,823,273	EBN산업뉴스	45,172	8,869,600
국제신문	26,547	5,742,610	부산일보	95,840	16,198,715
대구일보	17,776	3,100,030	영남일보	24,048	5,288,332
대전일보	51,148	7,794,688	울산매일	22,563	3,457,384
매일신문	54,811	8,879,059	전남일보	27,215	5,431,876
무등일보	20,521	4,004,708	전북도민일보	38,403	6,141,893

<표 4> 수집한 원문 데이터의 전체 기사 수 및 어절 수

전체 매체에서 수집한 원문 데이터의 기사 수 및 어절 수 통계를 그래프를 통해 살펴보면 다음과 같다.[1)]



<그림 4> 수집 데이터의 매체별 기사 수 및 어절 수

수집된 매체별 기사 원문의 기사 수와 어절 수 통계를 살펴본 결과, 전국 종합지의 기사 수가 많을 것으로 예상했던 바와는 달리 한국경제와 스포츠서울 순으로 기사 수가 가장 많았다.

한국경제와 스포츠서울은 경제, 스포츠와 같이 특정 분야에 대한 기사 수가 높긴 하였으나, 특정 분야의 기사 외에도 사회, 문화, 정치 등 다양한 분야의 기사가 함께 존재하기 때문에 전국 종합지 매체들보다 기사의 건수가 가장 많았다.

스포츠서울은 기사의 건수가 152,881 건으로 2 번째로 많았지만 어절 수의 경우 23,490,375 건으로 4 번째로 많았다. 또한 전자신문의 경우 전체 기사 건수 88,610 건에 비해 어절 수는 19,577,357 건으로 기사 건수에 비해 어절의 수가 높았다.

위와 같이 기사의 건수와 어절 수의 순위가 조금씩 차이를 보였는데, 이는 매체별로 기사의 길이가 다르기 때문에 나타나는 결과임을 확인할 수 있었다.

1) 좌측 세로축이 어절 수, 우측 세로축이 기사 수임

또한 조선일보를 제외한 나머지 34 개의 매체는 한국언론진흥재단을 통해 일괄적으로 수집하였다. 수집 후 원문 데이터를 검수하는 과정에서 기자 값이 none 값으로 비어있거나, 연합뉴스나 오마이뉴스 등 저작권 협의가 되지 않은 기사들이 다수 포함되어 있는 이슈를 확인하였다. 기자 값이 none 값인 데이터들은 한국언론진흥재단에 재요청하여 해당 값을 받기 위해 노력하였으나, 확인이 불가하다는 답변을 받았다. 따라서 기자 값이 공란인 기사는 본문에서 기자값을 확인하여 채워넣거나, 웹 검색을 통해 기자명을 찾아서 채워 넣거나, 기자명 확인이 되지 않는 경우에는 매체명으로 채워 넣었다. 그리고 저작권 협의가 되지 않은 기사들은 삭제하였다.

3. 1차 자동 정제 작업

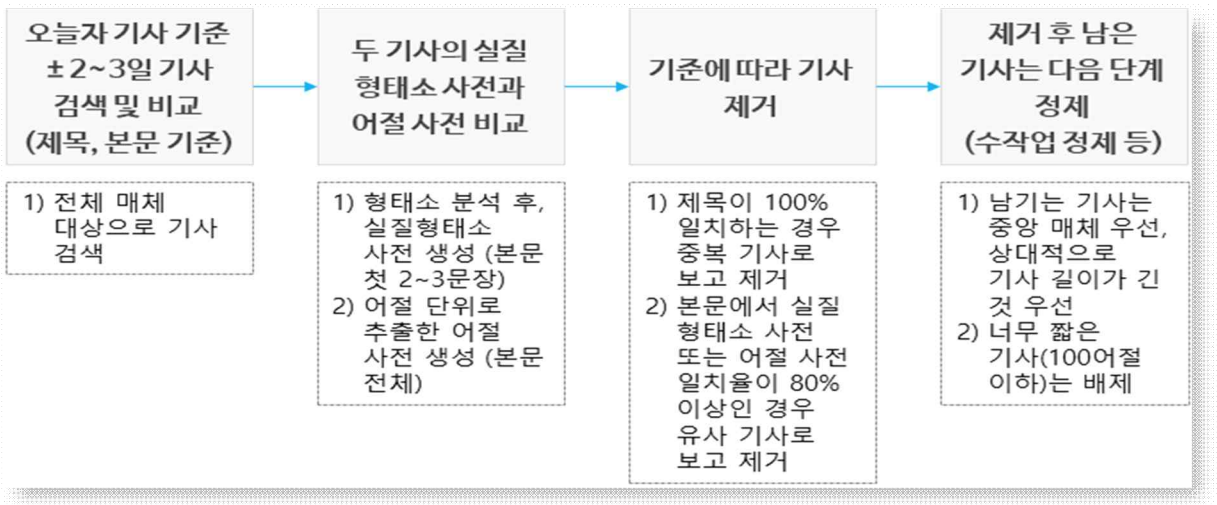
수집된 원문 기사를 대상으로 한 1 차 자동 정제 작업은 100 어절 이하의 짧은 기사, 매체별 중복 기사, 저작권 문제의 소지가 있는 기사를 제거하는 작업이다. 기사의 길이에 따른 제거 작업과 기사 내용 중복 기준에 따른 정제 작업으로 상세한 작업 기준은 다음과 같다.

○ 신문 기사 길이가 100 어절 이하인 기사는 배제했다. 한 줄짜리 속보 기사 등 지나치게 짧은 기사를 제외하도록 하였다.

○ 기사 제목이 100% 일치하는 경우는 동일한 기사로 보고 제거하였다.

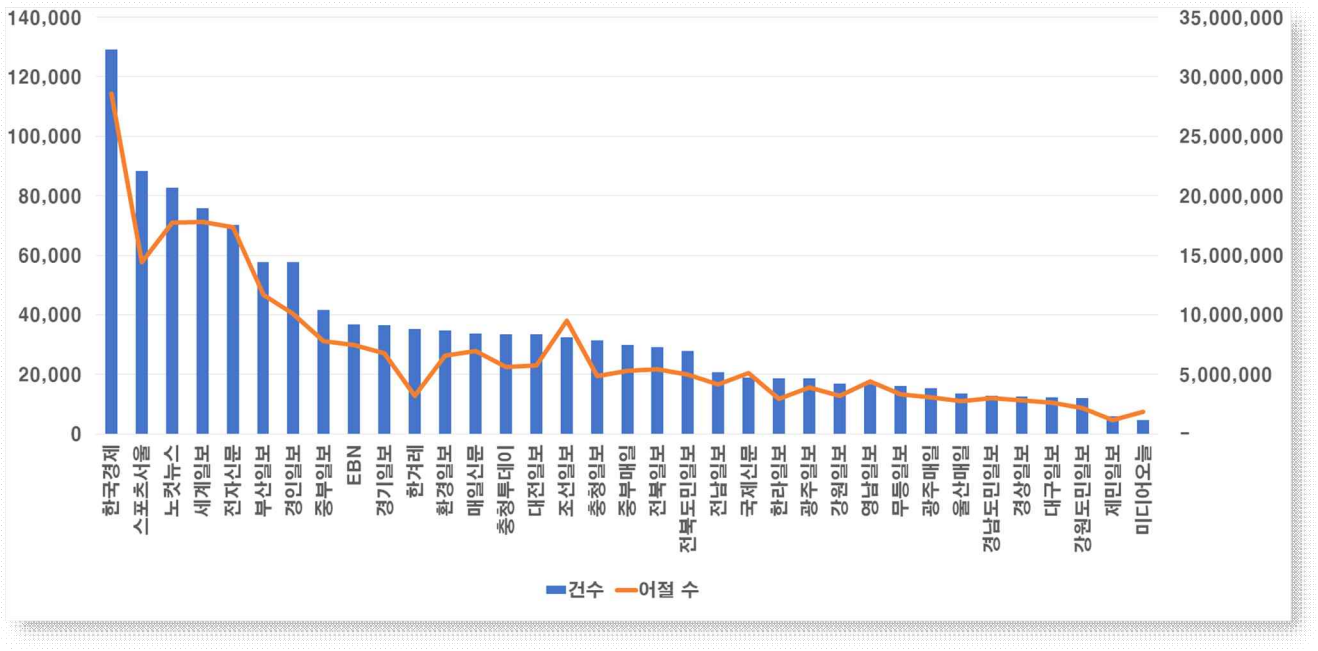
○ 기사 본문에서 형태소 사전 및 어절 사전을 만들어 비교한 후 일치율이 80% 이상인 경우 유사 내용 기사로 보고 제거하였다. 이때 비교 대상 중 남기는 기사는 중앙 일간지, 기사 본문이 더 긴 기사를 남기도록 하였다. 기사를 비교하는 대상은 동일 매체 내의 기사가 아니라 수집된 35 개 전체 기사를 대상으로 비교하였다.

○ 저작권이 해당 매체에 있지 않아 문제가 생길 수 있는 기사는 제외하였다. '연합뉴스', '뉴시스'의 경우 여러 매체에서 자주 전제하거나 그대로 보도하는 경우가 많은데 위의 두 매체는 이번 사업의 저작권 이용 허락 계약을 맺지 않았기 때문에 전제된 '연합뉴스', '뉴시스' 출처의 기사는 배제하였다.



<그림 5> 중복/유사 기사 제거

최초 수집 기사 총 1,839,277 건 및 328,431,587 어절에 대하여 1 차 자동 정제를 한 결과 1,213,575 건 및 251,579,628 어절의 결과가 도출되었다. 1 차 정제 후 기사 수 및 어절 수 통계는 다음과 같다.



<그림 6> 1 차 정제 후 매체별 기사 수 및 어절 수

전체 기사의 약 35%가 100 어절 이하, 중복된 기사, 저작권의 문제로 인해 자동 정제로 삭제되었다. 삭제된 기사의 수는 약 62 만건(625,702 건)으로 약 7 천 6 백만어절(76,851,959 어절)이 삭제되었다.

이 중 한국경제, 스포츠서울, 노컷뉴스 순으로 가장 많은 기사가 삭제되었는데 세 개의 매체의 경우 100 어절 이하의 짧은 기사와 중복된 기사의 수가 많았다. 특히 한국경제의 경우 연예 분야의 기사 중 상당수가 아래와 같이 한 줄짜리의 짧은 기사였다.

> 배우 장나라가 31 일 오후 서울 상암동 SBS 프리즘타워에서 열린 '2019 SBS 연기대상' 레드카펫 행사에 참석해 포토타임을 갖고 있다. ₩n ₩n 한경닷컴 기자 ₩n 기사제보 및 보도자료 newsinfo@hankyung.com

또한 이러한 형태의 한 줄짜리 연예기사의 대부분이 중복으로 들어가 있었기 때문에 35 개의 매체 중 가장 삭제된 기사의 비중이 높았다. 스포츠서울 역시 한국경제와 비슷하게 사진에 대한 한 줄짜리 설명만 들어간 기사들이 중복으로 들어가 있었다.

연예 기사와 스포츠 기사의 경우 기사 안에 사진이 들어가는 경우가 대부분이기 때문에 매체로부터 받은 원본 데이터에는 사진을 짧게 설명하는 100 어절 이하의 글들이 많았다. 연예 기사의 경우 다른 월에 비해 연초와 연말인 1 월과 12 월과 같이 연예 행사가 많은 달에 위와 같은 형식의 기사 건수가 높은 비중을 보였고, 스포츠 기사의 경우 다른 월에 비해 스포츠 경기가 많은 7~8 월에 사진을 설명하는 짧은 한 줄짜리 기사의 건수가 많았다.

반면 한겨레의 경우 기사 원문 대비 자동 정제로 삭제된 기사의 비중이 9%로 가장 낮은 매체였고, 자동 정제로 가장 적은 수가 삭제된 매체는 미디어오늘(840 건)이었다. 한겨레의 경우 삭제된 기사의 대부분이 부고, 별세, 승진 등의 기사였다. 미디어오늘은 100 어절 이하의 짧은 기사 건수가 많지 않아 가장 적은 기사 수가 삭제 되었다.

2019 년에 발행된 기사 중 요약봇이 작성한 기사들이 다수 존재했는데, 기존의 기사 내용을 요약봇이 자동으로 요약해주는 기사의 형식으로 주로 경제분야에서 확인할 수 있었다. 또한 주식 종목, 부동산 등의 단순 정보 나열 형식의 로봇기사도 많은 비중을 차지하고 있었다. 이러한 비문장 형태의 5 문장 이상의 기사는 기사의 형식을 유지하고 있지 않고 있으며, 신문 말뭉치에 적합한 데이터로 보기 어려우므로 자동 정제를 통해 삭제하였다.

또한 기사에 불필요한 텍스트나 기자 정보가 일정한 패턴([사진=노컷뉴스], 홍길동 기자=hongildong@메일주소 등)으로 들어가 있는 경우, 자동 정제하여 조금 더 수월하게 수작업 정제를 진행할 수 있도록 하였다.

4. 2차 수작업 정제 작업

1차 자동 정제 과정을 거쳐 선별된 신문 기사 데이터를 대상으로 국립국어원과 협의한 기준에 의해 2차 수작업 정제를 실시하였다. 상세한 수작업 기준은 다음과 같다.

○ 기사 본문 정제 시에 이미지, 표, 그래프 등의 캡션 정보를 삭제한다. 신문 기사에는 이런 이미지 정보들의 포함되어 있으나 본 사업에서는 순수하게 텍스트 정보만 수집한다. 캡션 정보들이 남아 있으면 전체 문맥에 혼란을 주기 때문이다.

구분	예시		
캡션 정보	사진제공= 사진| 사진= (사진출처: 왼쪽	[그래픽] <그래픽> 일러스트| 화면 캡처 {IMG:1}	화천= 홍길동 기자 /평택 [충청투데이 홍길동]

<표 5> 캡션 정보 삭제 예시

○ 해당 기사의 저작권 관련 정보는 메타 데이터로 작성하기로 하였으므로 기사 본문에서 삭제한다. 주로 이름 기자 메일주소로 시작하거나 'ⓒ미디어오늘 (http://www.mediatoday.co.kr)' 등의 형식을 띄고 있다.

○ 기사 본문 내 기자 정보도 저작권 관련 정보와 마찬가지 이유로 삭제하기로 하였는데 주로 기자의 이름과 전자우편 정보 등으로 구성되어 있다. 예를 들면 '홍길동 기자', '홍길동 hong@', '[스포츠서울 홍길동 기자]', 'hong@sinmun.co.kr' 등이다. 1차 자동 정제 과정에서 일정 패턴을 띄고 있는 기자 관련 정보를 삭제 처리 하였으나, 간혹 '서울='이나 '정리='처럼 불필요한 텍스트가 남는 경우가 있으므로 이것도 직접 수작업을 통해 삭제한다.

○ 기사 내용(맥락)과 관련 없는 정보는 삭제한다. 대부분의 매체에서 기사 데이터는 자체 홈페이지 등 인터넷 매체를 통한 서비스 기준으로 관리되고 있으므로 기사 말미에 특수문자와 함께 짧은 제목으로 적혀진 다른 기사로의 링크 정보나 광고

사이트로의 링크 정보가 붙어 있는 경우들이 존재한다. 이러한 형식의 문장(ex. ▶ 기자와 카톡 채팅하기 ▶ 노컷뉴스 영상 구독하기)은 삭제한다.

○ 저작권 문제의 가능성이 있는 타 매체의 기사를 완전히 삭제한다. 앞서 자동 정제 작업에서 명백히 '연합뉴스', '뉴시스'의 기사인 것을 삭제하였지만, 일부 매체의 경우 저작권이 타 매체에 있는 기사를 포함하고 있었다. 이 경우 수작업 정제 작업자가 확인 후 삭제한다.

○ 신문에는 기자 외에 외부 기고가가 작성한 기사들이 상당수 포함되어 있는데 이 또한 타 매체의 기사와 마찬가지로 저작권 관련 문제가 있어 삭제 대상이다. 매체별로 외부 기고가임을 표현하는 형식이 달라 수작업 정제 작업자가 일일이 확인 후 삭제하도록 하였다. 외부 기고가는 주로 영화평론가, 외부 기관 관계자 및 단체장, 소설가나 시인 등 다양하며 명시적으로 외부 기고가의 글이라고 표시되어 있지 않은 경우에도 문맥상 외부 기고가가 확실하다고 판단되는 경우는 삭제하도록 하였다. (ex. 나는 평범한 두 아이의 엄마이다. 등) 단, 해당 매체의 정책에 의해 운영되는 인턴 기자나 어린이 기자, 청소년 기자가 쓴 글은 기자에 준하는 위치에 있는 사람이 쓴 기사로 보고 삭제하지 않았다.

○ 기타 일반적인 신문 기사로 보기 어려운 내용을 삭제하였다. 주로 승진, 부고, 운세 등의 기사였는데 기자가 직접 작성한 내용이 아니라 처음부터 끝까지 승진자나 부고 명단, 띠별 오늘의 운세, 퀴즈, 스포츠 경기의 결과 수치만으로 구성된 기사나 기사의 대부분이 영어나 일어 등 다른 언어로 된 것도 있었다. 또한 뉴스 기사의 특성이 전혀 없는 시(詩)나 소설 등 문학작품은 삭제한다.

○ 마지막으로 구어체로 된 기사도 삭제한다. 본 사업은 신문 기사 원문 정보를 대상으로 하는 문어체 말뭉치 구축이 목적이기 때문에 '~했어요.', '~란다.', '~할까요?' 등 기사 전체가 인터뷰이거나 구어체로 이루어진 것은 삭제한다.

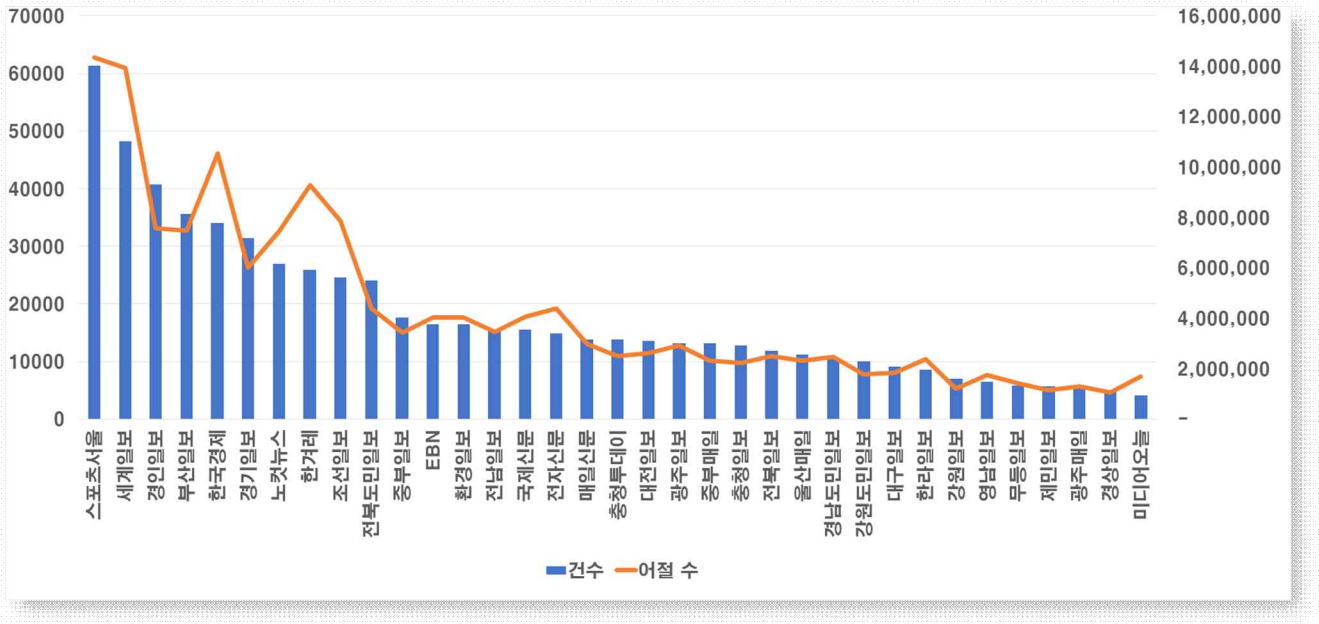
한편, 수작업 정제 중 기사의 중간에 사진에 대한 설명이나 캡션 정보가 들어가 있는 경우 쉽게 확인하기 어려우므로 작업 시 유의하였다.

재판부는 “구속 만기일(4 월 8 일)에 (항소심을) 선고한다고 해도 43 일밖에 남지 않았다”며 “재판부가 교체된 상황에서 만기일까지 충실한 심리를 마친 후 선고하기는 불가능하다”고 밝혔다. 그러면서 “구속 만료 후 석방되면 자유로운 상태에서 주거 제한이나 접촉 제한을 고려할 수 없어 증거 인멸의 염려가 높다”고 덧붙였다. 이 과정에서 재판부는 이 전 대통령의 건강문제가 아닌 충실한 재판을 위한 보석 석방이란 점을 재차 강조했다. **이명박 전 대통령이 항소심 속행 공판에서 보석 허가를 받고 동부구치소로 향하는 호송차에 오르고 있다.** 이날 재판부는 이 전 대통령의 석방을 허가하며 여러 조건을 내세웠다. 우선 이 전 대통령은 서울 강남구 논현동 자택에서만 생활해야 한다. 외출은 원칙적으로 금지된다.

예를 들어, 위의 예에서 굵은 글씨로 표시된 내용은 사진에 대한 설명으로 삭제 대상이나, 작업자가 주의 깊게 읽지 않으면 발견하기 어렵다. 이 경우 기사의 원문을 찾아보고 해당 내용이 사진에 대한 캡션이 맞다면 삭제하고, 그렇지 않다면 남겨두는 식으로 정제 작업을 진행했다. 특히 세계일보, 노컷뉴스, 환경일보의 경우 사진 위, 아래, 오른쪽, 왼쪽, 자료 출처 =, 예시> 등 다양한 유형의 캡션과 기사의 중간에 사진에 대한 설명이 많았기 때문에 정제의 난이도가 높아 다소 어려움이 많았다.

기자 정보가 기사 문단의 여러군데에 위치해 있는 매체들과 달리, 매일신문과 스포츠서울은 기자정보가 모두 기사 하단에 위치해 있어 정제 작업에 속도를 낼 수 있었으며, 캡션 정보가 비교적 적어 정제 양이 많지 않았다. 사설이나 외부기고가의 글을 작업 불가 처리하는 정도의 작업이 대부분이라 비교적 수월하게 정제 작업을 진행할 수 있었다.

1 차 자동 정제 기사 1,213,575 건 및 251,579,628 어절에 대하여 2 차 수작업 정제 및 검수 결과 630,095 건 및 150,669,174 어절이 남았다. 2 차 정제 후 기사 수 및 어절 수 통계 그래프는 다음과 같다.



<그림 7> 매체별 수작업 정제, 검수 완료 기사 수 및 어절 수

수작업 정제를 마친 후의 매체별 기사 및 어절 수는 스포츠서울, 세계일보, 경인일보 순으로 높았다. 스포츠서울은 1 차 자동 정제로 약 절반 가량의 기사가 삭제되었음에도 불구하고 사설의 수가 타 매체에 비해 적어서 수작업 정제 시 삭제할 기사 수가 적었기 때문에 가장 많은 어절 수가 남았다.

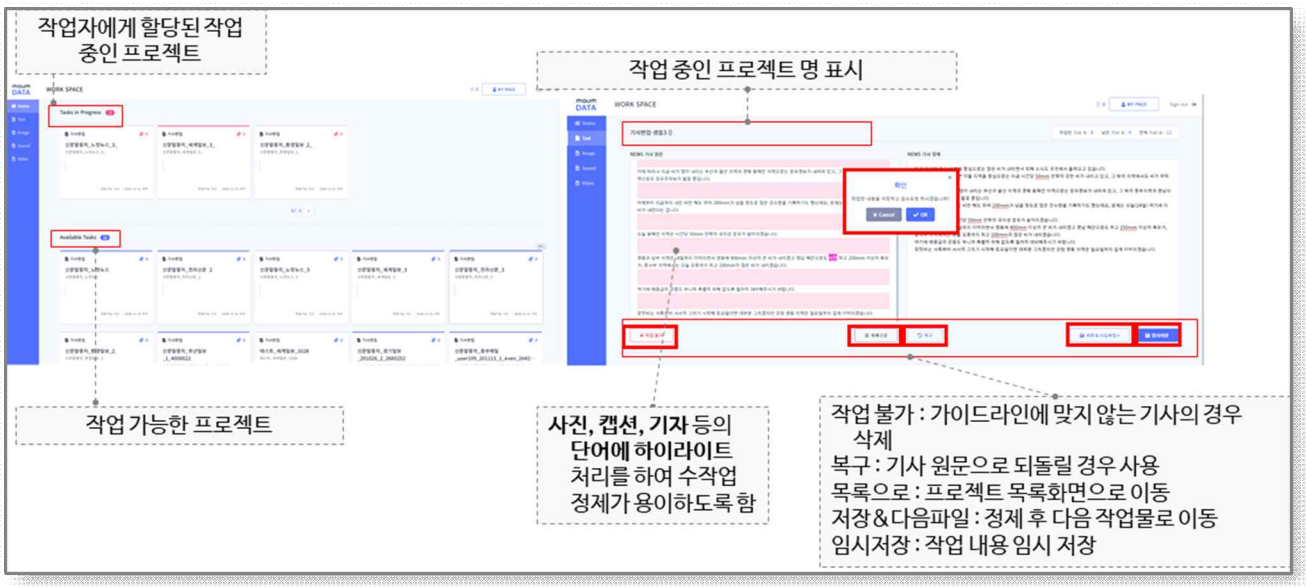
세계일보는 평균 기사의 길이가 타 매체에 비해 긴 편이었고 정제해야 하는 캡션들이 많아 작업에 어려움이 있었으나 두 번째로 많은 어절 수가 남았다.

5. 정제 도구

신문 기사의 수작업 정제 및 검수 작업을 위해 작업 도구를 개발하였다. 1 억 어절 이상의 데이터를 다수의 작업자들이 동시에 작업을 하는 데 무리가 없고, 작업 중 필요한 의사소통을 진행할 수 있도록 작업 플랫폼을 구축하였다.

정제 도구 화면은 작업자의 편의를 고려하고, 실수를 방지할 수 있도록 설계하였다. 작업 화면은 <그림 8>과 같이 작업자가 직관적으로 이해할 수 있도록 설계하였다.

왼쪽에는 기사의 원문을 띄우고, 오른쪽에는 자동 정제된 기사를 띄워 수작업 정제 및 검수를 진행했다. 기사의 원문에는 정제해야 하는 단어에 하이라이트 표시를 주고, 문단별로 구분을 두어 가독성을 높였다. 화면의 하단에 반려 사유를 작성할 수 있는 칸을 두어 작업자와 검수자 간의 의사소통이 가능하도록 하였다.



<그림 8> 프로젝트 정제 화면

또한 작업자의 실수를 줄이기 위해 단축키는 기능에 포함하지 않았다. 신문 정제 작업의 경우 삭제해야 하는 패턴이 반복적으로 등장하기 때문에 단축키 기능을 넣을 경우 작업자가 단축키를 연속적으로 누르면서 작업 완료로 넘어가는 등의 실수가 발생할 수 있기 때문이다.

플랫폼을 이용한 작업자와 검수자의 작업 절차는 <그림 9>와 같다.

<그림 9> 정제 및 검수 작업 순서

○ 작업자는 미작업 프로젝트를 할당받아 작업을 시작한다. 작업 지침에 맞춰 정제를 진행하고, 내용을 저장한 뒤 작업 완료로 넘긴다.

○ 정제가 완료된 기사는 검수자에게 넘어간다. 검수자는 내용을 확인하고 이상이 없다면 검수 완료로 넘긴다. 만약 기사에 정제되지 않은 정보가 있거나, 작업 불가처리 해야 하는 기사를 작업완료로 넘겼을 경우 사유를 작성하고 해당 기사를 반려한다.

○ 반려된 기사는 해당 기사를 작업한 작업자에게 돌아간다. 작업자는 검수자가 작성한 반려 사유를 확인하고 사유에 맞게 기사를 재작업한 뒤 작업 완료로 넘긴다.

○ 검수자는 재작업한 기사를 확인 후 문제가 없다면 최종적으로 검수를 완료한다.

○ 프로젝트 내 모든 기사의 수정이 끝나면 프로젝트는 자동으로 완료되고, 작업자는 다음 프로젝트를 배정받아 작업을 하게 된다.

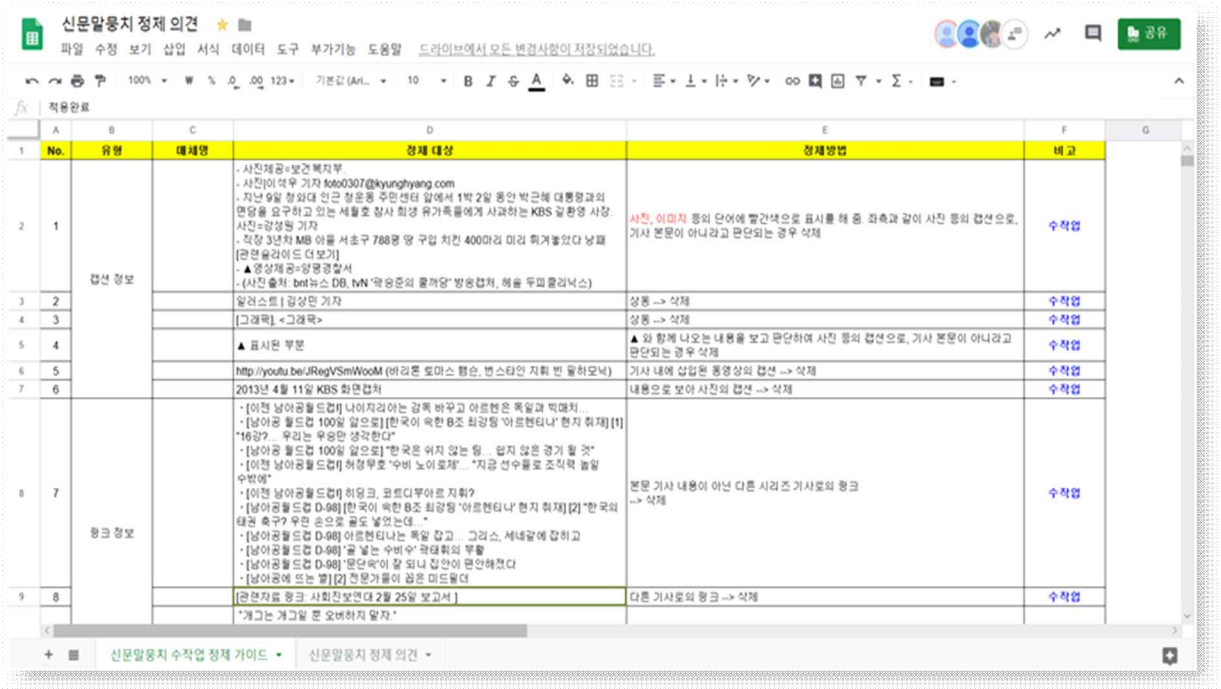
○ 단, 새로운 프로젝트를 배정 받기 전, 본인이 작업했던 프로젝트 중 반려된 기사가 있는지 우선 확인하고, 없다면 새로운 프로젝트를 배정받아 작업을 진행한다.

하나의 프로젝트는 3000 건 ~ 4000 건의 기사로 분배하였고, 1 명의 작업자가 1 일 80 만 ~ 90 만 어절 정제를 완료하는 것을 기준으로 일주일동안 작업할 수 있는 분량인 15,000 건 내외로 할당하였다.

6. 작업 지침 및 교육

정제 작업자 선발은 작년에 동일 사업을 진행했던 작업자 중 높은 퍼포먼스를 보여준 작업자를 우선으로 채용하여 작업을 진행하였다. 총 7명의 인원이 작업하였으며 이 인력들과 작업을 통제 감독할 사업 수행사 인력과의 의사소통 및 정제 원칙의 공지를 위해 구글 문서와 단체 대화방을 활용하였다.

작업에 대한 교육은 크게 작업 지침 교육과 플랫폼 교육 두 가지로 이루어졌다.



<그림 10> 말뭉치 정제 작업 지침 구글 문서

작업 지침 교육은 다음과 같이 진행하였다.

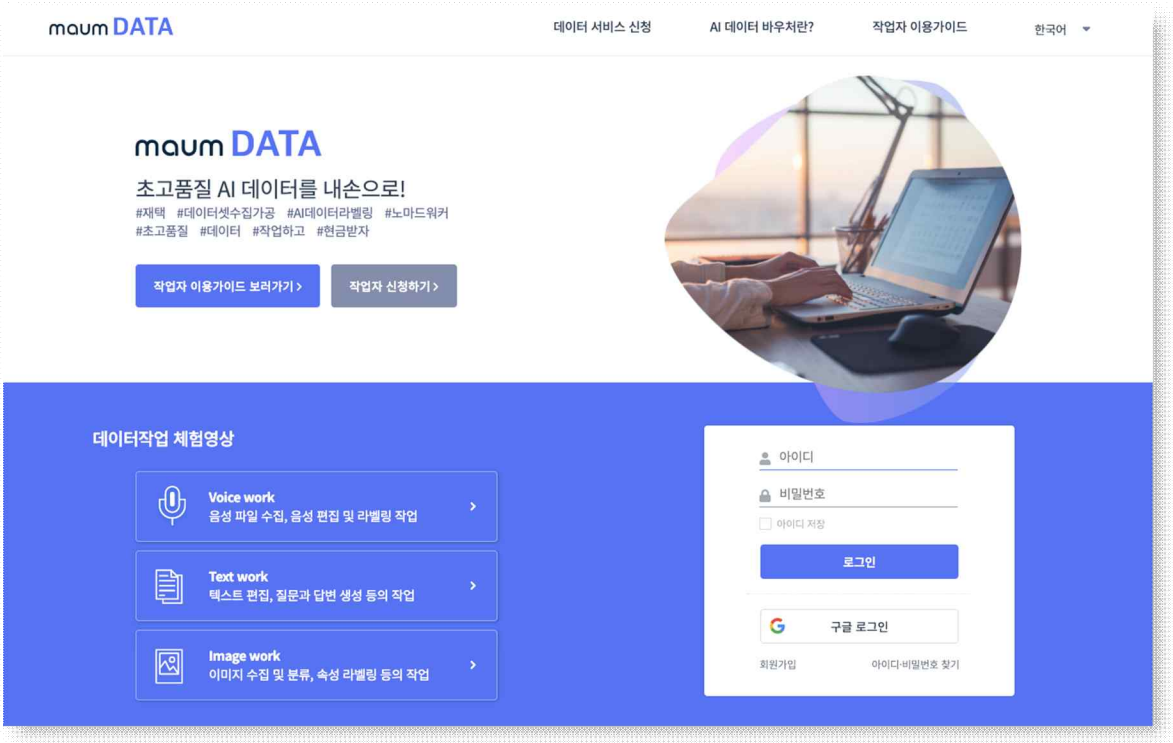
<그림 10>과 같이 구글 문서를 통해 정제 내용을 파악한 후 정제 방법에 따라 작업하고, 정제 지침에 없으나 확인할 필요가 있는 경우에는 정제 의견에 내용을 작성하거나 단체 대화방을 통해 실시간으로 정제 지침을 공유하였다.

균일한 데이터 품질을 위해 작년 데이터 정제 지침을 기반으로 지침을 작성하였으며, 인용 문구가 들어간 기사, 로봇 작성 기사, 영상의 인터뷰 내용이 그대로 작성된 구어체 기사 등 새롭게 발견되는 이슈들의 경우 내부 회의를 거친 뒤 1차로 판단을 하여 작업 지침에 반영하였다. 내부 회의로 결론이 나지 않는 이슈들은 국립국어원과 협의 후

정제 지침에 반영하였다. 예를 들어 로봇 작성 기사의 경우 현대 한국어 사용자의 일반적인 언어 사용 양상이라고 보기 어렵기 때문에 최종 말뭉치에는 포함하지 않되, 추후 유의미한 자료가 될 수 있으므로 별도 납품하라는 국립국어원의 지침에 따라 작업을 진행하였다.

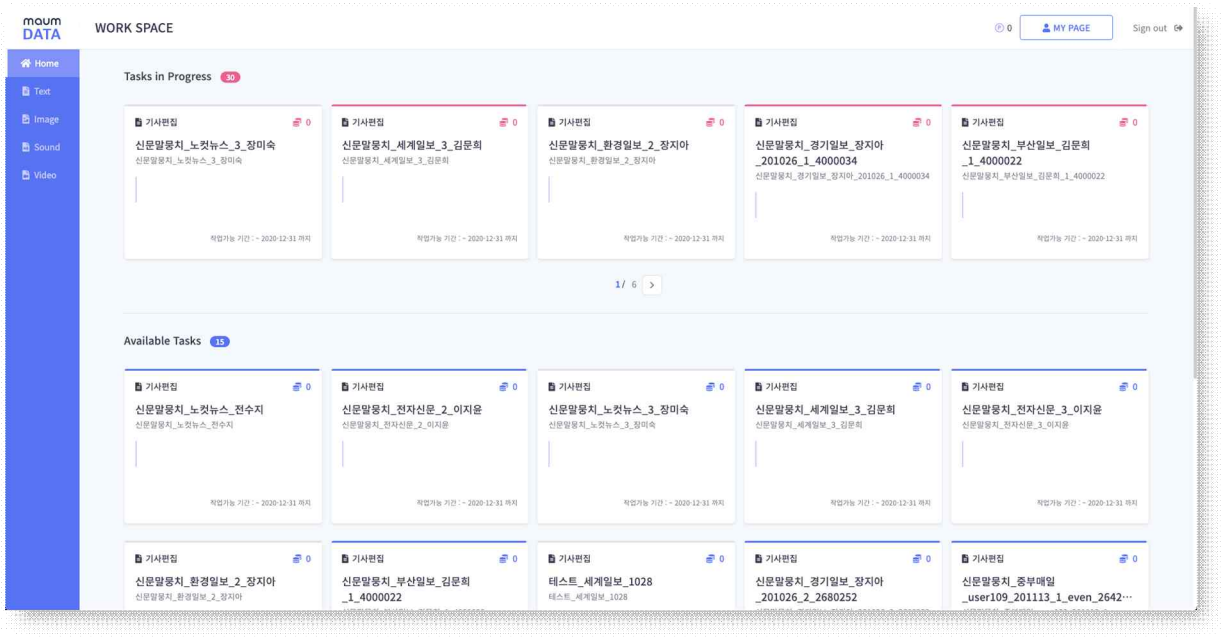
작업 플랫폼에 대한 교육은 아래의 내용으로 진행하였다.

○ 작업자는 접속 URL 로 들어와 각자 배정받은 ID 와 Password 로 로그인한다. 이때 반드시 자신이 가입한 구글 계정 혹은 배정받은 ID 로 로그인해야만 정확한 로그 정보에 의해 작업량 파악이 가능하므로 처음 작업을 시작하는 작업자에게 이 내용을 주지시킨다.



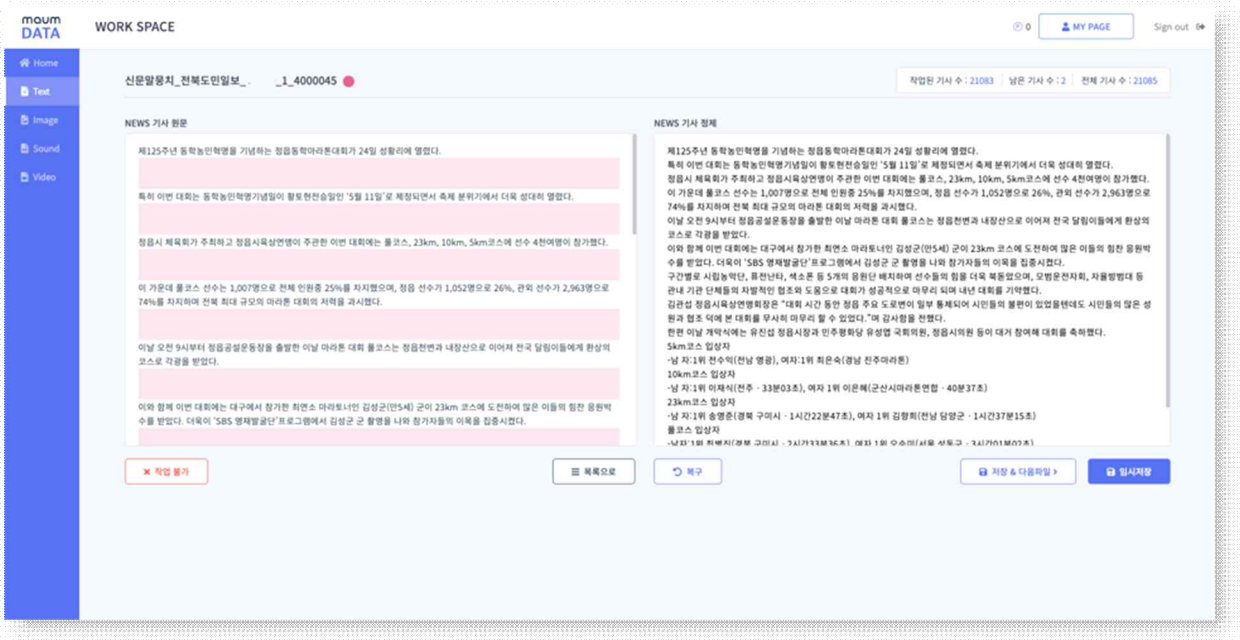
<그림 11> 정제 도구 로그인 페이지

○ 로그인한 작업자는 Tasks in Progress 혹은 Available Tasks 에서 본인의 이름(혹은 아이디)으로 지정된 프로젝트를 클릭한 다음 시작하기 버튼을 클릭하여 작업을 시작한다. 이미 선택된 프로젝트가 있다면 진행 중인 작업에서 선택하여 더블클릭한다.



<그림 12> 프로젝트 선택 페이지

○ 프로젝트 화면은 아래의 <그림 13>과 같다. 왼쪽이 기사의 원문, 오른쪽에는 정제해야 할 기사를 동시에 띄우고 왼쪽의 원문을 확인하며 오른쪽에서 삭제해야 하는 캡션, 기자 정보 등을 정제한다. 왼쪽 화면에는 사진, 캡션, 기자와 같이 빈도 수가 높게 나타나는 단어에 하이라이트 표시를 하여 삭제해야 하는 텍스트가 눈에 잘 띌 수 있게 하였다.



<그림 13> 정제 화면

○ 사설, 외부 기고자의 글은 작업불가 버튼을 클릭하여 삭제할 수 있다. ① 작업불가 버튼을 클릭하고 ⑥ 메시지 창에서 OK 버튼을 클릭한다. 작업 중 입력을 잘못하여 기사를 편집 전 상태로 다시 불러오고 싶다면 ③ 복구 버튼을 클릭한다.

○ 정제한 기사를 임시로 저장하고 싶다면 ⑤ 임시저장을 누른다. 기사의 정제가 완료된 뒤 ④ 버튼을 클릭하면 나오는 ⑥ 메시지 창에서 OK 버튼을 클릭하면 해당 기사는 저장 완료되고 다음 기사 본문을 작업할 수 있다.


maum DATA
WORK SPACE
MY PAGE
Sign out
Home
Text
Image
Sound
Video
기사편집-샘플3 ()
작업된 기사 수 : 8
남은 기사 수 : 4
전체 기사 수 : 12
NEWS 기사 원문
NEWS 기사 정제
확인
작업한 내용을 저장하고 검수요청 하시겠습니까?
Cancel
OK
1
2
3
4
5
6
작업 보가
목록으로
복구
저장 & 다음페이지 >
임시저장

<그림 14> 작업 도구 기능

7. 메타 정보 추가

정제를 거친 신문 기사 원문 자료에 국립국어원의 신문 말뭉치 구축 지침에 따라 최종 말뭉치 파일의 명명 규칙 및 인코딩 방식 등을 적용하였다.

○ 파일명의 총 자릿수는 14 자리로 영문자와 숫자로 구성된다.

○ 각 자리의 영문자 혹은 숫자는 각각 고유의 의미와 개수를 가지며, 그 각각의 의미 및 자릿수는 다음과 같다.

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 말뭉치	W: 전국 종합지 L: 지역 종합지 P: 전문지 I: 인터넷 기반 신문 Z: 기타	RW: 원시 자료		20: 2020년		00000001 ~ 99999999 (여덟 자리 일련번호)							
※ 예시: NWRW2000000001.json 신문 전국 종합지 매체의 기사 원시 말뭉치 1 번째 파일 json format														

<표 6> 파일 명명 규칙

○ 말뭉치 파일의 인코딩은 UTF-8 을 기본으로 한다.

말뭉치 파일 내 포함해야 하는 메타 데이터는 국립국어원과의 협의를 통해 다음과 같이 결정하였다.

○ 메타 데이터는 제목, 저자, 발행자, 연도, 기사번호, 분류, 기사 작성일, 기사 작성자로 구성되며 기사의 주제 분류의 경우 매체 자체 분류와 통합 분류의 2 가지를 포함한다. 수집한 원문 기사들 중에는 매체 자체의 분류 정보가 없거나 있다고 하더라도 서로 다른 분류 체계에 의한 것이어서 35 개 매체 모두에서 수집한 원문 기사의 주제를 모두 아우르는 통합 분류 체계가 필요하다. 따라서 통합 분류 체계에 따라 기사를 분류한 뒤 이 분류도 병기하기로 하였다.

○ 메타 데이터의 종류, 의미 및 예시는 다음과 같다.

. title: 말뭉치의 제목 (ex. '강원일보 2019 년 기사')

. author: 기자의 이름과 이메일 주소 등 정보 (ex. '홍길동 기자 abced@abcilbo.co.kr')

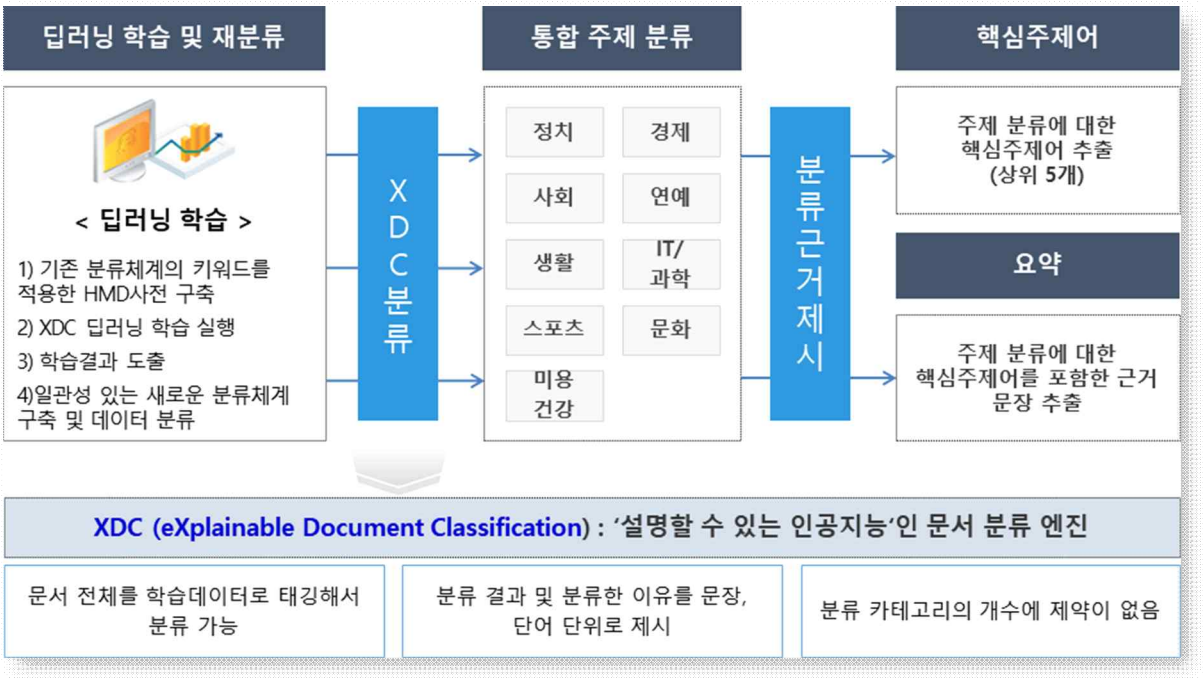
. publisher: 매체 (ex. '강원일보')

. date: 기사의 발행일자, 8 자리로 표시 (ex. '20190130')

. topic: 통합 분류 정보 (ex. '스포츠')

. original_topic: 매체에서 분류한 정보(만약 없으면 '미분류'로 작성)

모든 매체 기사에 공통으로 적용될 통합 분류는 총 9 가지로 결정했다. 이 통합 분류 체계는 사업 수행사가 이전에 수행한 다양한 자연어 관련 연구 및 신문 기사 관련 인공지능(AI) 데이터 사업을 바탕으로 자체적으로 분류한 것으로 ① 정치, ② 경제, ③ 사회, ④ 생활, ⑤ IT/과학, ⑥ 연예, ⑦ 스포츠, ⑧ 문화, ⑨ 미용/건강이다.



<그림 15> 추가 메타 정보 개요

그 밖에 해당 신문 기사의 핵심 주제어와 요약을 추가 정보로 부착하기로 하였다. 앞서 9 가지의 통합 분류의 근거가 되는 핵심 주제어를 5 개까지 선정하고 그 핵심 주제어를 포함한 근거 문장을 추출하여 요약을 작성하는데, 이는 사업 수행사가 보유한 XDC 기술 기반으로 심층학습(deep-learning)을 실행한 결과이다.

XDC 란 'eXplainable Document Classification'의 약자로, '설명할 수 있는 인공지능'이라는 뜻의 문서 분류 엔진이다. 대부분 인공지능 신경망에 의해 학습되어 나온 결과는 왜 그러한 결과가 나왔는지 아무도 모른다는 측면에서 흔히 '블랙박스'로 불리곤 했는데, 그에 비해 XDC 를 통해 얻어지는 분류 결과는 그 이유를 설명할 수 있다는 것이 특징이다. XDC 는 문서 전체를 학습 데이터로 사용하여 분류 결과 및 분류한 이유를 문장이나 단어 단위로 제시하는 것이 가능하다. 또한 분류 범주의 개수에 제약이 없어 10 개 이상의 범주로 분류하는 것도 가능하다.

8. 최종 말뭉치 제작

국립국어원의 신문 말뭉치 납품 포맷에 따라 최종 말뭉치 파일은 JSON 형식으로 변환하였다. JSON 의 상세 구조는 다음과 같다.

○ 신문 말뭉치 형식(JSON)

1수준	2수준	3수준	타입	설명
id			str	* 원시 말뭉치 ID 혹은 작업세트 파일 ID * 고유 ID로 중복이 없어야 함
metadata			obj	* 파일의 메타 정보
	title		str	* 원시 말뭉치 : 국립국어원 [말뭉치 유형 구분] [파일 ID] * 작업세트 : 국립국어원 [말뭉치 유형 구분] 추출 [파일ID]
	creator		str	* 생성자 : 국립국어원
	distributor		str	* 배포자 : 국립국어원
	year		str	* 말뭉치 구축년도
	category		str	* 분류 : 예) 신문>인터넷 신문
	annotation_level		arr(str)	* 분석층위 : 원시
	sampling		str	* 샘플링 방식 (본문 전체 / 부분 추출 - 임의 추출 / 부분 추출 - 특정 부분 추출)
document			arr(obj)	* 문서 정보
	id		str	* 문서 ID

				* '원시말뭉치파일 ID 파일 내 문서 순서' 로 구성
	metadata		obj	* 문서의 메타 정보
		title	str	* 문서 제목
		author	str	* 작성자, 게시자
		publisher	str	* 출판사, 신문사, 방송사 등
		date	str	* 작성일시, 게시일시, 크롤링 일시
		topic	str	* 주제
		original _topic	str	* 신문사에서 설정한 주제
	paragraph		arr (obj)	* 문단
		id	str	* 문단 ID. * '문서 ID.문서 내 문단 순서' 로 구성. 문서 내 문단 순서는 1부터 시작
		form	str	* 문단 정보
	keyword		**str**	*** 주제어**
	summary		**str**	*** 요약문**

<표 7> 신문 말뭉치 JSON 형식

JSON 형식의 전체 구성 및 말뭉치 예시는 다음과 같다.

```
{    "id": "NLRW2000000005",
    "metadata": {
        "title": "국립국어원 신문 말뭉치 NLRW2000000005",
```

```
        "creator": "국립국어원",
        "distributor": "국립국어원",
        "year": "2020",
        "category": "신문>지역 종합지",
        "annotation_level": [
            "원시"
        ],
        "sampling": "부분 추출-임의추출"
    },
    "document": [
        {
            "id": "NLRW2000000005.1",
            "metadata": {
                "title": "경상일보 2019년 기사",
                "author": "김창식",
                "publisher": "경상일보",
                "date": "20190120",
                "topic": "경제",
                "original_topic": "경제,국제경제|경제,산업_기업|경제,자원"
            },
            "paragraph": [
                {
                    "id": "NLRW2000000005.1.1",
                    "form": "현대중공업그룹, 1550억원 규모 원유운반선 2척 수주"
                },
                {
                    "id": "NLRW2000000005.1.2",
                    "form": "현대중공업그룹이 새해 첫 수주에 성공하며 올해 수주 회복세를 이어가고 있다. 원유 수송량이 늘면서 올해 선주들의 유조선 발주가 늘어나고 지난해에 이어 올해도 액화천연가스(LNG)운반선 시장도 호황을 이어갈 것으로 보이면서 조선업계의 수주 증대가 기대된다."
                },
                {
                    "id": "NLRW2000000005.1.3",
                    "form": "현대중공업그룹은 최근 유럽지역 선사로부터 1550억원 규모의 15만 8000t급 원유운반선 2척을 수주했다고 20일 밝혔다."
                },
                {
                    "id": "NLRW200000005.1.4",
                    "form": "이번에 수주한 선박은 길이 274m, 너비 48m로, 영암 현대삼호중공업에서 건조돼 2020년 하반기부터 순차적으로 인도될 계획이다."
```

```
            },
            {
                "id": "NLRW2000000005.1.5",
                "form": "현대중공업그룹은 올해 조선부문 수주목표를 지난해 대비 21% 높은 159억달러로 잡았다."
            },
            {
                "id": "NLRW2000000005.1.6",
                "form": "이는 지난 2014년 이후 가장 높은 수치로 본격적으로 회복세에 접어든 시황을 적극 반영해 수립한 계획이다."
            },
            {
                "id": "NLRW2000000005.1.7",
                "form": "실제로 영국 조선 해운·분석기관인 클락슨(Clarkson)은 올해 글로벌 발주량을 지난해(2859만 CGT) 대비 20% 이상 상승한 3440만 CGT로 전망했다. 글로벌 발주량은 향후 지속적으로 회복세를 유지해 2023년에는 4740만 CGT에 이를 것으로 기대된다."
            },
            {
                "id": "NLRW2000000005.1.8",
                "form": "현대중공업그룹 관계자는 ₩"새해부터 선주들의 발주문의가 이어지고 있다₩"며, ₩"조선 시황이 본격적인 회복세에 접어든 만큼 올해 수주목표 달성을 위해 수주에 역량을 집중할 것₩"이라고 밝혔다."
            },
            {
                "id": "NLRW2000000005.1.9",
                "form": "현대중공업그룹은 지난해 총 163척, 140억달러 어치를 수주하며 목표인 132억달러를 초과 달성하는 등 조선업 회복을 주도하고 있다."
            },
            {
                "id": "NLRW2000000005.1.10",
                "form": "한편 대우조선해양은 지난 14일 오세아니아 지역 선주로부터 초대형원유운반선(VLCC) 4척에 이어 18일에는 오만 국영해운회사인 OSC(Oman Shipping Company)로부터 VLCC 2척을 수주했다. 이 선박들은 옥포조선소에서 건조돼 2020년 4분기까지 선주 측에 인도될 예정이다."
            }
        ],
        "keyword": "올해,현대중공업그룹,수주,새해,선주",
        "summary": "현대중공업그룹이 새해 첫 수주에 성공하며 올해 수주 회복세를 이어가고 있다. 원유 수송량이 늘면서 올해 선주들의 유조선 발주가 늘어나고 지난해에 이어 올해도 액화천연가스(LNG)운반선 시장도 호황을 이어갈 것으로 보이면서 조선업계의 수주 증대
```

가 기대된다. 현대중공업그룹은 최근 유럽지역 선사로부터 1550억원 규모의 15만 8000t급 원유운반선 2척을 수주했다고 20일 밝혔다."

},

....

제 3 장

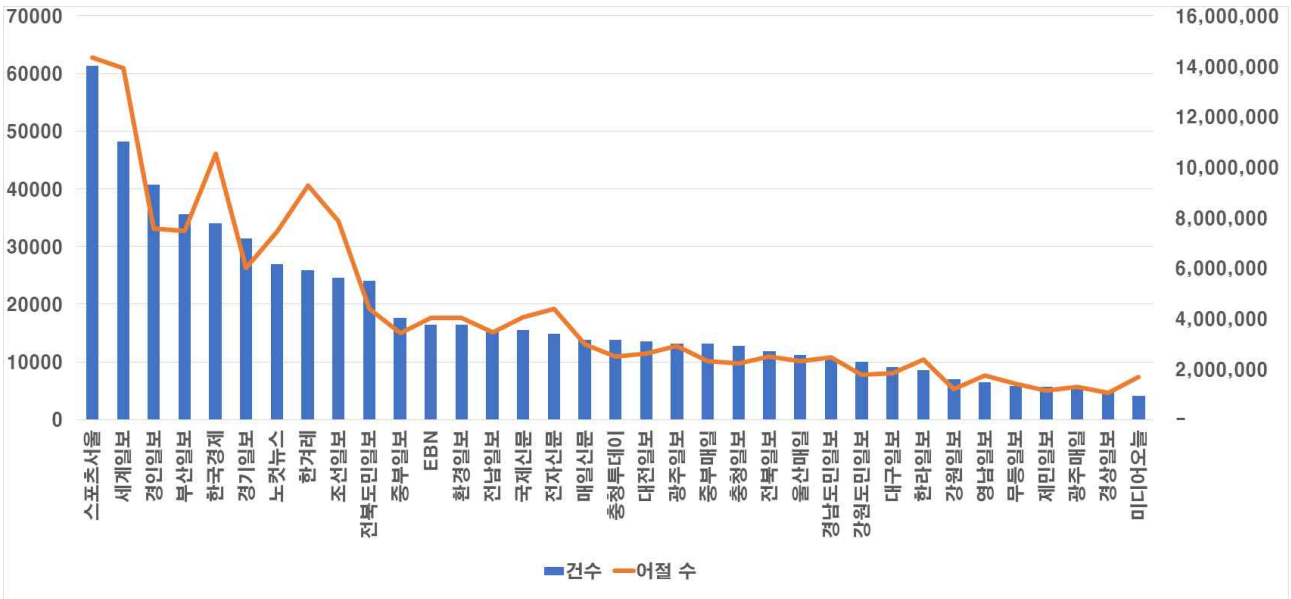
사업 수행 결과

1. 신문 기사 정제 결과

본 사업은 매체 선정부터 기사 수집, 정제, 메타 정보 작성 등 4 단계를 거쳐 수행하였다.

<그림 16> 사업 수행 내용

최종 정제 완료된 전체 35 개 매체의 기사 수 및 어절 수는 전체 630,095 건, 150,669,174 어절로 이에 대한 통계는 다음과 같다.

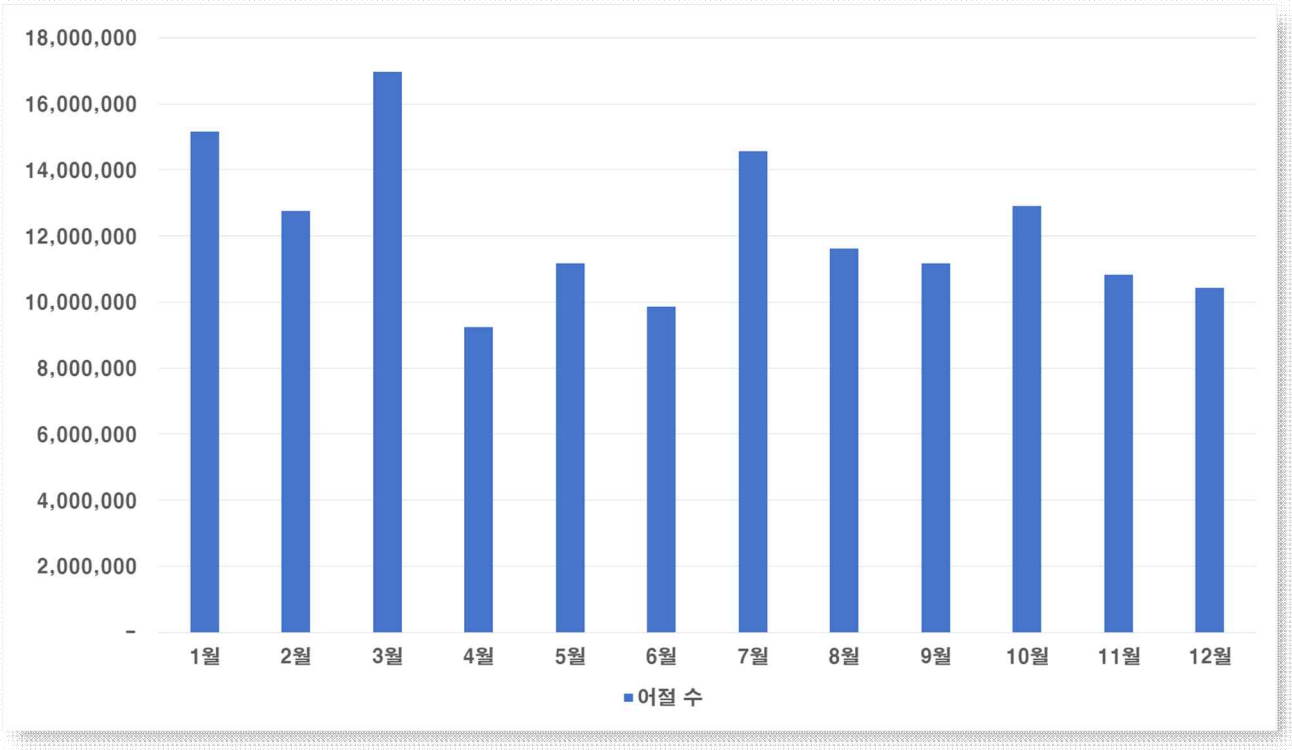


<그림 17> 최종 매체별 기사 및 어절 수 그래프

매체 Top 3			매체 Bottom 3	
순위	매체명	어절 수	매체명	어절 수
1	스포츠서울	14,356,329	경상일보	1,044,969
2	세계일보	13.928.771	제민일보	1,137,968
3	한국경제	10,540,402	강원일보	1,215,821

<표 8> 매체별 상위, 하위 어절 수

최종 구축한 매체 중 가장 많은 어절 수를 구축한 매체는 스포츠서울(14,356,329)였다. 강원일보의 경우 자동 정제 및 수작업 정제로 걸러낸 기사(100 어절 이하의 짧은 기사, 외부기고글, 사설 다수 차지)들이 많아 매체 중 최하위의 어절 수를 기록했다.



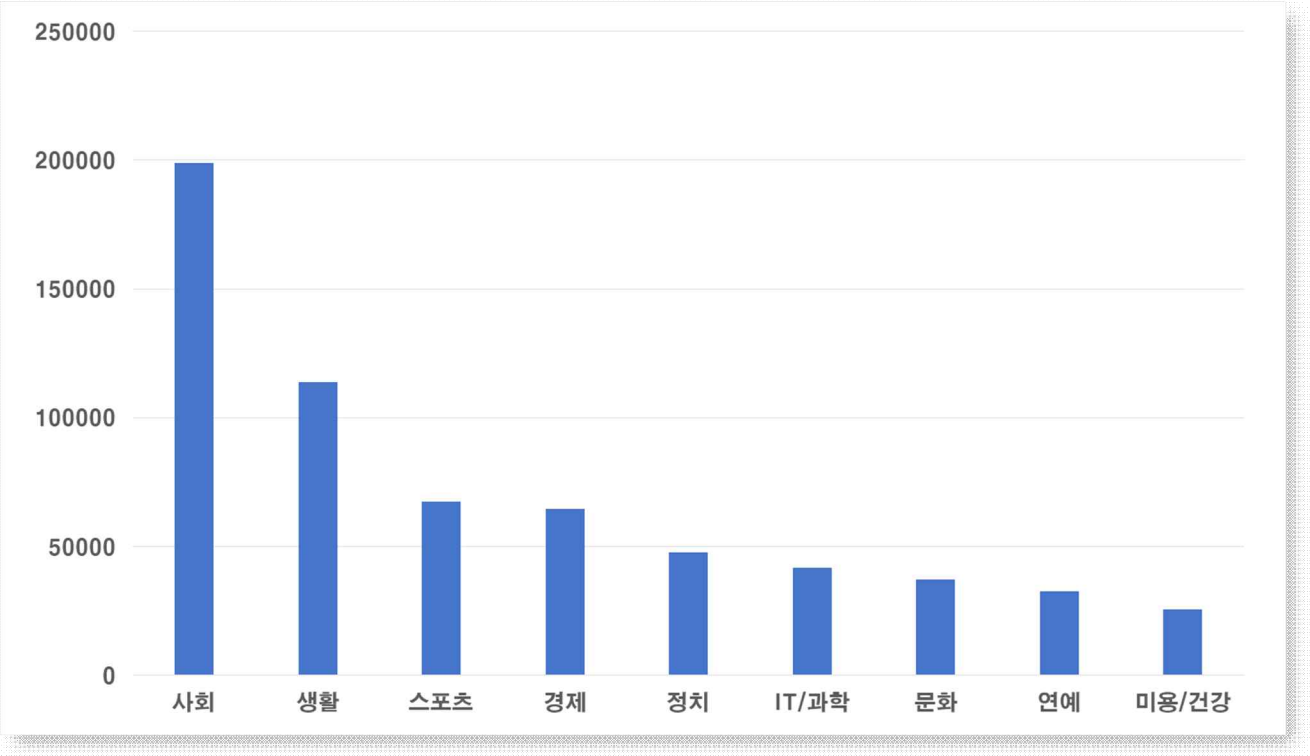
<그림 18> 최종 월별 어절 수 그래프

행사가 많은 연초와 스포츠 경기가 많은 여름 시즌에 기사의 어절 수가 높은 것을 확인할 수 있다.

최종 월별 어절 수	
월	어절 수
1월	15,159,468
2월	12,759,116
3월	16,968,224
4월	9,228,191
5월	11,162,161
6월	9,867,498
7월	14,565,390
8월	11,617,470
9월	11,164,304
10월	12,911,589
11월	10,815,576
12월	10,420,897

<표 9> 최종 월별 어절 수

9 개의 통합 주제 분류별 기사 수 통계를 보면 매우 많은 수의 사회 분야 및 생활 분야의 기사가 쓰였음을 알 수 있다. 반면 연예 기사의 경우 100 어절이 되지 않는 짧은 기사가 많아 최종 통합 주제별 기사 수에서는 적은 기사 수를 보였다.



<그림 19> 통합 주제별 기사 수 그래프

분야별 기사 수 상위 3개 매체		
순위	TOPIC	기사 수
1	사회	207,414 건
2	생활	117,169 건
3	스포츠	69,464 건

<표 10> 분야별 기사 수 상위 3개 매체

전자신문, 한국경제의 경우 로봇 기사가 작성한 기사들이 다수를 차지하고 있었는데, 이는 현대 한국어 사용자의 일반적인 언어 사용 양상이라고 보기 어려우나 유의미한 자료가 될 수 있으므로 국립국어원과 협의하여 별도로 분류, 납품하였다.

분야별 기사 수 상위 3개 매체		
매체명	어절 수	기사 수
한국경제	92,939	591
전자신문	4,341,604	10,564

<표 11> 로봇 작성 기사 및 어절 수

2. 향후 발전 방향

본 사업에서 최근 1년간 발행된 신문 기사 원문 수집과 이용권 확보를 통해 구축한 신문 원시 말뭉치는 현시대 언어생활을 반영하는 언어 자원으로서 다양한 국어 연구와 인공지능 등 산업 분야에서 활용할 수 있을 것으로 기대된다. 구체적인 성과는 아래와 같이 정리할 수 있다.

첫 번째는 2019년 1년동안 발행된 신문 기사 원문 자료와 함께 이용권을 확보한 것이다. 지금까지 신문 말뭉치의 경우 국어 연구나 산업 분야에서 자체적으로 크롤링 등의 방법을 통해 필요한 데이터를 생성해서 사용하고 있었다. 그러나 이렇게 구축한 데이터는 저작권이 해결되지 않은 자료로 활용에 제약이 있을 수밖에 없다. 본 사업을 통해 최소 2031년까지 이용 허락 계약을 맺음으로써 자유롭게 연구 및 개발 활동을 보장할 수 있게 되었다.

두 번째는 현시대의 다양한 언어생활을 반영하는 신문 기사 기반의 원시 말뭉치를 구축한 것이다. 여러 매체의 기자들이 작성한 기사를 표준 형식으로 디지털화하고, 정해진 기준에 의해 정제하여 신문 말뭉치 데이터로 구축하여 대한민국의 언어생활을 대표하는 문어체 원시 말뭉치 역할을 할 수 있게 되었다.

세 번째는 일관성 있고 유용한 메타 정보를 구축한 것이다. 수집된 수많은 기사 데이터 전체를 아우르는 통합 분류 체계를 정하고, 그 체계에 따라 일관성 있는 분류 정보를 부착했을 뿐 아니라 기사의 핵심어 및 요약 정보를 추가로 제공하여 국어 연구 및 산업 분야 활용에 도움을 줄 수 있게 되었다.

또한 작년 신문말뭉치 사업에 이어서 올해도 동일한 기준을 가지고 2019년 한 해 동안 발행된 신문 데이터를 구축함으로써 우리말을 꾸준히 보존한다는 것에 의의가 있다.

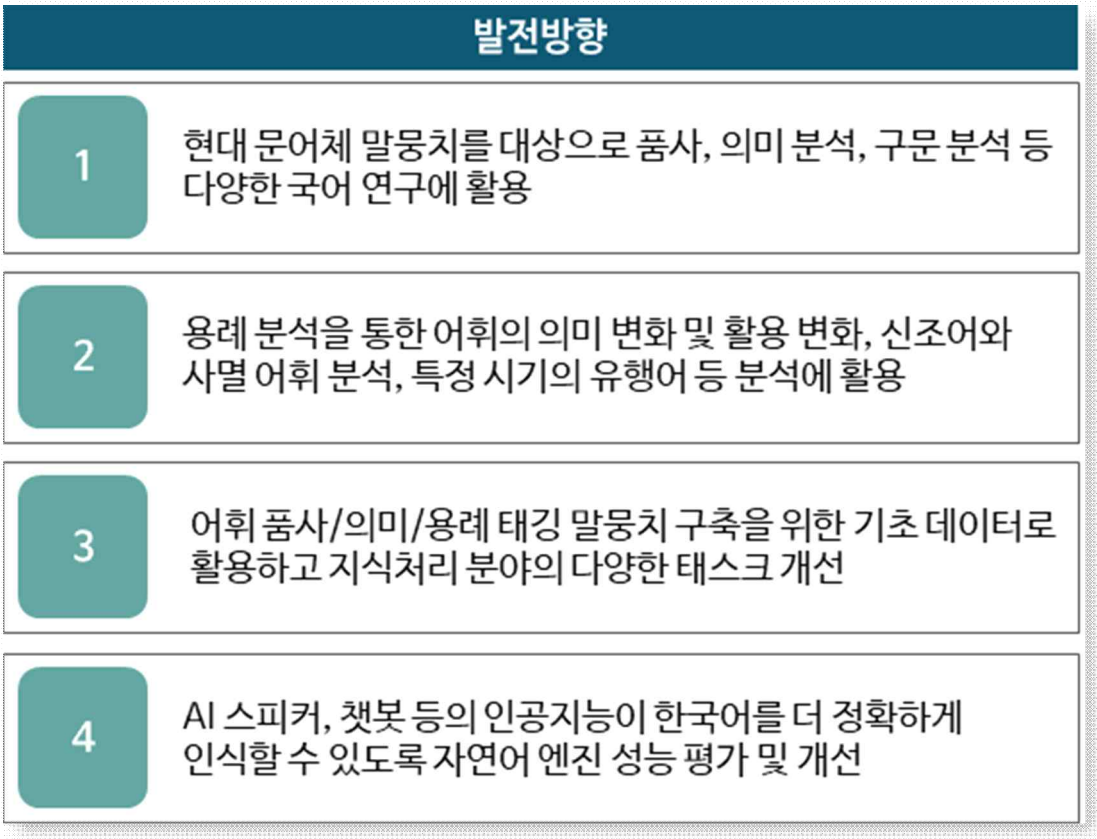
본 사업의 산출물인 말뭉치는 다음과 같은 방향으로 활용할 수 있을 것으로 기대한다.

○ 현대 문어체 말뭉치를 대상으로 어휘, 문장, 텍스트 등 다양한 단위의 국어 연구에 활용

○ 어휘의 의미 변화 및 활용 변화, 신조어와 사멸 어휘 분석, 특정 시기의 유행어 등 시기별 언어 자료 분석을 통한 사회 문화 연구에 활용

○ 다양한 언어 정보를 부착한 분석 말뭉치 구축의 기초 데이터로 활용하고 지식처리 분야의 다양한 과제의 기계 학습 자료로 활용

○ AI 스피커, 챗봇 등 인공지능 한국어 처리 응용 시스템들의 성능 개선 및 평가



<그림 20> 향후 발전 방향

<부록1> 국가 언어 자원(말뭉치) 구축 및 활용
저작권 이용 허락 계약서

국가 언어 자원(말뭉치) 구축 및 활용 저작권 이용허락 계약서

저작권 이용허락자 _____(이하 “권리자”이라 함)과 저작권 이용자 국립국어원(이하 “이용자”이라 함)은 아래 저작물에 관한 저작재산권 이용허락과 관련하여 다음과 같이 계약을 체결한다.

다 음

제1조 (계약의 목적)

본 계약은 국가 언어 자원(말뭉치) 구축 및 활용을 위한 저작재산권 이용허락과 관련하여 권리자와 이용자 사이의 권리관계를 명확히 하는 것을 목적으로 한다.

제2조 (정의)

본 계약에서 사용하는 용어의 뜻은 다음과 같다.

(1) ‘전체 기사’라 함은 권리자가 제공하는 2019년 1년 동안 생산된 신문 기사 원문 자료를 말한다.

(2) ‘수집 기사’ 라 함은 국립국어원 및 국립국어원이 발주한 용역 사업의 수행자(이하 “과업수행자”라 함)가 ‘전체 기사’에서 수집한 신문 기사 월별 1000만 어절 분량(총 1.2억 어절)에 포함된 기사를 말한다.

(3) ‘대상저작물’이라 함은 ‘수집 기사’ 중 국립국어원 및 과업수행자가 말뭉치 구축 대상으로 선정한 1억 어절 분량의 기사 원문을 말한다.

(4) ‘복제·변형물’이라 함은 국립국어원 및 과업수행자가 ‘대상저작물’에 편집 및 삭제, 언어 단위별 분리, 언어적·비언어적 정보 부착 등 처리를 더한 결과물인 원시 및 분석 말뭉치를 말한다.

제3조 (계약의 대상)

본 계약의 이용허락 대상이 되는 권리는 아래의 저작물에 대한 저작재산권 중 본 조에 명시한 이용허락 범위로 한다.

저작물: 2019년 1월 1일 ~ 2019년 12월 31일까지(1년 간)의 기사 중 권리자가 저작권 또는 저작권 재이용을 허락할 권리를 보유한 기사

매체명 :

저작권 이용 허락 범위

1. 국립국어원 및 과업수행자가 ‘수집기사’, ‘대상저작물’ 및 ‘복제·변형물’을 일정한 형식으로 전자적 기록 매체에 담아 보존하는 일
2. 국립국어원 및 과업수행자가 자모, 음절, 어휘, 어절, 구절, 문장 및 텍스트 단위의 국어 연구와 언어 정보 처리 분야에 응용하기 위해 ‘대상저작물’을 복제·변형(목차·머리말·도표·그림·각주 등의 편집 및 삭제, 언어 단위별 분리, 언어적·비언어적 정보 부착 등)하여 원시 및 분석 말뭉치로 구축하는 일
3. 국립국어원이 ‘대상저작물’ 및 ‘복제·변형물’을 국어 연구와 언어 정보 처리 분야 응용을 위하여 학계·연구기관·산업체 등이 이용할 수 있도록 홈페이지 등을 통해 제공하고 ‘복제·변형물’을 배포하는 일
4. ‘복제·변형물’을 제공·배포 받은 학계·연구기관·산업체 등이 국어 연구와 언어 정보 처리 분야 응용을 위하여 ‘복제·변형물’을 분석 및 처리하여 사용하는 것을 허락하는 일

제4조 (이용허락 기간)

(1) ‘전체 기사’ 및 ‘수집 기사’의 이용허락 기간은 계약체결일부터 2020년 12월 31일까지로 한다.

(2) ‘대상저작물’ 및 ‘복제·변형물’의 이용허락 최소 기간은 계약체결일부터 2031년 12월 31일까지로 한다. 최소 기간 만료 후 권리자 또는 저작자인 언론사가 이용허락 중지 의사를 밝히지 아니하면 이용허락이 1년 단위로 자동 갱신되며, 권리자 또는 저작자인 언론사가 이용허락 중지 의사를 밝히면 그 의사 내용에 따라 이용허락이 중지된다.

제5조 (권리자의 의무)

(1) 권리자는 이용자에게 본 계약서 제3조에 따른 저작재산권을 이용할 권리를 제4조의 기간 동안 비독점적으로 허락한다.

(2) 권리자는 이용자에게 계약 체결일로부터 20일 이내에 ‘대상저작물’의 이용을 위해 필요한 상당한 자료를 인도하여야 한다. 이때 자료를 인도하는 형식과 방법은 부속합의서에 따른다.

(3) 권리자는 ‘대상저작물’에 본 계약 이행에 지장을 주는 제3자의 이용허락권, 질권 등이 존재하는 경우, 이용자에게 그 사실을 사전에 알려야 한다.

제6조 (이용자의 권리 및 의무)

(1) 이용자는 ‘대상저작물’을 제4조의 이용허락 기간 동안 제3조의 이용 허락을 받은 범

위 내에서 비독점적으로 자유롭게 이용할 수 있다.

(2) 이용자는 과업수행자를 통해 별지 이용료를 지급하되 지급방법은 부속합의서로 정한다. 이용허락 기간 자동 갱신에 따른 추가적인 이용료는 발생하지 않는다.

(3) 이용자는 관례적으로 저작자 및 저작재산권자의 성명 등 표시를 허용하는 '대상저작물'을 이용하는 경우, 그 저작자 및 저작재산권자의 성명 등을 표시하여야 한다.

(4) 이용자는 '대상저작물'을 이용함에 있어서 저작인격권을 침해하지 아니한다. 다만, 본 계약의 목적에 따라 '대상저작물'의 본질적인 내용을 변경하지 않는 범위 내에서 변형 할 수 있다.

제7조 (확인 및 보증)

(1) 권리자는 이용자에게 다음 각 호의 사항을 확인하고 보증한다.

1. 본 저작권 이용허락 계약을 체결하는 데 필요한 권리 및 권한을 적법하게 보유하고 있다는 것
2. '대상저작물'에 대하여 이용자에게 사전에 알린 제3자의 권리 외에는 이용자의 이용을 제한할 수 있는 부담이 더 이상 존재하지 아니한다는 것

(2) 이용자는 권리자에게 다음 각 호의 사항을 확인하고 보증한다.

1. '대상저작물' 및 '복제·변형물'에 적용된 이용허락 조건에 의해서만 재이용을 허락할 것
2. '대상저작물' 및 '복제·변형물'을 제3자의 명예권을 비롯한 인격적 권리를 침해하는 방식으로 이용하지 아니할 것
3. '대상저작물' 및 '복제·변형물'의 제공·배포 시 이용허락 조건 및 재배포 금지, 목적 외 사용금지 등 주의사항을 고지할 것

제8조 (계약내용의 변경)

본 계약 내용 중 일부를 변경할 필요가 있는 경우에는 권리자와 이용자의 서면합의에 의하여 변경할 수 있으며, 그 서면합의에서 달리 정함이 없는 한, 변경된 사항은 그 다음날부터 효력을 가진다.

제9조 (계약의 해지)

(1) 당사자는 천재지변 또는 기타 불가항력으로 계약을 유지할 수 없는 경우에 본 계약을 해지할 수 있다.

(2) 당사자는 상대방이 정당한 이유 없이 본 계약을 위반하는 경우에 상당한 기간을 정

하여 상대방에게 그 시정을 최고하고, 상대방이 그 기간이 지나도록 이행하지 아니하는 경우에는 계약을 해지할 수 있다. 다만, 상대방이 명백한 시정 거부의사를 표시하였거나 위반 사항의 성격상 시정이 불가능하다는 것이 명백히 인정되는 경우에는 위와 같은 최고 없이 계약을 해지할 수 있다.

(3) 본 계약에 대한 해지권의 행사는 상대방에 대한 손해배상청구권 행사에 영향을 미치지 아니한다.

제10조 (손해배상)
당사자가 정당한 이유 없이 본 계약을 위반하는 경우, 그로 인하여 상대방에게 발생한 모든 손해를 배상할 책임이 있다. 다만, 제9조 1항의 사유로 본 계약을 이행하지 못한 경우에는 손해배상책임을 면한다.

제11조 (분쟁해결)
(1) 본 계약에서 발생하는 모든 분쟁은 권리자와 이용자가 상호 원만한 합의에 이르도록 노력하여야 하며, 분쟁이 원만히 해결되지 않는 경우에는 소제기에 앞서 한국저작권위원회에 조정을 신청할 수 있다.

(2) 제1항에 따라 해결되지 아니할 때에는 대한민국의 민사소송법 등에 따른 관할법원에서의 소송에 의해 해결토록 한다.

제12조 (비밀유지)
양 당사자는 본 계약의 체결 및 이행과정에서 알게 된 상대방에 관한 정보, 본 계약의 내용을 상대방의 서면에 의한 승낙 없이 제3자에게 공개하여서는 아니 된다. 다만, 계약의 내용을 저작자에게 알리는 경우는 예외로 한다.

제13조 (기타부속합의)
(1) 권리자와 이용자는 본 계약의 내용을 보충하거나, 이 계약에서 정하지 아니한 사항을 규정하기 위하여 부속합의서를 작성할 수 있다.

(2) 제1항에 따른 부속 합의는 본 계약의 내용과 배치되거나 위반하지 않는 범위 내에서 유효하다.

제14조 (계약의 해석 및 보완)
본 계약서에서 명시되어 있지 아니하거나 해석상 이견이 있을 경우에는 저작권법, 민법 등을 준용하고 사회 통념과 조리에 맞게 해결한다.

제15조 (계약 효력 발생일)

본 계약의 효력은 계약 체결일로부터 발생한다.

2020년 월 일

권리자 :	이용자 :
성명 (인)	성명 국립국어원장 (인)
주소	주소 서울특별시 강서구 금낭화로 154

<부록2> 유사도 구간별 기사 샘플

[중복 기사 제거 기준을 정하기 위해 살펴본 유사도 구간별 기사 샘플]

구분	기사 #1	기사 #2
한국경제 유사도 95%	그룹 방탄소년단이 미국 빌보드 ‘소셜 50’ 차트에서 17개월 연속 1위를 달성했다. 1일(현지시간) 빌보드가 발표한 최신 차트에 따르면, 방탄소년단은 ‘소셜 50’에서 77주째 1위를 기록했다. 이로써 방탄소년단은 이 차트가 생긴 이래 2017년 7월 29일자 차트부터 현재까지 17개월 연속 1위에 올라 최장 기간 연속 기록을 자체 경신했다. 또 2016년 10월 29일자 차트에서 1위로 첫 진입 후 현재까지 통산 107번째 1위를 기록하며 독보적인 존재감을 보여줬다. 한편 방탄소년단의 리패키지 앨범 LOVE YOURSELF 結 ‘Answer’는 ‘빌보드 200’ 77위에 올랐다. 지난해 9월 진입 첫 주 1위로 시작해 현재까지 18주 연속 진입 기록이다. 이어 ‘월드 앨범’ 1위, ‘인디펜던트 앨범’ 2위, ‘톱 앨범 세일즈’ 38위, '빌보드 캐나디안 앨범' 55위에 이름을 올렸다. 이외에도 LOVE YOURSELF 轉 ‘Tear’와 LOVE YOURSELF 承 'Her'는 ‘월드 앨범’ 2위와 3위, ‘인디펜던트 앨범’ 4위와 5위, ‘톱 앨범 세일즈’ 64위와 68위를 기록했다. 방탄소년단은 ‘아티스트	'대체 불가' 그룹 방탄소년단이 미국 빌보드 '소셜 50' 차트에서 17개월 연속 1위를 달성했다. 1일(현지시간) 빌보드가 발표한 최신 차트에 따르면, 방탄소년단은 ‘소셜 50’에서 77주째 1위를 기록했다. 이로써 방탄소년단은 이 차트가 생긴 이래 2017년 7월 29일자 차트부터 현재까지 17개월 연속 1위에 올라 최장 기간 연속 기록을 자체 경신했다. 또한 2016년 10월 29일자 차트에서 1위로 첫 진입 후 현재까지 통산 107번째 1위를 기록하며 독보적인 존재감을 보여줬다. 아울러 방탄소년단의 리패키지 앨범 LOVE YOURSELF 結 ‘Answer’는 ‘빌보드 200’ 77위에 올랐다. 지난해 9월 진입 첫 주 1위로 시작해 현재까지 18주 연속 진입 기록이다. 이어 ‘월드 앨범’ 1위, ‘인디펜던트 앨범’ 2위, ‘톱 앨범 세일즈’ 38위, '빌보드 캐나디안 앨범' 55위에 이름을 올렸다. 이외에도 LOVE YOURSELF 轉 ‘Tear’와 LOVE YOURSELF 承 'Her'는 ‘월드 앨범’ 2위와 3위, ‘인디펜던트 앨범’ 4위와 5위, ‘톱 앨범 세일즈’ 64위와 68위를

	100’ 2위도 차지했다. 기사제보 및 보도자료 hub@hankyung.com	기록했다. 방탄소년단은 ‘아티스트 100’ 2위도 차지했다. 한편 방탄소년단은 오는 12~13일 일본 나고야돔에서 월드투어를 재개한다. 19일 싱가포르 국립경기장, 2월 16~17일 일본 후쿠오카 야후오쿠돔, 3월 20~21일과 23~24일 홍콩 아시아월드 엑스포 아레나, 4월 6일 태국 방콕 라자망갈라 국립경기장에서도 공연한다. 기사제보 및 보도자료 newsinfo@hankyung.com
세계일보 유사도 90%	하반기부터 공연도 영화처럼 좌석 점유율 등의 성적표를 인터넷을 통해 알 수 있게 된다. 어느 공연이 인기인지 알 길 없던 깜깜한 시장 상황이 개선되는 셈이다. 문화체육관광부는 공연정보를 정확하게 공연예술통합전산망(KOPIS)에 전송하지 않으면 500만원 이하 과태료를 부과하는 내용의 공연법 일부 개정법률이 지난해 12월 24일 공포됐다고 3일 밝혔다. 이 법은 오는 6월 25일부터 시행된다. 개정된 공연법에 따르면 공연장 운영자, 공연기획·제작자, 입장권 판매자 등 공연 관계자는 문체부 장관이 정하는 공연정보를 누락·조작하지 않은 상태로 공연예술통합전산망에 전송해야 한다. 이를 위반하면 500만원 이하의 과태료가 부과된다. 문체부 관계자는 “전송해야 할 정보에 개별 공연 좌석예매율을	하반기부터 공연도 영화처럼 좌석 점유율 등의 성적표를 인터넷을 통해 알 수 있게 된다. 문화체육관광부는 공연정보를 정확하게 공연예술통합전산망(KOPIS)에 전송하지 않으면 500만원 이하 과태료를 부과하는 내용의 공연법 일부 개정법률이 오는 6월 25일부터 시행된다고 3일 밝혔다. 이 법은 지난해 12월 24일 공포됐다. 개정 공연법에 따르면 공연장 운영자, 공연기획·제작자, 입장권 판매자 등 공연 관계자는 문체부 장관이 정하는 공연정보를 누락·조작하지 않은 상태로 KOPIS에 전송해야 한다. 위반하면 500만원 이하의 과태료가 부과된다. 문체부 관계자는 “전송해야 할 정보에 개별 공연 좌석예매율을 포함시킬 예정”이라며 “이외 추가 공개할 정보는 상반기 중 공연계와

	포함시킬 예정”이라며 “이 외 추가 공개 정보는 공연계와 협의를 거쳐 상반기 중 시행령·시행 규칙으로 정할 계획”이라고 밝혔다. 국내 공연시장 규모는 2017년 처음으로 8000억원을 넘어선 것으로 추산되지만, 영화와 달리 박스오피스 정보를 알기 어려웠다. 2014년부터 운영된 공연예술 통합전산망은 데이터 수집율이 38%에 불과해 제 구실을 못했다. 기획·제작사가 자료 공개 여부를 자율 선택할 수 있었던 탓이다. 이번 법률 시행으로 공연 시장의 정확도와 투명성이 높아질 것으로 기대된다. 다만 공연계 일부에서 관객수 공개 등을 꺼리고 있어, 정보 공개 범위가 얼마나 될지는 미지수다. 문체부는 개정된 공연법 시행에 맞춰 소규모 공연장 등의 전산예매시스템 구축·운영을 지원할 방침이다. 개정된 공연법에는 공영장 폐업신고 조문도 신설됐다. 폐업신고를 해야 하는 사람이 폐업신고를 하지 않으면 지방자치단체가 폐업한 사실을 확인한 후 그 등록사항을 직권으로 말소할 수 있도록 명시적 근거 규정을 마련했다. 아울러 개정된 공연법은 기존에 ‘등록한 날로부터 3년이 경과한 경우’에만 무대시설에 대한 정기 안전검사를 받도록 했던 것을, ‘정기 안전검사를 받은 날부터 3년이 경과한 경우’, ‘자체 안전검사 결과 공연장운영자 또는 무대시설	협의를 거쳐 시행령·시행 규칙으로 정할 계획”이라고 밝혔다. 국내 공연시장 규모는 2017년 처음으로 8000억원을 넘은 것으로 추산되지만, 영화와 달리 박스오피스 정보를 알기 어려웠다. 2014년부터 운영된 KOPIS는 데이터 수집률이 38%에 불과해 제 구실을 못했다. 개정 공연법에는 공연장 폐업신고 조문도 신설됐다. 폐업신고를 해야 하는 사람이 신고하지 않으면 지방자치단체가 폐업 사실 확인 후 직권으로 등록사항을 말소할 수 있도록 근거 규정을 마련했다.

	안전진단 전문기관이 특별히 필요하다고 인정하는 경우'에도 받을 수 있도록 했다.	
세계일보 유사도 86%	새해 첫날 도쿄 번화가에서 차량이 행인들을 무더기로 치는 무차별 테러 사건을 일으킨 범인이 "옴진리교 사형 집행에 대한 보복"이라고 범행 동기를 밝혀 일본 사회에 충격을 주고 있다. 일본 정부는 1995년 도쿄 지하철역 사린가스 테러사건 등과 관련해 교주 아사하라 쇼코(麻原彰晃·본명 마쓰모토 지즈오·松本智津夫)등 옴진리교 관계자 13명에 대해 지난해 7월 두차례 걸쳐 사형을 집행했다. 사형 집행 후 일본 내부에서는 교주에 대한 사형 집행을 보복하기 위한 옴진리교 신도의 테러 가능성이 제기되기도 했다. 새해 첫날 일본 도쿄 도심에서 차량 테러를 일으킨 구사카베 가즈히로 용의자. NNN 캡처2일 TV아사히 계열 ANN은 1일 도쿄 시부야(澁谷)구 하라주쿠(原宿) 다케시타(竹下)거리에서 차량으로 행인들을 들이받아 체포된 구사카베 가즈히로(日下部和博·21) 용의자가 경찰에 "옴(진리교)사형에 대한 보복으로 (범행)했다"고 말했다고 보도했다. 구사카베 용의자는 전날 새해를 맞은 직후인 새벽 0시10분쯤 메이지진구(明治神宮) 인근으로 연말연시를 맞아 차량의 통행이 금지됐던 도로에서 행인 8명을	새해 첫날 일본 도쿄 번화가에서 차량이 행인을 무더기로 치는 무차별 테러 사건을 일으킨 용의자가 "옴진리교 사형 집행에 대한 보복"이라고 범행 동기를 밝혀 일본 사회에 충격을 주고 있다. 새해 첫날 일본 도쿄 도심에서 차량 테러를 일으킨 구사카베 가즈히로 용의자. NNN 캡처2일 TV아사히 계열 ANN은 지난 1일 도쿄 시부야(澁谷)구 하라주쿠(原宿) 다케시타(竹下)거리에서 경차로 행인 8명을 들이받아 체포된 구사카베 가즈히로(日下部和博·21) 용의자가 경찰에 "옴(진리교) 사형에 대한 보복으로 (범행)했다"고 말했다고 보도했다. 새해 첫날 일본 도쿄 경시청 소속 경찰관들이 차량 테러가 일어난 현장에 출동하고 있다. NNN 캡처일본 정부는 1995년 도쿄 지하철역 사린가스 테러사건과 관련해 교주 아사하라 쇼코(麻原彰晃·본명 마쓰모토 지즈오·松本智津夫) 등 옴진리교 관계자 13명에 대해 지난해 7월 두 차례에 걸쳐 사형을 집행했다. 사형 집행 후 일본 내부에서는 교주에 대한 사형 집행을 보복하기 위한 옴진리교 신도의 테러 가능성이 제기됐다. 구사카베 용의자는 1일 0시10분쯤 연말연시를 맞아 보행자전용 도로가

	차례로 들이받아 다치게 해 같은 날 살인미수 혐의로 체포됐다. 이 중 남자대학생(19)이 의식불명의 중태다. 이 차량은 도로를 100m 이상 역주행하면서 인명피해를 일으켰다. 새해 첫날 일본 도쿄 경시청 소속 경찰관들이 차량 테러가 일어난 현장에 출동하고 있다. NNN 캡처구사카베 용의자는 체포된 직후 자신의 행동을 테러라고 강조하며 범행 동기에 대해 “사형에 대한 보복”이라고 밝혀 옴진리교와의 관련성이 의심됐다. 구사케바 용의자는 이후 구체적으로 옴 진리교 사형수들에 대한 사형 집행이 범행의 이유라고 설명한 것이다. 옴진리교는 지난 1995년 도쿄 지하철역에서 13명을 숨지게 하고 6200명 이상을 부상하게 한 사린가스 테러사건을 일으킨 직후 해산됐다. 이후 일부 신자들은 아레후 등 새로운 단체를 만들어 활동하고 있다. 일본 경찰은 아레후가 아사하라 씨를 여전히 스승으로 모시고 있다며 경계를 늦추지 않고 있다. 새해 첫날 차량 테러가 일어난 일본 도쿄 도심에서 구급대원들이 부상자를 후송하고 있다. NNN 캡처구사카베 용의자는 스스로 옴진리교를 언급하기는 했지만, 그가 아레후 등 옴진리교 후속 단체와 관련이 있는지는 알려지지 않았다. 전날 사건과 관련해서는	된 메이지(明治)신궁 인근 도로를 100m가량 역주행하면서 행인 8명을 차례로 들이받아 다치게 해 같은 날 살인미수 혐의로 체포됐다. 중상자 4명 중 남자대학생(19)은 의식불명의 중태다. 지난해 7월 사형이 집행된 교주 아사하라 쇼코. ANN 캡처구사카베 용의자는 체포된 직후 자신의 행동을 테러라고 강조하며 범행 동기에 대해 “사형에 대한 보복”이라고 밝혀 옴진리교와의 관련성이 의심됐다.

	A씨가 범행에 사용한 차를 등유로 태우려고 계획했다는사실이 새로 드러나기도 했다. 지난해 7월 사형이 집행된 교주 아사하라 쇼코. ANN 캡처민영방송 TBS에 따르면 구사카베 용의자가 운전하던 차량 안에서는 등유 20ℓ가 든 기름통과고압 세정기가 발견됐다. 구사카베 용의자는 이와 관련해 경찰에 “등유로 차 전체를 태우려고 했다”고 말했다. 사고가 일어난 곳은 당시 새해가 되는 순간을 즐기려는 행인들로 북적였다. 구사카베 용의자가 차량을 태우는 범행도 실행에 옮겼다면 자칫 대량 인명 피해가 발생할 수 있었다.	
경인일보 유사도 76%	정부가 "청와대가 KT&G 사장교체를 지시하는 등 부당한 압력을 가했다"고 주장한 신재민(33·행정고시 57회) 전 기획재정부 사무관을 검찰에 고발할 예정이다. 2일 기획재정부에 따르면 신 전 사무관을 공무상비밀누설 혐의와 공공기록물 관리에 관한 법률 위반 혐의로 이날 오후 서울중앙지검에 고발할 계획이다. 앞서 기재부는 "공무원이었던 자가 직무상 취득한 비밀을 누설하는 것은 금지돼있다"며 "특히 소관 업무가 아닌 자료를 편취해 이를 대외 공개하는 것은 더욱 심각한 문제"라고 배경을 설명한 바 있다. 형법 127조는 공무원 또는 공무원이었던 자가 법령에 의한	정부는 청와대에서 기획재정부에게 KT&G 사장교체 지시 등의 압력을 가했다고 최근 폭로한 신재민(33·행정고시 57회) 전 기재부 사무관을 2일 검찰에 고발할 방침이다. 기재부 관계자는 이날 신 전 사무관을 공무상비밀누설 혐의와 공공기록물 관리에 관한 법률 위반 혐의로 이날 오후 서울중앙지검에 고발할 계획이라고 밝혔다. 앞서 기재부측은 "공무원이었던 자가 직무상 취득한 비밀을 누설하는 것은 금지돼있으며, 특히 소관 업무가 아닌 자료를 편취해 이를 대외 공개하는 것은 더욱 심각한 문제"라고 문제를 제기한 바 있다. 형법 127조에 의거, 공무원 또는

	직무상 비밀을 누설하면 2년 이하의 징역이나 금고 또는 5년 이하의 자격정지에 처하도록 규정하고 있다. 공공기록물관리에 관한 법률에 따르면 공공기록물을 무단 유출하는 경우 3년 이하의 징역 또는 2천만원 이하의 벌금형을 처할 수 있다. 앞서 신 전 사무관은 지난달 29일부터 유튜브와 고려대 인터넷 커뮤니티인 '고파스' 등에 올린 동영상과 글에서 청와대가 KT&G 사장을 교체하도록 압력을 넣었고 이에 정부가 기업은행을 동원해 영향력을 행사했다고 주장했다. 또 그는 2017년 11월 대규모 초과 세수입이 예상되는 상황에서 청와대가 적자 국채 발행을 요구하는 등 무리하게 개입했으며, 기재부는 문재인 정부의 정치적 부담을 고려해 1조원 규모의 국채매입을 갑자기 취소했다는 주장도 펼쳤다. 이와 관련, 기재부는 문건이 담배사업법상 정상적인 업무의 일환으로 KT&G 경영 현황 등을 파악한 결과물이며 사장 인사와 관련해 청와대의 지시가 있었다는 주장은 사실과 다르다고 밝혔다. 또 적자 국채 추가발행과 관련해서도 청와대도 의견을 제시했으나 강압적 지시는 전혀 없었다고 반박했다. 한편, 신 전 사무관은 기재부의 해명에 맞서 국채업무를 담당하던	공무원이었던 자가 법령에 의한 직무상 비밀을 누설하면 2년 이하의 징역이나 금고 또는 5년 이하의 자격정지에 처하도록 규정하고 있다. 공공기록물관리에 관한 법률에 따르면 공공기록물을 무단 유출할 경우 3년 이하의 징역 또는 2천만원 이하의 벌금형을 처할 수 있다. 신 전 사무관은 지난달 29일부터 유튜브와 고려대 인터넷 커뮤니티인 '고파스' 등에 올린 동영상과 글에서 청와대가 KT&G 사장을 교체하도록 압력을 넣었고 이에 정부가 기업은행을 동원해 영향력을 행사했다고 주장한 바 있다. 앞서 지난해 5월 MBC 뉴스는 "KT&G 사장 선임에 정부가 개입한 대응 문건이 확인됐다"며 기재부 내부에서 작성된 'KT&G 관련 동향 보고'라는 문서 내용을 보도했으며, 신 전 사무관은 자신이 해당 문건을 언론측에 제공했다고 자신의 유튜브 채널을 통해 밝힌 바 있다. 그러나 당시 KT&G 사장은 외국인 주주 등의 반대로 교체되지 않았다. 기재부는 문건이 담배사업법상 정상적인 업무의 일환으로 KT&G 경영 현황 등을 파악한 결과물이며 사장 인사와 관련해 청와대의 지시가 있었다는 주장은 사실과 다르다고 반박했다1. 신 전 사무관은 지난해 11월 대규모 초과 세수입이 예상되는 상황에서 청와대가 적자 국채 발행을 요구하는 등 무리하게

	조규홍 당시 재정관리관(차관보급)으로 추정되는 인물과 모바일 메신저 카카오톡으로 나눈 대화 화면을 일부 공개하기도 했다.	개입했으며, 기재부는 문재인 정부의 정치적 부담을 고려해 1조원 규모의 국채매입을 갑자기 취소했다는 주장도 펼쳤다. 기재부는 적자 국채 추가발행과 관련해 청와대도 의견을 제시했으나 강압적 지시는 전혀 없었고, 청와대와 협의를 거쳐 기재부가 적자 국채를 추가 발행하지 않기로 최종적으로 결정했다고 맞섰다. 기재부는 아울러 국채매입 취소는 적자 국채 추가발행 여부 논의 상황, 국채시장에 미치는 영향, 연말 국고자금 상황 등을 종합적으로 고려해 내린 어쩔 수 없는 결정이었다고 부연했다.
무등일보 유사도 64%	광주·전남 지역민 10명 가운데 6명 가량이 더불어민주당을 지지하는 것으로 나타났다. 무등일보와 뉴시스 광주전남본부, 사랑방닷컴이 여론조사전문기관 한국갤럽(Gallup Korea)에 의뢰해 지난달 27~28일 이틀간 광주와 전남에 거주하는 만 19세 이상 남녀 1천21명을 대상으로 실시한 정당 지지도 여론조사(표본오차 95% 신뢰수준 ±3.1%p)에 따르면 민주당 지지율이 58.4%로 나타났다. 정의당 9.7%, 민주평화당 4.1%, 바른미래당 3.7%, 자유한국당 1.6%로 지지도가 한자리 수를 면치 못했다. 텃밭경쟁에서 밀린 민주평화당은 정의당 보다 낮은 지지율을 기록했다. 정의당은 40~50대에서 18.2%와	올해 국내 경제 전망에 대해 광주·전남 지역민들은 지난해보다 나쁘거나 비슷한 수준이 될 것으로 내다봤다. 경기 불황이 장기화될 것으로 예상되는 가운데 젊은층일수록 경제 전망에 부정적인 것으로 나타났다. 무등일보와 뉴시스 광주전남본부, 사랑방닷컴이 여론조사전문기관 한국갤럽(Gallup Korea)에 의뢰해 지난달 27~28일 이틀간 광주와 전남에 거주하는 만 19세 이상 남녀 1천21명을 대상으로 2019년 국내 경제 전망 여론 조사(표본오차 95% 신뢰수준 ±3.1%p)를 실시한 결과 '좋아질 것'으로 예상한 응답자는 27.0%에 그쳤다. 반면 나빠질 것으로 전망한 응답자는 34.6%로 가장 많았으며 '비슷할 것'으로 응답한 경우도

	13.2% 등을 기록하는 고령층에서 지지를 얻었다. 민주평화당은 19~29세 3.2%, 30대 0%, 40대 3.9%, 50대 5.1%, 60세 이상 6.3%를 기록했다. 김현주기자 5151khj@srb.co.kr 어떻게 조사했나 광주·전남 2019년 새해를 맞아 무등일보와 뉴시스 광주전남본부, 사랑방닷컴은 공동으로 선거여론조사기관인 한국갤럽(Gallup Korea)에 의뢰해 대통령과 시장·지사, 시·도교육감에 대한 직무평가 등 지역 정치 현안을 알아봤다. 이번 조사는 지난달 27~28일 이틀간 광주·전남지역 만 19세 이상 지역민 1천21명을 대상으로 응답을 완료, 15.7%의 응답률을 나타냈다. 표본 추출은 통신사에서 제공한 휴대전화 가상번호 및 유선 무작위 전화걸기(RDD) 표본 프레임에서 무작위 추출했다. 조사방법은 무선 84%, 유선 16% 비율의 전화면접을 통해 실시했으며 표본오차는 95% 신뢰수준에 ±3.1%p다. 자세한 사항은 중앙선거여론조사심의위원회 홈페이지(www.nesdc.go.kr)를 참조하면 된다.	32.2%에 달했다. 성별로는 남성보다 여성이 경제 전망에 긍정적이었지만 표본오차 이내로 대체로 엇비슷한 수준이다. 남성은 응답자의 26.3%가 '좋아질 것'으로 예상했으며 여성은 이보다 1.3%p가 많은 27.6%가 긍정적으로 전망했다. 부정적으로 본 응답자도 남성(36.8%)이 여성(32.4%)보다 4.4%p가 높았으며 '비슷할 것'으로 응답 한 경우도 남성(32.9%)이 여성(31.5%)보다 높았다. 연령대별로는 19~29세, 30대는 경제 전망이 좋아질 것으로 기대한 이는 응답자의 20%에 못미치는 10%대 수준이다. 19~29세는 17.9%. 30대는 17.4%로 60대(35.9%)의 절반 수준에도 미치지 못했다. 19~29세 응답자 중 43.0%는 '나빠질 것', 37.1%는 '비슷할 것'으로 봤으며 30대 응답자는 37.5%가 '나빠질 것', 42.7%가 '비슷할 것'으로 전망했다. 자세한 조사 개요와 결과는 중앙선거여론조사심의위원회 홈페이지(www.nesdc.go.kr)를 참조하면 된다. 김현주기자 5151khj@srb.co.kr 어떻게 조사했나 광주·전남 2019년 새해를 맞아 무등일보와 뉴시스 광주전남본부, 사랑방닷컴은 공동으로 선거여론조사기관인 한국갤럽(Gallup Korea)에 의뢰해

		대통령과 시장·지사, 시·도교육감에 대한 직무평가 등 지역 정치 현안을 알아봤다. 이번 조사는 지난달 27~28일 이틀간 광주·전남지역 만 19세 이상 지역민 1천21명을 대상으로 응답을 완료, 15.7%의 응답률을 나타냈다. 표본 추출은 통신사에서 제공한 휴대전화 가상번호 및 유선 무작위 전화걸기(RDD) 표본 프레임에서 무작위 추출했다. 조사방법은 무선 84%, 유선 16% 비율의 전화면접을 통해 실시했으며 표본오차는 95% 신뢰수준에 ±3.1%p다. 자세한 사항은 중앙선거여론조사심의위원회 홈페이지(www.nesdc.go.kr)를 참조하면 된다.

<부록3> 신문기사 원문 자료 수집 및 정제 시 수작업 정제 지침

2020 신문말뭉치 작업 정제 및 검수 가이드라인

1. 신문말뭉치 작업 정제 기준

- 1) 정보를 전달하는 기사의 형식이어야 하며 2) 사진, 캡션 등의 정보는 삭제하고 3) 기자 정보가 없는 논설, 사설, 기고글, 광고, 부고, 승진 등의 글은 기사에 해당하지 않기 때문에 삭제합니다.
- 화면에 뜨는 두 개의 기사 중 왼쪽에 뜨는 기사는 기사 원문, 오른쪽에 뜨는 기사는 기계정제된 기사입니다. 왼쪽의 원문을 확인하고, 오른쪽에 있는 기사를 수작업 정제합니다.
- 정상적으로 작업을 진행 할 기사인지 아닌지 확인하는 기준 중 첫 번째는 기자 정보입니다.

2. 삭제 해야 하는 정보

- 사진, 캡션 정보 삭제
 예) <사진>, [일러스트 | 김상민기자], 2013년 4월 11일 KBS 화면캡처, [관련자료 링크],

- 사진을 설명하는 글 삭제

- 기자 정보(기자 이름, 이메일) 삭제
 예) 박보검 ㅇㅇ일보 기자 demian@chosun.com

- 사진 설명과 같은 캡션에 들어간 링크의 경우 삭제
 예) 이미지 출처 youtu.be/blahblahblah, 영상 출처 youtu.be/~~~~
 http://youtu.be/JRegVSmWooM (바리톤 토마스 햄슨, 번스타인 지휘 빈 필하모닉)

- 명단, 연도별 업적, 수상 내역 등의 내용 삭제
 예) '◇2019 오키나와 스프링캠프 명단(54명)' 과 그 이하 내용 삭제

2군 캠프 선수단은 내달 9일 대만으로 건너간다.
역시 캠프 초반 체력 및 기술, 전술 훈련을 소화하게 되는 대만 캠프단은 연습 경기 일정에 맞춰, 오키나와행에 도전하게 된다.
◇2019 오키나와 스프링캠프 명단(54명)
▲코칭 스태프(14명) = 김기태 김민호 강상수 코우조 김종국 홍세완 김민우 이대진 서재응 김상훈 박종하 배요한 정상옥 고영득 ▲투수(20명) = 윤석민 양현종 터너 윌랜드 김윤동 임기영 한승혁 이민우 황인준 김세현 문경찬 ▲내야수(11명) = ~~~~ ▲외야수(6명) = ~~~~

- 불필요한 줄바꿈 삭제

- 본문 아래 특정 내용 추가 해설을 위해 따로 존재하는 정보(혹은 전문 인용)의 경우 삭제

3. 작업 불가 처리
- 논설, 사설, 기고글, 광고, 부고, 운세, 승진 글은 작업불가 처리합니다.

- ~위원, ~연구소장, ~교수, ~논설위원, ~평론가 등 기자가 쓴 글이 아닌 다른 투고 글은 작업불가 처리
* 단, 해당 매체의 사회부 부장, 경제부 부장 등의 경우 유지

- 인터뷰 기사의 경우 기자정보가 없고, 구어체를 사용하는 경우 삭제

- 객원 기자의 기사 작업 불가 처리

4. 유지 및 남겨야 하는 기사

- 소속매체의 00부 부장 등이 쓴 기사, 논설 글의 경우 유지.
(OO 부 부장, OO 부 차장과 같이 부서와 직책이 나와있을 시 해당 언론사 내의 부서인 것으로 간주하고 유지함)

- 시민기자 및 학생기자가 쓴 글은 유지 (단, 구어체가 많을 경우 삭제)

- 기사 본문 내용에 해당되는 링크는 삭제하지 않음
예) 신청은 행사 페이지(www.event.com)에서 가능하다.

<Abstract>

The Sejong Corpus which was built as a "21st Century Sejong Project" was the alrgest in the world at that time, but has not been built continuously. Currently, the Sejong corpus is far behind the corpus construction of major countries such as the US, China, and Japan. Therefore, The Korean courpus construction project, which can be used as public goods for the development of artificial intelligence services and technological innovations in the 4th industrial revolution, was resumed.

This project is an extension of the newspaper corpus collection and refining project conducted in 2019 as a part of the construction of large-scale Korean language learning materials of the National Institute of the Korean Language that can be used as public goods in the artificial intelligence industry and related research institutions from 2018. The 2020 original newspaper articles data collection and refining project collected over 10 million words per month of original newspaper articles in various fields published during the year 2019, and built the latest corpus available for public use. The corpus of newspaper articles built through this project will be able to contribute to the development of various technologies and research in the industry and academia, including high-tech industries such as artificial intelligence industry.

The scope of this project can be divided into four parts: collect the original newspaper articles, media composition and securing secondary copyrights, refinement and normalization work, and tagging metadata. In addition, it was carried out in four steps in preparation for construction, selection of media, collection and digitization of original text data, removal and purification of duplicate articles, attachment of meta information, and creation of a list.

Collecting the original data of newspaper articles included 3 national comprehensive magazines and 2 internet media (10% or less of the total number of media), and 35 media were selected. After negotiations regarding permission to use copyright, a copyright license agreement and an annexed agreement were signed between the National Institute of the Korean Language and the media and between the project companies.

A total of 1,839,277 articles and 328,431,587 word segments were collected from the selected media, and developed a tool for workers to refine them. The refining tool was built as a system that allows multiple workers to work at the same time.

Workers who logged in to the website were able to work as a project unit with as few as 3,000 to 4,000 articles and as many as 15,000 articles based on the distributed manuals.

On the other hand, separate from this, the automatic purification work was first performed before the manual purification work for the workers. Depending on the length of the article, the main task was to exclude articles that were too short, articles that were overlapped by more than a certain level, and articles from media that did not sign a copyright agreement for this project. As a result of this first refinement, the number of articles was 1,213,575 and 251,579,628 word segments.

In the second manual refining, the work was carried out according to the standards agreed with the National Institute of the Korean Language, including caption information such as images, tables, and graphics, copyright-related information and article information of the article, and information not related to the article content (context), articles from other media with the possibility of copyright issues, articles written by external contributors, articles that are difficult to see as general newspaper articles, and articles in colloquial style. Workers worked by sharing detailed work manuals from online, resulting in 630,095 articles and 150,669,174 word segments.

After the first and second refining, the final corpus naming convention and encoding method were applied according to the newspaper corpus construction guidelines, and the meta data to be included in the corpus file was determined. Meta data is composed of title, article author, newspaper company, article creation date, subject, keyword, and article summary content, and in the case of article subject classification, two types of classification are included: media self classification and integrated classification.

The final corpus file was produced in JSON format, and the basic structure of JSON is composed of id, metadata, and document parts.

The original corpus of newspapers built by collecting and securing the right to use the original newspaper articles for one year in 2019 is a language resource that reflects the actual language life. It is expected to be used in various industrial fields such as performance improvement and evaluation of processing application systems.

사업 책임자 안준환(주식회사 마인즈랩 전무)
사업 참여자 임성모(주식회사 마인즈랩 이사)
서상원(주식회사 마인즈랩 팀장)
송혜원(주식회사 마인즈랩 매니저)
이원문(주식회사 마인즈랩 매니저)
박지원(주식회사 마인즈랩 매니저)
정소라(주식회사 마인즈랩 매니저)
손영효(주식회사 마인즈랩 매니저)
권영현(주식회사 마인즈랩 매니저)
이미수(주식회사 마인즈랩 매니저)
담당 연구원 이승재(국립국어원 언어정보과장)
김소희(국립국어원 언어정보과 학예연구사)

발행인: 국립국어원장
발행처: 국립국어원
서울시 강서구 금낭화로 154
전화 02-2669-9775, 전송 02-2669-9757
인쇄일: 2020년 12월 24일
발행일: 2020년 12월 24일
인 쇄: 비즈카피

※ “이 책은 국립국어원의 용역비로 수행한 ‘신문 기사 원문 자료 수집 및 정제’ 사업의 결과물을 발간한 것입니다.”