국립국어원 2019-01-33

발 간 등 록 번 호
11-1371028-000790-01

신문 기사 원문 자료 수집 및 정제

사업 책임자

황 이 규

제 출 문

국립국어원장 귀하

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

■ 사업기간: 2019년 06월 ~ 2019년 12월

2019 년 12 월 12 일

사업 책임자: 황 이 규 (주식회사 마인즈랩)

사업 수행 기관 주식회사 마인즈랩
사업 책임자 황이규
사업 참여자 안준환, 진영순, 서상원, 임성모,
정소라, 박영선, 박다솜, 송혜원,
윤서영, 송동훈, 김종범, 이재성,
이석준, 김마로, 이원문

〈사업 수행자〉 주식회사 마인즈랩

사업 책임자	황이규(주식회사 마인즈랩 전무이사)
사업 참여자	안준환(주식회사 마인즈랩 상무)
	진영순(주식회사 마인즈랩 이사)
	서상원(주식회사 마인즈랩 팀장)
	임성모(주식회사 마인즈랩 이사)
	정소라(주식회사 마인즈랩 매니저)
	박영선(주식회사 마인즈랩 매니저)
	박다솜(주식회사 마인즈랩 매니저)
	송혜원(주식회사 마인즈랩 매니저)
	윤서영(주식회사 마인즈랩 매니저)
	송동훈(주식회사 마인즈랩 상무)
	김종범(주식회사 마인즈랩 팀장)
	이재성(주식회사 마인즈랩 팀장)
	이석준(주식회사 마인즈랩 매니저)
	김마로(주식회사 마인즈랩 매니저)
	이원문(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 최근 10 년간의 다양한 분야의 신문 기사 원문을 수집하여 공공으로 사용 가능한 최신 말뭉치로 구축하기 위한 신문 기사 원문 자료 수집 및 정제 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.

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 중앙 종합지 5 개, 인터넷 매체는 4 개(전체 매체 수 대비 10% 이하)를 포함하여 최종적으로 42 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 18,369,901 건 및 3,351,131,155 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 8,000~9,000 건에는 많게는 20,000 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 짧거나 긴 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 5,029,926 건 및 1,656,947,078 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 3,991,282 건 및 1,003,899,229 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 데이터에 대해 결정하였다. 메타 데이터는 제목, 저자, 발행자, 연도, 기사번호, 분류, 기사 작성일, 기사 작성자, 어절 수로 구성되며, 기사의 주제 분류의 경우 매체 자체 분류와 통합 분류의 2 가지를 포함하였다.

최종 말뭉치 파일은 SJML 형식으로 제작하였으며, SJML 의 기본 구조는 header 와 text 부분으로 구성되어 있다.

본 사업에서 최근 10 년간의 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.

주요어: 문어 말뭉치, 신문 기사, 현대 한국어, 기사 말뭉치

차 례

제1장 서론

제2장 사업 수행 내용

제3장 사업 수행 결과

표 차례

그림 차례

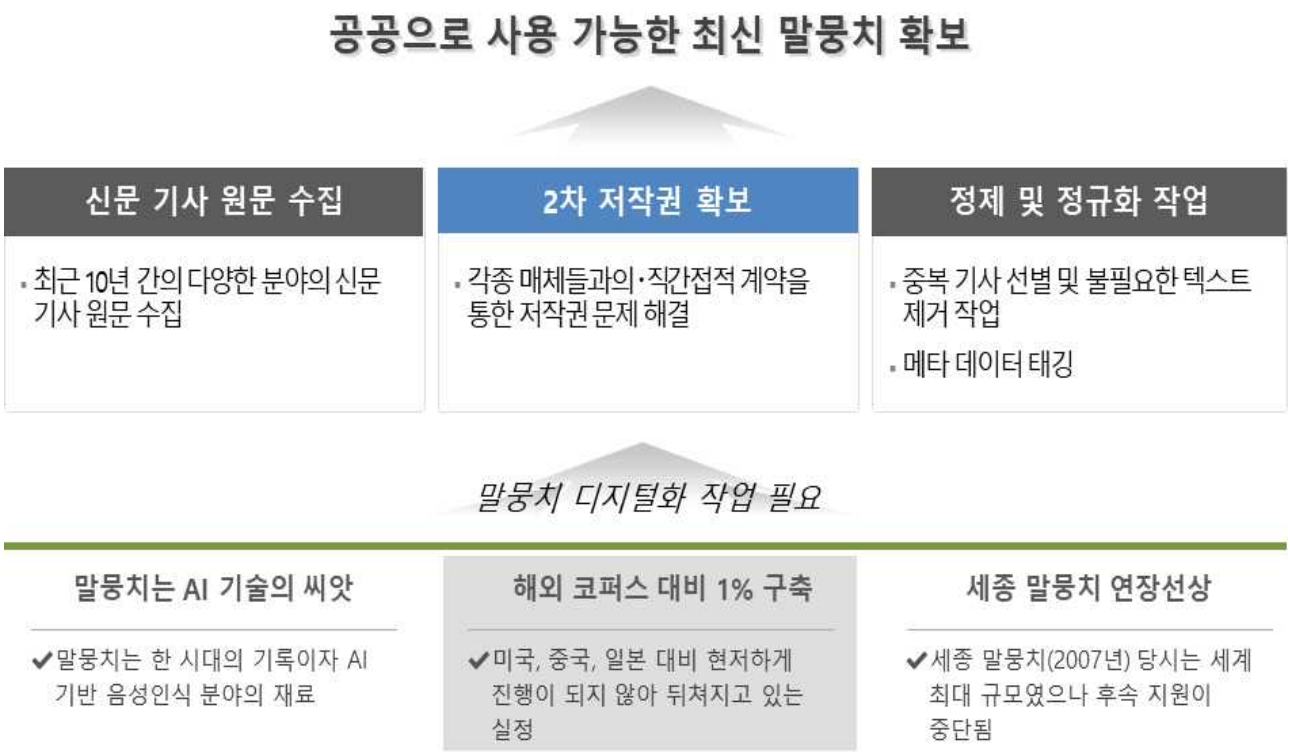
제 1 장

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.

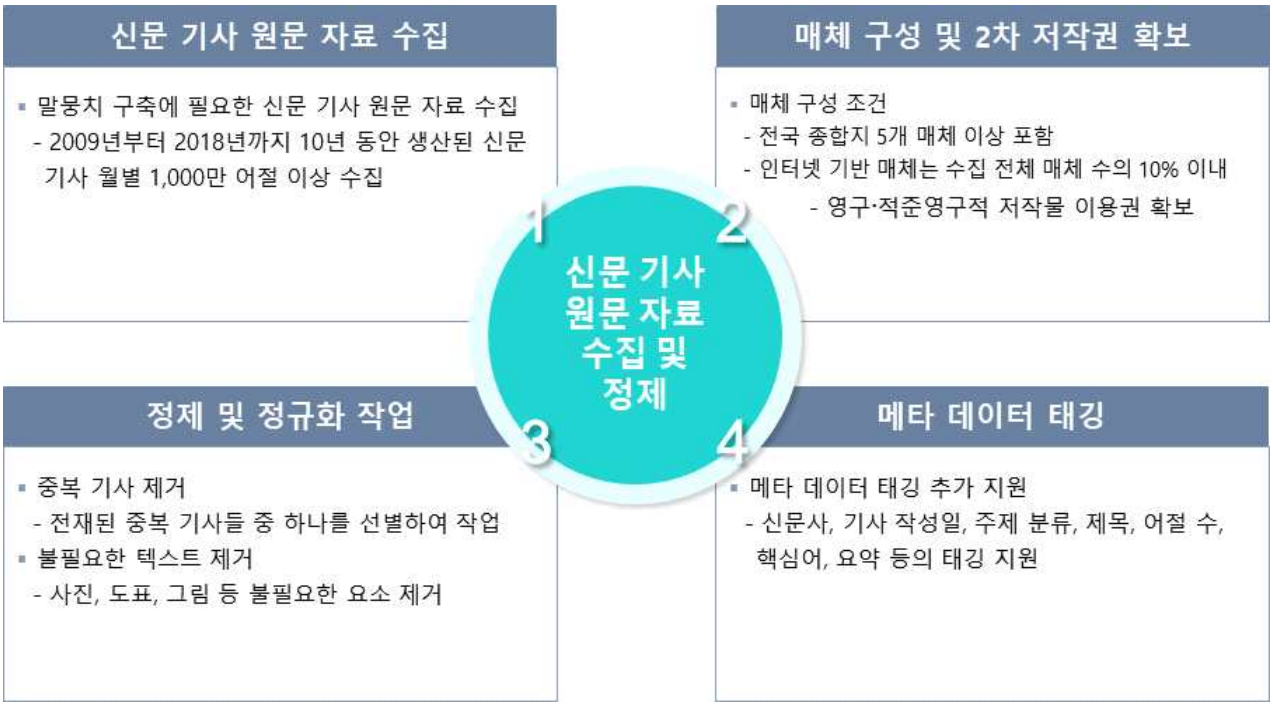
본 사업은 2018 년부터 시작된 “4 차 산업혁명 대비 국어 빅데이터(말뭉치) 구축” 사업의 일환으로 추진되는 사업으로, 최근 10 년간의 다양한 분야의 신문 기사 원문을 수집하여 공공으로 사용 가능한 최신 말뭉치로 구축하기 위한 신문 기사 원문 자료 수집 및 정제 사업이다. 본 사업을 통해 구축된 신문 기사 말뭉치는 인공지능 산업 등 첨단 산업을 비롯하여 산업계 및 학계에서 각종 기술 개발과 연구 발전에 이바지할 수 있을 것이다.



<그림 1> 사업의 배경 및 목적

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 2009 년부터 2018 년까지 10 년 동안 생산된 신문기사 월별 1,000 만 어절 이상 수집하는 것을 목표로 한다. 둘째, 매체 구성 및 저작권 확보에서는 전국 종합지 5 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 요소를 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 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.

첫 번째 매체 선정 단계에서는 중앙 종합지, 지역 종합지, 전문지, 인터넷 신문을 대상으로 총 40 개 이상 매체를 선정하고, 대상 매체의 원문 자료를 말뭉치 구축과 활용하는 데에 필요한 저작권 이용 허락 계약을 체결하는 것이다.

두 번째로는 저작권 이용 허락 계약을 맺은 각 신문사 및 한국언론진흥재단으로부터 신문 기사 데이터를 입수하고, 국립국어원이 제시한 표준 형식을 기반으로 원문 기사의 데이터베이스를 마련하는 것이다.

세 번째로 진행된 기사 정제는 2 단계에 걸쳐 진행된다. 1 차 자동 정제는 중복 기사 제거 단계로 제목 및 본문에서 실질 형태소와 어절 사전을 만들어 비교하여 이를 바탕으로 중복 여부를 확인하여 기준에 따라 제거하는 것과, 너무 길거나 짧은 기사(100 어절 이하, 1000 어절 이상)를 제거하는 것이다. 2 차 수작업 정제에서는 이미지, 표 등의 캡션 정보를 제거하고 저작권 문제가 있는 내용 및 본문 내 기자 정보를 제거하며 외부 기고가의 글을 제거한다.

마지막으로 메타 정보를 추가하는데 필수 메타 정보 항목에 핵심어, 요약 등을 추가하고 신문사, 기사 작성일, 주제 분류, 제목, 어절 수 등 수집 기사의 정보를 부착한 후 목록을 작성한다.

<그림 3> 사업 수행 절차

4. 사업 추진 경과

본 사업의 추진 경과는 다음과 같다.

단계	작업 내용	6월	7월	8월	9월	10월	11월	12월
준비	착수 보고							
매체 선정 및 데이터 확보	매체 선정							
	매체 계약							
	데이터 확보							
정제 도구 개발	정제 도구 개발							
정제	데이터 정제							
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 발행 부수, 매체의 종류 및 사업 요구 사항 등을 종합적으로 고려하여 최종적으로 42 개 매체를 선정하였다. 그중 중앙 종합지는 5 개를 선정하였으며, 인터넷 매체는 4 개로 전체 매체 수 대비 10% 이하로 선정하였다.

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

<표 2> 선정 매체 목록

대상 매체 선정 후 저작권 이용 허락에 관련한 협상 과정에서 가장 큰 문제가 되었던 사항은 이용 기간에 관한 것이었다. 제안요청서 상의 저작권 이용 허락 범위에 대한 요구 사항은 '이용 기간은 영구적, 또는 준영구적이어야 함'으로 규정되어 있으나 매체 측 의견은 영구적이거나 준영구적인 저작권 이용 허락에 대해 부정적인 입장이었다. 이에 국립국어원과 사업 수행사는 매체 관계자와의 3 자간 협의를 통해 다음과 같이 정리하였다.

- 본 사업에서 수집하는 저작물(신문 기사)은 말뭉치 구축 및 활용 목적으로, 산업계 및 학계의 기술 개발 및 연구에 활용하기 위함이며 별도의 콘텐츠 재판매 등 수익 사업을 위한 것이 전혀 아님을 확인하였다.

- 국립국어원과 매체가 맺는 2 자간 계약서에 다음과 같은 조항을 포함하여 저작권 이용 허락 기간에 관한 문제를 해소하였다.

> 대상저작물 및 복제·변형물의 이용허락 최소 기간은 계약체결일로부터 2030 년 12 월 31 일까지로 한다. 최소 기간 만료 후 권리자 또는 저작자인 언론사가 이용허락 중지 의사를 밝히지 아니하면 이용 허락이 1 년 (또는 5 년) 단위로 자동 갱신되며, 권리자 또는 저작자인 언론사가 이용 허락 중지 의사를 밝히면 그 의사 내용에 따라 이용 허락이 중지된다.

- 저작권 이용허락 계약과는 별도로 국립국어원과 매체, 사업 수행사 3 자간에 부속합의서를 체결하여 매체는 사업 수행사에게 10 년치의 신문 기사 데이터를 인도하고, 사업 수행사는 이에 대한 검수 후 저작권 이용료를 지급하기로 하였다.

2. 원문 자료 수집

선정된 매체로부터 수집된 기사는 총 18,369,901 건 및 3,351,131,155 어절로 기사 당 평균 182 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
건 수	노컷뉴스 (약 109만 건)	광주일보 (약 20만 건)
	충청투데이 (약 88만 건)	울산매일 (약 19.5만 건)
	경향신문 (약 85만 건)	무등일보 (약 16만 건)
	강원도민일보 (약 78만 건)	스포츠동아 (약 12만 건)
	전자신문 (약 77만 건)	미디어오늘 (약 6만 건)
어절 수	노컷뉴스 (약 214백만 어절)	한라일보 (약 34백만 어절)
	경향신문 (약 165백만 어절)	울산매일 (약 32백만 어절)
	오마이뉴스 (약 155백만 어절)	미디어오늘 (약 31.9백만 어절)
	전자신문 (약 147백만 어절)	무등일보 (약 31.6백만 어절)
	조선일보 (약 134백만 어절)	스포츠동아 (26백만 어절)

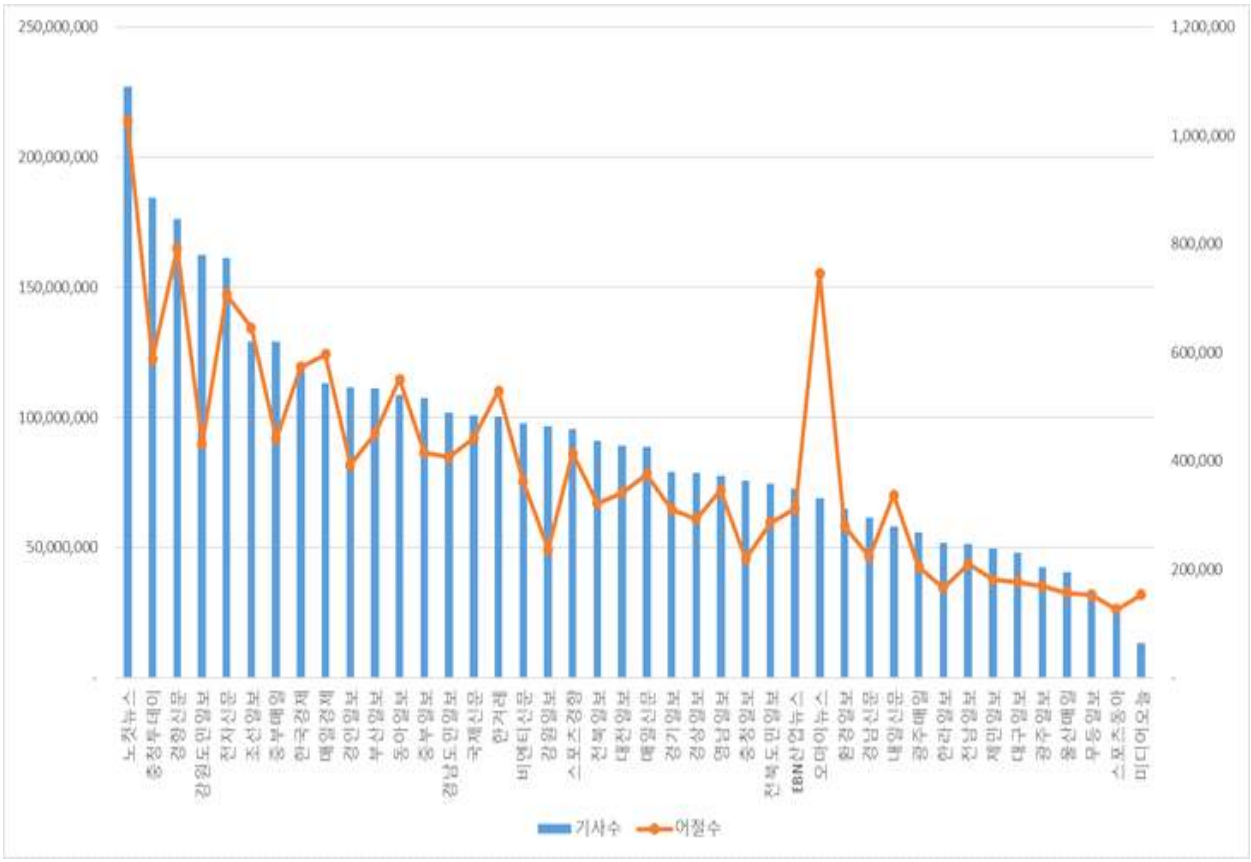
<표 3> 수집 원문 데이터의 건수 및 어절 수 요약

42 개 매체 전체에서 수집된 기사 수와 어절 수는 다음과 같다.

매체	기사 수	어절 수	매체	기사 수	어절 수
노컷뉴스	1,089,748	214,036,646	매일신문	426,008	78,187,699
충청투데이	885,897	122,442,647	경기일보	380,127	64,546,365
경향신문	846,644	165,137,364	경상일보	377,566	61,052,030
강원도민일보	779,972	89,816,988	영남일보	372,341	71,998,576
전자신문	774,292	147,324,182	충청일보	363,531	45,797,309
조선일보	620,356	134,466,041	전북도민일보	358,896	59,607,470
중부매일	620,268	91,642,899	EBN산업뉴스	349,896	65,061,170
한국경제	564,897	119,346,015	오마이뉴스	330,470	155,435,131
매일경제	542,379	124,183,689	환경일보	311,521	58,211,205
경인일보	535,609	81,733,649	경남신문	295,445	46,487,505
부산일보	534,174	93,949,156	내일신문	279,155	70,243,235
동아일보	520,896	114,771,316	광주매일	268,948	42,837,324
중부일보	516,457	86,456,525	한라일보	248,109	34,450,260
경남도민일보	488,412	84,728,034	전남일보	246,405	43,744,286
국제신문	484,548	92,077,374	제민일보	237,920	37,614,780
한겨레	481,730	110,202,861	대구일보	230,185	36,798,151
비엔티신문	469,942	75,325,215	광주일보	203,100	35,270,182
강원일보	463,429	49,200,333	울산매일	195,139	32,716,291
스포츠경향	458,649	86,225,965	무등일보	163,422	31,682,425
전북일보	436,336	67,057,000	스포츠동아	124,396	26,272,566
대전일보	428,577	71,092,651	미디어오늘	64,109	31,900,645

<표 4> 수집 원문 데이터의 기사 건수 및 어절 수 전체

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



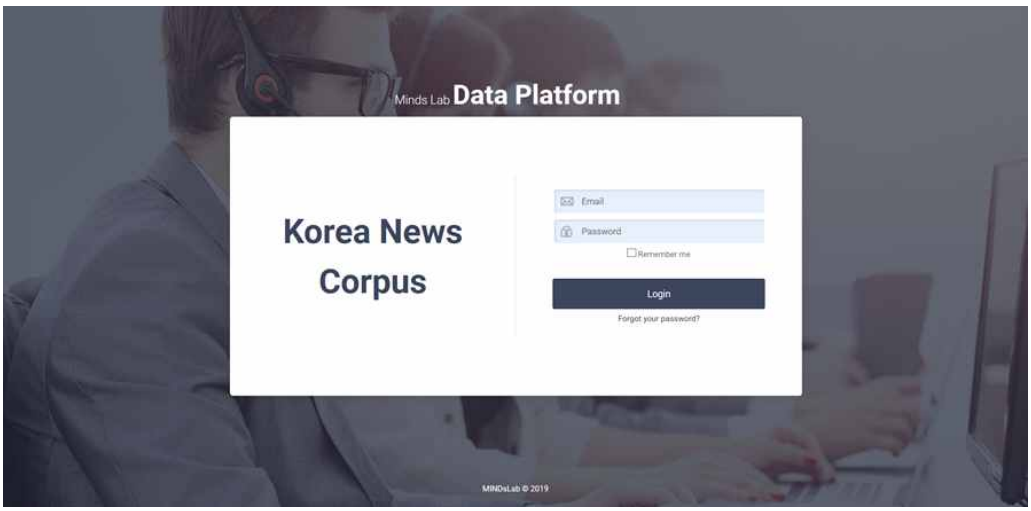
<그림 4> 원문 데이터 기준 매체별 기사 수 및 어절 수

수집된 매체별 기사 원문의 기사 수와 어절 수 통계를 살펴본 결과, 중앙 일간지의 기사 수가 많을 것으로 예상했던 바와는 달리 인터넷 뉴스 매체인 노컷뉴스의 기사 수가 가장 많았고, 대체로 지방지의 어절 수가 기사 수에 비하여 적었다. 특히 <그림 4>의 통계 그래프를 보면 '오마이뉴스'가 특이한 양상을 보이고 있는 것을 알 수 있는데, 기사 수 대비 어절 수가 평균 대비 2 배 정도 많은 것으로 나타났다. 이는 "모든 시민은 기자"라는 기치 아래 직업 기자에 의한 심층 취재 뉴스와 일반 시민 기자에 의한 생활 체험 뉴스를 5:5 비율로 편집하는 등의 '오마이뉴스'만의 특성에 근거한 것으로 보인다. 그 이외의 매체의 경우에는 약간의 차이는 있으나 기사 수와 어절 수가 비슷한 양상으로 수집된 것을 알 수 있다.

1) 좌측 세로축이 어절 수, 우측 세로축이 기사 수 임. 이하 동일

3. 정제 도구 개발

신문 기사의 수작업 정제 작업을 위해 정제 도구를 개발하였다. 목표 산출물이 10 억 어절 이상의 많은 양이므로 정제 도구는 다수의 작업자가 동시에 작업을 할 수 있는 시스템으로 구축하여야 한다.



<그림 5> 정제 도구 로그인 페이지

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

<그림 6> 프로젝트 선택 페이지

로그인한 작업자는 신규 프로젝트 창에서 본인의 이름(혹은 아이디)으로 지정된 프로젝트를 더블클릭한 다음 시작하기 버튼을 클릭하여 작업을 시작한다. 이미 선택된 프로젝트가 있다면 진행 중인 작업에서 선택하여 더블클릭한다.

<그림 7> 정제 작업 페이지

프로젝트 화면은 <그림 7>와 같이 구성된다. 왼쪽 창이 기사의 원문, 오른쪽 창이 정제해야 할 기사 본문이다. 기사 제목, 저작권자 등 메타 데이터는 말뭉치로 구축하는 과정에서 프로그램에 의해 자동으로 처리할 것이므로 작업자는 순수하게 본문만을 대상으로 작업한다.

<그림 8> 정제 기능

만약 해당 기사 전체가 삭제되어야 하는 것이라면 ① 비우기 버튼을 클릭하고 ② 수정 버튼을 클릭한다. 본문 중 일부가 삭제 대상이라면 오른쪽 창에서 직접 편집한 뒤 수정한다. 수정 버튼을 클릭하면 나오는 메시지 창에서 ③ OK 버튼을 클릭하면 해당 기사는 저장 완료되고 다음 기사 본문을 작업할 수 있다.

하나의 프로젝트는 적게는 8,000~9,000 건에서 약 20,000 건의 기사로 이루어져 있으며, 그 기준은 한 명의 작업자가 1 주일 동안 작업할 수 있는 분량을 목표로 하였다. 프로젝트 내 모든 기사의 수정이 끝나면 프로젝트는 자동으로 완결되고, 작업자는 다음 프로젝트를 배정받아 작업을 하게 된다.

4. 1차 자동 정제 작업

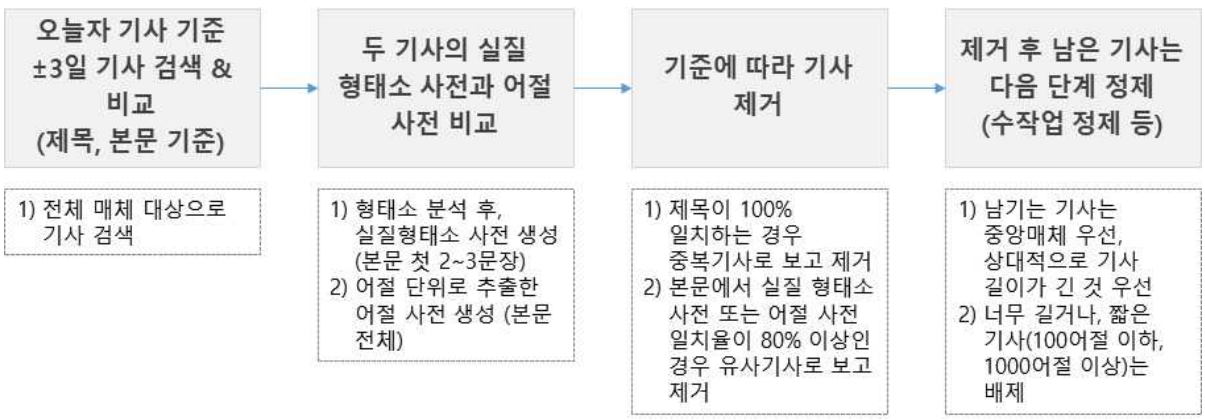
수집된 원문 기사를 대상으로 한 1차 자동 정제 작업은 기사 중에 말뭉치로 구축하기에 적절하지 않은 기사를 선별하여 제거하는 작업이다. 기사의 길이에 따른 제거 작업과 기사 내용 중복 기준에 따른 정제 작업으로 상세한 작업 기준은 다음과 같다.

○ 신문 기사 길이가 100 어절 이하 및 1000 어절 이상인 기사는 배제했다. 한 줄짜리 속보 기사 등 지나치게 짧은 기사와 지나치게 긴 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 42 개 전체 기사를 대상으로 비교하였다.

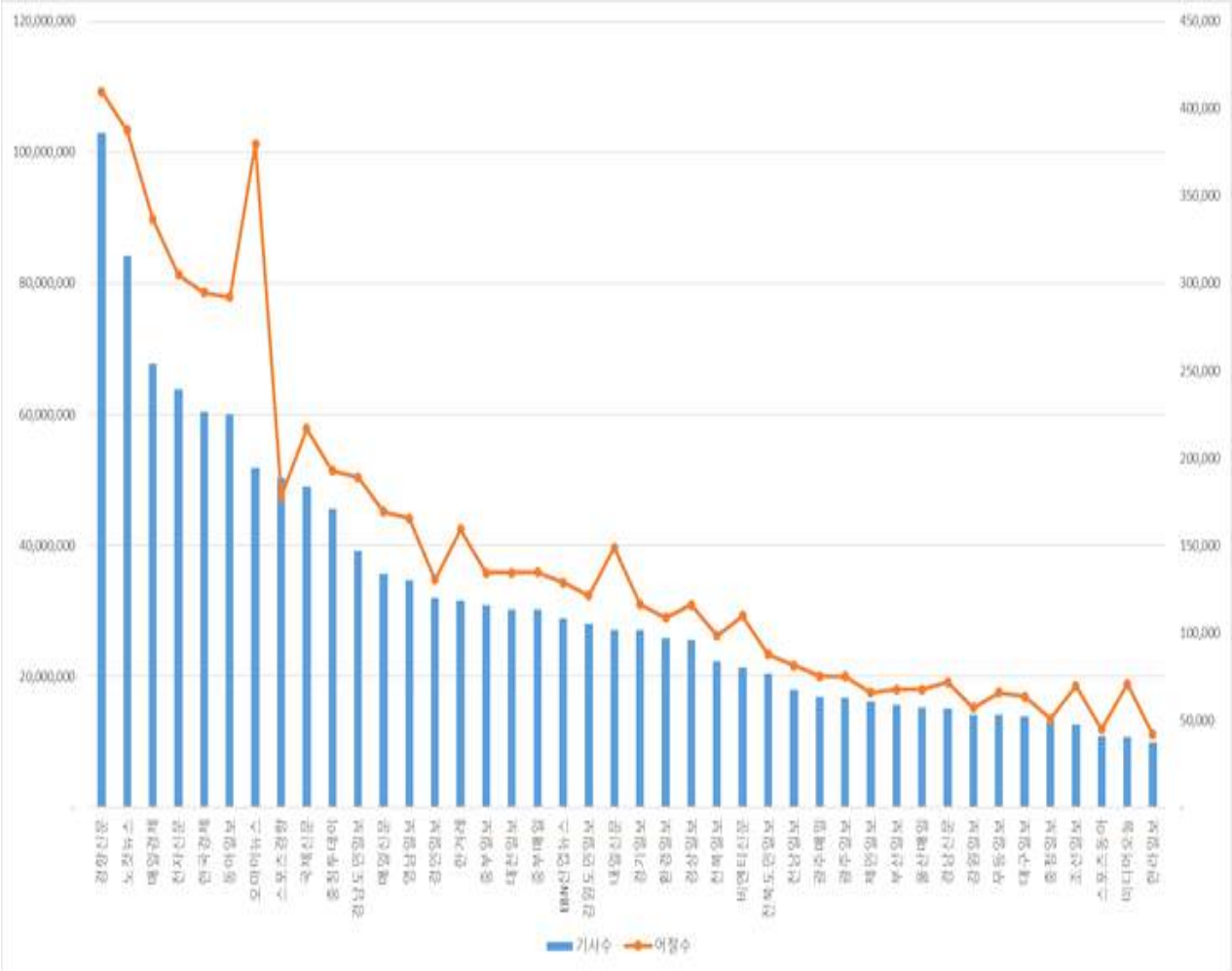
○ 저작권 이용 허락 계약을 맺은 매체에 실린 기사라 하더라도 저작권이 그 매체에 있지 않아 저작권 문제가 생길 수 있는 기사를 제외하였다. '연합뉴스'의 경우 여러 매체에서 자주 전제하거나 그대로 보도하는 경우가 많은데 '연합뉴스'는 이번 사업의 저작권 이용 허락 계약을 맺지 않은 매체이므로 전제된 '연합뉴스' 출처의 기사는 배제하였다.



<그림 9> 중복/유사 기사 제거

1 차 자동 정제 작업을 거친 후 어절 수 기준의 기초 통계를 내본 결과 '조선일보', '동아일보', '한겨레신문' 등 중앙 일간지의 경우 일치율 80%의 유사 내용 기사 제거 기준을 통해 많은 수의 기사가 제외되어 전체 정제 작업을 완료되었을 때 목표인 10 억 어절에 미달될 우려가 있었다. 그래서 국립국어원과의 협의를 통해 '조선일보', '동아일보', '경향신문', '한겨레신문' 등 4 개 중앙 일간지에 대해서는 유사 내용 기사 제거 기준을 완화하여 유사도 80~85% 구간의 기사도 2 차 수작업 정제 작업 대상에 포함하였다.

최초 수집 기사 총 18,369,901 건 및 3,351,131,155 어절에 대하여 1 차 자동 정제를 한 결과 5,029,926 건 및 1,656,947,078 어절의 결과가 도출되었다. 1 차 정제 후 기사 수 및 어절 수 통계는 다음과 같다.



<그림 10> 1 차 정제 후 매체별 기사 수 및 어절 수

최초 수집 데이터 기준 기사 수 1, 2 위였던 ‘노컷뉴스’와 ‘충청투데이’는 1 차 자동 정제 작업 후 각각 2 위와 10 위로 나타났으며, 29 위였던 ‘오마이뉴스’는 7 위로 순위 상승하였다. 이는 길이 기준 및 중복·유사 기준 중심으로 한 자동 정제가 매체별로 다른 결과를 가져온 것임을 알 수 있다.

5. 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 수집하는 것으로 캡션 정보들이 남아 있으면 전체 문맥에 혼란을 주기 때문이다.

구분	예시		
캡션 정보	사진제공= 사진| 사진= (사진출처:	[그래픽] <그래픽> 일러스트| 화면 캡처	▲영상제공= (특히 ▲와 함께 나오는 내용은 그 내용을 확인해서 캡션 정보로 판단되는 경우 삭제)

<표 5> 캡션 정보 삭제 예시

○ 해당 기사의 저작권 관련 정보는 메타 데이터로 작성하기로 하였으므로 기사 본문에서 삭제한다. 주로 'Copyright©'로 시작하거나 '저작권자 © 경남도민일보 무단전재 및 재배포 금지' 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@', 'hong@sinmun.co.kr' 등이다. 1차 자동 정제 과정에서 기자 관련 정보를 삭제 처리하기는 했지만, 간혹 '서울='이나 '정리='처럼 남는 경우도 있으므로 이것도 직접 수작업을 통해 삭제한다.

○ 기타 기사 내용(맥락)과 관련 없는 정보는 삭제한다. 대부분의 매체에서 기사 데이터는 자체 홈페이지 등 인터넷 매체를 통한 서비스 기준으로 관리되고 있으므로 기사 말미에 다른 기사로의 링크 정보나 광고 사이트로의 링크 정보가

붙어 있는 경우가 많은데 이런 내용은 수작업 정제 작업자가 개인적으로 판단하는 것보다 전체적인 기준으로 정제하고자 하였다.

○ 저작권 문제의 가능성이 있는 타 매체의 기사를 완전히 삭제한다. 앞서 자동 정제 작업에서 명백히 '연합뉴스'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.

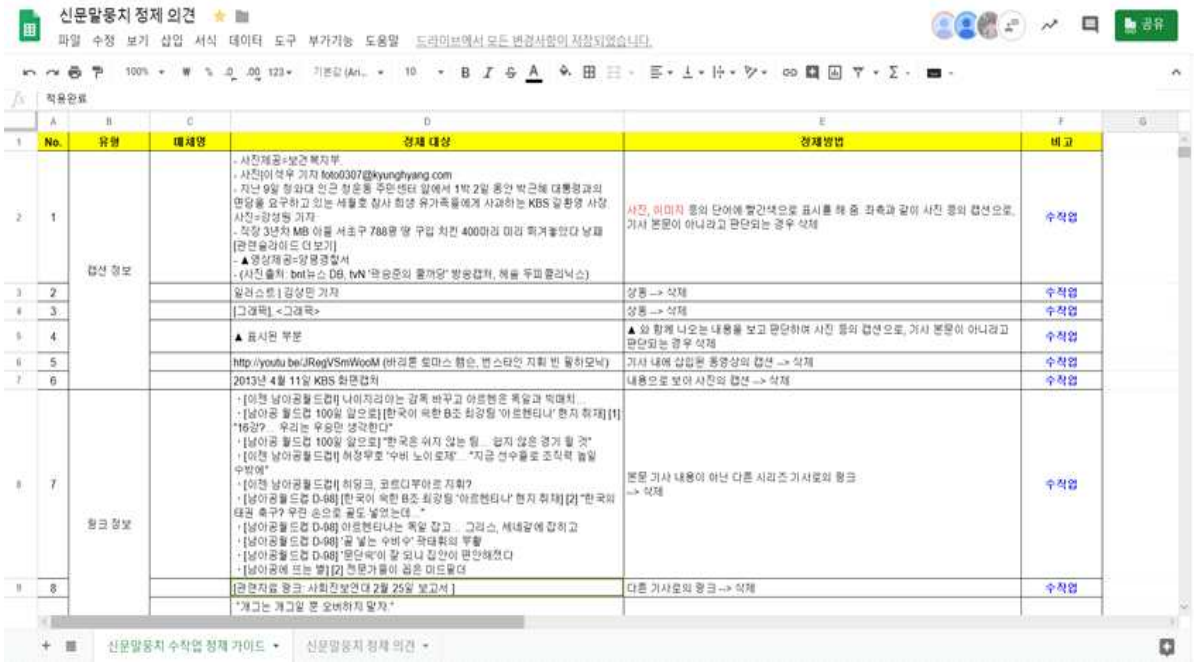
○ 기타 일반적인 신문 기사로 보기 어려운 내용을 삭제하였다. 주로 승진, 부고, 운세 등의 기사였는데 기자가 직접 작성한 내용이 아니라 처음부터 끝까지 승진자나 부고 명단, 띠별 오늘의 운세, 퀴즈, 스포츠 경기의 결과 수치만으로 구성된 기사나 기사의 대부분이 영어나 일어 등 다른 언어로 된 것도 있었다. 또한 뉴스 기사의 특성이 전혀 없는 시(詩)나 소설 등 문학작품은 삭제한다.

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

수작업 정제 작업자들은 구인구직 사이트와 고용노동부 워크넷 (www.work.go.kr)을 통해 출근해서 근무하는 단기 계약직(이하 아르바이트) 및 재택근무 아르바이트로 나누어 채용하고 작업을 진행하였다. 채용 인력은 출근 아르바이트는 34 명[2], 재택 아르바이트는 17 명으로 최종적으로 51 명이 작업을 하였으며 이 인력들과 작업을 통제

2) 작업 초기에는 40명 정도로 시작했으나, 전체적인 투입 M/M 기준으로 보면 34명으로 집계되었다.

감독할 사업 수행사 인력과의 의사소통 및 정제 원칙의 공지를 위해 구글 문서와 단체 대화방을 활용하였다.

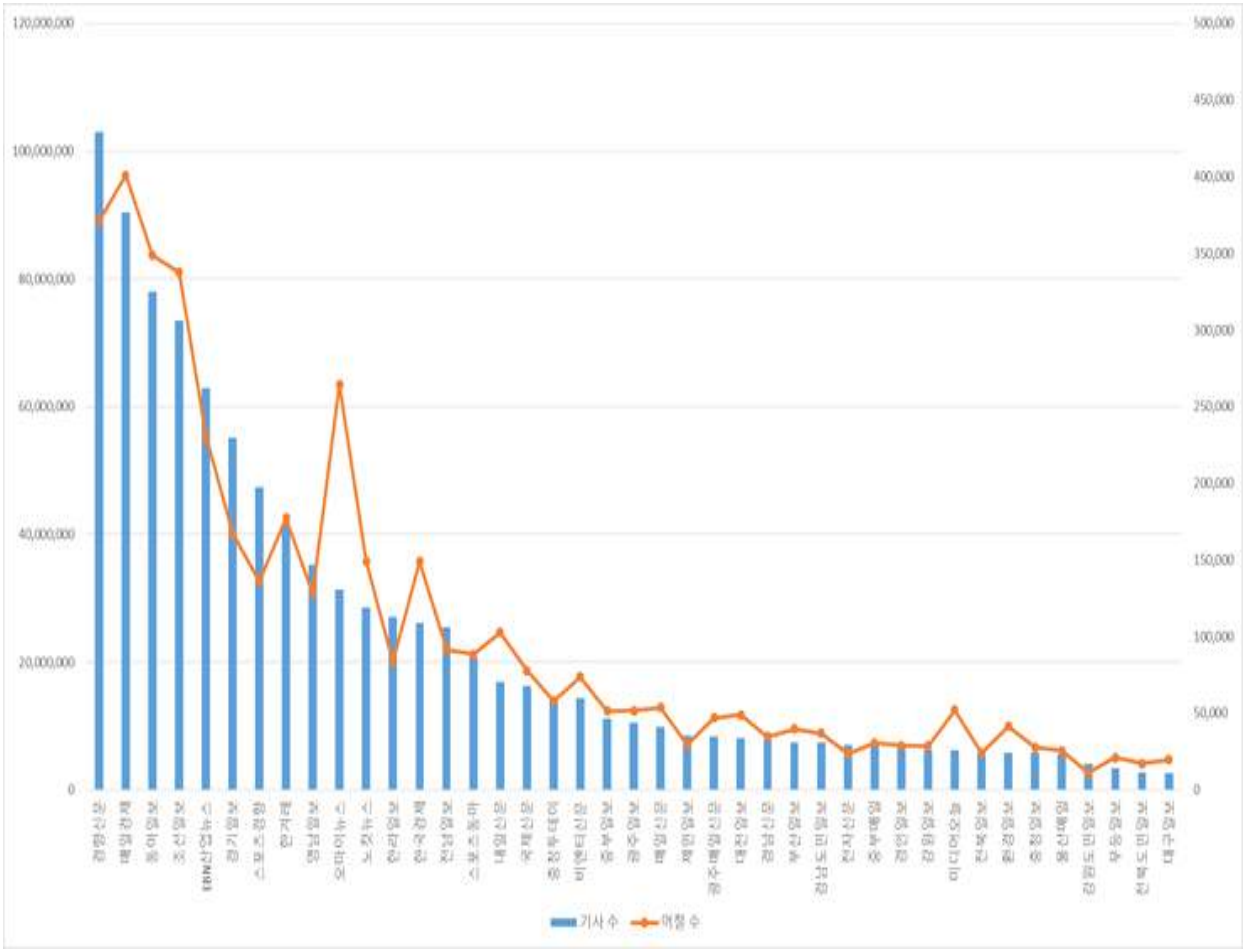

신문말뭉치 정제 의견

No.	유형	대체명	정제 대상	정제방법	비고
1	캡션 정보		- 사진제공=보건복지부 - 사진이 작우 기자 toto0307@kyunghyang.com - 지난 9일 청와대 인근 청운동 주민센터 앞에서 1박 2일 동안 박근혜 대통령과의 면담을 요구하고 있는 세월호 참사 희생 유가족들에게 사과하는 KBS 길환영 사장 사진=장성동 기자 - 작장 3년차 MB 아들 서초구 788평 땅 구입 차킨 400마리 미리 화겨놓았다 낭패 [관련슬라이드 더 보기] - ▲경상제공=양평경찰서 - (사진출처: bnt뉴스 DB, tvN '유승준의 롤까당' 방송캡처, 해물 두피클리닉스)	사진, 이미지 등의 단어에 빨간색으로 표시를 해 줌 좌측과 같이 사진 등의 캡션으로, 기사 본문이 아니라고 판단되는 경우 삭제	수작업
2			윌러스트] 강상민 기자	상동 --> 삭제	수작업
3			[그래픽] <그래픽>	상동 --> 삭제	수작업
4			▲ 표시된 부분	▲ 와 함께 나오는 내용을 보고 판단하여 사진 등의 캡션으로, 기사 본문이 아니라고 판단되는 경우 삭제	수작업
5			http://youtu.be/JRegVSmWooM (바리톤 토마스 햄슨, 빕스타인 지휘 빈 필하모닉)	기사 내에 삽입된 동영상의 캡션 --> 삭제	수작업
6			2013년 4월 11일 KBS 화면캡처	내용으로 보아 사진의 캡션 --> 삭제	수작업
7	링크 정보		· [이젠 남아공월드컵] 나이지리아는 감독 바꾸고 아르헨은 독일과 막매치... · [남아공 월드컵 100일 앞으로] [한국이 속한 B조 최강팀 '아르헨티나' 현지 취재] [1] "16강?... 우리는 우승만 생각한다" · [남아공 월드컵 100일 앞으로] "한국은 쉬지 않는 팀... 쉽지 않은 경기 할 것" · [이젠 남아공월드컵] 허정무호 '수비 노이로제'... "지금 선수들로 조직력 높일 수밖에" · [이젠 남아공월드컵] 허당크, 코트디부아르 지휘? · [남아공월드컵 D-98] [한국이 속한 B조 최강팀 '아르헨티나' 현지 취재] [2] "한국의 태권 축구? 우린 손으로 골도 넣었는데..." · [남아공월드컵 D-98] 아르헨티나는 독일 잡고... 그리스, 세네갈에 잡히고 · [남아공월드컵 D-98] '골 넣는 수비수' 작태희의 부활 · [남아공월드컵 D-98] "문단속'이 잘 되니 집안이 편안해졌다 · [남아공에 뜨는 별] [2] 전문가들이 꼽은 미드필더	본문 기사 내용이 아닌 다른 시리즈 기사로의 링크 --> 삭제	수작업
8			[관련자료 링크: 사회진보연대 2월 25일 보고서]	다른 기사로의 링크 --> 삭제	수작업
			"개그는 개그일 뿐 오버하지 말자."		

<그림 11> 말뭉치 정제 작업 지침 구글 문서

<그림 11>과 같이 구글 문서를 통해 정제 대상을 파악한 후 정제 방법에 따라 작업하고, 정제 지침에 없으나 확인할 필요가 있는 경우에는 정제 의견에 내용을 작성하거나 단체 대화방을 통해 실시간으로 정제 지침을 공유하였다.

이와 같이 1 차 자동 정제 기사 5,029,926 건 및 1,656,947,078 어절에 대하여 2 차 수작업 정제를 한 결과 3,991,282 건 및 1,003,899,229 어절의 결과가 도출되었다. 2 차 정제 후 기사 수 및 어절 수 통계 그래프는 다음과 같다.



<그림 12> 2 차 정제 후 매체별 기사 수 및 어절 수

6. 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)							
정의 값	N: 신문 말뭉치	W: 전국 종합지 L: 지역 종합지 P: 전문지 I: 인터넷 기반 신문 Z: 기타	OR: 원문 자료		19: 2019년		00000001 ~ 99999999 (여덟 자리 일련번호)							
※ 예시: NWOR1900000001.sjml 신문 전국 종합지 매체의 기사 원시 말뭉치 1 번째 파일 XML format														

<표 6> 파일 명명 규칙

○ 말뭉치 파일의 인코딩은 UTF-8 을 기본으로 한다.

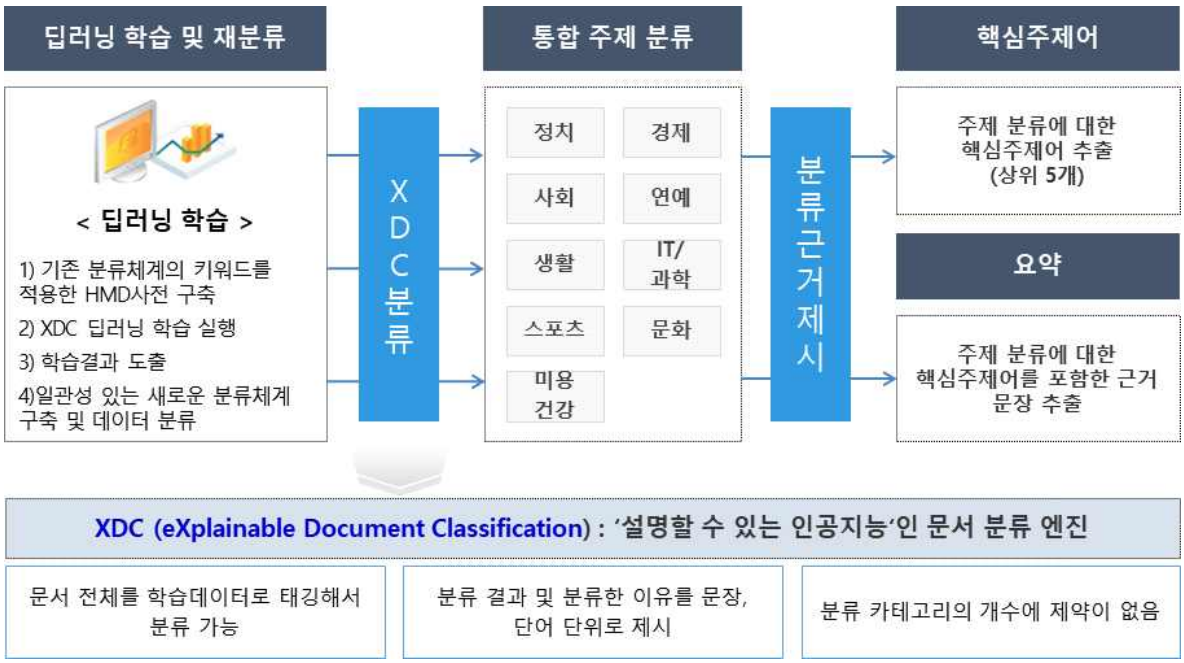
말뭉치 파일 내 포함해야 하는 메타 데이터에 대해서도 국립국어원과의 협의를 통해 다음과 같이 결정하였다.

○ 메타 데이터는 제목, 저자, 발행자, 연도, 기사번호, 분류, 기사 작성일, 기사 작성자, 어절 수로 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 42 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.

. 제목: 말뭉치의 제목 (ex. ‘한겨레 2004 년 기사’)
. 저자: 말뭉치의 저자 (ex. ‘한겨레’)
. 발행자: 기사의 발행자 (ex. ‘한겨레신문’)
. 연도: 기사의 발행연도 (ex. ‘2004’)
. 기사 번호: 6 자리수 일련번호 (ex. ‘000001’)
. 분류 1: 매체에서 분류한 정보(만약 없으면 공란으로 비움)
. 분류 2: 통합 분류 정보
. 기사 작성일: 기사 작성일자를 8 자리로 표시 (ex. ‘20040101’)
. 기사 작성자: 기사 작성자(기자)의 이름과 이메일 주소 등 정보 (ex. ‘○○○ 기자 abcde@abcilbo.co.kr’)
. 어절 수 : 해당 기사의 어절 수 (ex. ‘234’)

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 13> 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.

7. 최종 말뭉치 제작

국립국어원의 신문 말뭉치 구축 지침에 따라 최종 말뭉치 파일은 SJML 형식으로 제작하였다. SJML 의 기본 구조는 header 와 text 로 구성되어 있으며 그 상세 구조는 다음과 같다.

○ <header>: 파일의 메타 정보를 담는 요소

태그		설명 및 하위 태그	예시
<fileInfo>	<fileId>	파일의 고유 식별자 (말뭉치 파일명)	NWRW1900000012
	<annoLevel>	주석 수준	원시
	<sampling>	샘플링 방식 (본문 전체/부분 추출-임의 추출/부분 추출-특정 부분 추출)	부분 추출-임의추출
	<class>	구축 계획에 따른 장르 분류	전국 종합지
<sourceInfo>	<title>	제목(매체명 연도 표시)	○○일보 2015년 기사
	<author>	신문사	○○일보사
	<publisher>	매체명	○○일보
	<year>	작성 연도	2015

<표 7> SJML <header> 부분 구조

○ <text>: 본문

태그	설명	예시
<text id>	기사 구분 아이디 (<fileId>에 6자리 연번을 '-' 기호와 함께 부착)	NWRW1900000021-000001
<text date>	기사 작성일	20150103
<text topic_or>	매체별 자체 주제 분류	국제
<text topic>	9개 분류에 따른 조정된 분류	정치
<p>	문단 경계 정보	
<byline>	기사를 작성한 기자 정보 (기자 이름, 전자우편 주소, 소속, 지역 정보 등)	○○○기자 aaaa@bbbb.com
<keyword>	핵심 주제어	도시개발구역
<summary>	요약	은평뉴타운 조성사업이 도시개발구역 지정 승인을 받아 6월경 본격 착수된다.

<표 8> SJML <text> 부분 구조

SJML 형식의 전체 구성 및 말뭉치 예시는 다음과 같다.

SJML 의 형식은 XML 의 형식을 따르며, 문서의 첫 번째 행에 아래와 같이 문서 형식을 선언한다.

```
<?xml version="1.0" encoding="UTF-8"?>
```

들여쓰기(소프트탭: 스페이스 2 개)를 통해 요소의 계층을 시각화한다.

```
<SJML>
  <header>
    <fileInfo>
      <fileId>NWRW1900000012</fileId>
```

```
        ....
      </fileInfo>
    </header>
    <text>
     ....
    </text>
  </SJML>
```

마크업된 최종 신문 말뭉치 파일의 예시는 다음과 같다.

```
  <?xml version="1.0" encoding="UTF-8"?>
  <SJML>
    <header>
      <fileInfo>
        <fileId>NWRW1900000021</fileId>
        <annoLevel>원시</annoLevel>
        <sampling>부분 추출 - 임의 추출</sampling>
        <class>전국 종합지</class>
      </fileInfo>
      <sourceInfo>
        <title>동아일보 2004년 기사</title>
        <author>동아일보</author>
        <publisher>동아일보사</publisher>
        <year>2004</year>
      </sourceInfo>
    </header>
    <text id="NWRW1900000021-000001" date="20040103" topic_or="지역"
topic="경제">
      <p>[수도권]은평뉴타운 6월 공사 착수</p>
      <p>서울시가 은평구 진관내·외동과 구파발동 일대(108만여평)에 추진 중인 은평뉴타운 조성사업이 최근 건설교통부로부터 도시개발구역 지정 승인을 받아 본격 착수된다.</p>
      <p>2일 서울시에 따르면 최근 건교부는 은평뉴타운에 대한 도시개발구역 지정을 승인했으며 시는 1월 중순 도시계획위원회에서 그린벨트 해제 및 도시개발구역 지정을 거쳐 건축계획이 완료된 1구역(진관내동 일대)부터 6월경 공사에 착수할 방침이다.</p>
      <p>은평뉴타운은 도시개발 방식에 의한 공영개발로 사업이 진행되며 시행은 시 도시개발공사가 맡는다. 2008년 완공 목표이며 임대주택 4750가구와 일
```

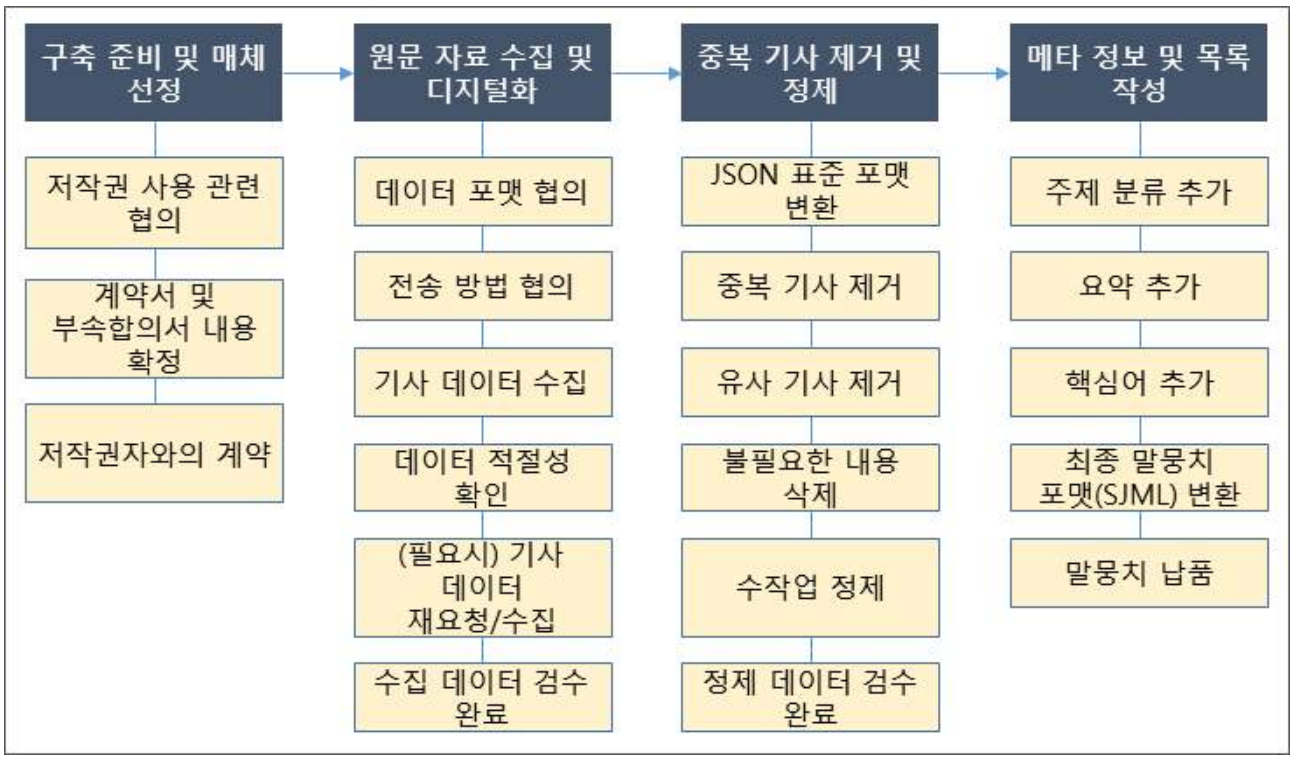
반분양 9250가구 등 총 1만4000가구가 공급될 예정이다. 한편 길음뉴타운은 최근 도시계획절차가 마무리돼 3월에 착공되며 왕십리뉴타운도 1월 중순 도시계획위원회에서 지구단위계획을 결정한 뒤 3월에 착공될 전망이다.</p>
<byline>채지영기자 yourcat@donga.com</byline>
<keyword>은평뉴타운; 도시개발구역</keyword>
<summary>은평뉴타운 조성사업이 도시개발구역 지정 승인을 받아 6월경 본격 착수된다.</summary>
</text>
<SJML>

제 3 장

사업 수행 결과

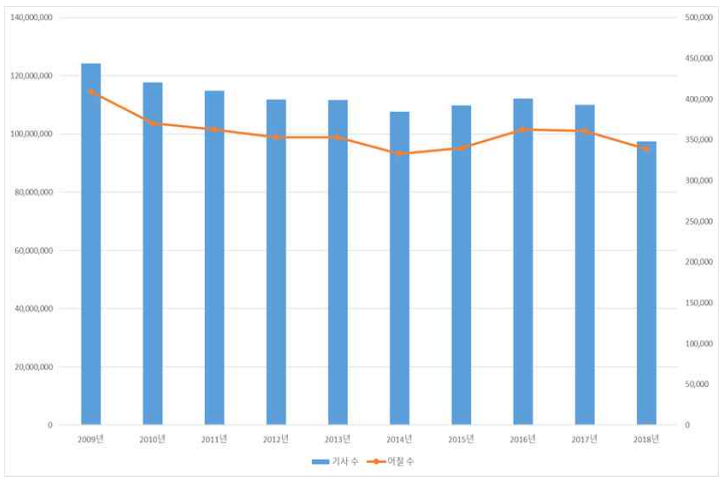
1. 신문 기사 정제 결과

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



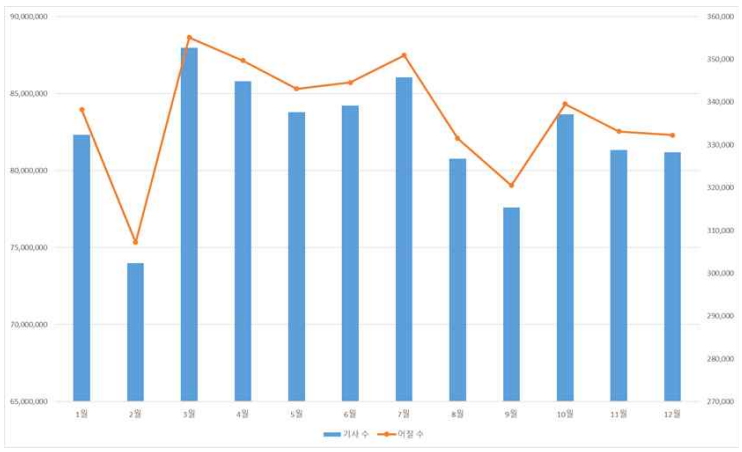
<그림 14> 사업 수행 전체 내용

이렇게 정제 완료된 전체 42 개 매체의 기사 수 및 어절 수에 대한 통계는 다음과 같다.



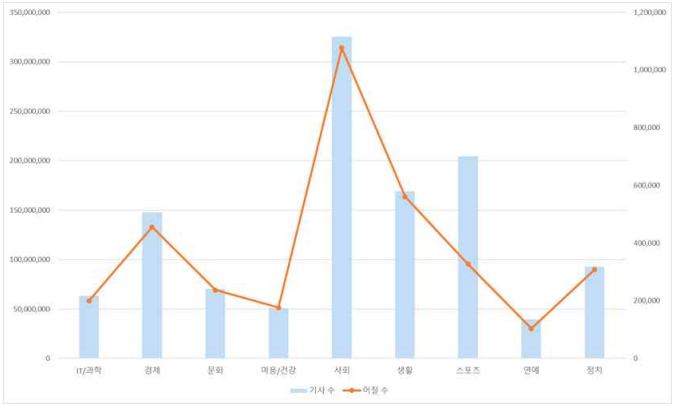
해당 연도	기사 수	어절 수
2009	443,726	114,599,998
2010	420,464	103,729,360
2011	410,351	101,604,087
2012	399,401	98,943,004
2013	399,123	98,914,491
2014	384,361	93,261,012
2015	392,206	95,188,518
2016	400,557	101,640,132
2017	392,791	101,045,690
2018	348,139	94,926,029
2019	163	46,908
총합	3,991,282	1,003,852,321

<그림 15> 연도별 기사 수 및 어절



해당 월	기사 수	어절 수
1	332,378	83,955,561
2	302,421	75,344,144
3	352,670	88,636,881
4	344,874	87,152,024
5	337,646	85,311,287
6	339,151	85,717,976
7	345,755	87,486,058
8	326,759	82,089,297
9	315,384	79,044,402
10	337,181	84,318,926
11	328,812	82,539,538
12	328,251	82,303,135
총합	3,991,282	1,003,899,229

<그림 16> 월별 기사 수 및 어절 수



유형	기사 수	어절 수
IT/과학	216,961	58,121,008
경제	507,232	132,720,774
문화	240,870	69,144,471
미용/건강	175,067	51,496,475
사회	1,115,034	313,955,822
생활	579,510	163,449,910
스포츠	701,750	95,265,121
연예	136,446	30,092,049
정치	318,412	89,652,990
총합	3,991,282	1,003,898,620

<그림 17> 통합 주제 분류별 기사 수 및 어절 수

9 개의 통합 주제 분류별 기사 수와 어절 수 통계를 보면 매우 많은 수의 사회 분야 기사가 쓰였음을 알 수 있고, 대부분의 분류에서 기사 수와 어절 수는 거의 비슷한 양상을 보이고 있는 반면 스포츠 분야는 기사 수는 많지만 상대적으로 어절 수가 적은 양상을 보였다. 이는 다양한 스포츠 행사에 대한 짧은 기사가 많이 게재되었다는 것을 알 수 있다.

2. 신문 말뭉치 납품 이후 계획

최종 산출물 납품 이후 국립국어원에서 검수 작업을 통해 받는 의견을 반영하는 것 외에 사업 수행사는 신문 말뭉치의 품질을 높이기 위해 자체적으로 무작위 샘플링을 통한 추가 정제 작업을 수행할 계획이다.

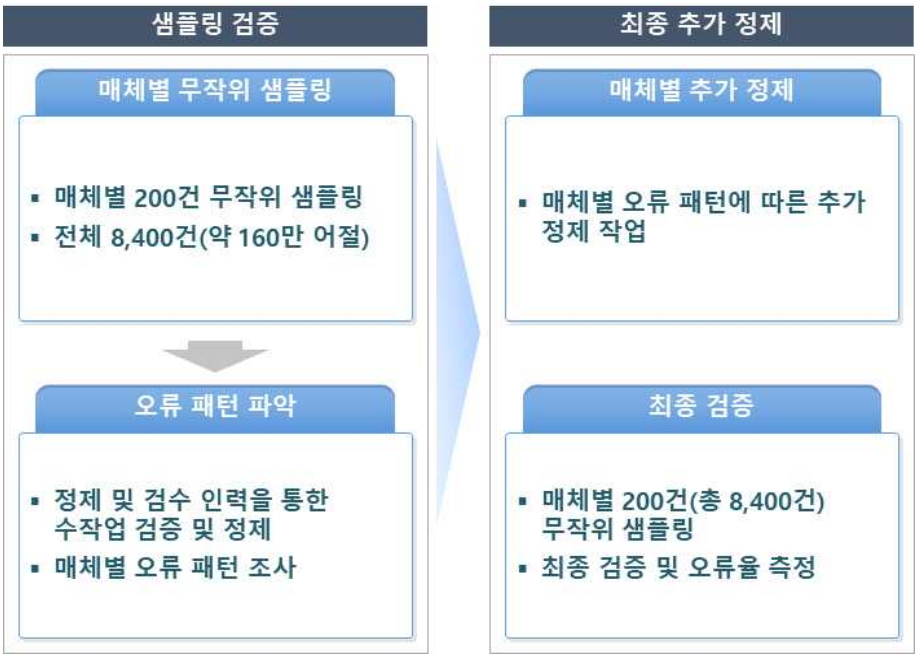
○ 각 매체별로 200 건의 기사를 무작위로 선별하여 추출한다. 42 개 매체에서 총 8,400 건을 추출하면 어절 수로는 약 160 만 어절에 해당한다.

○ 이 기사들에서 정제가 필요한 기사들을 선별한 후, 공통적인 문제를 가지고 있는 부분을 분석하여 매체별로 별도의 정제 작업을 진행한다.

○ 정제해야 할 부분이 형태상 유사성을 띠고 있다면 해당 기사는 자동 정제를 이용한다. 자동 정제 작업은 정규 표현식을 이용한 패턴 분석에 의한 작업으로 진행한다.

○ 정제해야 할 부분이 유사성이 거의 없거나 자동으로 정제하기 어려운 기사는 수작업으로 작업을 진행한다.

○ 추가 작업으로 정제가 된 최종 말뭉치는 국립국어원에서 검수 시에 요청한 사항과 함께 작업이 되어 전달한다.



<그림 21> 신문 말뭉치 추가 정제 계획

3. 향후 발전 방향

본 사업에서 최근 10년간의 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 세 가지로 정리할 수 있다.

첫 번째는 최근 10년간의 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 2030년까지 이용 허락 계약을 맺음으로써 자유롭게 연구 및 개발 활동을 보장할 수 있게 되었다.

두 번째는 현시대 언어생활을 반영하는 신문 기사 기반의 원시 말뭉치를 구축한 것이다. 다양한 매체 소속의 기자들이 작성한 10년 동안의 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.

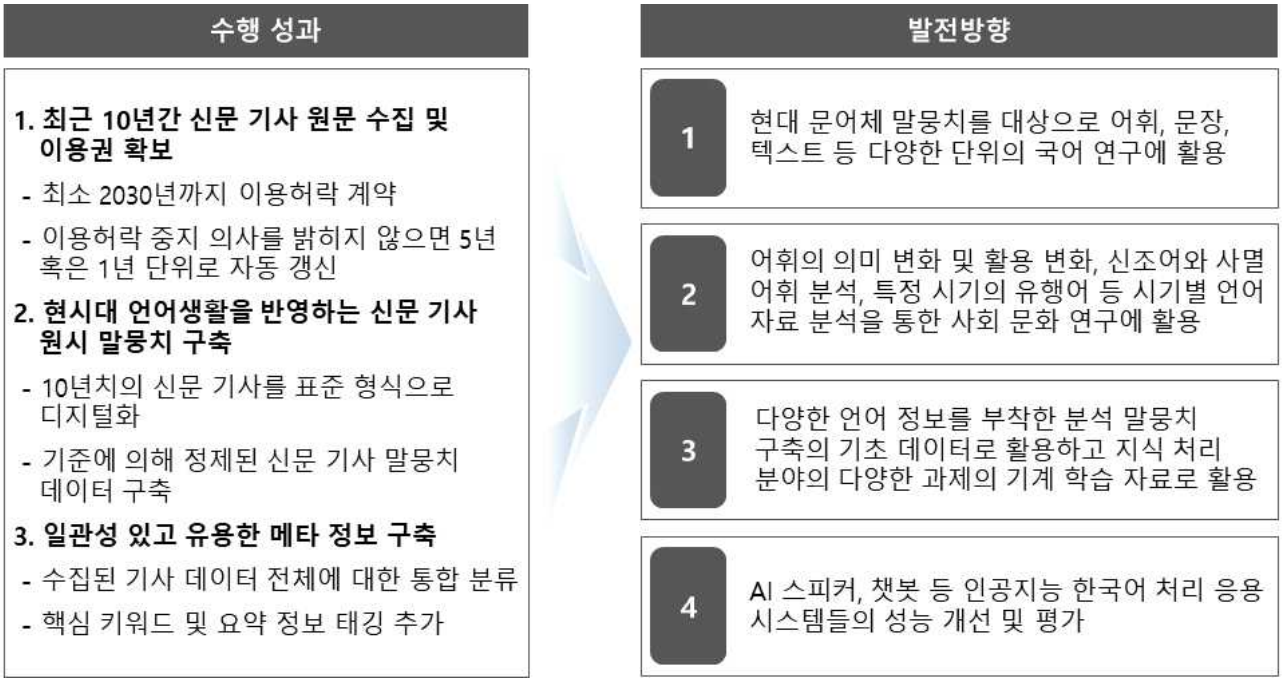
본 사업의 산출물인 말뭉치는 다음과 같은 방향으로 활용할 수 있을 것으로 기대한다.

○ 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 22> 수행 성과 및 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:

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 ‘복제·변형물’을 국어 연구와 언어 정보 처리 분야 응용을 위하여 학계·연구기관·산업체 등이 이용할 수 있도록 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 범위 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 변형할 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 다음 날부터 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 정하여 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	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
동아 일보 유사도 95%	11일 막 내리는 특별전… 안내설명인 3인의 소회\n“한동안 우리의 눈을 즐겁게 하던 페르시아 시대의 화려한 유물들이 며칠 뒤면 전시장을 떠난다고 생각하니 섭섭해요. 더 많은 분이 감상하도록 전시기간이 연장되면 좋을 텐데, 아쉽습니다….”\n 6일 오후 국립대구박물관 ‘황금의 제국 페르시아’ 특별전의 전시장 옆 휴게실.\n 관람객들에게 페르시아 유물의 특성과 이에 얽힌 이야기 등을 열정적으로 설명해 온 김미연(28·여) 박준우(24) 장수민(23·여) 씨 등 도슨트(안내설명인) 3명은 그동안의 소감과 에피소드 등을 밝혔다.\n 지난해 10월 7일 이곳에서 개막한 페르시아전이 11일로 막을 내린다.\n 현재 전시장에는 겨울방학을 맞아 지역 초중고교생 관람객이 몰려 막판 열기를 이어가고 있다. \n 개막일부터 도슨트로 활동해 온 이들은 “관람객들에게 페르시아 문화를 소개하고 한국과 이란의 가교 역할을 한 데 대해 자부심과 보람을 느낀다”고 입을 모았다.\n 이들은 “며칠 뒤면 전시회가 끝나는 만큼 아직 구경하지 않은 분들에게 서둘러 전시장을 찾아줄 것을 권유하고 싶다”고 말했다.\n 이들 중 연장자인 김 씨는 경북대 대학원 석사 과정에 재학 중이고 박 씨와 장 씨 등 2명은 대학생이다. 이들 모두 도슨트 활동은 이번이 처음이지만 매끄러운 진행으로 관람객들의 인기를 모았다. \n 사학을 전공한 김 씨는 “도슨트 근무가 전시회 개막 3일 전에 확정돼 전시 유물을 소개하는 엄청난 분량의 안내문을 외우느라 밤을 꼬박 새우기도 했다”고 말했다.\n 그는 “페르시아 유물전이 열린다는 소식을 접하고 ‘온몸으로 전시회를 느껴보자’는 취지로 도슨트에 지원했다”면서	11일 막 내리는 특별전… 안내설명인 3인의 소회\n“한동안 우리의 눈을 즐겁게 하던 페르시아 시대의 화려한 유물들이 며칠 뒤면 전시장을 떠난다고 생각하니 섭섭해요. 더 많은 분이 감상하도록 전시기간이 연장되면 좋을 텐데, 아쉽습니다….”\n 6일 오후 국립대구박물관 ‘황금의 제국 페르시아’ 특별전의 전시장 옆 휴게실.\n 관람객들에게 페르시아 유물의 특성과 이에 얽힌 이야기 등을 열정적으로 설명해 온 김미연(28·여) 박준우(24) 장수민(23·여) 씨 등 도슨트(안내설명인) 3명은 그동안의 소감과 에피소드 등을 밝혔다.\n 지난해 10월 7일 이곳에서 개막한 페르시아전이 11일로 막을 내린다.\n 현재 전시장에는 겨울방학을 맞아 지역 초중고교생 관람객이 몰려 막판 열기를 이어가고 있다. \n 개막일부터 도슨트로 활동해 온 이들은 “관람객들에게 페르시아 문화를 소개하고 한국과 이란의 가교 역할을 한 데 대해 자부심과 보람을 느낀다”고 입을 모았다.\n 이들은 “며칠 뒤면 전시회가 끝나는 만큼 아직 구경하지 않은 분들에게 서둘러 전시장을 찾아줄 것을 권유하고 싶다”고 말했다.\n 이들 중 연장자인 김 씨는 경북대 대학원 석사 과정에 재학 중이고 박 씨와 장 씨 등 2명은 대학생이다. 이들 모두 도슨트 활동은 이번이 처음이지만 매끄러운 진행으로 관람객들의 인기를 모았다. \n 사학을 전공한 김 씨는 “도슨트 근무가 전시회 개막 3일 전에 확정돼 전시 유물을 소개하는 엄청난 분량의 안내문을 외우느라 밤을 꼬박 새우기도 했다”고 말했다.\n 그는 “페르시아 유물전이 열린다는 소식을 접하고 ‘온몸으로 전시회를 느껴보자’는 취지로 도슨트에 지원했다”면서

	“보람을 느낀다”고 덧붙였다.\n 이어 그는 “이번 도슨트 근무경력이 인정돼 최근 경북대박물관 직원으로 채용됐다”며 “페르시아전이 나에게는 새로운 인생을 열어준 행운의 상징으로 오랫동안 기억될 것 같다”며 기뻐했다.\n 박 씨는 “도슨트 근무를 하면서 사람들을 많이 만난 게 무엇보다 좋았고, 수십 명의 관람객 앞에서 설명을 하다 보니 발표력도 아주 나아진 것 같다”고 밝혔다.\n 그는 특히 “도슨트 활동이 대학에서 전공하고 있는 경영학 공부에도 상당한 도움이 될 것 같다“며 ”한 달 전 관람 도중 끊임없이 질문을 하시던 70대 할아버지와 토론을 하는데 ‘목이나 축이고 하라’며 정겹게 음료수를 건네주시던 50대 아주머니의 표정이 지금도 눈에 선하다”고 말했다.\n 그는 “페르시아 유물에 깃든 높은 예술성과 실용성에 감탄했다”며 “전시 유물 중 ‘황금가면’은 다시 돌려보내기 아쉬울 정도로 마음에 든다”고 말했다.\n 또 장 씨는 “직원들이 전시관 내 유리를 너무 깨끗하게 닦아 관람객들이 머리를 부딪치는 일이 자주 있었다”며 “도슨트 활동 기간에 책임감을 갖고 완벽한 준비를 위해 노력하다 보니 인간적으로 성장하는 데도 도움이 됐다”고 밝혔다.\n 이들은 “전시회가 끝나면 가까운 곳에 함께 여행이라도 다녀오고 싶다”며 “황금만큼이나 값지고 소중한 체험을 한 게 오랫동안 추억으로 남을 것 같다”며 환하게 웃었다.\n정용균 기자 cavatina@donga.com\n▽기간=11일까지\n▽장소=국립대구박물관 \n▽관람료=성인 1만 원, 중고교생 9000원, 초등학생 8000원, 48개월～미취학 아동 5000원\n▽문의=1688-0577, 페르시아전 홈페이지(www.persia2008.com)	“보람을 느낀다”고 덧붙였다.\n 이어 그는 “이번 도슨트 근무경력이 인정돼 최근 경북대박물관 직원으로 채용됐다”며 “페르시아전이 나에게는 새로운 인생을 열어준 행운의 상징으로 오랫동안 기억될 것 같다”며 기뻐했다.\n 박 씨는 “도슨트 근무를 하면서 사람들을 많이 만난 게 무엇보다 좋았고, 수십 명의 관람객 앞에서 설명을 하다 보니 발표력도 아주 나아진 것 같다”고 밝혔다.\n 그는 특히 “도슨트 활동이 대학에서 전공하고 있는 경영학 공부에도 상당한 도움이 될 것 같다“며 ”한 달 전 관람 도중 끊임없이 질문을 하시던 70대 할아버지와 토론을 하는데 ‘목이나 축이고 하라’며 정겹게 음료수를 건네주시던 50대 아주머니의 표정이 지금도 눈에 선하다”고 말했다.\n 그는 “페르시아 유물에 깃든 높은 예술성과 실용성에 감탄했다”며 “전시 유물 중 ‘황금가면’은 다시 돌려보내기 아쉬울 정도로 마음에 든다”고 말했다.\n 또 장 씨는 “직원들이 전시관 내 유리를 너무 깨끗하게 닦아 관람객들이 머리를 부딪치는 일이 자주 있었다”며 “도슨트 활동 기간에 책임감을 갖고 완벽한 준비를 위해 노력하다 보니 인간적으로 성장하는 데도 도움이 됐다”고 밝혔다.\n정용균 기자 cavatina@donga.com\n▽기간=11일까지\n▽장소=국립대구박물관 \n▽관람료=성인 1만 원, 중고교생 9000원, 초등학생 8000원, 48개월～미취학 아동 5000원\n▽문의=1688-0577, 페르시아전 홈페이지(www.persia2008.com)
동아일보 유사도	평일인데도 사람이 굉장히 많네. 사람들이 이렇게 많은걸 보면 확실히 투자가치가 있기는 있나봐.\" \n 12일 오후 2시 인천	모델하우스 장사진… 10대 1 경쟁률도\n타지역은 한겨울… ‘묻지마 청약’ 금물\n“평일인데도 사람이 굉장히 많네.

86%	연수구 송도동에 있는 포스코건설의 '더 샵 하버뷰II' 모델하우스 앞. 이곳을 찾은 40대 주부는 평일에 모델하우스가 사람들로 붐비는 것을 보면서 놀랍다는 표정을 지었다.\n 지난달 말 시작된 '청라발(發) 훈풍'이 인천지역 분양시장을 계속 들뜨게 하고 있다. 청라지구는 일각에서 아파트 분양시장이 글로벌 금융위기 이전으로 복귀했다는 진단까지 하게 만든 진원지로 꼽힌다. 그러나 분양 훈풍은 아직 국지적 현상에 불과할 뿐이라고 전문가들은 지적하고 있다.\n ●시세차익 기대감이 훈풍 원인\n 2012년 5월 송도국제도시 D15블록에 들어설 예정인 이 아파트의 모델하우스는 분양 상담 코너와 평형별 구조를 둘러보는 곳에서 줄을 서야 했다. 또 모델하우스 주변에는 몰래 분양권 거래를 하려는 '떴다방'까지 등장했다. 8일 개관한 뒤 주말에는 평균 1만여 명, 평일에는 평균 4000여 명이 다녀갔다.\n 이규종 포스코건설 건축사업본부 차장은 \"처음 예상했던 방문자 숫자보다 훨씬 많은 고객들이 모델하우스를 찾아오는 바람에 기념품, 홍보책자, 안내 인력을 크게 늘렸다\"고 말했다.\n 11일 마감한 더 샵 하버뷰II의 특별 공급분은 10대 1의 높은 청약경쟁률을 보였다. 이 아파트보다 앞서 청약을 마감한 '청라 한화 꿈에그린'과 '청라 호반 베르디움'도 각각 1순위에서 청약을 마감하며 7.37대 1과 2.48대 1의 평균 경쟁률을 보였다.\n 이처럼 청라지구와 송도국제도시에서 분양되는 아파트들이 인기를 끄는 이유는 시세차익 효과를 확실히 볼 수 있을 것 같다는 기대감 때문이다. 부동산 정보업체들에 따르면 청라지구와 송도국제도시의 주변 시세는 3.3㎡당 1200만 원 대. 하지만 최근 이 곳에 분양된 아파트들의 3.3㎡당 분양가는 평균	사람들이 이렇게 많은 걸 보면 확실히 투자가치가 있기는 있나봐.” \n 12일 오후 2시 인천 연수구 송도동에 있는 포스코건설의 ‘더 샵 하버뷰II’ 모델하우스 앞. 이곳을 찾은 40대 주부는 평일에 모델하우스가 사람들로 붐비는 것을 보면서 놀랍다는 표정을 지었다.\n 지난달 말 시작된 ‘청라발(發) 훈풍’이 인천지역 분양시장을 계속 들뜨게 하고 있다. 청라지구는 일각에서 아파트 분양시장이 글로벌 금융위기 이전으로 복귀했다는 진단까지 하게 만든 진원지로 꼽힌다. 그러나 분양 훈풍은 아직 국지적 현상에 불과할 뿐이라고 전문가들은 지적하고 있다.\n○ 시세차익 기대감이 훈풍 원인\n 2012년 5월 송도국제도시 D15블록에 들어설 예정인 이 아파트의 모델하우스는 분양 상담 코너와 평형별 구조를 둘러보는 곳에서 줄을 서야 했다. 또 모델하우스 주변에는 몰래 분양권 거래를 하려는 ‘떴다방’까지 등장했다. 8일 개관한 뒤 주말에는 평균 1만여 명, 평일에는 평균 4000여 명이 다녀갔다.\n 이규종 포스코건설 건축사업본부 차장은 “처음 예상했던 방문자 수보다 훨씬 많은 고객이 모델하우스를 찾는 바람에 기념품, 홍보책자, 안내 인력을 크게 늘렸다”고 말했다.\n 11일 마감한 더 샵 하버뷰II의 특별 공급분은 10 대 1의 높은 청약경쟁률을 보였다. 이 아파트보다 앞서 청약을 마감한 ‘청라 한화 꿈에그린’과 ‘청라 호반 베르디움’도 각각 1순위에서 마감이 되면서 7.37 대 1과 2.48 대 1의 평균 경쟁률을 보였다.\n 이처럼 청라지구와 송도국제도시에서 분양되는 아파트가 인기를 끄는 이유는 시세차익을 확실히 볼 수 있을 것 같다는 기대감 때문이다. 부동산 정보업체들에 따르면 청라지구와 송도국제도시의 주변 시세는 3.3m²당

	1000만~1100만 원으로 시세보다 100만~200만 원 정도 싸다. 또 양도소득세가 5년간 100% 면제되고 전매제한 기간이 85㎡이상 크기면 1년, 85㎡ 미만이면 3년으로 단축돼 있다는 것도 장점으로 꼽힌다.\n 김규정 부동산114 부장은 \"인천지역 분양시장에 사람들이 몰리는 데는 단기간에도 시세차익을 확실히 볼 수 있고 중장기적으로는 국제화 관련 인프라가 대거 들어오는 등 개발호재로 계속 가격이 오를 것이란 기대감이 크게 작용한 때문\"이라고 말했다.\n ●다른 분양시장엔 아직 찬바람만\n 인천지역의 분양 아파트들이 인기를 끌자 일부에서는 지난해 하반기부터 얼어붙었던 아파트 분양시장이 본격적으로 살아나는 것 아니냐는 주장도 나오고 있다. 그러나 부동산 전문가들은 인천지역은 다른 곳과 확실히 차별화되는 메리트가 있기 때문에 관심을 받는 것일 뿐 분양시장 전체와는 거리가 있다고 강조한다.\n 부동산114에 따르면 지난달 서울 성동구 성수동에서 청약접수를 받은 '대명 루첸'은 '한강변 재개발'이라는 호재에도 불구하고 총 87채 분양에 청약자가 9명에 불과했다. 역시 지난달 청약접수를 받은 경기 파주 교하신도시의 '한양 수자인'도 총 8개 주택형 중 1개 주택형은 3순위 청약에서도 결국 미달됐다. 롯데건설이 3월 대구 서구에 분양한 '평리 롯데캐슬'도 청약률이 20% 정도 밖에 안됐다.\n 부동산 업계에서는 대명 루첸이 고급 아파트를 추구하면서 상대적으로 분양가를 높인 것이 결국 청약자가 미달된 결정적인 요인이 됐다고 보고 있다. 한양 수자인도 분양가가 주변시세보다 싸지 않아 청약자 모집에 어려움을 겪었다. 심지어 청라지구에서도 입지 조건이 상대적으로 떨어지는 '한일베라체'는 1순위에 마감이 안	1200만 원대. 하지만 최근 이곳에 분양된 아파트들의 3.3m²당 분양가는 평균 1000만~1100만 원으로 시세보다 100만~200만 원 싸다. 또 양도소득세가 5년간 100% 면제되고 전매제한 기간이 85m² 이상 크기면 1년, 85m² 미만이면 3년으로 단축돼 있다는 것도 장점으로 꼽힌다.\n 김규정 부동산114 부장은 “인천지역 분양시장에 사람들이 몰리는 데는 단기간에도 시세차익을 확실히 볼 수 있고 중장기적으로는 국제화 관련 인프라가 대거 들어오는 등 개발호재로 계속 가격이 오를 것이란 기대감이 크게 작용한 때문”이라고 말했다.\n○ 다른 분양시장엔 아직 찬바람만\n 인천지역의 분양 아파트들이 인기를 끌자 일부에서는 지난해 하반기부터 얼어붙었던 아파트 분양시장이 본격적으로 살아나는 것 아니냐는 주장도 나오고 있다. 그러나 부동산 전문가들은 인천지역은 다른 곳과 확실히 차별화되는 메리트가 있기 때문에 관심을 받는 것일 뿐 분양시장 전체와는 거리가 있다고 강조한다.\n 부동산114에 따르면 지난달 서울 성동구 성수동에서 청약접수를 한 ‘대명 루첸’은 ‘한강변 재개발’이라는 호재에도 불구하고 총 87채 분양에 청약자가 9명에 불과했다. 역시 지난달 청약접수를 받은 경기 파주 교하신도시의 ‘한양 수자인’도 총 8개 주택형 중 1개 주택형은 3순위 청약에서도 결국 미달됐다. 롯데건설이 3월 대구 서구에 분양한 ‘평리 롯데캐슬’도 청약률이 20% 정도밖에 안됐다.\n 부동산 업계에서는 대명 루첸이 고급 아파트를 추구하면서 상대적으로 분양가를 높인 것이 결국 청약자가 미달된 결정적인 요인이 됐다고 보고 있다. 한양 수자인도 분양가가 주변시세보다 싸지 않아 청약자 모집에 어려움을 겪었다. 심지어 청라지구에서도

	되고 3순위까지 신청을 받아 간신히 1대 1을 넘겼다.\n 이영진 닥터아파트 이사는 \"그동안 워낙 분양이 없었기 때문에 최근 관심이 과열되는 측면이 분명히 있다\"며 \"분양가, 브랜드, 입지조건 등을 따지지 않는 '묻지마 청약'을 해서는 안 된다\"고 말했다.\n인천=이세형기자 turtle@donga.com	입지 조건이 상대적으로 떨어지는 ‘한일베라체’는 1순위에서 마감이 안 되고 3순위까지 신청을 받아 간신히 1 대 1을 넘겼다.\n 이영진 닥터아파트 이사는 “그동안 워낙 분양이 없었기 때문에 최근 관심이 과열되는 측면이 분명히 있다”며 “분양가, 브랜드, 입지조건 등을 따지지 않는 ‘묻지 마 청약’을 해서는 안 된다”고 말했다.\n인천=이세형 기자 turtle@donga.com
동아일보 유사도 73%	지난달 31일 제109회 US아마추어골프챔피언십에서 역대 최연소로 우승한 안병훈(18)은 \"아빠와 엄마에게 잘 배운 덕분\"이라는 소감을 밝혔다. 그의 부모님은 널리 알려진 대로 한중 핑퐁 커플 안재형(44) 자오즈민(46)이다. 언뜻 보면 골프와 탁구는 공 크기가 비슷하다는 점을 빼면 별 공통점이 없어 보인다. 하지만 탁구인들은 \"두 종목은 닮은 구석이 많아 골프 시작할 때 도움이 된다\"고 입을 모은다. 안재형은 \"채를 잡고 작은 공을 다뤄야 하므로 둘 다 고도의 집중력이 필요하다. 탁구는 상대 심리 상태를 잘 파악해 공략해야 하는 데 병훈이가 우승한 이번 대회처럼 특히 골프 매치플레이에서는 심리전이 중요하다\"고 밝혔다.\n 탁구 대표팀에서 안재형을 지도했던 삼성생명 강문수 감독은 \"탁구와 골프 모두 연습량이 생명이다. 안재형은 선수 때 누구보다 땀을 많이 흘렸다\"고 칭찬했다. 핸디캡 6인 김택수 대우증권 감독은 \"하체를 고정하고 다리-복근-어깨로 연결되는 탁구 스윙이 골프와 흡사한 하다\"고 말했다.\n 이처럼 스포츠 세계에는 골프와 쉽게 친숙해지는 종목이나 포지션이 눈에 띈다. 섬세한 손 감각과 바람이 불면 일부러 오조준을 해야 하는 양궁을 한 궁사 출신들이 골프채를 잡으면 어프로치와 퍼팅에 강한 면모를 보인다. \n 야구에는 투수 중에 골프 고수가	골프와 다른 스포츠 상관관계 살펴보니\n지난달 31일 제109회 US아마추어골프챔피언십에서 사상 최연소로 우승한 안병훈(18)은 “아빠와 엄마에게 잘 배운 덕분”이라고 소감을 밝혔다. 그의 부모는 널리 알려진 대로 한중 탁구 커플 안재형(44), 자오즈민 씨(46)다.\n 골프와 탁구는 공 크기가 비슷한 것을 빼면 별 공통점이 없어 보인다. 공 무게는 탁구가 2.7g, 골프가 45g으로 크게 차이가 난다. 하지만 탁구인들은 “두 종목은 닮은 구석이 많아 골프를 시작할 때 도움이 된다”고 입을 모은다. 안 씨는 “작은 공을 다뤄야 하므로 둘 다 고도의 집중력이 필요하다. 탁구는 상대 심리 상태를 잘 파악해 공략해야 하는 것처럼 병훈이가 우승한 이번 골프 매치플레이 같은 골프 대회에서도 심리전이 중요하다”고 말했다.\n탁구 대표팀에서 안 씨를 지도했던 삼성생명 강문수 감독은 “탁구와 골프 모두 연습량이 생명이다. 안재형은 선수 때 누구보다 땀을 많이 흘렸다”고 칭찬했다. 핸디캡 6인 김택수 대우증권 감독은 “하체를 고정하고 다리-복근-어깨로 연결되는 탁구 스윙은 골프와 흡사하다”고 말했다.\n 이처럼 스포츠 세계에는 골프 친화적인 종목과 포지션이 눈에 띈다. 섬세한 손 감각과 바람이 불면 그에 맞춰 오조준을 해야 하는 양궁 선수 출신들은 어프로치에 강한 면모를

	많다. 선동렬 삼성 감독과 한화 송진우, 유백만 전 삼성 코치 등은 야구인 골프 모임에서 심심치 않게 우승을 한다. 베스트스코어가 66타인 선동렬 감독은 \"투수들은 대개 고교 때까지만 타석에 들어선다. 방망이로 공을 잘 다룰 줄 아는 데다 야구와 골프 스윙의 차이도 쉽게 파악할 수 있다. 마운드에서의 집중력은 골프를 칠때도 많이 도움이 된다\"고 말했다.\n 2007년 미국의 골프다이제스트가 보도한 '스포츠 스타의 골프 핸디캡'이란 기사에 따르면 메이저리그 피츠버그 파이리츠의 투수였던 릭 로든이 '+2.5(파 72인 코스에서 평균 69.5타를 친다는 의미)'로 1위에 올랐다. 20위 이내의 야구 선수 6명 중에는 존 스몰츠(+0.2·16위) 등 투수가 3명이었다. +0.7로 14위인 거포 마크 맥과이어도 대학 입학 당시에는 투수였다.\n 농구 선수중에는 대개 슈터나 가드들의 골프 실력이 뛰어난 데 섬세한 쇼트게임으로 스코어를 줄인 덕분이다. 명슈터였던 김영기 전 한국농구연맹 총재, 이충희 고려대 감독, 전창진 KT 감독과 임달식 신한은행 감독 등이 스코어카드에 '7'자를 자주 그린다. 골프 행사에 자주 등장하는 '농구황제'마이클 조든의 핸디캡은 1.2이다. 하지만 키가 2m 안팎인 센터들은 클럽을 따로 맞춰야 하는 핸디캡까지 있어 골프 입문을 망설이기도 한다.\n '스포츠 스타들은 골프를 해도 끝을 본다'는 얘기가 있다. 자존심이 걸려 있고 주위의 이목도 집중되므로 그만큼 노력과 투자를 아끼지 않기 때문이다.\n김종석 기자 kjs0123@donga.com	보인다. 장영술 현대제철 감독은 “그린이 과녁이라면, 홀은 엑스텐(10점 만점 중에서도 지름 6.1cm의 정중앙)이다. 아이언 샷과 화살은 포물선을 그리며 날아가는 궤적이 똑같다”고 말했다.\n 야구에는 투수 중에 골프 고수가 많다. 선동렬 삼성 감독과 한화 송진우, 유백만 전 삼성 코치 등은 야구인 골프 모임에서 심심치 않게 우승을 한다. 베스트스코어 66타에 파5 홀을 2타 만에 홀아웃해 앨버트로스까지 작성한 선 감독은 “투수들은 대개 고교 때까지만 타석에 선다. 방망이로 공을 잘 다룰 줄 아는 데다 야구와 골프 스윙의 차이도 쉽게 파악할 수 있다. 마운드에서의 집중력은 골프를 할 때 많은 도움이 된다”고 말했다.\n 2007년 미국의 골프다이제스트가 보도한 ‘스포츠 스타의 골프 핸디캡’이란 기사에 따르면 메이저리그 피츠버그 파이리츠의 투수였던 릭 로든이 ‘+2.5(파72인 코스에서 평균 69.5타를 친다는 의미)’로 1위에 올랐다. 20위 이내에 든 야구 선수 중에 존 스몰츠(+0.2·16위) 등 투수가 3명이었다. +0.7로 14위인 거포 마크 맥과이어도 대학 입학 당시에는 투수였다.\n 농구 선수 중에는 대개 슈터나 가드들의 골프 실력이 뛰어난 데 섬세한 쇼트게임으로 스코어를 줄인 덕분이다. 명슈터였던 김영기 전 한국농구연맹 총재, 이충희 고려대 감독, 전창진 KT 감독과 임달식 신한은행 감독 등이 스코어카드에 ‘7’자를 자주 그린다. 골프 행사에 자주 등장하는 ‘농구 황제’ 마이클 조든의 핸디캡은 1.2이다. \n ‘스포츠 스타들은 골프를 해도 끝을 본다’는 얘기가 있다. 주위의 이목이 집중되는 만큼 노력과 투자를 아끼지 않는다. 게다가 일반인에 비해 유리한 신체조건과 운동신경을 갖췄고 종목별로 비슷한 특성까지 겸비해 빨리 고수의 반열에 오르는 것으로 풀이된다.\n김종석 기자

		kjs0123@donga.com
동아일보 유사도 60%	불법으로 인터넷에 유포된 영화 '해운대'의 동영상 파일은 시사회전인 7월 초 제작된 편집본으로 밝혀졌다. 경찰청 사이버테러대응센터는 2일 \"유출된 동영상은 상영관용으로 만들어진 것은 아니고 시사회를 열흘 가량 앞둔 7월 초 만들어진 편집본\"이라며 \"유출된 동영상은 완성본과 비교할 때 일부 내용 및 자막이 다르다\"고 밝혔다. 유출된 파일은 VOB(DVD Video OBject) 형식으로, 이 파일에는 DVD 타이틀에 기록된 영화의 실제 동영상 데이터가 저장된다.\n 이에 따라 경찰은 1일 영화의 컴퓨터 그래픽과 음향 효과를 맡았던 업체 직원을 조사한 데 이어 이날 동영상 유출 전 편집본 관리와 관련된 이들을 불러 조사했다. 경찰은 또 유출 진원지를 확인하기 위해 파일공유 사이트 24곳을 조사해 접속 기록 등을 확보한다는 계획이다.\n 한편 대검찰청 형사부(부장 소병철)는 동영상 불법 유출자를 신속하게 검거해 엄중하게 처벌할 방침이라고 밝혔다. 검찰 관계자는 \"최근 해외업체가 제작한 음란영상물을 유통시킨 이들에 대해 영업성이 크고 범행 횟수나 동종 전과가 많은 경우 구속영장을 적극적으로 청구하는 등 저작권 침해사범에 대한 처벌을 강화하고 있다\"며 \"'해운대' 유출자에 대한 처벌도 같은 맥락\"이라고 말했다.\n 검찰은 문화관광부와 협조해 동영상의 불법유통을 적극 차단하는 한편 앞으로도 문광부, 경찰청 등 관련기관과 긴밀한 협조체계를 구축해 저작권 위반 사범에 대해 꾸준하게 단속과 처벌을 해나가기로 했다.\n유덕영 기자firedy@donga.com|\n전성철기자 dawn@donga.com	불법으로 인터넷에 유포된 영화 ‘해운대’의 동영상 파일은 시사회 전인 7월 초 제작된 편집본으로 밝혀졌다. 경찰청 사이버테러대응센터는 2일 “유출된 동영상은 상영관용으로 만들어진 것은 아니고 시사회를 열흘가량 앞둔 7월 초 만들어진 편집본”이라며 “유출된 동영상은 완성본과 비교할 때 일부 내용 및 자막이 다르다”고 밝혔다. 유출된 파일은 VOB(DVD Video OBject) 형식으로, 이 파일에는 DVD 타이틀에 기록된 영화의 실제 동영상 데이터가 저장된다.\n 이에 따라 경찰은 1일 영화의 컴퓨터 그래픽과 음향 효과를 맡았던 업체 직원을 조사한 데 이어 이날 동영상 유출 전 편집본 관리와 관련된 이들을 불러 조사했다. 경찰은 또 유출 진원지를 확인하기 위해 파일공유 사이트 24곳을 조사해 접속 기록 등을 확보한다는 계획이다.\n 한편 대검찰청 형사부(부장 소병철)는 동영상 파일 불법 유출자를 신속하게 검거해 엄중하게 처벌할 방침이라고 밝혔다.\n유덕영 기자 firedy@donga.com\n전성철 기자 dawn@donga.com
오마이뉴스 유사도	KBS·MBC·SBS 등 지상파방송 3사가 티브로드·CJ헬로비전·씨앤앰·HCN·CMB 등 5대 복수종합유선방송사업자(MSO) 쪽에	KBS·MBC·SBS 등 지상파방송 3사가 티브로드·CJ헬로비전·씨앤앰·HCN·CMB 등 5대 복수종합유선방송사업자(MSO) 쪽에

92%	지상파 재송신과 관련한 저작권 법적대응 공문을 지난 11일과 12일 각각 보낸 것으로 확인됐다. 이들 지상파 3사는 공문에서 저작권 침해 후속조치에 대한 회신기한을 오는 19일 오후 6시로 못박으며, '전향적이고 유의미한 회신'이 없을 경우 오는 22일 법적조치를 취하겠다고 밝혔다. 지상파 3사는 \"MSO 쪽에서 디지털케이블 상품을 통해 지상파 텔레비전방송 채널을 아무런 동의절차없이 재송신하고 있어 지상파방송사의 저작권을 침해하고 있다\"며 \"수차례 문제를 제기하고 이를 해결하기 위한 노력을 기울였음에도 합의에 이르지 못하고 있다\"고 밝혔다. 이들은 이어 \"이에 불가피하게 MSO 쪽을 저작권법 제16조(복제권), 18조(공중 송신권), 84조(복제권), 85조(동시중계 방송권) 위반 혐의로 의법조치하지 않을 수 없다\"며 \"법적 조치에 앞서 MSO 쪽이 적절한 조치를 취하기 바란다\"고 촉구했다. 지상파 쪽에 따르면, 지상파와 케이블 쪽은 각각 한국방송협회와 한국케이블TV방송협회를 내세워 지난 2008년 5월부터 같은 해 9월까지 관련협의를 진행했으나 결론을 내리지 못했다. 이어 지난 3월부터 이 달까지 각 지상파와 MSO간 협상을 벌였지만 이 역시 진전이 없었다. MSO 쪽은 그동안 난시청 해소라는 명분 등으로 지상파 재송신을 해왔으나, 지상파 쪽은 디지털 전환 이후까지 재산권 침해문제를 눈감아 줄 수 없다는 입장을 이번 공문에서 분명히 한 셈이다. 한편 지난 4월 KT·SK브로드밴드·LG데이콤 등 IPTV제공사업자들이 MSO 경우와 비교하며 지상파 채널 전체를 필수설비로 규정해달라거나 무료로 내보내게 해달라고 정부여당에 건의하기도 했으나, 지상파 쪽은 관련주장을 부당하다고 일축한 바 있다.	지상파 재송신과 관련한 저작권 법적대응 공문을 보낸 가운데 관련 논란이 IPTV 쪽에 영향을 미칠지 주목된다. 지상파 3사는 지난 11일과 12일 MSO 쪽에 보낸 공문에서 저작권 침해 후속조치에 대한 회신기한을 오는 19일 오후 6시로 못박으며, '전향적이고 유의미한 회신'이 없을 경우 오는 22일 법적조치를 취하겠다고 밝혔다. 지상파 3사는 MSO 쪽에서 디지털케이블 상품을 통해 지상파 텔레비전방송 채널을 아무런 동의절차없이 재송신하고 있어 지상파방송사의 저작권을 침해하고 있다며 수차례 문제를 제기하고 이를 해결하기 위한 노력을 기울였음에도 합의에 이르지 못하고 있다고 밝혔다. 이들은 이어 이에 불가피하게 MSO 쪽을 저작권법 제16조(복제권), 18조(공중 송신권), 84조(복제권), 85조(동시중계 방송권) 위반 혐의로 의법조치하지 않을 수 없다며 법적 조치에 앞서 MSO 쪽이 적절한 조치를 취하기 바란다고 촉구했다. 지상파 쪽에 따르면, 지상파와 케이블 쪽은 각각 한국방송협회와 한국케이블TV방송협회를 내세워 지난 2008년 5월부터 같은 해 9월까지 관련협의를 진행했으나 결론을 내리지 못했다. 이어 지난 3월부터 이 달까지 각 지상파와 MSO간 협상을 벌였지만 이 역시 진전이 없었다. MSO 쪽은 그동안 난시청 해소라는 명분 등으로 지상파 재송신을 해왔으나, 지상파 쪽은 디지털 전환 이후까지 재산권 침해문제를 눈감아 줄 수 없다는 입장을 이번 공문에서 분명히 한 셈이다. 남은건 IPTV제공사업자들이다. 이들은 지난 4월 MSO 경우와 비교하며 지상파 채널 전체를 필수설비로 규정해달라거나 무료로 내보내게 해달라고 정부여당에 건의하기도 했으나, 지상파 쪽은 관련주장을 부당하다고 일축한 바 있다.

		당시 지상파 쪽은 MSO 쪽에도 법적조치를 취할 것이라고 밝혔으며, 이번에 그 예고대로 칼을 빼든 것이다. 먼저 MSO 쪽은 재송신 유료화를 받아들이기 쉽지 않을 것으로 전해져, 이 사안은 법적 소송으로 비화될 가능성이 높다. 소송이 대법원까지 갈 경우 짧게는 2년에서 길게는 3년 이상의 시간이 걸릴 것으로 MSO와 지상파 쪽은 예상하고 있다. MSO 쪽은 그 2~3년 동안 유료방송 경쟁사업자인 IPTV 쪽과 싸워 이긴다면 지상파 쪽에 재전송대가를 지불할 여유가 있으나, 만약 경쟁에서 밀린다면 사업 자체가 위기인 마당이라 전송대가 문제는 크게 중요하지 않게 된다. 그렇다면 그 기간 동안 IPTV제공사업자들이 주요 '표적'이 될 것 이라는 게 전문가 전망이다. IPTV제공사업자들은 지난해 10월 이후 지상파 방송사와의 합의에 따라 수백 억 원 규모의 콘텐츠제작펀드를 조성하고 주문형비디오(VOD)의 사용료도 지불하기로 한 상태나, IPTV 상용화 이후 입장을 바꿔 재협상을 요구하고 있다. 하지만 한국방송협회 방송통신특별위원회는 이러한 주장을 일축하며 최악의 경우 재전송을 허용하지 않겠다는 입장이어서, MSO 법적조치까지 협상 테이블에 올려놓으면 IPTV제공사업자들은 더욱 수세에 몰릴 것으로 전문가들은 내다보고 있다.
오마이뉴스 유사도 81%	유선통신의 시장지배적 사업자인 KT(대표이사 이석채)가 2위 이동통신사업자 KTF와 합병한다. KT는 20일 이동통신 자회사인 KTF와의 합병을 이사회에서 결의하고 방송통신위원회에 합병인가를 신청할 예정이라고 밝혔다. KTF도 조만간 이사회를 열어 합병을 결의할 예정이어서, 지난 1996년 한국통신프리텔로 갈라선 뒤 13년 만에 KT와 다시 합치는 수순을 밟게 됐다. KT와 KTF 합병이 승인되면 매출 19조 원, 순익 1조2000억 원, 자산 25조	유선통신의 시장지배적 사업자인 KT(대표이사 이석채)가 2위 이동통신사업자 KTF와 합병한다. KT는 20일 이동통신 자회사인 KTF와의 합병을 이사회에서 결의하고 방송통신위원회에 합병인가를 신청할 예정이라고 밝혔다. KTF도 조만간 이사회를 열어 합병을 결의할 예정이어서, 지난 1996년 한국통신프리텔로 갈라선 뒤 13년 만에 KT와 다시 합치는 수순을 밟게 됐다. KT와 KTF 합병이 승인되면 매출 19조원, 순익 1조2000억 원, 자산 25조

	원대의 거대 통신기업이 탄생하게 된다. KT는 KTF와의 합병으로 유·무선통신 컨버전스(융합)산업을 선도하는 글로벌 사업자로 변화해 IT산업이 재도약하는 계기가 되도록 할 것이라고 합병추진 배경을 밝혔다. KT는 이탈리아, 스위스 등 11개국에서 단일기업이 유·무선통신서비스 모두를 제공하고 있으며, 미국 등 11개국에서는 유선통신 모회사가 이동통신 자회사의 지분을 100% 소유하고 있다는 점을 강조하기도 했다. KT의 고민은 유선통신 분야의 성장 정체가 뚜렷하고 이동통신은 투자 인센티브가 부족하다는 데 있었고, 이를 극복하기 위한 처방으로 합병법인을 선택했다. 합병 KT는 오는 2011년에 약 20조7000억 원의 매출을 올린다는 계획을 공개했다. 통합법인의 조직은 독립경영체제를 도입해 개인·홈(Home)·기업 고객부문 등으로 나누고, KTF는 개인고객 부문으로 독립 운영될 계획이다. 이석채 KT 사장은 합병은 KT 한 회사의 문제라기보다는 대한민국 IT산업의 동맥경화를 막는다는 차원이라며 합병을 통해 산업 내 리더십을 회복해 IT산업의 재도약을 이끌겠다고 밝혔다. 그러나 KT를 제외한 SK텔레콤과 SK브로드밴드, LG텔레콤, 그리고 케이블사업자까지 KT의 지배력 전이를 우려해 합병을 반대하고 있다. 합병 KT의 시장점유율은 현재 시내전화 91%, 초고속인터넷 43%, 이동통신 31.5%에 이르기 때문이다. 특히 결합상품으로 서비스가 급속도로 옮겨가는 상황이라, 시장지배력 전이는 곧바로 현실화될 것이라는 지적이다. SK텔레콤은 KT와 KTF가 합병하면 전체 통신 가입자의 51.3%, 매출액의 46.4%(이상 2007년말 기준)를 독식하는 거대 통신사업자가 등장해 공정	원대의 거대 통신기업이 탄생하게 된다. KT는 KTF와의 합병으로 유무선 통신 컨버전스(융합) 산업을 선도해 글로벌 사업자로 변화하고, 이를 통해 IT산업이 재도약하는 계기가 되도록 할 것이라고 합병추진배경을 밝혔다. KT는 이탈리아, 스위스 등 11개국에서 단일기업이 유·무선통신서비스 모두를 제공하고 있으며, 미국 등 11개국에서는 유선통신 모회사가 이동통신 자회사의 지분을 100% 소유하고 있다는 점을 강조하기도 했다. KT의 고민은 유선통신 분야의 성장 정체가 뚜렷하고 이동통신은 투자 인센티브가 부족하다는 데 있었고, 이를 극복하기 위한 처방으로 합병법인을 선택했다. 합병 KT는 오는 2011년에 약 20조7000억 원의 매출을 올린다는 계획을 공개했다. 통합법인의 조직은 독립경영체제를 도입해 개인·홈(Home)·기업 고객부문 등으로 나누고, KTF는 개인고객부문으로 독립 운영될 계획이다. 이석채 KT 사장은 합병은 KT 한 회사의 문제라기보다는 대한민국 IT산업의 동맥경화를 막는다는 차원이라며 합병을 통해 산업 내 리더십을 회복해 IT산업의 재도약을 이끌겠다고 밝혔다. 그러나 KT를 제외한 SK텔레콤과 SK브로드밴드, LG텔레콤, 그리고 케이블사업자까지 KT의 지배력 전이를 우려해 합병을 반대하고 있다. 합병 KT의 시장점유율은 현재 시내전화 91%, 초고속인터넷 43%, 이동통신 31.5%에 달하기 때문이다. 특히 결합상품으로 서비스가 급속도로 옮겨가는 상황이라, 시장지배력 전이는 곧바로 현실화될 것이라는 지적이다. SK텔레콤은 KT와 KTF가 합병하면 전체 통신 가입자의 51.3%, 매출액의 46.4%(이상 2007년말 기준)를 독식하는 거대 통신사업자가 등장해 공정

	경쟁이 사실상 불가능해진다고 반발했다. SK브로드밴드도 양사 합병에 따른 시장 지배력은 유무선 통신시장은 물론 IPTV, 인터넷전화 등 컨버전스 시장으로까지 확산, 고착화될 수 밖에 없다며 이로 인해 야기될 시장에서의 독점적인 지위는 투자 노력 감퇴 및 요금인하 여력을 소진시켜 궁극적으로 이용자 후생에도 역행하는 결과를 낳을 것이라고 지적했다. 결국 방통위가 이를 어떻게 판단하느냐가 관건이나, 미디어의 산업적 대형화를 기치로 내건 현 정부의 정책에 따라 인가될 가능성이 높다는 전망이다. 다만 공정거래위원회가 지난해 SKT의 SK브로드밴드(옛 하나로텔레콤) 인수 당시 지배력 전이를 막기 위한 조건을 내걸었던 사례가 있어 주목된다. 당시 공정위는 SKT가 독점하고 있는 800㎒ '황금 주파수'에 대한 회수 재분배와 결합상품 판매 시 경쟁업체에 대한 차별이나 경쟁업체 가입자를 끌어오는 행위는 금지해야 한다는 의견을 제시했었다. 한편 KT는 KTF의 2대 주주로 10.7%를 보유한 NTT도코모를 대상으로 5년 만기로 2억5000만 달러 규모의 교환사채(EB)를 발행하기로 결정했다. 교환사채 발행대금은 NTT도코모가 보유하고 있는 KTF 주식의 60%를 양도하는 방법으로 이뤄진다. 이는 합병을 위한 주식교환 시 자사주를 최대한 활용하고 외국인 지분한도를 고려한 신주 발행 물량을 최소화함으로써 주주가치를 높이기 위한 것이라고 KT는 밝혔다.	경쟁이 사실상 불가능해진다고 반발했다. 결국 방통위가 이를 어떻게 판단하느냐가 관건이나, 미디어의 산업적 대형화를 기치로 내건 현 정부의 정책에 따라 큰 문제없이 인가될 가능성이 높다는 전망이다. 다만 공정거래위원회가 지난해 SKT의 SK브로드밴드(옛 하나로텔레콤) 인수 당시 지배력 전이를 막기 위한 조건을 내걸었던 사례가 있어 주목된다. 당시 공정위는 SKT가 독점하고 있는 800㎒ '황금 주파수'에 대한 회수 재분배와 결합상품 판매 시 경쟁업체에 대한 차별이나 경쟁업체 가입자를 끌어오는 행위는 금지해야 한다는 조건을 달았었다. 한편 KT는 KTF의 2대 주주로 10.7%를 보유한 NTT도코모를 대상으로 5년 만기로 2억5000만 달러 규모의 교환사채(EB)를 발행하기로 결정했다. 교환사채발행대금은 NTT도코모가 보유하고 있는 KTF 주식의 60%를 양도하는 방법으로 이뤄진다. 이는 합병을 위한 주식교환 시 자사주를 최대한 활용하고 외국인 지분한도를 고려한 신주 발행 물량을 최소화함으로써 주주가치를 높이기 위한 것이라고 KT는 밝혔다.
오마이뉴스 유사도 75%	YTN이 사장도 공정방송위원회(이하 공방위) 심의 대상이 될 수 있도록 한 노사 협약을 체결했다. 공방위가 노사 양쪽 중 한쪽의 거부로 열리지 않을 경우에도 제재할 수 없었던 현행 규정을 이번 협약을 통해 실질적으로 담보하고 의결에 따른 사후 제재도 할 수 있게 개정했다. YTN 노사가 '공정방송을 위한 YTN 노사 협약'(이하	YTN 노사가 사장도 공정방송위원회(이하 공방위) 심의 대상이 될 수 있도록 협약을 체결했다. 또 노사 양쪽 중 한 쪽의 거부로 공방위가 열리지 않을 경우에도 제재할 수 없었던 현행 규정을 보완했으며 표결의 실효성을 높이고 의결에 따른 사후 제재도 가능하게 했다. YTN노사는 이 같은 내용을 담은'공정방송을 위한 YTN 노사 협약'(이하

	공정방송 협약)을 지난 10일 체결했다. △공방위의 개최와 표결의 실효성을 담보했고 △공추위 간사를 상근직으로 전환했으며 공추위 간사의 편집회의 참석 보장했다. 또 △경영진을 포함한 모든 구성원이 협약의 규율 대상임을 명시했으며 △윤리강령 위배 여부도 공방위 심의 대상이 될 수 있게 했다. 이번 공정방송 협약은 전국언론노동조합 YTN지부(지부장 노종면)가 지난해부터 벌여온 '공정방송 투쟁'의 성과물로 평가된다. 공방위 열리지 않을 경우 보도국장 신임투표 그간 공방위가 노사 양쪽 중 한쪽의 거부로 열리지 않을 경우에도 제재할 수 있는 규정이 없었으나 이번 협약으로 공방위 정례회의 2회, 임시회의 3회 이상 열지 않을 경우 회사쪽 공방위 대표인 보도국장 신임투표를 실시할 수 있게했다. 공방위 제재의 실효성을 위해 인사위원회에 징계 심사를 요구하거나 사장에 보직 변경을 요구할 수 있으며 인사위와 사장은 이를 '존중'하도록 했다. 공추위 간사 상근직…편집회의 참석 보장 공방위의 회사 대표는 보도국장으로 특정하고 노조 대표인 공정방송추진위원회 간사는 상근직화했다. 공추위 간사의 편집회의 참석도 보장했다. 공추위와 노조 집행위와는 분리 운영된다. 윤리강령 위배 여부도 공방위 심의 대상으로 규정해 보도 내용 뿐 아니라 부적절한 취재 및 보도 관행을 바로 잡는 계기를 마련했다. 사장도 공방위 심의 대상 이번 협약에는 현재까지는 규정에 없던 경영진의 책임이 명시돼 있다. 사장의 공정방송 실현 책무와 경영진의 의무 등을 규정하고 이를 어길 경우 사장도 공방위 심의 대상이 될 수 있도록 했다. 시청자위원회가 'YTN 보도의 공정성 평가' 기능을 수행할 수 있도록 했으며 시청자위를 구성할 때에는 노사 의견을 균형 있게	공정방송 협약)을 지난 10일 체결했다. 이번 협약은 전국언론노동조합 YTN지부(지부장 노종면)가 지난해부터 벌여온 '공정방송 투쟁'의 성과물로 평가된다.이번 협약에는 현재까지 규정에 없던 경영진의 책임이 명시돼 있다. 사장의 공정방송 실현 책무와 경영진의 의무, 보도국장의 공정보도 준수 공표 의무를 규정했으며 이를 어길 경우 사장도 공방위에 회부될 수 있게 했다. 공정방송 협약 3조2항은 사장을 비롯한 모든 YTN 구성원들은 방송강령과 윤리강령을 준수해야 하며 이에 위배되는 사안이 발생할 경우 공방위에서 다루도록 한다고 규정했다.공방위 개최의 실효성을 담보하기 위해 정례회의 2회, 임시회의 3회 이상 열지 않을 경우 회사쪽 공방위 대표인 보도국장 신임투표를 실시할 수 있게 했다. 문책요구권도 명시했다. 인사위원회에 징계 심사를 요구하거나 사장에 보직 변경을 요구할 수 있으며 인사위와 사장은 이를 '존중'하도록 했다. 안건이 가부동수로 부결됐을 경우 '동일인'이 '6개월 내' 공방위에 회부돼 가부동수가 나오면 가결로 본다는 내용도 삽입했다. 공방위가 회사와 노조쪽 인사 5명씩으로 구성돼 표결을 통한 안건 처리가 힘들다는 구조적인 문제를 풀기위해서다.공방위 회사 대표는 보도국장으로 특정하고, 노조 대표인 공정방송추진위원회 간사는 상근직화했다. 공추위 간사의 편집회의 참석도 보장했다. 공추위와 노조 집행위와는 분리 운영된다. 심의대상을 기존 '방송강령' 위배 사례뿐 아니라 '윤리강령' 위배 사항으로 넓혀 부적절한 취재 및 보도 관행을 바로 잡는 계기를 마련했다. 이에 대해 한국기자협회 YTN지회(지회장 김기봉)는 15일 자정결의를 통해 관행으로 용인돼온 부적절한 취재행태가 있었다면 이를 과감히 배척하고

	반영하기로 했다. 외부 기관 등에 YTN 보도 모니터 보고서 작성을 의뢰해 이를 공방위 논의자료로 활용하기로 했다. YTN노조 현존 사례 중 가장 실질적인 공정보도 보장제도 전국언론노동조합 YTN지부(지부장 노종면)는 이번 협약에 대해 노조의 투쟁이 궁극적으로 지향하고 있는 '공정방송'의 가치에 부합하는 성과물이라며 현존 사례 중 가장 실질적이고 합리적인 공정보도 보장제도라고 평가했다. YTN노사는 지난 4월1일 공정방송 제도화를 위해 노력하기로 합의했으며 4월10일 협의를 시작해 지난 5월23일 협의를 끝냈다. 지난 10일 노사가 '공정방송 협약'에 서명했으며 노조는 오는 18일 '노무현 전 대통령 서거 보도'와 관련해 임시 공방위 개최를 요구해 놓은 상태다.	취재원 관리라는 명분으로 합리화돼온 취재원과의 유착에도 경계선을 그어야 한다고 밝혔다. 시청자위원회가 'YTN 보도의 공정성 평가' 기능을 수행할 수 있도록 했으며 시청자위를 구성할 때에는 노사 의견을 균형 있게 반영하기로 했다. 외부 기관 등에 YTN 보도 모니터 보고서 작성을 의뢰해 이를 공방위 논의자료로 활용하기로 했다. YTN노조는 이번 협약에 대해 노조의 투쟁이 궁극적으로 지향하고 있는 '공정방송'의 가치에 부합하는 성과물로 현존 사례 중 가장 실질적이고 합리적인 공정보도 보장제도라고 평가하며 잘 운영해 YTN 보도의 질을 높이는 것이 남은 과제라고 말했다.YTN노사는 지난 4월1일 합의에 따라 공정방송 제도화를 위해 노력하기로 했으며 5월8일 실무 협의 기구를 구성해 5월23일 논의를 끝냈다. 지난 10일 노사가 '공정방송 협약'에 서명했으며 노조는 오는 18일 '노무현 전 대통령 서거 보도'와 관련해 회사에 임시 공방위 개최를 요구해 놓은 상태다.

<Abstract>

The Sejong corpus, which was built in the 21st Century Sejong Project, was the largest 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. The Korean corpus 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 to collect original texts of newspaper articles from various fields over the recent 10 years and to build up the publicly available corpus. Newspaper articles made up through this project will be able to contribute to the development of various technologies and research in the industry and academia as well as high-tech industries such as artificial intelligence industry.

The scope of the project can be divided into four parts: collection of original articles for newspaper articles, media composition and secondary copyright acquisition, refinement and normalization, and metadata tagging. In addition, the process was carried out in four steps: preparation for construction, selection of media, collection and digitization of original data, elimination and purification of duplicate articles, attachment of meta information, and cataloging.

According to amount of issueance, media type and requirements, 42 newspapers were selected to collect the original text articles, and which include 5 nationwide daily newspapers and 5 internet newspapers. Subsequently, after negotiating copyright permission, National Institute of Korean Language, media companies and project operator signed a copyright license agreement and a subsidiary agreement.

Article data of 18,369,901 articles and 3,351,131,155 words were collected from the selected media, and a tool for workers to refine them was developed. The refining tool was built as a system that allows multiple operators to work simultaneously. Workers logged on to the website and worked on a project basis, with as many as 8,000~20,000 articles, based on the manuals distributed.

On the other hand, a separate automatic refining process was conducted before manual refining by workers. The main focus was on excluding articles that were too short or long and too similar, and articles from media that did not have a license to use. The first refinement resulted in 5,029,926 articles and 1,656,947,078 words.

In the second manual refining process, the work was conducted in accordance with the criteria agreed with National Institute of Korean Language, which includes caption information such as images, tables, and graphs, copyright information, article information, and information not related to the article. Articles from other media with potential copyright issues, articles written by external contributors, articles that are difficult to see in general newspaper articles, and articles in full colloquial language. Workers shared the detailed work criteria online, resulting in 3,991,282 articles and 1,003,899,229 words.

After the 1st and 2nd refining, the final corpus naming convention and encoding method were applied according to the newspaper corpus construction guidelines, and the metadata to be included in the corpus file was determined. Metadata consists of title, author, publisher, year, article number, classification, article creation date, article author, and word count. In the case of article subject classification, media self classification and subject classification according to contents using artificial intelligence technology are included.

The final corpus file is produced in SJML format, and the basic structure of SJML consists of header and text parts.

In this project, the original newspaper corpus, which has been constructed through the collection of the original texts of newspaper articles and securing the right to use, is a language resource reflecting the language life of the present age and is expected to be utilized in various Korean language research and artificial intelligence.

사업 책임자 황이규(주식회사 마인즈랩 전무)
사업 참여자 안준환(주식회사 마인즈랩 상무)
진영순(주식회사 마인즈랩 이사)
서상원(주식회사 마인즈랩 팀장)
임성모(주식회사 마인즈랩 이사)
정소라(주식회사 마인즈랩 매니저)
박영선(주식회사 마인즈랩 매니저)
박다솜(주식회사 마인즈랩 매니저)
송혜원(주식회사 마인즈랩 매니저)
윤서영(주식회사 마인즈랩 매니저)
송동훈(주식회사 마인즈랩 상무)
김종범(주식회사 마인즈랩 팀장)
이재성(주식회사 마인즈랩 팀장)
이석준(주식회사 마인즈랩 매니저)
김마로(주식회사 마인즈랩 매니저)
이원문(주식회사 마인즈랩 매니저)
담당 연구원 이승재(국립국어원 언어정보과장)
이현주(국립국어원 언어정보과 학예연구관)

발행인: 국립국어원장
발행처: 국립국어원
서울시 강서구 금낭화로 154
전화 02-2669-9775,
전송 02-2669-9757
인쇄일: 2019년 12월 12일
발행일: 2019년 12월 12일
인 쇄: 비즈카피

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

(책 등)(표지 이미지 참조)

국립국어원 2019-01-33

신문 기사 원문 자료 수집 및 정제

국립국어원

(겉표지 뒷면)