

발 간 등 록 번 호

11-1371028-000907-01

# 2022년 신문 기사 원문 자료 수집 및 정제

사업 책임자 | 윤 종 웅

국립국어원 2022-01-33

발간등록번호
11-1371028-000907-01

## 2022년 신문 기사 원문 자료 수집 및 정제

사업책임자

윤 중 응



국립국어원



## 제 출 문

국립국어원장 귀하

국립국어원과 체결한 연구용역 계약에 따라 ‘2022년 신문 기사 원문 자료 수집 및 정제’에 관한 연구 보고서를 작성하여 제출합니다.

■ 사업 기간: 2022년 4월 18일 ~ 2022년 10월 18일

2022년 10월 18일

사업책임자: 윤종웅(주)윤즈정보개발)

연구 기관: (주)윤즈정보개발

사업 책임자: 윤종웅

사업 참여자: 남가운, 박지영, 서경찬  
안소연, 윤종성, 이승철  
임순영, 최원수





## 〈국문 요약〉

# 2022년 신문 기사 원문 자료 수집 및 정제

인공 지능 학습에는 양질의 대량 데이터가 필수적이다. 하지만 개인이나 기업, 학계에서는 학습에 필요한 말뭉치를 확보하는 데 어려움이 있다. 국립국어원에서는 이러한 문제를 해결하기 위해서 신문 기사 말뭉치를 구축하여 누구나 자유롭게 활용할 수 있도록 제공하고 있다. 국립국어원은 2018년부터, '21세기 세종 계획' 이후 단절되었던 말뭉치 구축 사업을 재개하여 인공 지능 산업계와 관련 연구 기관 등에서 공공재로 활용할 수 있는 대규모 한국어 학습 자료 구축 사업을 이어가고 있다. 그 연장선상에 있는 본 사업 또한 올해로 4년 차를 맞는다.

본 사업의 수행 범위는 신문 기사 원문 자료 수집(월별 1,000만 어절 이상), 저작권리자와의 이용 허락 계약을 통한 저작권 해결, 중복 기사 제거 및 정제, 신문 기사 원시 말뭉치 구축, 기사별 메타 정보 작성 및 목록 작성으로 구분되어 있다. 실제 언어 사용 양태를 반영한 신문 기사 말뭉치를 구축하기 위하여 최근 1년에 해당하는 2021년 신문 기사 원문을 수집, 다양한 분야에서 활용할 수 있는 데이터 생산에 목적을 두고 인용 부호 수정, 문장 분할 등의 정제를 진행하였다.

수집 대상은 국립국어원과 협의하여 총 34개 매체를 선택하였고 33개 매체의 저작권 신탁 기관인 한국언론진흥재단과 조선일보, 두 기관과 계약서 및 부속합의서를 작성하고, 해당 내용을 공증하여 저작권 해결을 진행하였다. 34개 매체로부터 확보한 원시 자료는 총 511,384,792개의 어절로 이루어진 2,656,468건의 기사이다.

기존 사업의 결과물과 같은 방식으로 구축되는 신문 기사 말뭉치와 더불어 추가 제안으로 인용 부호를 수정한 인용 부호 수정 말뭉치, 문단 내 문장으로 분할한 문장 말뭉치 총 3벌의 말뭉치 구축을 통해 활용성을 높이하고자 하였다.

신문 기사에는 인용 부호의 사용 코드가 동일 매체 내에서도 통일이 되지 않고, 열고 닫는 인용 부호가 알맞게 사용되지 않는 등, 정제되지 않은 기사들이 대부분을 이루고 있다. 또한 기사 내에 오타가 포함되어 있는 경우가 있다. 이러한 데이터를 학습하게 되면 말뭉치 활용에 효율이 크게 제약을 받을 수밖에 없다.

또, 대부분의 인공 지능 학습은 문장을 기본 단위로 하고 있다. 특히 형태소 분석과 기계 번역이 그렇다. 여러 개의 문장으로 이루어진 긴 단락을 단위로 말뭉치를 학습하는 것은 효율이 떨어질 수밖에 없다. 이에 단락을 문장으로 세분한 문장 말뭉치

구축을 추가하였다.

문장 말뭉치 구축 과정은 인용 부호 수정 말뭉치 데이터에서 단락으로 되어 있는 정보를 종결부호를 기준으로 하여 문장으로 나누었다. 피인용문 내 종결부호는 나누지 않았으며 평균 1개의 단락이 약 1.6개의 문장으로 분할되었다.

최종적으로 데이터 구축 후 한 번 더 저작권이 문제가 있는 기사가 있는지 확인하는 절차를 거쳤다. 기자 정보 전체를 확인하여 문제의 소지가 조금이라도 있는 기사는 각 매체에 확인 절차를 거쳐 제거하는 방법으로 진행하였다. 최종적으로 구축한 기사는 978,344개이고 208,320,912어절이다.

본 사업을 통해 구축한 말뭉치는 실제 언어 사용을 반영하고 있는 최신 말뭉치로서 3종의 말뭉치를 함께 이용하게 된다면 4차 산업혁명 대비 인공 지능 기술 개발 및 학계 연구 등 여러 분야에서 활용, 인공 지능 학습에 유용한 데이터가 될 것으로 기대한다.

**주요어:** 신문 말뭉치, 인공지능, 신문 기사, 학습용 데이터, 현대 한국어, 인용 부호 수정

<Abstract>

## Collection and Refinement of Data from Original Newspaper Articles in 2022

A large amount of high-quality data is necessary for artificial intelligence learning. However, it is difficult for individuals, companies, and academics to secure the corpus needed for learning. In order to combat this problem, National Institute of the Korean Language (NIKL) has built and provided the Newspaper Corpus that is widely available for anyone to use. In fact, NIKL, since 2018, has resumed the corpus building project, which had been discontinued since the 21st Century Sejong Project, and NIKL is continuing the project of building large-scaled Korean language learning materials that can be used as public resources in the artificial intelligence industry and other related researching institutes. Now, this corpus project, as an extension of the project by NIKL, has also been carried out over the last four years.

The scope of this project is divided into collecting the original text data from newspaper articles, amounting to more than 10 million *Eojeol* (word-spacing unit) per month, addressing copyright-related issues through usage agreements with copyright holders, removing and refining duplicate articles, building raw corpora from newspaper articles, and creating and listing metadata for each article. To build a Newspaper Corpus that reflects actual language usage patterns, original newspaper articles from 2021 were collected over a one year period, and then refined by correcting quotation marks and segmenting sentences, all with the aim of producing data that can be used in various fields.

In consultation with NIKL, a total of 34 media were selected for collection. The Korea Press Foundation which are the copyright management organizations for 33 media, and the Chosun Ilbo concluded the contract and annexed schedules with NIKL, as well as notarized the content to resolve copyright issues. The raw data secured from the 34 media included 2,656,468 articles with a total of 511,384,792 *Eojeol*.

We intended to increase usability by building a total of three corpora: a newspaper article corpus built in the same way as existing project output; a corpus which quotation marks were corrected; and a corpus in which paragraphs are split into sentences.

For newspaper articles, most of the articles are unrefined, including instances of quotation marks not being unified even within the same newspaper and improper use of

opening and closing quotation marks. Additionally, there are sometimes typos within the articles. Learning unrefined data inevitably causes the reverse effect (learning result) of corpus utilization.

In addition, sentences are used as the most efficient unit for most artificial intelligence learning basically. This is especially true for morphological analysis and machine translation. It is more efficient for AI to learn a corpus as a single sentences rather than long paragraph unit. Therefore, in order to compensate, the construction of a sentence corpus, subdivided paragraph into sentences, has been added.

During the sentence corpus-building process, the paragraphed information in the corpus which quotation marks were corrected was divided into sentences based on the ending period marks. The period marks within the quotes were not divided, and on average, one paragraph was divided into approximately 1.6 sentences.

After we finished building the data, we once more thoroughly checked if there were any articles with copyright issues. We conducted this process by checking all reporter information and removing articles that had even the slightest potential for problems through a confirmation procedure for each newspaper. The final build features 978,344 articles with 208,320,912 *Eojeol*.

The corpus built through this project is the latest corpus that reflects actual language use. If used together, the three corpora can be used in various fields, such as the development of artificial intelligence technology and academic research in preparation for the 4th Industrial Revolution, and will be useful data for artificial intelligence learning.

**Key-words:** corpus from newspaper, artificial intelligence, newspaper articles, AI training data, contemporary Korean, correction of quotation marks



# 차례

## 제 1장 서론

1. 사업 목적 .....	1
2. 사업 수행 범위 .....	1
3. 사업 수행 절차 .....	4
4. 사업 추진 경과 .....	5

## 제 2장 사업 수행 내용

1. 매체 선정 .....	7
2. 데이터 수집 .....	8
3. 데이터 1차 정제 .....	18
4. 데이터 2차 정제 .....	23
5. 메타데이터 작성 .....	38
6. 인용 부호 수정 말뭉치 .....	39
7. 문장 말뭉치 구축 .....	49

## 제 3장 사업 수행 결과

1. 신문 기사 정제 결과 .....	53
2. 매체별 납품 파일명 .....	57

<부록 1> 국가 언어 자원(말뭉치) 구축 및 활용 저작권 이용 허락 계약서 .....	58
<부록 2> 데이터 정제 작업 지침 .....	64
<부록 3> 말뭉치 종류별 구축 예시 .....	72

## 표 차례

<표 1> 사업공정표 .....	5
<표 2> 선정된 매체 구분 .....	7
<표 3> 최초 수집 기사 수,와 어절 수 .....	8
<표 4> 한국언론진흥재단 제공 원시 데이터 예시 .....	9
<표 5> 데이터 특징 .....	16
<표 6> 저작권 이용 문제로 인해 사용하지 않는 기사의 특징 .....	22
<표 7> 불필요한 요소 제거 내용 .....	29
<표 8> 원시 데이터와 정제된 데이터 비교 1 .....	30
<표 9> 원시 데이터와 정제된 데이터 비교 2 .....	30
<표 10> 원시 데이터와 정제된 데이터 비교 3 .....	33
<표 11> 2022년 신문 기사 주제별 통계 .....	37
<표 12> 인용 부호 치환 표 .....	39
<표 13> 인용 부호 수정 데이터 정제 전 후 .....	40
<표 14> 인용 부호 수정 데이터 정제 전 후 2 .....	40
<표 15> 최종 선정 기사 수 .....	41
<표 16> 최종 선정 기사 ‘한·중·일 호환용 한자 영역’의 한자 수(이하 생략) .....	41
<표 17> ‘한·중·일 호환용 한자 영역’ 한자 치환 표 .....	43
<표 18> 치환 코드 리스트 .....	44
<표 19> 오타류 후보 목록 글자 수정 전 후 .....	46
<표 20> 문장 말뭉치 데이터 정제 예 .....	49
<표 21> 신문 기사 정제 총괄표 .....	53
<표 22> 구축 연도별 기사와 어절 수 .....	54
<표 23> 월별 구축 어절 수 .....	55
<표 24> 주제별 기사 및 구축 어절 수 .....	55
<표 25> 말뭉치 파일명 .....	56



## 그림 차례

<그림 1> 구축 공정별 내용 .....	4
<그림 2> 연도별 매체 비율과 상위, 하위 기사수 매체 .....	8
<그림 3> 오류 예시 .....	16
<그림 4> 원본 데이터와 정제된 데이터의 예 .....	29
<그림 5> 작업 편집 화면 .....	35
<그림 6> 작업 프로그램 화면 .....	35
<그림 7> 데이터 정제 2차 검수 공정 .....	36
<그림 8> 인공지능인공 지능을 활용한 주제 분류 .....	37
<그림 9> 연도별 기사 주제 통계 .....	38
<그림 10> 문장 말뭉치 개념 .....	49
<그림 11> 문단 내 문장 분할 수(상/하위 5개 매체) .....	50
<그림 12> 구축 공정별 내용 .....	52
<그림 13> 매체별 최종 기사 수 및 월별 구축 어절 수 .....	54





## 제 1 장

# 서 론



## 제 1장 서론

### 1. 사업 목적

언어 처리 인공지능 기술 개발에는 다양한 형태의 언어 자료가 대량으로 필요하다. 또한, 언어는 끊임없이 변화하므로 인공지능이 학습할 자료가 꾸준히 새로 제공되어야 한다. 이를 위해 국립국어원은 ‘21세기 세종계획’을 통해 2억 어절의 자료를 구축하였으며, ‘모두의 말뭉치’에 매년 새로운 자료들을 추가로 구축하는 노력을 기울이고 있다.

여기에 맞추어 본 사업은 2021년에 생산된 신문 기사 말뭉치 구축 및 활용에 필요한 저작권을 확보하고, 원문 자료를 수집, 정제하여 신문 기사 원시 말뭉치를 구축하는 것을 목표로 한다.

신문 기사 원시 말뭉치의 구축량은 월별로 약 1,000만 어절 이상, 총 1.2억 어절 이상이다. 효과적으로 활용할 수 있는 말뭉치 구축을 위해 본 사업에서 구축될 말뭉치는 단순히 신문 기사를 수집만 한 것이 아니라, 가공을 거쳐 불필요한 정보를 제거한 말뭉치가 될 것이다.

### 2. 사업 수행 범위

본 사업의 범위는 네 부분으로 나눌 수 있다. 첫 번째는 **신문 기사의 원문 자료 수집**이다. 원문 자료 수집 대상은 2021년에 작성된 기사이며, 월별 1,000만 어절 이상을 목표로 한다. 매체는 25개 이상 선정해야 하며, 이 중 인터넷 기반 매체는 전체 매체 수의 10% 이내로 해야 한다.

두 번째는 **해당 매체 기사의 저작권을 확보**하는 것이다. 저작권 침해를 방지함으로써 사업 수행 결과물을 누구나 자유롭게 이용할 수 있도록 하는 과정이다. 언론진흥재단 및 언론사와 저작권 협약을 통해 말뭉치 구축 및 활용에 필요한 저작권을 확보한다. 협약을 통해 저작권이 확보된 매체의 기사 중에서도 작성자 정보가 없어 저작권이 불분명한 기사나 언론사 외부 기고자, 언론사 공동취재단이 작성한 기사, 대학생 기자, 연합 뉴스 제공 기사 등은 저작권 관리를 위해 말뭉치에서 삭제하였다.

세 번째는 **기사 데이터의 정제**이다. 데이터 정제의 주 내용은 기사 내 불필요한 요소(이미지, 도표, 문장으로 볼 수 없는 정보 등)를 제거하는 것이다. 이 작업을 통해 인공지능 학습 및 학계에서 활용할 수 있는 데이터를 생성해야 한다. 본 수행사는 기존 사업(2019년~2020년)과 동일한 형태의 데이터를 생산하고, 이에 더해 해당 데이터를 효율적으로 활용할 수 있

는 인용 부호 수정 말뭉치와 문장 말뭉치를 추가 생산하였다.<sup>1)</sup>

현재까지 공개된 신문 기사 말뭉치는 기사의 단락을 최소 단위로 하고 있다. 그러나 대부분의 인공 지능 학습은 문장을 기본 단위로 하고 있으며, 형태소 분석과 기계 번역은 대부분 문장을 기본 단위로 하고 있다. 따라서 단락을 최소 단위로 하는 말뭉치가 아니라, 문장을 최소 단위로 하는 말뭉치 구축이 필요하다. 본 사업에서는 단락을 문장으로 세분하여 문장 말뭉치를 구축하였다.

기술 협상 단계에서 체결된 내용은 다음과 같다. 기사 내 불필요한 요소를 제거하고 저작권에 위배될 수 있는 기사들을 제외한 신문 기사 말뭉치 중 인용 부호를 통일하지 않은 말뭉치 1종, 인용 부호를 수정한 인용 부호 수정 말뭉치 1종, 인용 부호 수정 말뭉치에서 문장 단위로 분할한 문장 말뭉치 1종, 총 3종의 데이터를 납품하기로 하였다.

이 3종의 말뭉치는 인공 지능 학습을 통해서 문장을 자동으로 분할하거나 기계 번역에 사용할 병렬 말뭉치 구축에 활용하는 등, 기존 말뭉치에 비해 훨씬 다양하게 활용할 수 있다.

마지막으로 구축된 기사 데이터의 기자 정보, 어절 수, 주제 분류, 기사 작성일 등의 **메타 데이터를 작성**하는 것이 사업의 범위이다.

## 가. 신문 기사 원문 자료 수집(2021년 작성 기사, 1억 어절 이상)

- ❖ 신문 기사 말뭉치 구축에 필요한 신문 기사 원문 자료를 수집.
- ❖ 대상은 2021년 기사로 월별 1,000만 어절 이상.
- ❖ 전국 종합지는 3개 이상의 매체를 포함하고, 인터넷 기반 매체는 수집하는 전체 매체 수의 10% 이내로 한정(매체 25개 이상).
- ❖ 현재 한국어 사용자의 일반적인 사용 양상이 반영된 신문 기사 원시 말뭉치는 매체별, 월별, 기사 주제별로 균형을 갖춰 1억 어절 이상 구축.
- ❖ 파일명과 표지의 종류 및 부착 형식 등은 국립국어원의 지침을 따름.

---

1) 2019년, 2020년 신문 기사 원문 자료 수집 및 정제 보고서 참조  
([https://www.korean.go.kr/front/reportData/reportDataList.do?mn\\_id=207](https://www.korean.go.kr/front/reportData/reportDataList.do?mn_id=207)), 2021년 사업에서는 인용부호를 통일한 데이터 제공

## 나. 신문 기사 저작 권리와 저작권 이용 허락 계약 체결

- ❖ 국립국어원 및 사업 수행자가 수집한 기사 원문 자료 전체 활용에 필요한 저작권을 확보.
- ❖ 수집한 기사 원문 자료 중 국립국어원에서 말뭉치 구축 대상으로 선정하는 매체의 기사 원문에 대해서 저작권자와 저작물 이용 허락 계약을 체결.
- ❖ 계약과 관련해 법률적인 검토를 받은 후 주관 기관이 제공한 계약서 양식에 따라 국립국어원과 협의하여 계약을 체결.
- ❖ 저작권 이용 허락 대상 권리는 신문 기사 원문 자료 및 신문 기사 말뭉치의 저장, 복제, 전송, 배포, 2차적 저작물 작성권을 포함.
- ❖ 이용 허락 기간은 계약일로부터 최소 2033년 12월 31일까지로 함.

## 다. 기사 데이터의 정제

- ❖ 수집된 기사 중에 동일 매체 내에서 기사 내용이 동일한 기사는 제거해야 함.
- ❖ 신문 기사 내에 삽입되어 있는 사진, 표, 그래프, 그림 및 캡션, 불필요한 태그 등 기사 원문 외의 요소들을 제거하고, 기사 내용과 관련 없는 텍스트 및 저작권 침해 요소가 포함된 기사나 외부 인원의 논설 등도 제거.
- ❖ 중복 기사, 길이가 너무 짧거나 긴 기사 등 말뭉치로 구축하기에 부적절한 기사 원문은 대상에서 제외하고, 정제된 신문 기사 원문을 대상으로 헤더 정보 부착 등의 표지 부착을 수행하여 원시 말뭉치 형태로 가공해야 함.

## 라. 메타 데이터 작성

- ❖ 국어원이 지정하는 9가지 분류 체계로 신문 기사 주제 재분류
- ❖ 신문사명, 기사 작성일, 주제 분류, 기사 제목, 어절 수 등 국립국어원이 지정하는 항목과 형식으로 기사별 메타 정보 입력 및 수집 기사 목록 작성

### 3. 사업 수행 절차

공정은 크게 7단계로 구분된다. 먼저 한국언론진흥재단(조선일보를 제외한 33개 매체의 저작권 계약 신탁 기관), 조선일보와 계약을 체결하여 기사를 확보하고 저작권을 해결하였다. 원시 데이터를 분석하고 가공하여 쓸 수 있는 데이터를 구분하는 1차 정제를 하였고, 해당 데이터를 가공하여 신문 기사 말뭉치 데이터를 생성하였다. 메타 데이터는 최종 선정된 기사를 바탕으로 작성하였다.

데이터 2차 정제까지 완료된 기사는 1종인 신문 기사 말뭉치로 납품되었고, 인용 부호 통일 공정까지 완료된 데이터는 인용 부호 수정 말뭉치로, 문장 단위 분할 공정까지 완료된 데이터는 문장 말뭉치로 총 3종의 데이터를 납품하였다.



<그림 1> 구축 공정별 내용

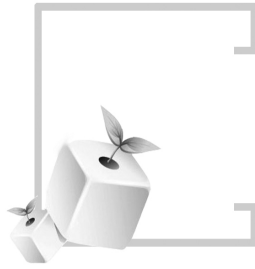
#### 4. 사업 추진 경과

본 사업의 추진 경과는 다음과 같다.

단 계	내 용	4 월	5 월	6 월	7 월	8 월	9 월	10 월
준 비	계약 및 착수 보고							
수 집	매체 선정							
	매체 계약 및 공증 진행							
	데이터 확보							
정 제	데이터 1차 정제							
	데이터 2차 정제							
	인용 부호 수정 말뭉치							
	문장 말뭉치							
메타데이터 생성	통계 추출							
	검수 및 반영							
납품 및 종료	샘플 데이터 납품							
	완료 보고							
	최종 데이터 납품							

<표 1> 사업공정표





## 제 2 장

# 사업 수행 내용



## 제 2장 사업 수행 내용

### 1. 매체 선정

매체를 선정하기 전 한국언론진흥재단과 접촉하여 선정된 매체의 기사 수를 먼저 확보하고 최소 25개 이상의 매체를 선정하여야 한다는 조건에 따라 국립국어원과 협의를 통하여 최종적으로 34개의 매체를 선정하였다. 사업비의 약 45%를 저작권 계약에 사용해야 한다는 의견 또한 수렴되었다. 수행사는 작년 사업을 수행하면서 얻은 경험으로 최소 200만 건 이상의 기사와 5억 어절 이상의 데이터를 확보해야 계약 어절 수를 충족하는 데이터를 생성할 수 있다고 판단하였고, 사전에 한국언론진흥재단이 가지고 있는 매체별로 기사 수를 문의하여 다양한 매체를 선정하였다. 전국 종합지 5개 매체를 선정하였고, 인터넷 매체는 전체 매체 수 대비 10% 이하로 선정하였다.

저작물은 2021년 1월 1일부터 2021년 12월 31일까지의 기사였으며, 계약 대상자는 한국언론진흥재단과 조선일보로, 저작권 이용 허락 계약을 통해 최대한 많은 기사 수와 어절 수를 확보하였다.

본 사업은 작년 사업과 동일하게 국립국어원과 매체 2자 간 저작권 이용 허락 계약과 국립국어원, 매체, 사업 수행사 3자 간의 부속합의서 계약으로 진행하였다. 계약서와 부속합의서에 대해 공증 절차도 진행되었다. 금액과 기간을 제외하고는 큰 쟁점 없이 계약을 체결하였다. 이용 허락 최소 기간은 2033년 12월 31일까지로 하였고, 저작자인 언론사가 이용 허락 중지 의사를 밝히지 않으면 이용 허락이 1년 단위로 자동 갱신되도록 하였다.

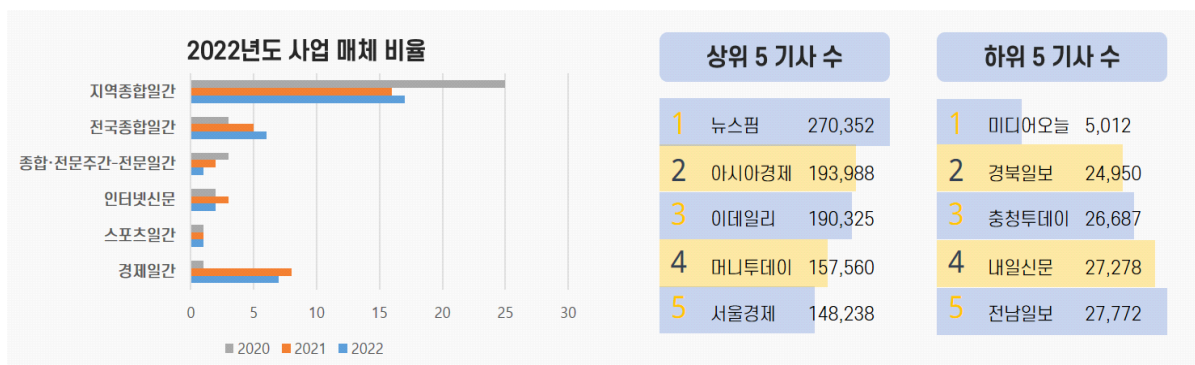
구분	매체명
전국종합일간	<ul style="list-style-type: none"> <li>국민일보, 내일신문, 서울신문, 조선일보, 한겨레, 한국일보</li> </ul>
지역종합일간	<ul style="list-style-type: none"> <li>강원일보, 경기일보, 경북일보, 경인일보, 기호일보, 남도일보, 대구신문, 대전일보, 동양일보, 매일신문, 부산일보, 전남일보, 전북도민일보, 중도일보, 충북일보, 충청일보, 충청투데이</li> </ul>
경제일간	<ul style="list-style-type: none"> <li>머니투데이, 서울경제, 아시아경제, 아주경제, 이데일리, 이투데이, 헤럴드경제</li> </ul>
스포츠일간	<ul style="list-style-type: none"> <li>스포츠서울</li> </ul>
종합전문주간	<ul style="list-style-type: none"> <li>미디어오늘</li> </ul>
인터넷신문	<ul style="list-style-type: none"> <li>노컷뉴스, 뉴스핌</li> </ul>

<표 2> 선정된 매체 구분

## 2. 데이터 수집

34개의 매체에서 수집된 기사와 어절 수는 각각 2,656,468건, 511,384,792개이고 한 기사의 평균 어절 수는 192개로 집계되었다. 최초 수집된 기사의 문장은 정제가 되기 전의 정보이므로 캡션 정보 등이 포함될 수 있다. 어절 수는 문장의 공백과 줄바꿈 수로 집계하였다.

수집한 데이터는 최초 목표인 기사 2백만 건 이상, 5억 어절 이상의 조건을 충족하였다. 매체 수는 2021년도 사업과 비교하여 한 매체가 줄었지만, 최초 수집된 기사는 조건을 충족하였다. 매체별 기사와 어절 수에 대한 내용을 정리하면 다음과 같다.



<그림 2> 연도별 매체 비율과 상위, 하위 기사수 매체

매체명	기사 수	어절 수	매체명	기사 수	어절 수
강원일보	37,910	4,118,201	서울경제	148,238	31,538,120
경기일보	29,741	5,693,024	서울신문	110,873	24,241,181
경북일보	24,950	5,052,512	스포츠서울	79,877	12,756,461
경인일보	35,522	6,549,123	아시아경제	193,988	36,561,962
국민일보	98,630	21,841,538	아주경제	102,500	24,500,654
기호일보	46,919	7,397,577	이데일리	190,325	33,177,908
남도일보	31,543	6,080,505	이투데이	104,123	20,518,842
내일신문	27,278	7,237,421	전남일보	27,772	5,533,794
노컷뉴스	142,024	25,841,804	전북도민일보	36,908	5,929,208
뉴스핍	270,352	35,191,030	조선일보	50,305	11,150,683
대구신문	31,012	6,024,116	중도일보	51,224	9,308,780
대전일보	47,114	7,793,680	충북일보	33,045	4,874,304
동양일보	31,639	4,442,671	충청일보	64,405	9,289,659
매일신문	57,870	11,265,352	충청투데이	26,687	4,920,738
머니투데이	157,560	35,552,827	한겨레	44,529	14,322,341
미디어오늘	5,012	2,701,758	한국일보	81,404	21,465,509
부산일보	87,073	17,275,077	헤럴드경제	148,116	31,236,432
계			총 합	2,656,468	511,384,792

<표 3> 최초 수집 기사와 어절 수

## 가. 원시 데이터 엑스엠엘(XML) 특징 분석

한국언론진흥재단에서 구매하여 제공받는 데이터는 한 기사가 하나의 엑스엠엘(XML) 파일로 되어 있다.

```
<?xml version="1.0" encoding="UTF-8"?>
<NewsML>
  <Catalog Href="http://newsml.or.kr/topicset/MasterCatalog.xml" />
  <NewsEnvelope>
    <DateAndTime>20210303T000626+0900</DateAndTime>
  </NewsEnvelope>
  <NewsItem>
    <Identification>
      <NewsIdentifier>
        <ProviderId>kmib.co.kr</ProviderId>
        <DateId>20210303</DateId>

<NewsItemId>01100201.20210303000626001</NewsItemId>
        <RevisionId PreviousRevision="0"
Update="N">1</RevisionId>

<PublicIdentifier>kmib.co.kr:20210303:01100201.20210303000626001:1</PublicIdentifier>
      </NewsIdentifier>
    </Identification>
    <NewsManagement>
      <NewsItemType FormalName="News" />
      <FirstCreated>20210303T000626+0900</FirstCreated>

<ThisRevisionCreated>20210303T000555+0900</ThisRevisionCreated>
      <Status FormalName="Usable"/>
      <StatusWillChange>
        <FutureStatus FormalName="Usable" />
        <DateAndTime>20210303T000626+0900</DateAndTime>
      </StatusWillChange>
      <Urgency FormalName="6" />
    </NewsManagement>
    <NewsComponent>
      <Role FormalName="Main" />
      <NewsLines>
        <HeadLine><![CDATA[여론조사·기호2번·참여경선... 야권
```

‘단일화’에 놓인 난제들]]></HeadLine>

<SubHeadLine><![CDATA[국힘 “2번 아니면 못 도와줘” 압박... 安 “통합선대위 꾸리자”  
정면돌파]]></SubHeadLine>

<ByLine>김동우, 이상현</ByLine>

<DateLine>20210303T000555+0900</DateLine>

<CreditLine></CreditLine>

<CopyrightLine>Copyright(C) 국민일보. All Right  
Reserved.</CopyrightLine>

<RightsLine></RightsLine>

<SeriesLine></SeriesLine>

<SlugLine></SlugLine>

<KeywordLine></KeywordLine>

</NewsLines>

<AdministrativeMetadata>

<FileName>01100201.20210303000626001.xml</FileName>

<Provider>

<Comment xml:lang="ko">국민일보</Comment>

<Party FormalName="01100201" />

</Provider>

<Creator>

<Party FormalName="001" >

<Property FormalName="Name"  
Value="김동우, 이상현" />

<Property FormalName="Title" Value="" />

<Property FormalName="Post" Value="" />

<Property FormalName="Email"

Value="love@kmib.co.kr," />

<Property FormalName="Blog" Value="" />

</Party>

</Creator>

</AdministrativeMetadata>

<DescriptiveMetadata>

<Language FormalName="ko" />

<Genre FormalName="Current" />

<SubjectCode>

```

        <Subject FormalName="" />
        <SubjectMatter FormalName="" />
        <SubjectDetail FormalName="" />
    </SubjectCode>
    <DateLineDate>20210303</DateLineDate>
    <Location HowPresent="Origin">
        <Property FormalName="Country" Value="" />
    </Location>
</DescriptiveMetadata>
<Metadata>
    <Property FormalName="PublishDate" Value="20210303" />
    <Property FormalName="PublisherCode" Value="" />
    <Property FormalName="PaperEdition" Value="1" />
    <Property FormalName="PageCategory" Value="정치" />
    <Property FormalName="PageCategoryId" Value="" />
    <Property FormalName="PrintingPage" Value="11" />
    <Property FormalName="PrintingPageNo" Value="0" />
    <Property FormalName="GenreInfo" Value="" />
    <Property FormalName="KsOrgan" Value="" />

    <Property FormalName="KsCompany" Value="" />

    <Property FormalName="KsPeople" Value="" />

    <Property FormalName="UciCode"
Value="G703:RA101-01100201.20210303000626001:1" />
        <Property FormalName="ModifyInfo" Value="" />
        <Property FormalName="CharacterCounter" Value="1574"
/>

        <Property FormalName="NewsItemOrgId"
Value="0924180858" />

        <Property FormalName="LinkPage"
Value="http://news.kmib.co.kr/article/view.asp?arcid=0924180858&code=11121900" />
        <Property FormalName="LinkmPage" Value="" />
        <Property FormalName="SubjectInfo" Value="정치" />
        <Property FormalName="SubjectInfo1" Value="선거" />
        <Property FormalName="SubjectInfo2" Value="11121900"
/>

        <Property FormalName="SubjectInfo3" Value="" />

```

```

        <Property FormalName="SubjectInfo4" Value="" />
        <Property FormalName="MapSubjectInfo" Value="" />
        <Property FormalName="AutoSubjectInfo"
Value="정치,국회_정당|정치,선거|" />
        <Property FormalName="AutoSubjectCode"
Value="001000000,001004000|001000000,001002000|" />
        <Property FormalName="AutoKeywordInfo"
Value="단일화,여론조사,야권 단일화,야권 단일,야권 단일 후보,시장 선거,야권 단일화 경선
승자,서울,3지대 단일화 후보,3지대 단일화" />
        <Property FormalName="Latitude" Value="" />
        <Property FormalName="Longitude" Value="" />
        <Property FormalName="PageCoordinate" Value="" />
        <Property FormalName="GroupCoordinate" Value="" />
        <Property FormalName="ScrapPage" Value="" />
        <Property FormalName="ScrapCategoryId" Value="" />
        <Property FormalName="ScrapCoordinate" Value="" />
        <Property FormalName="ScrapPdfFileName" Value="" />

```

</Metadata>

<NewsComponent>

<Role FormalName="Main" />

<ContentItem>

<MediaType FormalName="Text"/>

<MimeType FormalName="text/plain" />

<DataContent><![CDATA[야권 단일화가 순탄치

않다. 국민의힘은 당원 등 20만명이 단일화 최종 후보를 결정짓는

국민참여경선(오픈프라이머리)까지 꺼내 들었다. 각종 여론조사에서 밀리는 현 상황을 ‘우회’  
돌파하겠다는 얘기다. 제3지대 단일화 후보로 올라선 안철수 국민의당 대표는 “단일화 과정에서  
정권교체를 바라는 국민의 뜨거운 열망에 찬물을 끼얹는 그 어떤 행동도 조심해야 한다”고  
당부하면서도 뾰족한 수를 제시하지 못하고 있다.

사전투표를 한 달여 앞둔 2일 국민의힘과 국민의당 간 본격적인 살바싸움이 시작됐다. 김근식  
국민의힘 비전전략실장은 “여론조사 문항 하나를 놓고 기싸움하기보다 정권교체를 바라는 많은  
사람이 참여하고 결집하는 시너지를 만들 수 있을지를 고민해야 한다”고 말했다. 여론조사는  
물론 당원과 일반인 구분 없이 시민 20만명 정도가 야권 단일 후보를 고르는 오픈프라이머리 등  
야권 단일화 경선 승자를 가를 다양한 방안을 염두에 두겠다는 발언으로 해석된다. 여론조사를  
한정해 ‘적합도 조사’나 ‘경쟁력 조사’냐를 따지는 안 대표의 프레임에 빠져들지 않겠다는 뜻이다.

여기엔 김종인 국민의힘 비상대책위원장의 의중이 담겨 있다. 김 위원장은 “제3지대 후보로  
단일화돼서는 시장 선거에서 이길 수 없다”며 기호 2번(국민의힘)이 아니면 선거운동을 해줄 수

없다"며 안 대표에게 연일 선을 긋고 있다. 김 위원장은 "국민의당 4번으로 선거에 이기는 것을 확신할 수 있다. 나는 그런 확신이 없다. 안 대표가 단일화 협상을 하는 과정에서 장애적인 여파가 돼서는 안 된다"고 직격탄을 날렸다.

이에 안 대표는 통합선거대책위원회를 꾸리자는 제안으로 돌파하려 하고 있다. 국민의힘 입당·합당은 받아들일 수 없지만 통합선대위 구성만큼은 양보할 수 있다는 얘기다. 오픈프라이머리에 대해 안 대표 측은 국민일보와의 통화에서 "국민의힘 측에서 조직이 있으니 동원령을 내리겠다는 건데 서울시민의 보편적 정서를 대변하는 것은 선거인단이 아니다"며 불쾌한 반응을 보였다.

결론은 본선에서 국민의힘 지지층이 역할을 발휘하느냐, 중도층이 투표장으로 발걸음을 할 수 있느냐로 판가름날 것으로 보인다. 나경원 전 의원과 오세훈 전 서울시장 중 승자가 결정되는 4일 국민의힘으로서는 '컨벤션 효과'를 누릴 수 있는 타이밍이 마련된다. 그 후로도 여론조사에서 국민의힘이 승기를 잡지 못한다면 안 대표의 제안에 끌려다닐 수밖에 없는 상황이다.

서울시장을 탈환해 정권교체의 교두보로 삼겠다는 열망이 큰 만큼 야권이 극적으로 단일대오를 이룰 수 있는 여지는 있다. 국민의힘 관계자는 "국민의당에서도 기호 2번이 유리한지, 4번이 유리한지 객관적 데이터를 갖고 선택하지 않겠느냐"며 "결국 힘을 실어주는 주체는 국민이며 향후 여론조사 결과에 따라 향배가 결정될 수 있다"고 말했다.

중앙선거관리위원회는 4·7 재보궐선거 지역 21곳을 이날 확정했다. 서울·부산시장 광역단체장 2곳과 울산 남구청장, 경남 의령군수 등 기초단체장 2곳, 광역의원 8곳, 기초의원 9곳이 대상이다. 후보자 등록기간은 오는 18~19일이고, 사전투표는 다음 달 2~3일 이틀간 오전 6시부터 오후 6시까지 실시하기로 했다.

김동우 이상현 기자 love@kmib.co.kr]]></DataContent>

</ContentItem>

</NewsComponent>

<NewsComponent>

<Role FormalName="OriginMain" />

<ContentItem>

<MediaType FormalName="Text"/>

<MimeType FormalName="text/plain" />

<DataContent><![CDATA[야권 단일화가 순탄치 않다. 국민의힘은 당원 등 20만명이 단일화 최종 후보를 결정짓는 국민참여경선(오픈프라이머리)까지 꺼내 들었다. 각종 여론조사에서 밀리는 현 상황을 '우회' 돌파하겠다는 얘기다. 제3지대 단일화 후보로 올라선 안철수 국민의당 대표는 "단일화 과정에서 정권교체를 바라는 국민의 뜨거운 열망에 찬물을 끼얹는 그 어떤 행동도 조심해야 한다"고 당부하면서도 뾰족한 수를 제시하지 못하고



있다.&lt;br&gt;&lt;br&gt;사전투표를 한 달여 앞둔 2일 국민의힘과 국민의당  
 간 본격적인 살바싸움이 시작됐다. 김근식 국민의힘 비전전략실장은 “여론조사 문항 하나를 놓고  
 기싸움하기보다 정권교체를 바라는 많은 사람이 참여하고 결집하는 시너지를 만들 수 있을지를  
 고민해야 한다”고 말했다. 여론조사는 물론 당원과 일반인 구분 없이 시민 20만명 정도가 야권  
 단일 후보를 고르는 오픈프라이머리 등 야권 단일화 경선 승자를 가를 다양한 방안을 염두에  
 두겠다는 발언으로 해석된다. 여론조사를 한정해 ‘적합도 조사’나 ‘경쟁력 조사’냐를 따지는 안  
 대표의 프레임에 빠져들지 않겠다는 뜻이다.&lt;br&gt;&lt;br&gt;여기엔  
 김종인 국민의힘 비상대책위원장의 의중이 담겨 있다. 김 위원장은 “제3지대 후보로  
 단일화해서는 시장 선거에서 이길 수 없다”며 기호 2번(국민의힘)이 아니면 선거운동을 해줄 수  
 없다”며 안 대표에게 연일 선을 긋고 있다. 김 위원장은 “국민의당 4번으로 선거에 이기는 것을  
 확신할 수 있다. 나는 그런 확신이 없다. 안 대표가 단일화 협상을 하는 과정에서 장애적인  
 여파가 돼서는 안 된다”고 직격탄을 날렸다.&lt;br&gt;&lt;br&gt;이에 안  
 대표는 통합선거대책위원회를 꾸리자는 제안으로 돌파하려 하고 있다. 국민의힘 입당·합당은  
 받아들일 수 없지만 통합선대위 구성만큼은 양보할 수 있다는 얘기다. 오픈프라이머리에 대해 안  
 대표 측은 국민일보와의 통화에서 “국민의힘 측에서 조직이 있으니 동원령을 내리겠다는 건데  
 서울시민의 보편적 정서를 대변하는 것은 선거인단이 아니다”며 불쾌한 반응을  
 보였다.&lt;br&gt;&lt;br&gt;결론은 본선에서 국민의힘 지지층이 역할을  
 발휘하느냐, 중도층이 투표장으로 발걸음을 할 수 있느냐로 판가름날 것으로 보인다. 나경원 전  
 의원과 오세훈 전 서울시장 중 승자가 결정되는 4일 국민의힘으로서는 ‘컨벤션 효과’를 누릴 수  
 있는 타이밍이 마련된다. 그 후로도 여론조사에서 국민의힘이 승기를 잡지 못한다면 안 대표의  
 제안에 끌려다닐 수밖에 없는 상황이다.&lt;br&gt;&lt;br&gt;서울시장을  
 탈환해 정권교체의 교두보로 삼겠다는 열망이 큰 만큼 야권이 극적으로 단일대오를 이룰 수  
 있는 여지는 있다. 국민의힘 관계자는 “국민의당에서도 기호 2번이 유리한지, 4번이 유리한지  
 객관적 데이터를 갖고 선택하지 않겠느냐”며 “결국 힘을 실어주는 주체는 국민이며 향후  
 여론조사 결과에 따라 향배가 결정될 수 있다”고  
 말했다.&lt;br&gt;&lt;br&gt;중앙선거관리위원회는 4·7 재보궐선거 지역  
 21곳을 이날 확정했다. 서울·부산시장 광역단체장 2곳과 울산 남구청장, 경남 의령군수 등  
 기초단체장 2곳, 광역의원 8곳, 기초의원 9곳이 대상이다. 후보자 등록기간은 오는  
 18~19일이고, 사전투표는 다음 달 2~3일 이틀간 오전 6시부터 오후 6시까지 실시키로  
 했다.&lt;br&gt;&lt;br&gt;김동우 이상현 기자  
 love@kmib.co.kr&lt;br&gt;&lt;br&gt;GoodNews paper © &lt;a  
 href=&quot;http://www.kmib.co.kr&quot;  
 target=&quot;\_blank&quot;&gt;국민일보(www.kmib.co.kr)&lt;/a&gt;, 무단전재  
 및 수집, 재배포금지]]></DataContent>

</ContentItem>

</NewsComponent>

<NewsComponent>

<Role FormalName="Photo" />

```

<NewsComponent>
  <Role FormalName="Preview" />
  <ContentItem
Href="01100201.20210303000626001.01.jpg">
    <MediaType FormalName="Photo" />

    <Format FormalName="JPEG Baseline" />
    <MimeType FormalName="image/jpeg" />

    <Characteristics>
        <PhotoItemId></PhotoItemId>

<SizeInBytes>131300</SizeInBytes>
        <Property FormalName="Width"
Value="600" />
        <Property FormalName="Height"
Value="301" />
        <Property
FormalName="ColorSpace" Value="" />
        <Property
FormalName="ICCProfile" Value="" />
    </Characteristics>
  </ContentItem>
</NewsComponent>
<NewsComponent>
  <Role FormalName="Caption" />
  <ContentItem>
    <MediaType FormalName="Text" />
    <MimeType FormalName="text/plain" />
    <DataContent>
      <![CDATA[연합뉴스TV 캡처]]>
    </DataContent>
  </ContentItem>
</NewsComponent>
</NewsComponent>
</NewsItem>
</NewsML>

```

<표 4> 한국언론진흥재단 제공 원시 데이터 예시

한국언론진흥재단으로부터 제공받은 데이터를 분석해 보면 데이터가 메타 정보와 본문으로 구분되어 있다. 본문 내용(DataContent)으로 태그된 것이 수집 대상이 되는데, 이 데이터의 특징은 다음과 같다.

• 하나의 기사를 한 개의 XML 파일로 제공함.
• XML 문서 내의 본문 내용(DataContent)은 기사의 구조 정보가 없는 단순한 텍스트 형태임.
• 저자, URL, UCI, 제목, 분류 등의 메타 정보를 제공함.
• XML 파일은 &lt; &gt; 등 엔티티가 그대로 남아 있음.
• XML 문서에 소제목 등의 구조 마크업이 누락되어서 다음 단락과 붙어 버리는 문제가 있음.
• 원시 데이터에서 서명 기호 안의 글자가 누락된 것이 발견됨.
• 인용 부호로 ' , ' , " , " 등을 사용해 표준에 맞지 않음. 인용 부호의 열고 닫는 짝이 맞지 않음.
• 같은 의미로 사용되는 가운뎃점, 마침표, 쉼표 등이 여러 가지 코드로 일관성 없이 사용됨.
• 이(李), 리(李)와 같은 한자 호환용 코드가 사용되어 데이터의 공유와 유통에 문제를 일으킴.

<표 5> 데이터 특징

한국언론진흥재단의 데이터 중 일부 자료에서는 캡션 정보가 본문과 구분되지 않는 경우가 존재한다. 캡션 정보가 마치 본문인 것처럼 등장하는데, 캡션 정보는 불필요한 요소로 삭제 대상이다. 이러한 오류 유형은 웹 페이지의 데이터를 확인하면서 캡션 정보를 삭제하는 방식으로 처리하였다. 오류의 유형은 다음과 같다.

☒ **신문 매체별로 다양한 패턴의 오류가 숨어 있음**

XML 문서 구조 분석

**전북도민일보**      부안군 줄포만갯벌생태공원에 백만송이 해바라기꽃 장관

줄포만갯벌생태공원은 약15만명의 웅활한 편익에 매년 계절별로 다채로운 대안 관광을 ...  
 성해 8월에는 마치 빈센트 반 고흐가 노랑꽃밭을 뿌려놓은듯한 햇노랑 해바라기 ...  
 지친 군민과 관광객을 맞이하고 있다.

**전북도민일보**      코로나세데-졸업, 실감나지 않아요"

△ 장수민 기자    △ 부안 2021.02.21 15:04    △ 1면 2단



졸업식도 못 가는데 학사모 쓰고 사탕이라도 남가에 대학생활 마무리 하는 기분인 날 것 같아 씁쓸  
 나랑"

시끌벅적한 분위기 속 친구들과 학사모를 던지며 대학  
 생활을 마무리하는 졸업식 풍경이 사라졌다. 코로나19  
 가 대학 졸업식의 풍경까지 변화시켜버린 것이다.

매체 별로 정밀하게 분석하고  
HTML 문서를 참조해서 해결함

데이터 특징

**코란19에** 지친 군민과 관광객을 맞이하고 있다. '태  
 양의 전 아폴로에게 한눈에 반한 물의 요정이 한자리에  
 서 아폴로를 기다리다 해바라기가 됐다'는 그리스 로마  
 신화처럼 태양을 그리다 얼굴마저 태양을 닮아버린 해  
 바라기꽃이 줄포만갯벌생태공원을 노랗게 물들이고 있  
 자. 줄포만갯벌생태공원은 자생하는 꽃말인 '당신만  
 을 사랑한다'코로나19와 폭염을 이겨달라"고 말했다.  
 부안=방선동 기자

한국언론진흥재단 제공 데이터

**2021년** 코로나시대 전북대학교 졸업식 장면 / **최기웅**  
**수습기자** 졸업식도 못 가는데 학사모 쓰고 자신이라  
 닮게야 대학생들 마무리 짓는 기분이 날 것 같아 왔습  
 니다."

<그림 3> 오류 예시

원시 데이터에서 데이터가 소실된 유형도 발견하여 아래와 같이 처리하였다. '< >'와 같이 서명 기호가 사용된 경우 실제 데이터 자체에서 해당 기호 안 글자가 누락된 경우가 있었다. 유형에 대한 예시는 다음과 같다.

**경기일보 실제 웹 화면**

**XML 문서 구조 분석**

정자연 기자 jiy84@kyeonggi.com  
기자페이지 >

현실이 고단할 때 사람들은 문학에서 힘을 얻는다. 책에서 울고 웃으며 현실에서 도피하기도, 새로운 삶을 살아갈 희망을 품기도 한다. 지난해 한국소설이 2012년 이후 가장 많은 판매량을 기록한 것만 봐도 알 수 있다. 문학의 힘은 2021년에도 이어질 것으로 보인다. 한국 문단을 이끄는 베스트셀러 작가들이 신작을 잇달아 내놓는다.

■스타 여성 작가들 대거 컴백

2015년 표절 시비 이후 문단을 떠났던 신경숙 작가는 장편소설 '아버지에게 갔었어'(창비)를 통해 공식 복귀한다. 지난해 찬자기비평 웹 매거진에 연재한 글을 엮었다. 2013년 짧은 소설집 <달에게 들려주고 싶은 이야기>를 낸 이후 8년 만의 신작이다. 고통을 참으며 자리를 지켜내는 아버지의 목소리를, 나와 아버지의 삶을 교차하며 풀어낸다.

맨부커상 수상 작가 한강은 제주 4·3 사건의 상흔을 다룬 신작 <작별하지 않는다>(문학동네)로 돌아온다. 계간 <문학동네>에 쓴 글을 엮어 올 상반기 출간한다. 악몽에 시달리며 괴로워하는 소설가 k를 통해 현대사의 비극인 제주 4·3사건을 비춘다. 한강 작가 특유의 소재를 통한 이미지화가 돋보인다. 소설 곳곳에 내리는 눈은 고통으로 다가온다. 조남주 작가는 상반기 출간 예정인 <오기>(민음사)를 통해 '82년생 김지영'에 쏟아진 질문에 답한다.

**한국언론진흥재단으로부터 받은 원시 데이터**

**데이터 특징**

```

1
2 "news_id": "01200101.20210107141620001",
3 "title": "신경숙, 한강부터 도스토예프스키까지, 2021년 출간계 키워드",
4 "content": "현실이 고단할 때 사람들은 문학에서 힘을 얻는다. 책에서 울고 웃으며 현실에서 도피하기도, 새로운 삶을 살아갈 희망을 품기도 한다. 지난해 한국소설이 2012년 이후 가장 많은 판매량을 기록한 것만 봐도 알 수 있다. 문학의 힘은 2021년에도 이어질 것으로 보인다. 한국 문단을 이끄는 베스트셀러 작가들이 신작을 잇달아 내놓는다. ■스타 여성 작가들 대거 컴백 2015년 표절 시비 이후 문단을 떠났던 신경숙 작가는 장편소설 '아버지에게 갔었어'(창비)를 통해 공식 복귀한다. 지난해 찬자기비평 웹 매거진에 연재한 글을 엮었다. 2013년 짧은 소설집 <달에게 들려주고 싶은 이야기>를 낸 이후 8년 만의 신작이다. 고통을 참으며 자리를 지켜내는 아버지의 목소리를, 나와 아버지의 삶을 교차하며 풀어낸다. ■맨부커상 수상 작가 한강은 제주 4·3 사건의 상흔을 다룬 신작 <작별하지 않는다>(문학동네)로 돌아온다. 계간 <문학동네>에 쓴 글을 엮어 올 상반기 출간한다. 악몽에 시달리며 괴로워하는 소설가 k를 통해 현대사의 비극인 제주 4·3사건을 비춘다. 한강 작가 특유의 소재를 통한 이미지화가 돋보인다. 소설 곳곳에 내리는 눈은 고통으로 다가온다. 조남주 작가는 상반기 출간 예정인 <오기>(민음사)를 통해 '82년생 김지영'에 쏟아진 질문에 답한다. ■노벨문학상 작가의 귀환부터 도스토예프스키까지 ■신경숙의 작품 중에는 노벨문학상을 수상한 거장들의 역작이 출간을 기다리고 있다. 2017년 노벨문학상을 수상한 영국 소설가 가즈오 이시구로의 장편 <만유사>이 4월에 나올 예정이다. 2006년 노벨문학상을 수상한 오르한 파무크의 도 같은 출판사에서 7월 한국을 찾는다. 두 책은 모두 편대역을 소개로 한다. 도스토예프스키 탄생 200주년을 맞아 출판사 앞선책들은 대표작들을 새롭게 단장해 선보인다. ■한강의 신작 '자'."
5 "published_at": "2021-01-07T00:00:00.000+09:00",
6 "thumbnail_url": "2021-01-07T00:00:00.000+09:00"

```

위 사례의 경우, 특정 기사에서 서명 기호(<, >) 안 텍스트가 원시 데이터에서는 사라져 있는 것을 확인할 수 있다. 위와 같은 경우에는 웹페이지의 데이터와 일일이 비교하여 해당 기사는 사용하지 않는 방법으로 진행하였다.

### 3. 데이터 1차 정제

#### 가. 중복 기사, 유사한 데이터 제거

원시 데이터를 수령한 뒤 중복 기사와 유사 기사를 제거한다. 중복 기사 판단은 전체 매체의 기사를 대상으로 하며, 유사 기사는 같은 매체에서 기사별 전후 14일 내의 기사들이 대상이다. 이후 불필요한 요소를 제거하는 데이터 2차 정제 공정을 진행한 뒤 중복, 유사 데이터 제거 공정을 한 차례 더 진행한다. 불필요한 부분을 삭제하는 공정 후 기사에 포함된 문장이 달라지면서 기사 내용이 중복되거나 유사도가 변하는 경우가 존재하기 때문이다. 작업 기준은 다음과 같다.

- ❖ 중복 체크를 통해 모든 매체의 기사들 중에서 내용이 일치하는 데이터는 최초 등장한 기사를 사용하고, 나머지 기사는 삭제함.
- ❖ 같은 매체 기사들 중 전후 14일 내의 기사들을 비교하여 85% 이상 유사도를 보이는 기사는 제외함.

○○일보 기사 중 제목은 다르나 내용이 중복된 기사의 예	
<p>제목: 정착하려니 “농구장 신축해라” 농구단 짜내는 지자체</p> <p>프로농구의 연고지 문제를 둘러싸고 지방자치단체와 구단의 갈등이 접점을 찾지 못하고 있다. 구단이 바라는 지원과 지자체가 하려는 지원이 서로 어긋나면서 애꿎은 팬심만 상처받고 있다.</p> <p>프로농구는 9일 연고지와 관련해 두 가지 이슈로 떠들썩했다. kt가 부산에서 수원으로 옮긴다는 것과 한국가스공사의 새 연고지가 확정되지 않았다는 사실이다. 두 사안 모두 지자체가 구단에 신축 구장을 요구하면서 갈등의 골이 깊어졌다.</p> <p>한국가스공사 관계자는 10일 “구장 신축에 대해서는 우리도 필요성을 공감한다”면서 “현실적으로 누가 어떤 방식으로 짓느냐인데 그 부분에 대해 협의가 아직 안 끝났다”고 설명했다. 대구시는 가스공사가 많은 비용을 부담해 신축구장을 건설해주기를 바라는 것으로 알려졌다. 가스공사는 대구체육관의 개</p>	<p>제목: “현 집 줄게 새 집 다오”... 지자체의 요상한 농구장 셈법</p> <p>프로농구의 연고지 문제를 둘러싸고 지방자치단체와 구단의 갈등이 접점을 찾지 못하고 있다. 구단이 바라는 지원과 지자체가 하려는 지원이 서로 어긋나면서 애꿎은 팬심만 상처받고 있다.</p> <p>프로농구는 9일 연고지와 관련해 두 가지 이슈로 떠들썩했다. kt가 부산에서 수원으로 옮긴다는 것과 한국가스공사의 새 연고지가 확정되지 않았다는 사실이다. 두 사안 모두 지자체가 구단에 신축 구장을 요구하면서 갈등의 골이 깊어졌다.</p> <p>한국가스공사 관계자는 10일 “구장 신축에 대해서는 우리도 필요성을 공감한다”면서 “현실적으로 누가 어떤 방식으로 짓느냐인데 그 부분에 대해 협의가 아직 안 끝났다”고 설명했다. 대구시는 가스공사가 많은 비용을 부담해 신축구장을 건설해주기를 바라는 것으로 알려졌다. 가스공사는 대구체육관의 개</p>

보수를 통해 다음 시즌을 준비하기를 원하지만 대구시는 개보수를 신축구장과 연계해 요구하면서 협상이 난항을 겪고 있다.

kt가 연고지 이전을 결정한 이유 역시 부산시의 신축구장 요구가 결정적 원인이 됐다. kt 관계자는 10일 “부산시에 사직체육관 내 2개 보조경기장 중 하나를 연습구장으로 쓰게 해달라고 건의했는데 시민들이 쓰니 아예 안 된다고 했다”면서 “4일에 시와 협상했는데 이 자리에서 부지를 마련할 테니 경기장을 지으라고 했다”고 설명했다. 부산시는 국비 30% 정도를 받을 수 있게 해보겠다고 논의했지만 신규 건립이 부담스러웠던 kt는 결국 수원 이전을 결정했다.

지자체의 프로스포츠단에 대한 열악한 지원은 이미 다른 종목에서도 심심치 않게 볼 수 있다. 대기업이 프로 구단을 운영하는 한국 프로스포츠의 특성상 지자체는 구단에게 지나친 희생을 요구하는 경향이 있다.

가스공사 관계자는 “대구체육관의 개보수마저 안 된다면 최후에 연고지를 다른 곳으로 알아봐야 하는 상황”이라고 말했다. 서울신문은 대구시와 부산시의 입장을 듣고자 수차례 통화를 시도했지만 연락이 닿지 않았다.

○○○기자 ---@---.---.---

보수를 통해 다음 시즌을 준비하기를 원하지만 대구시는 개보수를 신축구장과 연계해 요구하면서 협상이 난항을 겪고 있다.

kt가 연고지 이전을 결정한 이유 역시 부산시의 신축구장 요구가 결정적 원인이 됐다. kt 관계자는 10일 “부산시에 사직체육관 내 2개 보조경기장 중 하나를 연습구장으로 쓰게 해달라고 건의했는데 시민들이 쓰니 아예 안 된다고 했다”면서 “4일에 시와 협상했는데 이 자리에서 부지를 마련할 테니 경기장을 지으라고 했다”고 설명했다. 부산시는 국비 30% 정도를 받을 수 있게 해보겠다고 논의했지만 신규 건립이 부담스러웠던 kt는 결국 수원 이전을 결정했다.

지자체의 프로스포츠단에 대한 열악한 지원은 이미 다른 종목에서도 심심치 않게 볼 수 있다. 대기업이 프로 구단을 운영하는 한국 프로스포츠의 특성상 지자체는 구단에게 지나친 희생을 요구하는 경향이 있다.

가스공사 관계자는 “대구체육관의 개보수마저 안 된다면 최후에 연고지를 다른 곳으로 알아봐야 하는 상황”이라고 말했다. 서울신문은 대구시와 부산시의 입장을 듣고자 수차례 통화를 시도했지만 연락이 닿지 않았다.

○○○기자 ---@---.---.---

○○일보 기사 중 유사도 비교를 통해 사용하지 않는 기사의 예(유사도 85%)

<p>제목: 김태호, 국민의힘 복당 임박...부산시장 보선 역할 기대</p> <p>‘더 좋은 세상으로’ 포럼에 참가한 무소속 김태호 의원. 연합뉴스</p> <p><b>무소속</b> 김태호(경남 산청함양거창합천) 의원이 <b>이르면</b> 7일 국민의힘에 복당<b>할 것으로 알려졌다</b>.</p> <p>이에 따라 김 의원이 부산시장 보궐선거를 포함한 4월 부산·울산·경남(PK) 재보선에서 국민의힘 지원유세에 상당한 역할을 할 것으로 보인다. 이와 함께 내년 3월 차기 대선에서도 부울경 보수진영의 대표주자로 출마할 가능성이 높다.</p> <p>국민의힘 <b>핵심 관계자는 7일 오전 “현재 비상대책위에서 김 의원 복당 문제가 논의중이다”고 말했다. 다른 관계자는 “김 의원이 국민의힘에 복당할 가능성이 높다”고 덧붙였다</b>.</p> <p>3선의 김 의원은 지난해 4월 총선을 앞두고 경남 산청함양거창합천에 공천을 신청했지만 국민의힘 공천관리위원회가 ‘혐지 출마론’을 들어 부정적인 반응을 보이자 무소속으로 출마해 당선됐다. 이후 지난해 9월 17일 복당을 신청하고 지도부의 판단을 기다려왔다. 김 의원이 <b>복당하게 되면</b> 권성동(강원 강릉)에 이어 <b>무소속 출신으로 2번째 국민의힘으로 돌아오게 된다</b>.</p> <p><b>김 의원이 정식</b> <b>복당하게 되면</b> 4월 PK 재보선에 의미있는 역할을 할 것으로 보인다. 국민의힘 PK 정치권에선 “4월 재보선에서 승리하기 위해선 김 의원이 빨리 돌아와야 한다”며 여러차례에 걸쳐 당 지도부에 김 의원의 조속한 복당을 촉구해왔다. 국민의힘 소속 부산의 모 의원은 “김종인 비상대책위원장과 주호영 원내대표의 영향력이 극히 미진한 상황에서 김 의원 만한 확실한 지원세력이 없다”며 “김 의원 복당으로 국민의힘이 부산시장 선거에서 승리할 가능성이 매우 높아졌다”고 했다.</p> <p>김 의원은 차기 대선에서도 부울경 대표 주</p>	<p>제목: 김태호, 국민의힘 복당...4월 PK 재보선 의미있는 역할할 듯</p> <p>‘더 좋은 세상으로’ 포럼에 참가한 무소속 김태호 의원. 연합뉴스</p> <p>경남도지사를 지낸 3선의 김태호(경남 산청함양거창합천) 의원이 7일 국민의힘에 복당했다.</p> <p>이에 따라 김 의원이 부산시장 보궐선거를 포함한 4월 부산·울산·경남(PK) 재보선에서 국민의힘 지원유세에 상당한 역할을 할 것으로 보인다. 이와 함께 내년 3월 차기 대선에서도 부울경 보수진영의 대표주자로 출마할 가능성이 높다.</p> <p>국민의힘은 이날 비상대책위원회를 열어 김태호 의원의 복당을 허용했다.3선의 김 의원은 지난해 4월 총선을 앞두고 경남 산청함양거창합천에 공천을 신청했지만 국민의힘 공천관리위원회가 ‘혐지 출마론’을 들어 부정적인 반응을 보이자 무소속으로 출마해 당선됐다. 이후 지난해 9월 17일 복당을 신청하고 지도부의 판단을 기다려왔다. 김 의원의 복당은 권성동(강원 강릉)에 이어 2번째이다.</p> <p>이날 복당으로 김 의원은 4월 PK 재보선에 의미있는 역할을 할 것으로 보인다. 국민의힘 PK 정치권에선 “4월 재보선에서 승리하기 위해선 김 의원이 빨리 돌아와야 한다”며 여러차례에 걸쳐 당 지도부에 김 의원의 조속한 복당을 촉구해왔다. 국민의힘 소속 부산의 모 의원은 “김종인 비상대책위원장과 주호영 원내대표의 영향력이 극히 미진한 상황에서 김 의원 만한 확실한 지원세력이 없다”며 “김 의원 복당으로 국민의힘이 부산시장 선거에서 승리할 가능성이 매우 높아졌다”고 했다.</p> <p>김 의원은 차기 대선에서도 부울경 대표 주</p>
--	---

<p>이 부산시장 선거에서 승리할 가능성이 매우 높아졌다”고 했다.</p> <p>김 의원은 차기 대선에서도 부울경 대표 주자로 나설 전망이다. 현재 PK 보수진영에선 차기 대선에 출마할 유력 인사가 전무한 실정이다. 아직 차기 주자 선호도 조사에서 김 의원의 지지도가 낮은 것은 사실이지만 대중성 높은 그가 대선전에 본격 가세할 경우 ‘태풍(김태호 바람)’처럼 막강한 바람을 일으킬 수 있다는 관측이다. 16대 대선 때 <b>지지도 1%</b>로 시작해 대권을 거머쥔 노무현 전 대통령과 비슷한 ‘돌풍’을 기대하는 사람도 많다.</p>	<p>자로 나설 전망이다. 현재 PK 보수진영에선 차기 대선에 출마할 유력 인사가 전무한 실정이다. 아직 차기 주자 선호도 조사에서 김 의원의 지지도가 낮은 것은 사실이지만 대중성 높은 그가 대선전에 본격 가세할 경우 ‘태풍(김태호 바람)’처럼 막강한 바람을 일으킬 수 있다는 관측이다. 16대 대선 때 1%대의 지지율로 시작해 대권을 거머쥔 노무현 전 대통령과 비슷한 ‘돌풍’을 기대하는 사람도 많다.</p>
---	---



## 나. 기사 선택

이전 단계에서 원시 데이터의 유사도 비교를 통해 중복 기사와 유사도가 높은 기사를 먼저 제거함으로써 이후 단계에서 작업의 효율성을 높일 수 있었다. 이 단계에서는 유사도 비교를 통해 한 차례 걸러낸 기사들 중에서 유사도 이외의 이유로 사용할 수 없는 기사를 원시 데이터의 메타 정보를 활용하여 선별하고, 구축 대상 기사에서 제외하는 작업을 진행한다. 기준은 아래와 같다.

- ❖ 기사 길이 1,000어절 이상, 100어절 이하는 제외함(길이로 인한 기사 선택은 불필요한 요소를 제거한 후의 본문 어절 수를 확인하여 마지막에 제외.).
- ❖ 단순 광고, 떠벌 오늘의 운세, 퀴즈 등 기사로 보기 어려운 것은 제외함.
- ❖ 승진자나 부고 명단, 스포츠 경기의 결과 수치만으로 구성된 기사는 제외함.
- ❖ 기사의 대부분이 영어나 일어 등 다른 언어로 된 것은 제외함.
- ❖ ‘~했어요.’, ‘~란다.’, ‘~할까요?’ 등 기사 전체가 구어체로 이루어진 기사는 제외함.
- ❖ 인공 지능 로봇이 작성한 기사는 제외함.
- ❖ 저작권 이용에 문제가 될 소지가 있는 기사는 제외함.
  - 대학생 기자나 리포터, 같은 계열사이나 저작권을 따로 가지고 있는 매체, 타 기관 소장, 부장, 의사 등 매체에 속하지 않은 외부 기고가 및 전문가가 작성한 기사 등.
  - 기자 정보가 없는 데이터는 제외함(한국언론진흥재단 측에 문의한 결과 해당 기자 정보를 얻을 수 없다고 답변받음.).
  - 기자 정보가 공동취재단의 경우 해당 기사는 제외함.
  - 번역된 기사는 사용하지 않음(기관 협의).
  - 뉴스 기사의 특성이 전혀 없는 시(詩)나 소설 등 문학 작품은 제외함.

매체명	기자명 정보	사용여부	내용
각 매체	교수	삭제	해당 언론사 소속 기자 이외의 작성자가 쓴 기고문 (교수, 원장, 의사, 대표, 의원, 작가 등)
각 매체	명예기자	삭제	해당 언론사 소속 기자 이외의 작성자가 쓴 기사 (명예기자, 대학생 기자, 시민기자, 학생기자, 어린이기자 등)
각 매체	연합뉴스	삭제	연합뉴스가 출처인 기사, 또는 제공받은 기사
각 매체	공동취재단	삭제	국방부 공동취재단, 올림픽 공동취재단, 대선공동취재단 등 공동취재단의 경우 해당 매체와 계약 등을 일일이 확인 불가.
각 매체	특별취재팀	삭제	대선평별취재팀 등
각 매체	전국종합	삭제	
각 매체	대담	삭제	인터뷰가 아니라 대담임을 밝히고 있는 경우. 국립국어원 답변에 의함
각 매체	전문기자	삭제	분야별 전문가 작성 (에디터, 이코노미스트 등)

OO매체	○T○	삭제	해당 매체 미디어 그룹에 속한 별도의 법인
각 매체	아나운서	삭제	라디오 방송 또는 유튜브 영상을 그대로 옮겨 적음 (아나운서, PD, 프로듀서, 진행, 등)
OO매체	○뉴스팀	삭제	날씨 전문 매체 기사를 제공받은 기사
OO매체	교육○○기자	삭제	해당 매체의 별도 교육주간지
각 매체	리포터	삭제	모집 프리랜서 기자
각 매체	객원기자	삭제	모집 프리랜서 기자
OO매체	헬스○○기자	삭제	해당 매체 미디어 그룹에 속한 별도의 법인
OO매체	○○에듀기자	삭제	해당 매체 미디어 그룹에 속한 별도의 법인
OO매체	더나○○기자	삭제	해당 매체 미디어 그룹에 속한 별도의 법인
OO매체	○○고 기자	삭제	부동산 플랫폼 광고

<표 6> 저작권 이용 문제로 인해 사용하지 않는 기사의 특징

## 4. 데이터 2차 정제

데이터 1차 정제를 마친 기사는 데이터 총괄 관리자가 각 매체별로 에이치티엠엘(HTML) 정보를 활용하여 오류 등을 1차로 수정 및 정제하였다. 최종적으로 작업자가 직접 기사를 읽으며 불필요한 요소를 제거하고, 사용하지 않는 기사들은 불용 표시를 하여 작업을 진행하였다.

### 가. 웹 페이지 데이터 확인

데이터 수집 과정에서 각 매체별 특징 분석이 끝난 후 불필요한 요소를 삭제하거나 사용하지 않는 기사를 표시하는 등, 데이터로만 내용을 파악하여 작업 진행을 할 수 없는 경우가 발생한다. 각 매체별로 다양한 오류들이 존재하기 때문에 웹을 참조하여 작업을 진행하였다.

기사의 중간 소제목과 다음 단락이 붙어 버리는 오류의 경우에는 작업자가 기사를 정독하면서 해당 기사의 유알엘(URL)을 확인하지 않으면 발견하기가 어렵다. 또한 캡션 정보에 아무런 표기가 되어 있지 않다면 캡션 정보를 본문의 일부처럼 인식하여 처리해 버리는 오류도 발생한다. 이런 오류 등은 작업자가 기사를 웹에서 직접 확인하고 해결해야 한다. 웹페이지 데이터에는 한국언론진흥재단으로부터 받은 데이터에는 없는 정보가 표시되어 있어 이를 활용하여 오류를 바로잡을 수 있다.

❖ 중간 제목이 본문 사이에 들어간 경우

웹에서 확인한 실제 기사 내용

다음달 전기요금이 오를 예정인 가운데 가스요금, 대중교통요금 등 각종 공공요금도 인상될 가능성이 커 올해 물가상승률이 2%를 넘어설 것이라는 전망이 나온다. 정부 목표인 올해 연간 소비자물가 상승률 1.8%를 훌쩍 넘어설 것이라는 예상이다.



■ 인상 대기 중인 공공요금

26일 통계청 소비자물가 동향에 따르면 지난 달 물가는 1년 전보다 2.6% 올라 5개월째 2%대 상승률을 보였다. 올 들어 소비자물가 상승률은 1월(0.6%)은 0%대였지만, 4월(2.3%)부터 2%대로 올라섰다. 정부는 하반기부터는 지난해 상반기 국제유가 폭락 등에 따른 기저효과가 완화돼 물가가 안정세를 찾을 것이라고 예

상했지만 빗나갔다. 농·축·수산물 가격이 8월 7.8% 올라 상반기 보다는 둔화했지만 여전히 높고, 외식물가(2.8%)나 외식의 물가(2.7%)도 상승에 힘을 보탤다. 최근에는 서울우유협동조합이 10월부터 우윳값을 올린다고 밝힌 가운데 다른 업체들도 인상에 동참할 전망이다.

한국언론진흥재단으로부터 받은 데이터 내용

다음달 전기요금이 오를 예정인 가운데 가스요금, 대중교통요금 등 각종 공공요금도 인상될 가능성이 커 올해 물가상승률이 2%를 넘어설 것이라는 전망이 나온다. 정부 목표인 올해 연간 소비자물가 상승률 1.8%를 훌쩍 넘어설 것이라는 예상이다. ■ **인상 대기 중인 공공요금** 26일 통계청 소비자물가 동향에 따르면 지난달 물가는 1년 전보다 2.6% 올라 5개월째 2%대 상승률을 보였다. 올 들어 소비자물가 상승률은 1월(0.6%)은 0%대였지만, 4월(2.3%)부터 2%대로 올라섰다. 정부는 하반기부터는 지난해 상반기 국제유가 폭락 등에 따른 기저효과가 완화돼 물가가 안정세를 찾을 것이라고 예상했지만 빗나갔다. 농·축·수산물 가격이 8월 7.8% 올라 상반기 보다는 둔화했지만 여전히 높고, 외식물가(2.8%)나 외식의 물가(2.7%)도 상승에 힘을 보탤다. 최근에는 서울우유협동조합이 10월부터 우윳값을 올린다고 밝힌 가운데 다른 업체들도 인상에 동참할 전망이다.

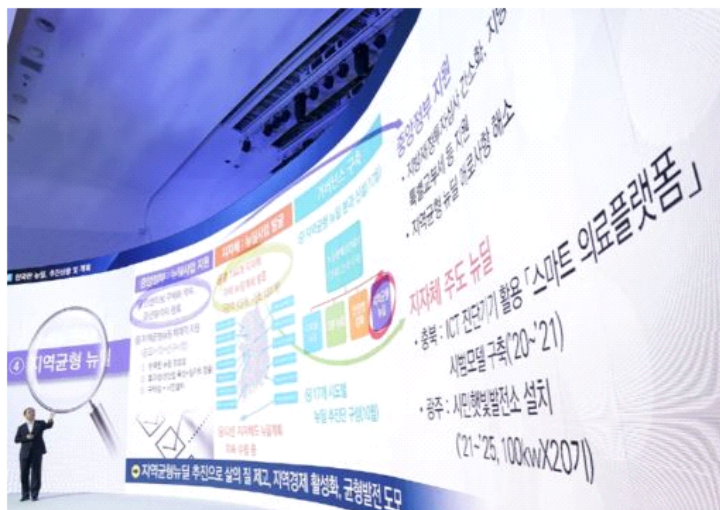
❖ 캡션 정보가 본문과 구분되지 않아 본문처럼 보이는 내용

## 웹에서 확인한 실제 기사 내용

HOME > 정치 > 도정/대통령실

# “한국판 뉴딜-지역균형 뉴딜 성공 위해 추진체계 개편·지방재정

☞ 청와대=이태영 기자 | Ⓞ 승인 2021.01.13 17:38 | 💬 댓글 0



문재인 대통령이 지난해 11월 16일 청와대에서 제3차 한국판 그린뉴딜 전략회의를 개최한 가운데 홍남기 부총리가 지역균형뉴딜과 관련 브리핑을 하고 있다. /청와대 제공

한국판 뉴딜과 지역균형 뉴딜의 성공적인 추진을 위해 무엇보다 추진체계를 개편하고, 지방재정 확충 방안 마련 방안이 최우선 과제로 지적됐다.

## 한국언론진흥재단으로부터 받은 데이터 내용

문재인 대통령이 지난해 11월 16일 청와대에서 제3차 한국판 그린뉴딜 전략회의를 개최한 가운데 홍남기 부총리가 지역균형뉴딜과 관련 브리핑을 하고 있다. /청와대 제공 한국판 뉴딜과 지역균형 뉴딜의 성공적인 추진을 위해 무엇보다 추진체계를 개편하고, 지방재정 확충 방안 마련 방안이 최우선 과제로 지적됐다.

## 나. 불필요한 요소 제거

이 단계는 데이터 정제 2차 작업 중에서 작업자들이, 선택된 기사를 읽어가며 불필요한 요소를 삭제하고 사용하지 않을 기사를 선별하는 공정이다. 데이터 1차 정제를 통해 사용하지 않는 기사를 걸러냈지만 내용을 전부 파악하고 선별한 것이 아니기 때문에 작업자는 내용을 읽으며 사용하지 않고 제외할 기사를 구별하는 표시를 한다. 이 과정에서 불필요한 요소도 확인한다.

### 1) 제외 대상 기사

- ❖ 문장이 도중에 잘렸거나 오류가 많은 기사
- ❖ 기사의 제목을 스크랩한 기사
- ❖ 기사로 볼 수 없는 문장이 나열되는 기사
- ❖ 불필요한 요소를 제거한 뒤 기사 내용이 극히 적은 기사

### 2) 불필요한 본문 내용 삭제

이 공정은 작업자가 직접 기사를 읽어가며 불필요한 요소를 삭제하는 작업이 진행된다. 기사 내에는 표, 그림, 기자 정보, 그래프와 같이 기사와 무관한 정보들이 그대로 남아 있다. 이 정보들은 전체 맥락을 해치므로 제거해 주어야 한다. 또한 연설문, 입장문, SNS 게시글 등 외부 전문이 실린 경우, 기자가 작성한 기사로 보기 어려우므로 신문 기사 말뭉치를 구축하는 본 사업의 목적에 맞지 않다. 따라서 이를 제거하되 다음에 전문이 존재함을 알리는 기사의 문장 또한 기사 텍스트의 완결성에 유의하며 제거한다.

아래 예시의 ‘전문’, ‘문장으로 볼 수 없는 정보’, ‘문장의 오류’ 항목에서 굵은 붉은색 글꼴이 삭제해야 할 대상이다.

삭제 정보	예시	
표, 그림, 그래프 등의 캡션 정보	[사진제공=서울시] 사진제공   크래프톤 사진제공= JTBC 사진/영화공간 주안 제공 [영상=시너지영상팀] (포스터) 일러스트	[표] 2월 연령별 확장실업 표=국가정보원 사진=cctv 캡처 사진· 창원시제공 사진=FNC엔터테인먼트 제공 출처  빅히트뮤직 홈페이지 사진설명)

삭제 정보	예시
기자의 이름, ID 등의 정보	[스포츠서울 용인 = 김○○기자] 2020 도쿄올림픽 남자 축구 클·사진 김○○ 기자 -----@seoul.co.kr 박○○ 기자 -----@hani.co.kr, 사진 고양시 제공
‘Copyright©’ 등 저작권 관련 내용	<저작권자(c) 연합뉴스, 무단 전재-재배포 금지> -----@yna.co.kr/2019-08-29 10:14:05/<저작권자 © 1980-2019 (주)연합뉴스. 무단 전재 재배포 금지.>
전문	<p><b>[서울=뉴스핌] 정○○ 기자 =</b> 일본 정부가 역사왜곡 교과서를 검정 통과시킨 데 대해 교육부가 강한 유감을 나타냈다.</p> <p>교육부는 30일 대변인 설명서를 통해 "일본이 역사 왜곡을 반복하는 교과서를 검정 통과시킨 것에 대해 크게 실망하지 않을 수 없다"며 "독도 영토 주권을 침해하고 강제동원, 일본군 '위안부' 등 전쟁 범죄를 축소·은폐한 교과서를 합격시켰다는 사실에 엄중한 우려를 표명한다"고 규탄했다.</p> <p>이어 "그릇된 역사관이 반영된 초·중·고 교과서로 학습한 일본의 미래세대는 왜곡된 역사관으로 세상을 바라보며 성장할 것이다. 이는 동북아시아의 평화와 공존을 크게 저해할 것이며 일본은 국제 사회로부터 더욱 고립될 수밖에 없다"며 강한 유감을 나타냈다.</p> <p>이날 일본 문부과학성은 극우적 입장에서 기술한 역사왜곡 교과서에 대한 검정을 통과시켰다.</p> <p>이 교과서에는 일본(야마토왜)이 4세기 후반에 한반도 남부지역에 진출해 백제와 신라, 가야 등을 지배했다는 '임나일본부설'을 비롯해 독도를 일본땅이라고 주장하는 역사왜곡이 포함된 것으로 알려졌다.</p> <p>교육부는 관계기관 및 민간·사회단체 등과 함께 일본 역사왜곡 교과서에 대해 적극 대응할 방침이다.</p> <p><b>◆다음은 교육부 대변인 성명 전문이다.</b></p> <p><b>일본은 오늘 고등학교 사회과 교과서 검정 결과를 발표하였다.</b></p> <p><b>대한민국 대통령이 3·1절 기념사에서 일본정부에 '한·일관계의 미래지향적 발전'을 함께 모색하자고 제안한 지 한 달도 채 지나지 않은 시점에, 일본이 역사 왜곡을 반복하는 교과서를 검정 통과시킨 것에 대해 크게 실망하지 않을 수 없다.</b></p> <p><b>그동안 그릇된 역사관이 반영된 일본 교과서 검정 결과가 있을 때마다 대한민국 정부는 일본 정부에 강력히 항의하고, 왜곡 내용의 시정을 촉구하였다. 그러나 이번 고등학교 교과서 검정 발표에도 시정되지 않았다.</b></p> <p><b>대한민국 정부는 독도 영토 주권을 침해하고 강제동원, 일본군 '위안부' 등 전쟁 범죄를 축소·은폐한 고등학교 교과서를 일본 정부가 검정 합격시켰다는 사실에 엄중한 우려를 표명한다.</b></p> <p><b>그릇된 역사관이 반영된 초·중·고 교과서로 학습한 일본의 미래세대는 왜곡된 역사관으로 세상을 바라보며 성장할 것이다. 이는 동북아시아의 평화와 공존</b></p>



삭제 정보	예시
	<p>을 크게 저해할 것이며 일본은 국제 사회로부터 더욱 고립될 수밖에 없다. 과거사에 대한 진정한 사과와 반성은 일본과 일본국민에게 자존심의 상처를 내는 것이 아니라 국제 사회의 당당한 일원으로 돌아올 수 있는 용기 있는 행동임을 인식해야 한다.</p> <p>한·일 관계의 옹졸한 매듭을 푸는 첫걸음은 왜곡된 역사를 바로잡는 것에서 시작된다. 일본 정부는 다음 세대를 위해 왜곡된 교과서 내용을 스스로 시정하라.</p> <p>대한민국 정부는 일본 정부의 영토 주권 침해와 역사 왜곡을 바로 잡기 위해 지속적으로 독도교육을 강화할 것이며, 관계기관, 민간·사회단체 등과 협력하여 적극적으로 대응해 나갈 것임을 분명히 밝힌다.</p> <p>-----@-----.com</p>
문장으로 볼 수 없는 정보	<p>◆정당 지지율...국민의힘 40.5%, 민주당 35.2%</p> <p>정당 지지율은 국민의힘(40.5%)이 민주당(35.2%)을 오차범위 내에서 앞섰다. 지난 5일 국민의힘 대선 최종 후보로 윤석열 전 검찰총장이 선출되자 이른바 ‘컨벤션 효과’가 반영된 것으로 분석된다.</p> <p>국민의힘은 60대 이상(55.5%), 대구·경북(57.5%), 부산·울산·경남(48.6%), 보수(68.2%)에서 지지율이 높았고, 민주당은 30대(45.0%)와 40대(42.9%), 호남권(71.2%), 인천·경기(39.4%) 등에서 상대적으로 지지율이 높았다.</p> <p>◆어떻게 조사했나</p> <p>△조사기관: 한길리서치 △조사의뢰: 아주경제신문 △일시: 2021년 11월 5~7일, 공표 8일 △대상: 전국 만 18세 이상 남녀 1005명 △조사방법:유선 전화 면접 17%, 무선 자동응답시스템(ARS) 83% △응답률: 6.6% △오차 보정 방법: 2021년 8월 말 행정안전부 주민등록 인구 기준(성별·연령별·지역별·가중값 부여) △표본오차: 95% 신뢰수준, ±3.1%포인트 △내용: 20대 대통령선거 등 (중앙선거여론조사심의위원회 홈페이지 참조)</p> <p>이와 관련 국토부는 이달 초 가덕도 신공항을 관문공항으로 만들려면 부산시가 주장하는 7조5000억원이 아닌 28조6000억원 상당의 사업비가 소요되며 시공성과 운영성, 안정성이 떨어진다는 보고서를 국회에 제출했다. (관련기사☞ [단독]국토부 "가덕신공항 막아달라...7.5조 아닌 28.6조원 소요") 이는 단군 이래 최대 토목사업이던 MB정권의 4대강 사업비 22조원보다도 더 많은 액수다.</p> <p>저자는 “암호화폐 코인 시장은 지금까지 존재하던 자산 시장과는 전혀 다른 특징을 가진 신흥 자산 시장”이라며 “24시간 시장이 열리며 전 세계에서 같은 종목을 동시에 사고팔 수 있기에 주류 시장에 진입 후 바뀔 사회는 상상 그 이상”이라고 말했다.</p> <p>◇비트코인 1억 칸다 2=신의두뇌 지음. 솔트앤씨드 펴냄. 340쪽/1만7000원.</p> <p>이날 금통위 시절 금리 인상 소수의견을 내는 등 매파적 성향이 금융위 정책에 영향을 미치는 게 아니냐는 취재진의 질문에 “소수의견은 통화정책 관련해서 소수의견인 것이고, 가계부채 관리 관련해서 거시건정성 정책은 금융위에서 수행해 왔다”고 답했다.</p>

삭제 정보	예시
	<p>이어 “지금 (금융위에서) 여러 가지 정책을 수립했고, 총부채원리금상환비율(DSR) 등 새로 추진해 온 정책들도 있다. 말씀드린 대로 철저하게 관리해 나가겠다”고 덧붙였다.</p> <p>그는 다음 달 종료되는 소상공인 채무 만기 연장과 이자 상환 유예에 대한 3차 연장 여부에 대해서는 “실물경제·방역상황과 밀접하게 관련돼 있다고 생각한다”며 “9월까지니까 좀 더 상황을 보면서 방안을 만들어나가도록 하겠다”고 말했다.</p> <p><b>아래는 한국은행 홈페이지에 소개된 ○ 위원회의 약력이다.</b></p> <p><b>19--년 재무부 국제금융국, 재정경제부 경제정책국 등 근무</b></p> <p><b>20--년 금융감독위원회 감독정책1국 은행감독과장</b></p> <p><b>20--년 금융감독위원회 감독정책1국 감독정책과장</b></p> <p><b>20--년 금융감독위원회 기획행정실장</b></p> <p><b>20--년 금융위원회 금융서비스국장</b></p> <p><b>20--년 금융위원회 금융정책국장</b></p> <p><b>20--년 금융위원회 사무처장</b></p> <p><b>20--년 금융위원회 상임위원</b></p> <p><b>20--년 한국은행 금융통화위원회 위원(금융위원회 위원장 추천)</b></p> <p><b>20--년 한국은행 금융통화위원회 위원(연임, 한국은행 총재 추천)</b></p>
문장의 오류	<p>경기도 화성시는 반입총량(4천551t)의 97.4%인 4천434t을 이미 반입해 <b>반입해</b> 총량 초과가 코 앞이다.</p> <p>이준석 <b>이준석</b> 국민의힘 대표는 3일 페이스북에서 국민의당을 겨냥해 “오픈 플랫폼, 플러스 통합 등 국민들이 알아들을 수 없는 자신들만의 용어로 시간을 끌려고 한다”고 비판했다.</p>

<표 7> 불필요한 요소 제거 내용

문 대통령 '백신 사수' 합동 모의훈련 현장 방문 "빈틈없는 준비"

△ 연합뉴스·이태원 기자 △ 승인 2021.02.25 11:53 △ 댓글 0



문재인 대통령이 이일부터 시작되는 신종 코로나바이러스 감염증(코로나19) 백신 운송과 방역상황 등에 대하여 대대적인 모의훈련에 임하고 있다.

△ 연합뉴스·이태원 기자 △ 승인 2021.02.16 16:49 △ 댓글 0



김승수 전주지사가 16일 오후 전주시청에서 열린 '코로나19 예방접종 지역사회 협업체 기관장 간담회'에 참석해 발언하고 있다. 최기용 수습기자

전주시가 코로나19 종식을 통한 시민들의 일상 복귀를 위해 의료계·경찰·소방 등과 유기적 협력 체계를 구축, 백신 접종 준비에 총력을 기울여 나가기로 했다.

16일 전주시는 김승수 전주시장과 서난이 전주시의회 복지환경위원장, 완산·덕진 경찰서장과 소방서장, 병원장 등 20여 명이 참석한 가운데 '코로나19 백신 접종 지역사회협업체 기관장 회의를 가졌다.

#### 원본 데이터

3일 오전 영종도 인천국제공항에서 신종 코로나바이러스 감염증(코로나19) 백신 운송과 돌발상황 등에 대비한 모의훈련이 진행되고 있다. 문재인 대통령이 이일부터 시작되는 신종 코로나바이러스 감염증(코로나19) 백신 접종을 앞두고 백신 수송 모의훈련을 참관했다.

#### 정제 데이터

문재인 대통령이 이일부터 시작되는 신종 코로나바이러스 감염증(코로나19) 백신 접종을 앞두고 백신 수송 모의훈련을 참관했다.

#### 원본 데이터

김승수 전주지사가 16일 오후 전주시청에서 열린 '코로나19 예방접종 지역사회 협업체 기관장 간담회'에 참석해 발언하고 있다. 최기용 수습기자전주시가 코로나19 종식을 통한 시민들의 일상 복귀를 위해 의료계·경찰·소방 등과 유기적 협력 체계를 구축, 백신 접종 준비에 총력을 기울여 나가기로 했다. 16일 전주시는 김승수 전주시장과 서난이 전주시의회 복지환경위원장, 완산·덕진 경찰서장과 소방서장, 병원장 등 20여 명이 참석한

#### 정제 데이터

전주시가 코로나19 종식을 통한 시민들의 일상 복귀를 위해 의료계·경찰·소방 등과 유기적 협력 체계를 구축, 백신 접종 준비에 총력을 기울여 나가기로 했다. 16일 전주시는 김승수 전주시장과 서난이 전주시의회 복지환경위원장, 완산·덕진 경찰서장과 소방서장, 병원장 등 20여 명이 참석한

<그림 4> 원본 데이터와 정제된 데이터의 예



데이터 정제 전	정제 데이터
<p>[서울경제]  25일 서울 중구 남산스퀘어 KPR에서 열린 제18회 KPR 대학생 PR 아이디어 공모전 시상식에서 대상을 수상한 성균관대 최은호(왼쪽부터), 구연재, 서윤재, 함동규 학생팀이 기념촬영을 하고 있다. KPR 공모전은 국내 최대 규모의 PR 공모전으로, 제18회 공모전은 최근 트렌드를 반영하여 기존 PR기획 부문 외에 영상 부문을 신설, 두 개 부문으로 진행했다. 이번 공모전에는 84개 학교에서 총 1,184명, 333개 팀이 기획서를 접수하는 등 지난 해보다 약 3배나 많은 인원이 참가해 공모전에 대한 높은 관심을 입증했다.</p> <p>(중략)</p> <p>이번 공모전은 코로나19 상황을 고려하여 접수부터 심사까지 전 과정 온라인 비대면 방식으로 진행됐으며, 시상식은 최소 인원으로 간소하게 진행됐다. 김주호 사장은 “최근 커뮤니케이션 트렌드에 발맞춰 이번 대회부터 새롭게 영상부문을 신설했는데 참가자들의 관심이 높았다”며 “젊은 PR인재들이 코로나19 상황에서 기술발달이 가져온 커뮤니케이션의 변화를 볼 수 있는 계기였다”고 밝혔다. /사진제공=KP종합  성○○ 기자  foru82@sedaily.com</p>	<p>KPR 공모전은 국내 최대 규모의 PR 공모전으로, 제18회 공모전은 최근 트렌드를 반영하여 기존 PR기획 부문 외에 영상 부문을 신설, 두 개 부문으로 진행했다. 이번 공모전에는 84개 학교에서 총 1,184명, 333개 팀이 기획서를 접수하는 등 지난 해보다 약 3배나 많은 인원이 참가해 공모전에 대한 높은 관심을 입증했다.</p> <p>(중략)</p> <p>이번 공모전은 코로나19 상황을 고려하여 접수부터 심사까지 전 과정 온라인 비대면 방식으로 진행됐으며, 시상식은 최소 인원으로 간소하게 진행됐다. 김주호 사장은 “최근 커뮤니케이션 트렌드에 발맞춰 이번 대회부터 새롭게 영상부문을 신설했는데 참가자들의 관심이 높았다”며 “젊은 PR인재들이 코로나19 상황에서 기술발달이 가져온 커뮤니케이션의 변화를 볼 수 있는 계기였다”고 밝혔다.</p>

<표 8> 원시 데이터와 정제된 데이터 비교 1

데이터 정제 전	정제 데이터
<p>“자연은 언제나 미술창작의 원천이다. 인간의 생(生)은 자연에서 파생되어 자연과 닮아 있다. 자연으로의 여정을 통해 내면의 시선으로 바라본 심상을 캔버스 위에 표현하고 싶다.”</p> <p>자연과 인간의 공존을 테마로 일상에 무심히 스쳤을 사람, 공간, 시간 등을 그리는 송선희 작가의 ‘자연으로의 여정’展이 10월 1일(금)부터 8일(금)까지 서울신문사 1층 서울신문·서울갤러리 특별전시장에서 열린다.</p> <p>▶ <b>송선희, 비상, 97x138cm, mixed media</b></p> <p>송 작가는 “늦가을의 어느 날 담벼락에 무심히 시들어가는 들꽃을 바라보며, 흡사 인간 삶의 일부와 중첩됨을 느꼈다”며, “자연과의 교감을 통해 조금은 메마른 일상에서 치유받기를 바라는 마음으로 전시를 기획하게 되었다”고 밝혔다.</p> <p>송 작가는 이번 전시에서 총 18점의 유화 및 혼합재료 작품을 선보인다. 작업의 소재는 일상에서 만나면 소소한 감동을 주는 모든 풍경, 자연이다. 그의 작품 속 빛바랜 꽃, 나무, 바다, 파도 등의 피사체는 과거와 현재의 이면 속에 비추어진 ‘작가 자신’의 모습을 형상화 한 것이다.</p> <p>▶ <b>(좌) 송선희, 산책A, 30x30cm, oil on canvas / (우) 송선희, 산책B, 30x30cm, oil on canvas</b></p> <p>그의 작품은 젤 스톤과 모델링 페이스트를 바탕으로 유화와 아크릴 작업을 반복해 완성된다. 작가만의 독특한 마티에르 기법으로 혼합재료를 믹싱하여 사용하는데 이것은 ‘오래된 거친 자연의 질감’을 표현하기 위함이다.</p> <p>송 작가는 “하얀 캔버스 위에 여러 재료를</p>	<p>“자연은 언제나 미술창작의 원천이다. 인간의 생(生)은 자연에서 파생되어 자연과 닮아 있다. 자연으로의 여정을 통해 내면의 시선으로 바라본 심상을 캔버스 위에 표현하고 싶다.”</p> <p>자연과 인간의 공존을 테마로 일상에 무심히 스쳤을 사람, 공간, 시간 등을 그리는 송선희 작가의 ‘자연으로의 여정’展이 10월 1일(금)부터 8일(금)까지 서울신문사 1층 서울신문·서울갤러리 특별전시장에서 열린다.</p> <p>송 작가는 “늦가을의 어느 날 담벼락에 무심히 시들어가는 들꽃을 바라보며, 흡사 인간 삶의 일부와 중첩됨을 느꼈다”며, “자연과의 교감을 통해 조금은 메마른 일상에서 치유받기를 바라는 마음으로 전시를 기획하게 되었다”고 밝혔다.</p> <p>송 작가는 이번 전시에서 총 18점의 유화 및 혼합재료 작품을 선보인다. 작업의 소재는 일상에서 만나면 소소한 감동을 주는 모든 풍경, 자연이다. 그의 작품 속 빛바랜 꽃, 나무, 바다, 파도 등의 피사체는 과거와 현재의 이면 속에 비추어진 ‘작가 자신’의 모습을 형상화 한 것이다.</p> <p>그의 작품은 젤 스톤과 모델링 페이스트를 바탕으로 유화와 아크릴 작업을 반복해 완성된다. 작가만의 독특한 마티에르 기법으로 혼합재료를 믹싱하여 사용하는데 이것은 ‘오래된 거친 자연의 질감’을 표현하기 위함이다.</p> <p>송 작가는 “하얀 캔버스 위에 여러 재료를 중첩하여 시간의 잔상을 표현하고자 노력했다.”며 “중첩된 재료들은 때로는 거칠게, 때로는 스며들듯 부드럽게 발현되어 또 다른 ‘그리움’의 형태로 생성된다.”고 말했다.</p>

중첩하여 시간의 잔상을 표현하고자 노력했다.”며 “중첩된 재료들은 때로는 거칠게, 때로는 스며들듯 부드럽게 발현되어 또 다른 ‘그리움’의 형태로 생성된다.”고 말했다.

▶ 송선희, 침잠의 바다, 60x120cm, oil on canvas

송선희 작가는 4번의 개인전을 개최하였고, ‘공간, 스며들다전’(서경갤러리, 2020년), ‘봄을 보다!’전(P for Y갤러리, 2019년), ‘Saion des inderendants em Coree 2019’ 전 등 다수의 단체전에 참여했다. 현재 전시기획 및 작품에 대한 끊임없는 고찰과 애정으로 신념있는 자신만의 작품활동을 펼치고 있다.

▶ 송선희, 2020장마, 60x120cm, oil on canvas

송선희 작가는 자신의 작품에 대해 “스치듯 지나는 일상의 풍경들과 한 사람의 일생을 기록하듯 그리움의 시선으로 자연의 사계를 돌아보았다. 화려하지는 않지만 누구나의 기억, 추억에 존재하는 풍경을 담기위해 노력했다”며, “이번 전시를 통해 코로나로 암울한 요즘 삭막한 현시대를 살아가는 사람들에게 제 작품이 한편의 위로와 평안을 드렸으면 한다”고 밝혔다.

자세한 전시내용은 서울갤러리 홈페이지([www.seoulgallery.co.kr](http://www.seoulgallery.co.kr))에서 확인할 수 있다. 서울갤러리는 서울신문이 운영하는 미술 전문 플랫폼으로, 다양한 전시를 소개하고 국내 작가들의 작품을 온라인으로 감상할 수 있다.

박○○ -----@seoul.co.kr

송선희 작가는 4번의 개인전을 개최하였고, ‘공간, 스며들다전’(서경갤러리, 2020년), ‘봄을 보다!’전(P for Y갤러리, 2019년), ‘Saion des inderendants em Coree 2019’ 전 등 다수의 단체전에 참여했다. 현재 전시기획 및 작품에 대한 끊임없는 고찰과 애정으로 신념있는 자신만의 작품활동을 펼치고 있다.

송선희 작가는 자신의 작품에 대해 “스치듯 지나는 일상의 풍경들과 한 사람의 일생을 기록하듯 그리움의 시선으로 자연의 사계를 돌아보았다. 화려하지는 않지만 누구나의 기억, 추억에 존재하는 풍경을 담기위해 노력했다”며, “이번 전시를 통해 코로나로 암울한 요즘 삭막한 현시대를 살아가는 사람들에게 제 작품이 한편의 위로와 평안을 드렸으면 한다”고 밝혔다.

자세한 전시내용은 서울갤러리 홈페이지([www.seoulgallery.co.kr](http://www.seoulgallery.co.kr))에서 확인할 수 있다. 서울갤러리는 서울신문이 운영하는 미술 전문 플랫폼으로, 다양한 전시를 소개하고 국내 작가들의 작품을 온라인으로 감상할 수 있다.

<표 9> 원시 데이터와 정제된 데이터 비교 2

데이터 정제 전	정제 데이터
<p>고용노동부가 주최하고 도장애인고용안정 협회(회장:김○○)가 주관한 이번 대회는 지난달 29일 춘천기계공고, 춘천한샘고, 송곡대에서 진행됐다. 총 14개 직종 96명의 선수가 참가해 그동안 갈고닦은 기량을 발휘했다.</p> <p>이날 대회는 PCB설계에 출전한 한○○(철원)씨, 가구제작 부문의 구○○(춘천)씨를 비롯해 14명이 금상 수상의 영예를 안았다. 또 12명이 은상을 차지했으며, 8명이 동상을 받는 등 총 34명이 입상했다. 각 직종별 금상을 차지한 선수들은 올 9월 경북 경주에서 개최되는 '제38회 전국장애인기능경기 대회'에 도 대표로 참가하게 된다. 한편 코로나19 확산 방지 차원에서 개·폐회식 등은 생략했다. 시상은 각 입상자에게 개별 우편발송된다.</p> <p>※금상 수상자 명단=△PCB설계 한○○(철원) △Word Processor 김○○(춘천) △가구제작 구○○(춘천) △그림(수채화+유화) 김○○(동해) △목공예 이○○(양양) △바리스타 김○○(춘천) △시각디자인 조○○(홍천) △양장 홍○○(춘천) △워드프로세서 허○○(양구) △자전거조립 양○○(춘천) △점역교정 한○○(강릉) △컴퓨터수리 김○○(양양) △컴퓨터활용능력 김○○(원주) △화훼장식 황○○(양구)</p> <p>김○○기자</p>	<p>고용노동부가 주최하고 도장애인고용안정 협회(회장:김○○)가 주관한 이번 대회는 지난달 29일 춘천기계공고, 춘천한샘고, 송곡대에서 진행됐다. 총 14개 직종 96명의 선수가 참가해 그동안 갈고닦은 기량을 발휘했다.</p> <p>이날 대회는 PCB설계에 출전한 한○○(철원)씨, 가구제작 부문의 구○○(춘천)씨를 비롯해 14명이 금상 수상의 영예를 안았다. 또 12명이 은상을 차지했으며, 8명이 동상을 받는 등 총 34명이 입상했다. 각 직종별 금상을 차지한 선수들은 올 9월 경북 경주에서 개최되는 '제38회 전국장애인기능경기 대회'에 도 대표로 참가하게 된다. 한편 코로나19 확산 방지 차원에서 개·폐회식 등은 생략했다. 시상은 각 입상자에게 개별 우편발송된다.</p>

<표 10> 원시 데이터와 정제된 데이터 비교 3(본문 기사와 상관없는 내용 삭제)

### 사용하지 않는 기사 예

- 방송 : CBS 라디오 <김현정의 뉴스쇼="">> FM 98.1 (07:20~09:00)
- 진행 : 김현정 앵커
- 대담 : 허석 (전남 순천시 시장)

코로나 때문에 우리가 못 하는 일이 참 많죠. 대표적인 게 5인 이상 모임 금지, 또 밤 9시 이후에 음식점 영업 금지. 어려워도 불가피한 일이다 생각하면서 참고 따르고 있는데, 그런데 이건 어떨까요? 순천시가 전국 최초로 낫술 금지령을 내렸습니다. 새벽 5시부터 오후 4시까지 식당에서 술을 팔면 처벌을 받습니다. 어제부터 시행이 됐는데 찬반 논란이 뜨거워요. 공감하는 의견도 있습니다. 마는 일부 자영업자들 사이에서는 과한 금지령 아니냐, 이런 반발의 목소리가 들립니다. 낫술 금지령을 내린 순천시의 입장 직접 확인을 해 보죠. 허석 순천시장 연결이 돼 있습니다. 시장님 안녕하세요.

◆ 허석> 안녕하세요.

◇ 김현정> 2주간 낫술금지 행정명령. 구체적으로 어떤 겁니까?

◆ 허석> 글자 그대로 낮에 술을 판매하는 것을 금지하는 것인데요. 구체적으로 식당에서 새벽 5시부터 오후 4시까지 주류 판매를 금지하는 것입니다.

◇ 김현정> 순천은 이미 유흥주점은 낮이고 밤이고 영업금지죠?

방송을 그대로 옮겨 적어 문어체 문장의 집합이 아닌 기사는 사용하지 않는 기사로 표기하였다.

<newsitemid>01100611.20211001182808001</newsitemid>  
 <HeadLine>송선희 개인전, '자연으로의 여정' 전 열려</HeadLine>  
 <ByLine></ByLine>  
 <url>http://www.seoul.co.kr/news/newsView.php?id=20211001500171</url>  
 <used>Y</used>  
 <content>  
 "자연은 언제나 미술창작의 원천이다. 인간의 생(生)은 자연에서 파생되어 자연과 닮아 있다. 자연으로의 여정을 통해 내면의 시선으로 바라본 심상을 캔버스 위에 표현하고 싶다."  
 자연과 인간의 공존을 테마로 일상에 무심히 스쳤을 사람, 공간, 시간 등을 그리는 송선희 작가의 '자연으로의 여정'전이 10월 1일(금)부터 8일(금)까지 서울신문사 1층 서울신문·서울갤러리 특별전시장에서 열린다.  
 송선희, 비상. 97x138cm, mixed media  
 송 작가는 '늦가을의 어느 날 담배라에 무심히 시들어가는 들꽃을 바라보며, 흡사 인간 삶의 일부와 중첩됨을 느꼈다'며, "자연과의 교감을 통해 조금은 메마른 일상에서 치유받기를 바라는 마음으로 전시를 기획하게 되었다"고 밝혔다.  
 송 작가는 이번 전시에서 총 18점의 유화 및 혼합재료 작품을 선보인다. 작업의 소재는 일상에서 만나면 소소한 감동을 주는 모든 풍경, 자연이다. 그의 작품 속 빛바랜 꽃, 나무, 바다, 파도 등의 피사체는 과거와 현재의 이면 속에 비추어진 '작가 자신'의 모습을 형상화 한 것이다.  
 송선희, 산책A. 30x30cm, oil on canvas / (우) 송선희, 산책B. 30x30cm, oil on canvas  
 그의 작품은 젤 스펀과 모델링 페이스트를 바탕으로 유화와 아크릴 작업을 반복해 완성된다. 작가만의 독특한 마티에르 기법으로 혼합재료를 믹싱하여 사용하는데 이것은 '오래된 거친 자연의 질감'을 표현하기 위함이다.  
 송 작가는 "하얀 캔버스 위에 여러 재료를 중첩하여 시간의 잔상을 표현하고자 노력했다"며 "중첩된 재료들은 때로는 거칠게, 때로는 스며들듯 부드럽게 발현되어 또 다른 '그라움'의 형태로 생성된다"고 말했다.  
 송선희, 침잠의 바다. 60x120cm, oil on canvas  
 송선희 작가는 4번의 개인전을 개최하였고, '공간, 스며들다전'(서경갤러리, 2020년), '봄을 보다'전(P for Y갤러리, 2019년), 'Saion des inderendants em Coree 2019'전 등 다수의 단체전에 참여했다. 현재 전시기획 및 작품에 대한 끊임없는 고찰과 애정으로 신념있는 자신만의 작품활동을 펼치고 있다.  
 송선희, 2020장마. 60x120cm, oil on canvas  
 송선희 작가는 자신의 작품에 대해 "스치듯 지나가는 일상의 풍경들과 한 사람의 일생을 기록하듯 그리움의 시선으로 자연의 사계를 돌아보았다. 화려하지는 않지만 누구나의 기억, 추억에 존재하는 풍경을 담기 위해 노력했다"며, "이번 전시를 통해 코로나 암울한 요즘 삭막한 현시대를 살아가는 사람들에게 제 작품이 한편의 위로와 평안을 드렸으면 한다"고 밝혔다.  
 자세한 전시내용은 서울갤러리 홈페이지(www.seoulgallery.co.kr)에서 확인할 수 있다. 서울갤러리는 서울신문이 운영하는 미술 전문 플랫폼으로, 다양한 전시를 소개하고 국내 작가들의 작품을 온라인으로 감상할 수 있다.  
 <d>박현갑 eagleduo@seoul.co.kr</d>

<그림 5> 작업 편집 화면

불필요한 요소는 리스트를 활용하여 처리함으로써 확인 요소임을 분명히 하였다. 작업은 수행사가 가지고 있는 프로그램을 사용하였으며 모든 데이터는 기사 단위로 데이터 베이스 관리 시스템(DBMS)에서 처리하였다. 작업자들은 해당 기사를 엑스엠엘(XML) 데이터로 받아 확인 목록을 이용하여 직접 삭제가 아닌 마크업을 부여하는 방식으로 작업하였다. 이때 사용하지 않는 기사는 기사 사용 여부를 나타내는 속성인 'used' 항목에 사용하지 않음을 표기하여 작업하였다.

새로운 유형이 나오는 경우 작업자들이 구글 시트를 활용하여 해당 유형을 공유, 축적하였고 불필요한 요소 제거 작업을 마친 데이터는 소실 비교를 통해 문자 데이터의 누락 등을 검수하였다.

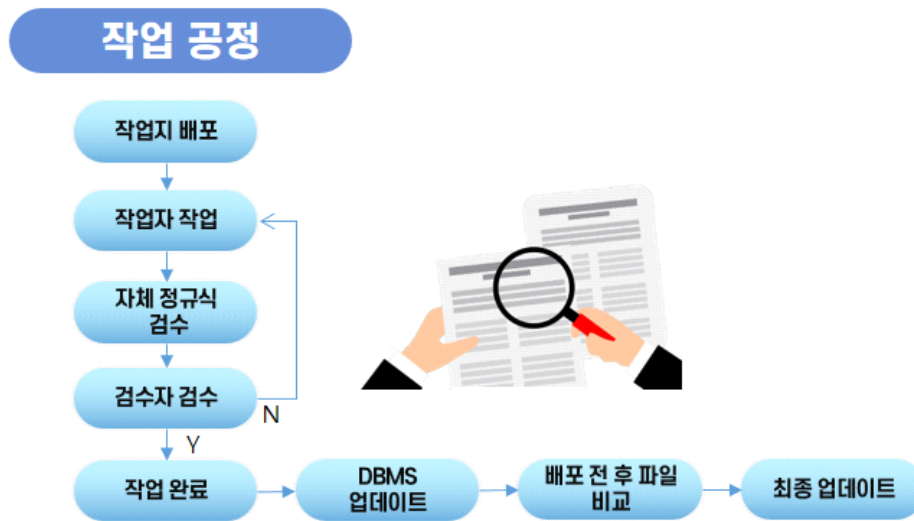
1. 아래와 같이 패턴을 지정
 

^.\*?사진=..\*?\$

^.\*?사진제공=..\*?\$
2. 문자열이 달라도 패턴에 일치하면 빠뜨리지 않고 색상으로 표현해 줌
3. 새로운 유형이 발견되면 DB에서 패턴 규칙을 추가하여 전체를 대상으로 쉽게 확인하고 정제할 수 있음

<그림 6> 작업 프로그램 화면





<그림 7> 데이터 정제 2차 검수 공정

데이터 검수는 할당된 작업을 완료한 후 검수자가 만들어 놓은 패턴을 활용하여 1차로 자체 검수를 실시하였다. 사진, 출처, 전문, 이메일로 끝나는 문장 등 작업 완료된 내용을 작업자 스스로 1차 검수를 진행한 후, 검수 폴더에 업로드하면 검수자가 2차로 해당 파일을 전수 검수하였다.

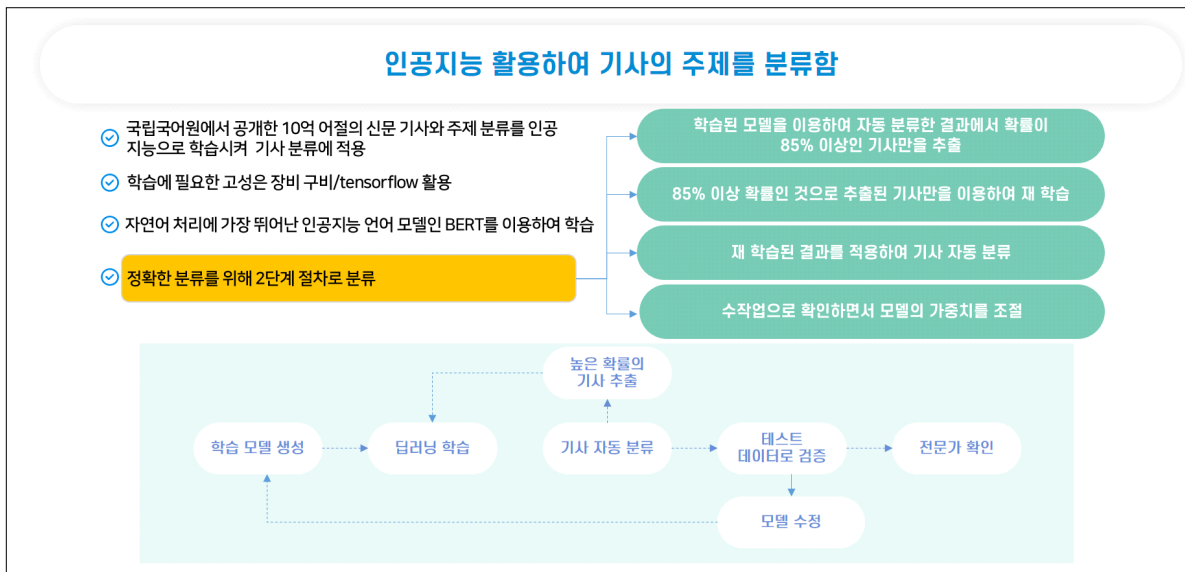
오류 유형을 찾는 패턴은 계속 업데이트되어 작업자들에게 공유되었으며 검수 도중 오류가 많이 발견된 경우에는 파일을 반려한 뒤 오류 유형에 대해 피드백하며 교육을 실시하였다.

## 5. 메타데이터 작성

메타데이터 작성은 기사의 제목, 저자, 발행자, 작성일, 원 주제, 국어원에서 제시한 9가지 주제, 어절 수 등을 작성하는 공정이다.

신문사별로 기사 범주를 분류하게 되는데 이는 매체마다 구분하는 방법도 다르고 범주명도 달라서 메타 정보에는 신문사에서 분류한 원 주제와 함께 이를 통합하여 관리하기 위한, 국립국어원이 제시한 9가지 분류 주제<sup>2)</sup>도 작성된다.

기사 분류는 수행사가 가지고 있는 인공 지능 모델을 신문 기사 학습에 최적화시켜 진행하였으며 기존에 공개된 약 350만 개의 기사와 주제 분류를 학습시켜 정확도 85% 이상인 데이터만을 선별하였다.



<그림 8> 인공 지능을 활용한 주제 분류

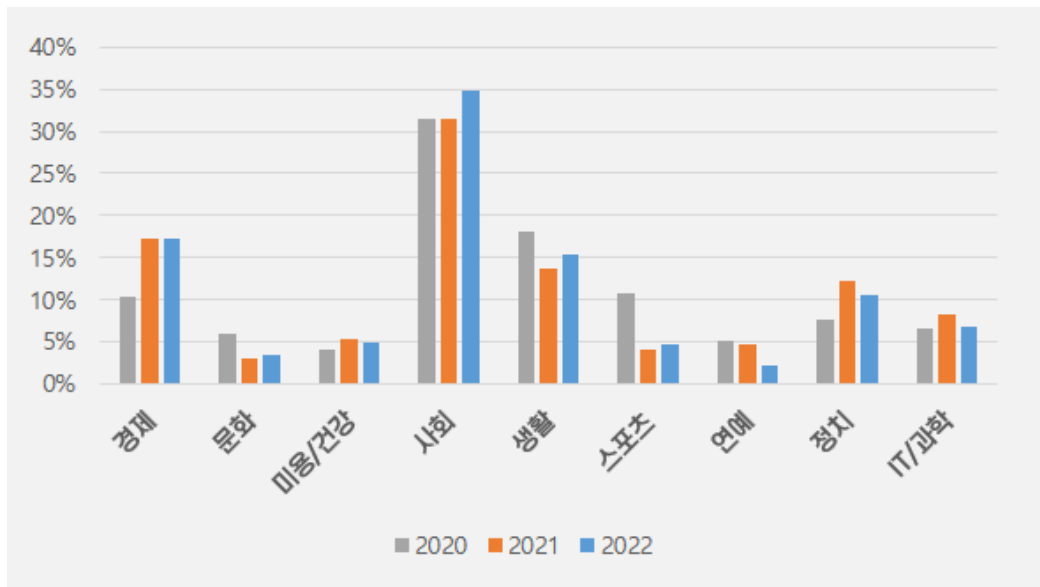
사회	경제	생활	정치	IT/과학	미용/건강	스포츠	문화	연예
34.8%	17.3%	15.3%	10.6%	6.8%	5%	4.6%	3.3%	2.1%

<표 11> 2022년 신문 기사 주제별 통계

기사 수 978,344

2) 정치, 경제, 사회, 생활, IT/과학, 연예, 스포츠, 문화, 미용/건강의 통합 분류 체계로 최종 선정된 기사를 재분류하였다.





<그림 9> 연도별 기사 주제 통계

## 6. 인용 부호 수정 말뭉치

데이터 2차 정제가 끝난 데이터는 ‘신문 말뭉치’ 1종으로, 국립국어원의 요청에 따라 인용 부호 등을 수정하지 않은 데이터로 구축되었다. 수정 전후의 데이터를 활용하여 인공지능에 오류 유형을 학습시키는 효과를 얻을 수 있을 것으로 기대한다. 인용 부호 수정 말뭉치 2종은 기사 내 인용 부호의 통일, 부호의 통일, 오타 수정 등을 거친 데이터로 모두의 말뭉치 공개 대상이다.

### 가. 인용 부호의 통일

최초 제안은 인용 부호의 통일을 진행하여 최종 2종(문단과 문장 단위 2종)의 데이터를 납품하기로 하였으나 기술 협상 과정에서 인용 부호가 통일되지 않은 데이터도 납품을 요청함에 따라 인용 부호 수정 전 말뭉치를 생성하여 총 3종(인용 부호 수정 전 말뭉치, 인용 부호 수정 말뭉치, 문장 분할 말뭉치)의 데이터를 납품하기로 하였다. 현재 대부분의 매체에서는 표준이 아닌 인용 부호로 '(0027)'와 "(0022)"를 사용하는 경우가 많고 실제 데이터에서도 대부분 위와 같은 부호가 사용되었다. 각 매체별 편집기에서 문서를 작성할 때 열고 닫는 인용 부호를 사용하기 번거로워서 간편한 '와 "'를 사용한 것으로 보인다.

인용 부호의 통일 작업은 생각보다 쉽지 않았다. 부호가 열리고 닫히지 않은 기사, 다르게 열고 닫힌 부호, 부호가 통일되지 않은 기사들이 다수 존재하였다.

아래 예시는 부호 짝이 맞지 않은 경우이다. 예시와 같이 짝이 맞지 않은 경우가 상당히 많이 존재하였으며 큰따옴표로 열리고 작은따옴표로 닫힌 경우, 큰따옴표로 열리고 닫히지 않은 경우, 작은따옴표로 열리고 닫히지 않은 경우, 닫는 부호만 있는 경우 등 다수의 사례가 존재하였다. 인용 부호에서만 해당 내용을 수정하였으며, 영어에 등장하는 ‘Apostrophe’의 경우에는 그대로 살려 주었다.

수행사는 데이터 베이스 관리 시스템(DBMS)을 이용하여 부호의 수가 맞지 않는 단락을 찾아내어 패턴 등을 통해 해당 문제를 해결하였다.

엔터 옆 기호를 사용한 부호를 표준 기호에 맞게 수정하였다.

코드	문자	치환 코드	치환 문자	비고
0027	'	2018	‘	여는 내용
0027	'	2019	’	닫는 내용
0022	"	201C	“	여는 내용
0022	"	201D	”	닫는 내용
02B9	/	2019	,	닫는 내용
2032	/	2019	,	닫는 내용
0060	`	2018	‘	여는 내용
02BB	‘	2018	‘	여는 내용
02BC	,	2019	,	닫는 내용
201B	‘	2018	‘	여는 내용
02D9	·	2018	‘	여는 내용
FF07	'	2019	,	닫는 내용
2033	”	201D	”	닫는 내용
02DD	”	201D	”	닫는 내용

<표 12> 인용 부호 치환 표

데이터 정제 전	데이터 정제 후
<p>시의회 한 관계자는 “시의회가 오 시장과의 관계를 설정하고 새로운 방향을 모색하는 그런 임시회가 될 것”이라며 “특히 다수당을 차지하고 있는 민주당 시의원들과 오 시장과의 기싸움이 있을 것”이라고 전망했다.</p> <p>(중략)</p> <p>시의회 다른 관계자는 “민주당 시의원들이 다수의 힘으로 밀어붙인다면 또 다른 횡포로 비칠 수 있다”고 우려했다.</p>	<p>시의회 한 관계자는 “시의회가 오 시장과의 관계를 설정하고 새로운 방향을 모색하는 그런 임시회가 될 것”이라며 “특히 다수당을 차지하고 있는 민주당 시의원들과 오 시장과의 기싸움이 있을 것”이라고 전망했다.</p> <p>(중략)</p> <p>시의회 다른 관계자는 “민주당 시의원들이 다수의 힘으로 밀어붙인다면 또 다른 횡포로 비칠 수 있다”고 우려했다.</p>

<표 13> 인용 부호 수정 데이터 정제 전 후

데이터 정제 전	데이터 정제 후
<p>한국투명성기구 부산본부 황○○ 상임대표는 “제보자 신원 보호를 철저히 하기 위해 담당 공무원의 사전 교육이 필요하다”라고 말했고 국민권익위원회 임○○ 신고자보호과장은 “포상금은 당사자의 신청이 필요한 보상금과 달리 부서추천으로 가능한 만큼 각 부서단위의 신고성민원에 대한 처분등이 부서 포상추천으로 이어질 수 있도록 적극적인 발굴이 필요하다”고 지적했다.</p> <p>부산시민운동지원센터 서○○ 팀장과 부산시민재단 방성애 사무국장은 시민들이 쉽게 공익제보를 할 수 있는 방안을 마련하고 시민단체와의 협력 네트워크 구축이 필요하다”고 강조했다.</p> <p>부산시 류○○ 감사위원장은 “위원회에서 제시해주신 의견을 바탕으로 심의의결된 활성화 계획을 책임있게 이행해 부산시 공익제보가 안착될 수 있도록 최선을 다하겠다”고 전했다.</p>	<p>한국투명성기구 부산본부 황○○ 상임대표는 “제보자 신원 보호를 철저히 하기 위해 담당 공무원의 사전 교육이 필요하다”라고 말했고 국민권익위원회 임○○ 신고자보호과장은 “포상금은 당사자의 신청이 필요한 보상금과 달리 부서추천으로 가능한 만큼 각 부서단위의 신고성민원에 대한 처분등이 부서 포상추천으로 이어질 수 있도록 적극적인 발굴이 필요하다”고 지적했다.</p> <p>부산시민운동지원센터 서○○ 팀장과 부산시민재단 방성애 사무국장은 “시민들이 쉽게 공익제보를 할 수 있는 방안을 마련하고 시민단체와의 협력 네트워크 구축이 필요하다”고 강조했다.</p> <p>부산시 류○○ 감사위원장은 “위원회에서 제시해주신 의견을 바탕으로 심의의결된 활성화 계획을 책임있게 이행해 부산시 공익제보가 안착될 수 있도록 최선을 다하겠다”고 전했다.</p>

<표 14> 인용 부호 데이터 정제 전후 2

매체명	기사 수	어절 수	매체명	기사 수	어절 수
강원일보	11,257	1,963,555	서울경제	81,698	17,850,080
경기일보	16,824	3,247,859	서울신문	28,702	7,432,116
경북일보	9,687	2,004,032	스포츠서울	36,780	7,831,088
경인일보	12,721	2,768,395	아시아경제	81,389	16,828,008
국민일보	36,180	8,427,331	아주경제	62,383	15,101,136
기호일보	23,686	4,001,184	이데일리	42,015	9,709,286
남도일보	21,109	3,825,611	이투데이	36,114	8,157,622
내일신문	11,572	2,579,348	전남일보	12,984	2,416,334
노컷뉴스	42,898	7,855,828	전북도민일보	20,929	3,456,314
뉴스핌	38,035	7,530,246	조선일보	13,593	3,604,488
대구신문	18,531	3,676,483	중도일보	18,550	3,449,634
대전일보	24,302	4,309,993	충북일보	15,331	2,708,803
동양일보	5,756	947,393	충청일보	22,641	3,706,618
매일신문	28,902	5,803,478	충청투데이	5,410	1,247,005
머니투데이	31,288	8,202,371	한겨레	21,912	5,618,983
미디어오늘	2,270	924,323	한국일보	13,297	3,004,110
부산일보	43,346	9,163,624	헤럴드경제	86,252	18,968,233
총 합				978,344	208,320,912

<표 15> 최종 선정 기사 수

1차 데이터 정제와 2차 데이터 정제를 통해 도출된 기사와 어절 수는 위와 같다.

## 나. 한·중·일 호환용 한자 영역(F900-FAFF) 한자의 통일

인공 지능 학습과 데이터 유통에 있어 통일되지 않은 코드는 크고 작은 문제를 일으킬 수 있다. 따라서 불필요한 요소를 제거한 후 해당 한·중·일 호환용 한자 영역에 대한 코드 통일을 진행하였다. 해당 데이터에서는 아래와 같은 ‘한·중·일 호환용 한자 영역’의 한자가 등장하였다.

데이터의 통일성을 위해 해당 한자의 표준 유니코드를 통일하는 작업을 진행하였다. 李(이, UF9E1), 李(리, U674E) 등 완전히 동일한 의미이면서 문자 코드가 다른 한자는 모두 통일시켜 주었다.

최종 선정된 기사에서 사용된 ‘한·중·일 호환용 한자 영역’의 한자 수는 아래와 같다. 약 3000건의 한자가 표준 유니코드로 통일되었다.

코드	한자	사용횟수	코드	한자	사용횟수	코드	한자	사용횟수
F9E1	李	600	F9BD	尿	26	F937	路	13
F90A	金	266	F9DA	栗	24	F95F	寧	13
F967	不	186	F980	呂	23	F98A	力	12
F978	兩	180	F94C	樓	21	F9DD	利	12
F981	女	131	F9B6	禮	21	F972	沈	11
F95C	樂	129	F90F	羅	20	F97A	梁	11
F934	老	105	F9E4	理	20	F99C	列	11
F9FE	茶	86	F9CA	流	19	F9D3	陸	11
F92F	勞	60	F9E3	泥	19	F91B	亂	10
F9AE	瑩	51	F9F6	臨	19	F92E	冷	10
F9C4	龍	51	F9B3	靈	18	F940	鹿	10
F98E	年	49	F914	樂	16	F961	率	9
F9C7	劉	47	F918	落	16	F97E	量	9
F9F7	立	44	F9EA	離	16	F983	旅	9
F99A	連	43	F9FA	狀	16	F9C9	柳	9
F933	盧	41	F9AA	寧	15	F901	更	8
F997	聯	41	F9B5	例	14	F938	露	8
F9F4	林	39	F9C3	遼	14	F93D	綠	8
F9D1	六	28	F9E0	易	14	F990	戀	8

<표 16> 최종 선정 기사 ‘한·중·일 호환용 한자 영역’의 한자 수(이하 생략)

❖ 기존 ‘한·중·일 호환용 한자 영역’의 한자는 아래 표의 정보로 치환함.

코드	한자	치환	코드	한자	치환	코드	한자	치환	코드	한자	치환
F978	兩	5169	F9F3	麟	9E9F	F91C	卵	5375	F9A1	說	8AAA
F90A	金	91D1	F98C	歷	6B77	F92A	浪	6D6A	F9AA	寧	5BE7
F967	不	4E0D	F9E1	李	674E	F94F	累	7D2F	F9CE	硫	786B
F981	女	5973	FA02	拓	62D3	F97C	良	826F	F9F7	立	7ACB
F95C	樂	6A02	F9D7	輪	8F2A	F983	旅	65C5	FA04	宅	5B85
F92F	勞	52DE	F9B0	聆	8046	F90E	癩	7669	F996	練	7DF4
F934	老	8001	F9B4	領	9818	F922	濫	6FEB	F9A8	令	4EE4
F933	盧	76E7	F9B3	靈	9748	F937	路	8DEF	F9B5	例	4F8B
F91B	亂	4E82	F9A0	裂	88C2	F939	魯	9B6F	F9B9	惡	60E1
F941	論	8AD6	F9C2	蓼	84FC	F93C	祿	797F	F9BA	了	4E86
F93D	綠	7DA0	F9BD	尿	5C3F	F95F	寧	5BE7	F9D8	律	5F8B
F97E	量	91CF	F9FA	狀	72C0	F966	復	5FA9	F9E0	易	6613
F914	樂	6A02	F99A	連	9023	F905	串	4E32	F989	黎	9ECE
F91F	蘭	862D	F9A3	念	5FF5	F912	裸	88F8	F999	蓮	84EE
F94C	樓	6A13	F9CA	流	6D41	F915	洛	6D1B	F99B	鍊	934A
F902	車	8ECA	F988	麗	9E97	F916	烙	70D9	F99C	列	5217
F940	鹿	9E7F	F9C1	療	7642	F91A	駱	99F1	F99F	烈	70C8
F90F	羅	7F85	F997	聯	806F	F91D	欄	6B04	F9A2	廉	5EC9
F92E	冷	51B7	F9AE	瑩	7469	F949	雷	96F7	F9C9	柳	67F3
F972	沈	6C88	F9E7	裏	88CF	F955	凌	51CC	F9D1	六	516D
F92D	來	4F86	F9AB	嶺	5DBA	F976	略	7565	F9F1	隣	96A3
F97A	梁	6881	F9F6	臨	81E8	F90D	懶	61F6	F990	戀	6200
F918	落	843D	F99D	劣	52A3	F923	藍	85CD	F9A9	囹	56F9
F932	爐	7210	F9B2	零	96F6	F942	壘	58DF	F9C3	遼	907C
F984	濾	6FFE	FA06	暴	66B4	F943	弄	5F04	F9C4	龍	9F8D
F973	拾	62FE	F9E9	里	91CC	F94E	漏	6F0F	F9CD	留	7559
F980	呂	5442	F9FE	茶	8336	F960	怒	6012	F9DA	栗	6817
F901	更	66F4	F987	驪	9A6A	F962	異	7570	F9DD	利	5229
F907	龜	9F9C	F98A	力	529B	F965	便	4FBF	F9DE	吏	540F
F938	露	9732	F9B6	禮	79AE	F96D	省	7701	F9E3	泥	6CE5
F945	龔	807E	F9C7	劉	5289	F974	若	82E5	F9EA	離	96E2
F90C	奈	5948	F98E	年	5E74	F975	掠	63A0	F9EE	燐	71D0
F961	率	7387	F9DB	率	7387	F979	涼	51C9	F9F4	林	6797
F96B	參	53C3	F9E2	梨	68A8	F985	礪	792A	FA08	行	884C
F986	閭	95AD									

<표 17> ‘한·중·일 호환용 한자 영역’ 한자 치환 표

## 다. 문장 부호 등 통일

신문 기사 내에는 일관성 없이 사용된 문자 등이 있어 인공 지능 학습에 나쁜 영향을 준다.

‘A, B, C, a 등’ 전각 알파벳, ‘[, ?, @, ;, (, ', & 등’ 전각 부호, ‘0, 1, 2 등’ 전각 숫자는 데이터의 일관성 및 정보 처리 효율성을 위해 모두 반각 문자로 치환하였다.

가운뎃점도 ‘·(MIDDLE DOT)’는 ‘·(318D), ·(22C5), ·(30FB), •(2219), •(2022), ·(0387), ·(1427), ·(2024), ·(2027), •(2981), ·(FF65) 등’과 같이 다양하게 쓰이고 있어 ‘·(00B7)’로 치환하였다.

대상 코드	대상 문자	대상 코드	치환 문자	비고
FF01	!	0021	!	
FF07	'	0027	'	
FF02	"	0022	"	
FF03	#	0023	#	
FF0A	*	002A	*	
FF0B	+	002B	+	
FF0C	,	002C	,	
FF0D	—	002D	—	
FF0E	.	002E	.	
FF0F	/	002F	/	
FF10	0	0030	0	
FF11	1	0031	1	
FF12	2	0032	2	
FF13	3	0033	3	
FF14	4	0034	4	
FF15	5	0035	5	
FF16	6	0036	6	
FF17	7	0037	7	
FF18	8	0038	8	
FF19	9	0039	9	
FF1B	;	003B	;	
FF1C	<	3008	<	
FF1D	=	003D	=	
FF1E	>	3009	>	
FF3F	—	005F	—	
FF5E	~	007E	~	

대상 코드	대상 문자	대상 코드	치환 문자	비고
FF65	·	00B7	·	
FFE5	₩	00A5	₩	
FFE6	₩	20A9	₩	
FFEB	→	2192	→	
FF62	「	300C	「	
FF63	」	300D	」	
3000		0020		공백
0009		0020		공백
00a0		0020		공백
2002		0020		공백
2003		0020		공백
2009		0020		공백
318D	·	00B7	·	
22C5	·	00B7	·	
30FB	·	00B7	·	
2219	•	00B7	·	
2022	●	00B7	·	
0387	·	00B7	·	
1427	·	00B7	·	
2024	·	00B7	·	
2027	·	00B7	·	
2981	•	00B7	·	
FF65	·	00B7	·	

<표 18> 치환 코드 리스트



## 라. 오타 후보 문자 수정

정제가 완료된 인용 부호 수정 말뭉치에서 사용된 글자를 전부 조회하여 오타 확률이 높은 글자를 추출한다. 그리고 해당 글자의 기사 내 쓰임을 확인하고 수정하는 방식으로 진행하였다.

오류 후보 글자	해당 내용	교정 내용
옴	낚시하는 6시간 내내 배에서는 여기 "왔어요" 라는 얘기를 계속해서 들었고 뜰채에 담아 물칸에 <b>옴기는</b> 선장의 발걸음이 바빴다.	낚시하는 6시간 내내 배에서는 여기 "왔어요" 라는 얘기를 계속해서 들었고 뜰채에 담아 물칸에 <b>옴기는</b> 선장의 발걸음이 바빴다.
권	특수학교에 모두 보·차도 분리를 추진하는 계획을 세웠지만, 사립유치원은 대상에서 빠져 있다. <b>권련법상</b> ‘사립학교’인 사립유치원에 재정 지원은 어렵다는 이유다.	특수학교에 모두 보·차도 분리를 추진하는 계획을 세웠지만, 사립유치원은 대상에서 빠져 있다. <b>관련법상</b> ‘사립학교’인 사립유치원에 재정 지원은 어렵다는 이유다.
환	현재 상류시설로 <b>분류된</b> 것을 물류시설로 전환하는 물류단지개발지침 개정, 물류단지 조성사업 내 기반시설 국비지원(약 2천600억원) 등도 요청했다.	현재 상류시설로 <b>분류된</b> 것을 물류시설로 전환하는 물류단지개발지침 개정, 물류단지 조성사업 내 기반시설 국비지원(약 2천600억원) 등도 요청했다.
장	에스24에서도 올 상반기 경제·경영 분야 도서의 전년 대비 판매 <b>성장률</b> 이 전 분야를 통틀어 가장 높은 52.2%로 나타났다.	에스24에서도 올 상반기 경제·경영 분야 도서의 전년 대비 판매 <b>성장률</b> 이 전 분야를 통틀어 가장 높은 52.2%로 나타났다.
경	그는 이번 판결이 1965년의 한일 <b>경구권협정과</b> 2015년의 한일 외교 장관 간 ‘위안부 합의’에도 어긋난다고 주장했다.	그는 이번 판결이 1965년의 한일 <b>청구권협정과</b> 2015년의 한일 외교 장관 간 ‘위안부 합의’에도 어긋난다고 주장했다.
체	안동시체육회와 <b>읍면동체육회장협의회</b> 는 28일 논의를 통해 하반기 개최	안동시체육회와 <b>읍면동체육회장협의</b> 회는 28일 논의를 통해 하반기 개최

오류 후보 글자	해당 내용	교정 내용
	예정이던 '제61회 안동시민체육대축전'을 취소하기로 결정했다.	예정이던 '제61회 안동시민체육대축전'을 취소하기로 결정했다.
되	구체적으로 개발행위허가가 <b>제한되는</b> 지역은 SRT 평택지제역 서쪽 268만여㎡와 서해선 안중역 반경 약 1km 이내인 518만여㎡ 등이다.	구체적으로 개발행위허가가 <b>제한되는</b> 지역은 SRT 평택지제역 서쪽 268만여㎡와 서해선 안중역 반경 약 1km 이내인 518만여㎡ 등이다.
좌	변이 바이러스가 확산한 오사카부에선 역대 <b>최다인</b> 1220명의 신규 확진자가 쏟아지며 20일 연속으로 도쿄보다 많은 확진자가 나왔다.	변이 바이러스가 확산한 오사카부에선 역대 <b>최다인</b> 1220명의 신규 확진자가 쏟아지며 20일 연속으로 도쿄보다 많은 확진자가 나왔다.
활	현재 속초 생활치료센터에는 3개, 고성 <b>생활치료센터에는</b> 67개의 병상 여유가 있다.	현재 속초 생활치료센터에는 3개, 고성 <b>생활치료센터에는</b> 67개의 병상 여유가 있다.
맞	벨기에 국왕내외의 국민방문에 <b>맞춰</b> 열린 1회 '남북 글로벌 해양프로젝트'에 이어 열리는 이번 2회 국제심포지엄은 우리나라 갯벌의 가치를 재발견하고	벨기에 국왕내외의 국민방문에 <b>맞춰</b> 열린 1회 '남북 글로벌 해양프로젝트'에 이어 열리는 이번 2회 국제심포지엄은 우리나라 갯벌의 가치를 재발견하고
방	이 최첨단 <b>양방향</b> 혈관조영장비는 고해상도의 실시간 영상 화질을 지원해 보다 세밀한 영상을 구현하고, 시술 상황에 맞춘	이 최첨단 <b>양방향</b> 혈관조영장비는 고해상도의 실시간 영상 화질을 지원해 보다 세밀한 영상을 구현하고, 시술 상황에 맞춘
젊	디스크 퇴행이 있는 사람들을 비교했을 때 코어 근육이 안 좋더라는 연구 결과는 있지만, 코어 운동으로 근육을 강화시켜도 퇴행된 디스크가 다시 <b>젊어지지</b> 는 않다는 것이다.	디스크 퇴행이 있는 사람들을 비교했을 때 코어 근육이 안 좋더라는 연구 결과는 있지만, 코어 운동으로 근육을 강화시켜도 퇴행된 디스크가 다시 <b>젊어지지</b> 는 않다는 것이다.

<표 19> 오타 후보 목록 글자 수정 전후

## 7. 문장 말뭉치 구축

대부분의 자연어 처리 분야에서 인공 지능을 활용한 학습은 문장을 기본 단위로 하고 있다. 특히 형태소 분석과 기계 번역은 대부분 문장을 기본 단위로 한다. 그렇기 때문에 단락을 정확한 문장 단위로 분할하는 것은 중요한 의미를 지닌다.

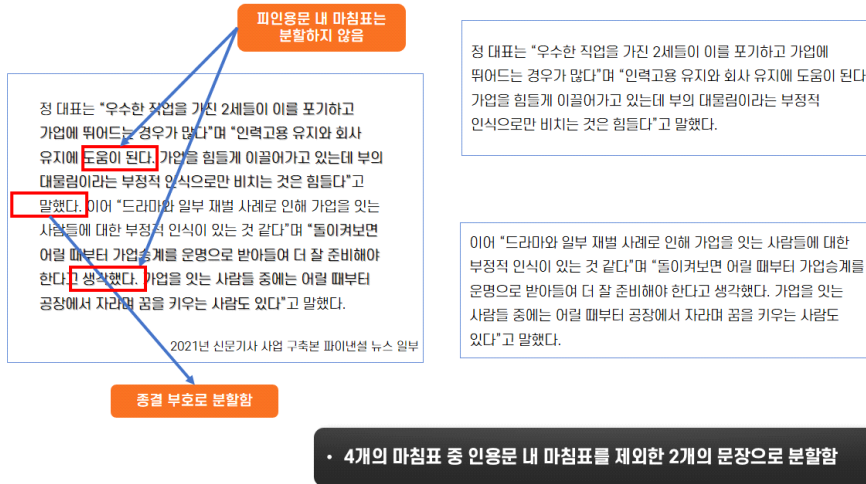
현재 문장 분할은 대부분 사람이 직접 작업하는 방식으로 진행하고 있다. 문장을 기계적으로 분할할 때 예외의 변수가 많아 자동 분할된 문장을 수정하는 데 오히려 시간이 더 소요되기 때문이다. 이번 구축을 통해 단락과 문장으로 구성된 말뭉치 세트가 갖춰지면 인공 지능을 이용하여 문장 자동 분할에 활용할 수 있을 것이다.

하나의 문장은 보통 마침표(.), 느낌표(!), 물음표(?) 등의 문장 부호를 기본 단위로 한다. 그러나 문장이 끝나는 부분이 아닌 곳에 사용되는 예가 많기 때문에 반드시 예외 처리를 해주어야 한다. 예컨대 피인용문 내부에 사용된 마침표(.), 느낌표(!), 물음표(?) 등의 문장 부호에서는 분할하지 않아야 하기 때문이다.

문장 말뭉치는 인용 부호 수정 말뭉치와 함께 활용할 수 있도록 하기 위해 단락 구분을 표시하는 ‘<p>’ 태그에 더해 ‘<s>’ 태그를 삽입하여 문장을 구분하였다.

### 가. 문장 분할

- ❖ 문장의 분할은 수행사가 가지고 있는 문장 분할 프로그램을 이용하여 진행함.
- ❖ 하나의 문장은 보통 마침표(.), 느낌표(!), 물음표(?) 등의 문장 부호를 기본 단위로 함.
- ❖ 자동으로 문장을 분할하면 반드시 그 결과를 다시 확인하는 검수 절차를 진행함.
- ❖ 피인용문 내 마침표(.), 느낌표(!), 물음표(?) 등의 문장 부호에서는 분할하지 않음.



<그림 10> 문장 말뭉치 개념

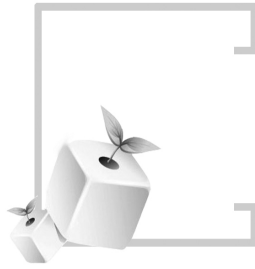
인용 부호 수정 말뭉치	문장 말뭉치
이에 곧장 라이올라가 전면 반박에 나섰다. 그 역시 '아스'를 통해 "명백히 잘못된 가짜 뉴스다. 저도 그렇고 홀란드도 단 한 명의 바르셀로나 회장 후보와 이야기 나눈 적이 없다. 홀란드를 제외한 나머지 선수들과 연계해서도 논의된 바 없다"고 홀란드의 바르셀로나 이적설을 강력하게 부인했다. 그러면서 "바르셀로나 새 회장이 되면 나와 통화할 수는 있겠으나, 지금은 전혀 이야기된 것이 없다"고 강조했다.	이에 곧장 라이올라가 전면 반박에 나섰다.  그 역시 '아스'를 통해 "명백히 잘못된 가짜 뉴스다. 저도 그렇고 홀란드도 단 한 명의 바르셀로나 회장 후보와 이야기 나눈 적이 없다. 홀란드를 제외한 나머지 선수들과 연계해서도 논의된 바 없다"고 홀란드의 바르셀로나 이적설을 강력하게 부인했다.  그러면서 "바르셀로나 새 회장이 되면 나와 통화할 수는 있겠으나, 지금은 전혀 이야기된 것이 없다"고 강조했다.

<표 20> 문장 말뭉치 데이터 정제 예

상위 5 문장분할수			하위 5 문장분할수		
1	조선일보	2.8	1	충청일보	1.1
2	한겨레	2.5	2	충북일보	1.1
3	한국일보	2.2	3	전국도민일보	1.1
4	서울경제	2.0	4	동양일보	1.1
5	미디어오늘	1.9	5	경북일보	1.2

<그림 11> 문단 내 문장 분할 수(상/하위 5개 매체)

<그림 11>은 한 문단(매체가 나눈 단락) 내에서 수행사가 문장으로 나눈 평균 분할 수이다. 조선일보의 경우 문장 분할 평균이 한 단락당 2.8개로 가장 높았으며, 충청일보의 경우 1.1개로 가장 낮은 수치를 보였다. 평균 문장 분할 수는 1.6개이다.



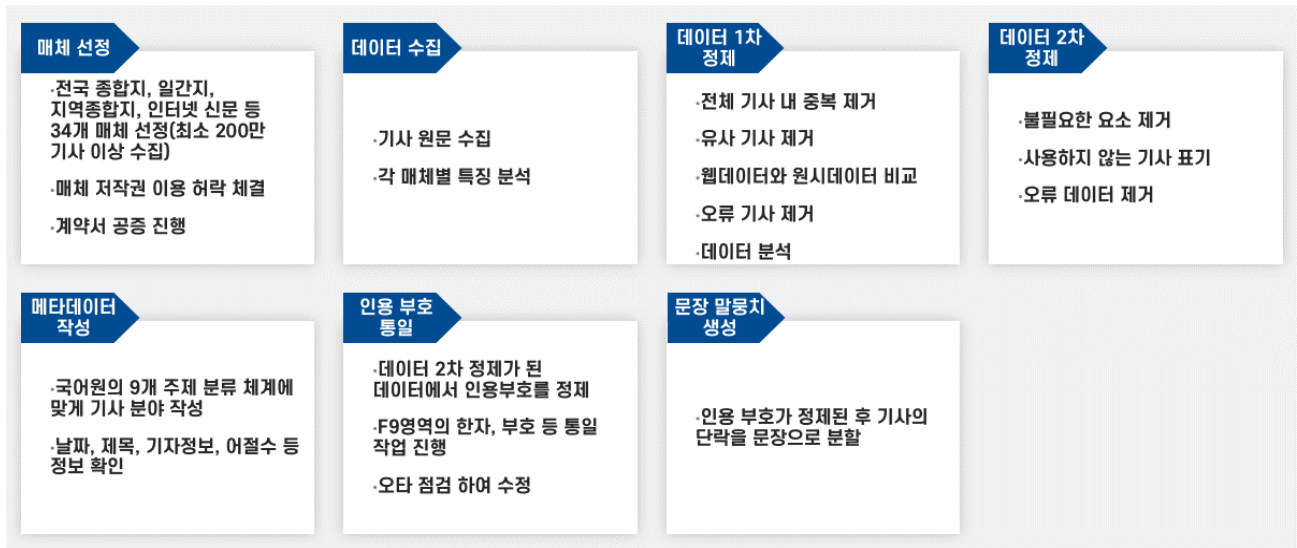
## 제 3 장

# 사업 수행 결과



## 제 3장 사업 수행 결과

### 1. 신문 기사 정제 결과



<그림 12> 구축 공정별 내용

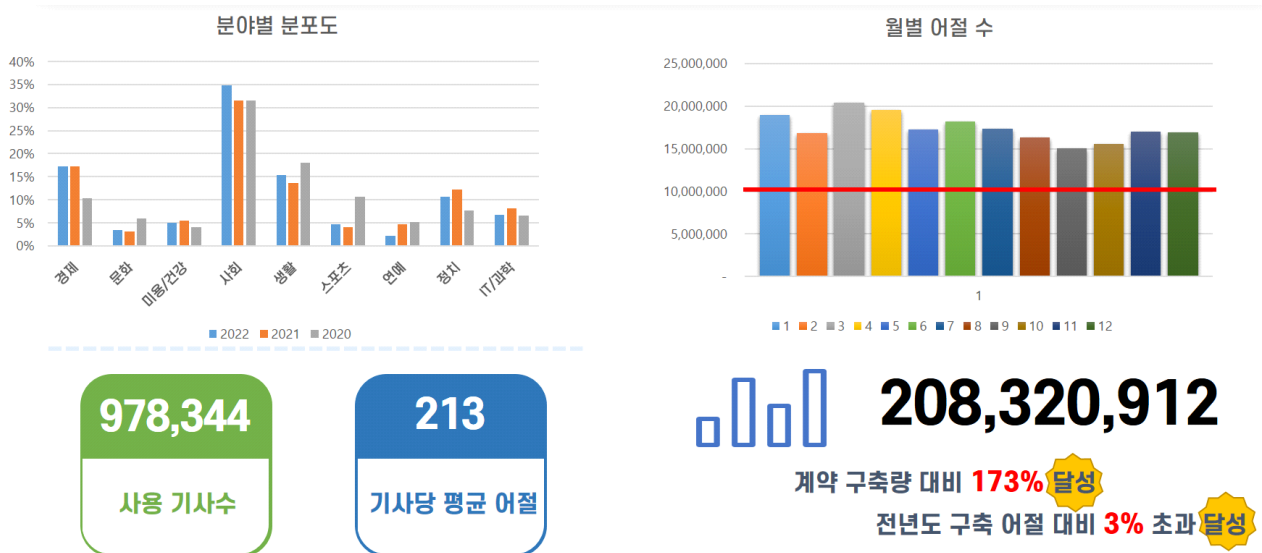
본 사업은 매체 선정부터 문장 말뭉치 작업까지 총 7단계의 과정을 거쳐 수행되었다. 정제가 완료된 원시 말뭉치는 978,344건의 기사와 208,320,912개의 어절로 구축되었으며, 가장 많은 기사와 어절 수를 차지한 매체는 헤럴드경제였고, 미디어오늘은 가장 적은 기사와 어절 수를 보였다.

매체명	최초 수집 기사 수	최초 수집 어절 수	정제 수집 기사 수	정제 수집 어절 수
강원일보	37,910	4,118,201	11,257	1,963,555
경기일보	29,741	5,693,024	16,824	3,247,859
경북일보	24,950	5,052,512	9,687	2,004,032
경인일보	35,522	6,549,123	12,721	2,768,395
국민일보	98,630	21,841,538	36,180	8,427,331
기호일보	46,919	7,397,577	23,686	4,001,184
남도일보	31,543	6,080,505	21,109	3,825,611
내일신문	27,278	7,237,421	11,572	2,579,348
노컷뉴스	142,024	25,841,804	42,898	7,855,828
뉴스핌	270,352	35,191,030	38,035	7,530,246
대구신문	31,012	6,024,116	18,531	3,676,483
대전일보	47,114	7,793,680	24,302	4,309,993
동양일보	31,639	4,442,671	5,756	947,393
매일신문	57,870	11,265,352	28,902	5,803,478
머니투데이	157,560	35,552,827	31,288	8,202,371
미디어오늘	5,012	2,701,758	2,270	924,323
부산일보	87,073	17,275,077	43,346	9,163,624
서울경제	148,238	31,538,120	81,698	17,850,080
서울신문	110,873	24,241,181	28,702	7,432,116
스포츠서울	79,877	12,756,461	36,780	7,831,088
아시아경제	193,988	36,561,962	81,389	16,828,008
아주경제	102,500	24,500,654	62,383	15,101,136
이데일리	190,325	33,177,908	42,015	9,709,286
이투데이	104,123	20,518,842	36,114	8,157,622
전남일보	27,772	5,533,794	12,984	2,416,334
전북도민일보	36,908	5,929,208	20,929	3,456,314
조선일보	50,305	11,150,683	13,593	3,604,488
중도일보	51,224	9,308,780	18,550	3,449,634
충북일보	33,045	4,874,304	15,331	2,708,803
충청일보	64,405	9,289,659	22,641	3,706,618
충청투데이	26,687	4,920,738	5,410	1,247,005
한겨레	44,529	14,322,341	21,912	5,618,983
한국일보	81,404	21,465,509	13,297	3,004,110
헤럴드경제	148,116	31,236,432	86,252	18,968,233
<b>총 합</b>	<b>2,656,468</b>	<b>511,384,792</b>	<b>978,344</b>	<b>208,320,912</b>

<표 21> 신문 기사 정제 총괄표



'22년 신문 말뭉치는 월별 1,000만 어절 이상의 데이터를 구축해야 하는 목표를 초과 달성하였다. 월평균 약 1,700만 어절의 데이터를 구축하였으며, 총 2억 어절 이상의 말뭉치를 구축하였다. 한 기사당 평균 어절 수는 213개이다. 올해 구축 분이 최대 어절 수를 기록했음을 아래 표로 알 수 있다.



<그림 13> 매체별 최종 기사 수 및 월별 구축 어절 수

내용/연도	2019년도 <sup>3)</sup>	2020년도	2021년도	2022년도
기사 기간	10년치 기사	1년치 기사	1년치 기사	1년치 기사
매체 수	42	35	35	34
기사 수	3,991,282	630,095	730,017	978,344
어절 수	1,003,852,321	150,669,174	203,585,743	208,320,912

<표 22> 구축 연도별 기사와 어절 수

3) 2019년도 사업에서는 10년치의 기사를 수집하고 정제하는 과제였으며, 예산도 현재(1년치 기사) 수준의 약 10배 수준이었다.

월	어절 수
1월	18,872,057
2월	16,738,087
3월	20,343,616
4월	19,422,431
5월	17,204,271
6월	18,063,733
7월	17,261,575
8월	16,246,602
9월	14,982,971
10월	15,449,783
11월	16,889,237
12월	16,846,549
합계	208,320,912

<표 23> 월별 구축 어절 수

일 년치의 기사가 다양한 주제로 선정되어야 한다는 과업 내용에 맞추어 구축된 주제별 분포는 아래와 같다.

주제별	기사 수	어절 수	기사당 평균 어절 수
경제	169,640	37,990,364	224
문화	32,530	6,592,709	203
미용/건강	48,806	10,558,965	216
사회	340,222	69,306,510	204
생활	150,109	30,664,287	204
스포츠	45,317	9,400,933	207
연예	21,021	4,456,170	212
정치	103,927	23,976,626	231
IT/과학	66,772	15,374,348	230

<표 24> 주제별 기사 및 구축 어절 수

## 2. 매체별 납품 파일명

말뭉치 유형 구분	매체 및 장르 분류	분석 층위 구분	구축 연도	매체일련번호	매체명
N	W	RW	22	0000001	국민일보
N	W	RW	22	0000002	내일신문
N	W	RW	22	0000003	서울신문
N	W	RW	22	0000004	조선일보
N	W	RW	22	0000005	한겨레
N	W	RW	22	0000006	한국일보
N	L	RW	22	0000001	강원일보
N	L	RW	22	0000002	경기일보
N	L	RW	22	0000003	경북일보
N	L	RW	22	0000004	경인일보
N	L	RW	22	0000005	기호일보
N	L	RW	22	0000006	남도일보
N	L	RW	22	0000007	대구신문
N	L	RW	22	0000008	대전일보
N	L	RW	22	0000009	동양일보
N	L	RW	22	0000010	매일신문
N	L	RW	22	0000011	부산일보
N	L	RW	22	0000012	전남일보
N	L	RW	22	0000013	전북도민일보
N	L	RW	22	0000014	중도일보
N	L	RW	22	0000015	충북일보
N	L	RW	22	0000016	충청일보
N	L	RW	22	0000017	충청투데이
N	P	RW	22	0000001	머니투데이
N	P	RW	22	0000002	서울경제
N	P	RW	22	0000003	스포츠서울
N	P	RW	22	0000004	아시아경제
N	P	RW	22	0000005	아주경제
N	P	RW	22	0000006	이데일리
N	P	RW	22	0000007	이투데이
N	P	RW	22	0000008	헤럴드경제
N	I	RW	22	0000001	노컷뉴스
N	I	RW	22	0000002	뉴스핌
N	Z	RW	22	0000001	미디어오늘

<표 25> 말뭉치 파일명

<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 저작권 이용허락과 관련하여 다음과 같이 계약을 체결한다.

## 다 음

### 제1조 (계약의 목적)

본 계약은 국가 언어 자원(말뭉치) 구축 및 활용을 위한 저작권 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 2021년 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 저작권 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 2022년 12월 31일까지로 한다.

(2) ‘대상저작물’ 및 ‘복제·변형물’의 이용허락 최소 기간은 계약체결일부터 2033년 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 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
  1. ‘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
  1. ‘대상저작물’ 및 ‘복제·변형물’을 제3자의 명예권을 비롯한 인격적 권리를 침해하는 방식으로 이용하지 아니할 것
  1. ‘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 정하여 상대방에게 그 시정을 최고하고, 상대방이 그 기간이 지나도록 이행하지 아니하는 경우에는 계약을 해지할 수 있다. 다만, 상대방이 명백한 시정 거부 의사 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 사



항을 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.

2022년    월    일

권리자 :

성명

주소

이용자 :

성명    국립국어원장 (인)

주소 서울특별시 강서구 금남화로 154

제안요청서에 삽입된 계약서 형태

<부록2>

데이터 정제 작업 지침

## 데이터 정제 작업 지침

□ 사용하지 않는 기사의 표시

삭제 기사 구분	내용
저작권 관련 검토 필요 기사	<ul style="list-style-type: none"> <li>- 연합뉴스발 기사</li> <li>- 외부 기고가가 작성한 기사(~위원, ~교수, 영화평론가 등)</li> <li>- 명예기자, 객원기자, 시민기자가 작성한 기사 (기자 이름에 명예기자나, 객원기자라고 표기되지 않고, 이름만 나오는 경우에는 그대로 사용함.)</li> <li>- 외국 기사를 번역한 기사</li> <li>- 다른 매체의 헤드라인을 모아 놓은 기사</li> <li>- 공동취재단이 작성한 기사</li> <li>- 기자명이 비어있는 기사</li> </ul>
구어체 기사	<ul style="list-style-type: none"> <li>- 대부분이 구어체로 이루어진 기사는 사용하지 않는다,</li> <li>- 구어체의 경우 ‘~입니다’ 등과 같은 기사는 사용해도 무방하다.</li> </ul>
불필요한 정보를 삭제한 후 기사 내용이 짧은 기사	<ul style="list-style-type: none"> <li>- 불필요한 요소를 삭제하고 남은 기사가 최소 어절 수에 못 미치는 경우, 기사를 사용하지 않는다.</li> </ul>
불완전하게 종료되는 기사	<ul style="list-style-type: none"> <li>- ‘관계자는...라고 말’, ‘정부는...예정’ 처럼 ‘했다.’, ‘다.’가 빠진 것으로 유추할 수 있는 경우에는 사용한다. 누락된 한두 글자를 유추하여 기사의 마지막 문장을 완성을 시킬 수 있는 경우에는 사용하지만, 문장의 대부분, 또는 주요성분이 누락되어 유추가 어려운 수준으로 불완전한 기사는 사용하지 않는다.</li> </ul>
한글이 모아쓰기 되지 않은 기사	<ul style="list-style-type: none"> <li>- 한글이 모아쓰기 되지 않았거나, 원래의 단어를 확정할 수 있으면 사용하고, 수정 후보가 많으면 사용하지 않는다.</li> </ul>
명확한 광고 기사	<ul style="list-style-type: none"> <li>- 기사 내용 안에 명확히 광고라고 표기하는 경우에는 사용하지 않는다.</li> </ul>
단순 기사	<ul style="list-style-type: none"> <li>- 날씨, 승진, 부고, 운세, 전보, 임용, 스포츠 득점 정보, 여론 조사 결과, 출구 조사 결과, 어록 모음</li> <li>- 매체의 기사 제목을 모아 놓은 기사</li> </ul>

## □ 기사 본문 내 불필요한 정보의 삭제

예시 안의 붉은 붉은색 글꼴과 같은 내용들은 불필요한 정보로 기사 정제 시 삭제한다.

삭제 정보	예시
표, 그림, 그래프 등의 캡션 정보는 삭제함	<p>(사진제공=건국대학교) (사진제공=SBA) 사진제공=tvN 사진=CJ엔터테인먼트 제공 [그래픽] &lt;그래픽&gt; 일러스트  화면 캡처</p> <p>표&gt; 공정위 망 이용대가 불공정 조사 쟁점 &lt;표&gt; 한상혁 후보자 주요 ICT 정책 현안 입장 ▲영상제공= 사진=FNC엔터테인먼트 제공 출처: 라디오타임스 / 굿모닝브리튼, 사진=스타쉽 제공</p>
기자의 이름, ID 등의 정보는 제거함	<p>[서울경제TV=배요한기자] 동양네트웍스(030790)가 강세다. 글·사진=양○○ 기자 -----@kmib.co.kr 김OO -----@kmib.co.kr. 사진=인터파크 제공</p>
‘Copyright©’ 등 저작권 관련 내용은 제거함	<p>Copyright © 한국경제. All rights reserved. 무단 전재 및 재배포 금지. &lt;저작권자 © ‘돈이 보이는 리얼타임 뉴스’ 머니투데이, 무단전재 및 재배포 금지&gt;</p>
전문	<p>대검찰청 정책관 등 중간간부들이 26일 윤석열 검찰총장 직무배제가 부당하다며 추미에 법무부 장관에게 재고를 요청하는 성명을 냈다. 손준성 수사정보정책관, 이창수 대검 대변인 등 대검 중간간부 27명은 이날 검찰게시판에 ‘대검찰청 중간 간부들의 입장’이라는 제목의 글을 올렸다. 이들은 “검찰총장에 대한 직무집행정지는 적법절차를 따르지 않고, 충분한 직상확인 과정도 없이 이뤄진 것으로 위법부당하다”며 “이는 검찰의 중립성은 물론이고 검찰개혁, 나아가 소중하게 지켜온 대한민국의 법치주의 원칙을 크게 훼손하는 것”이라고 강하게 비판했다. 이어 “검찰이 헌법과 법률에 따라 책임과 직무를 다 할 수 있도록 (윤 총장에 대한) 징계청구와 직무집행 정지를 재고해주시길 것을 간곡히 요청드린다”고 적었다. 아래는 대검 중간간부들의 성명서 전문이다.</p> <p>&amp;lt;대검찰청 중간 간부들의 입장&amp;gt;</p> <p>○ 코로나19 등으로 인한 국가적 위기 상황 속에서 검찰과 관련된 각종 논란으로 국민들께 심려를 끼치고 있어 송구스럽게 생각합니다.</p> <p>○ 검찰이 변화해야 한다는 국민의 뜻에 부응하기 위해 노력하고 있으나, 여전히 많이 부족하다는 것을 잘 알고 있습니다.</p> <p>○ 다만, 최근 검찰을 둘러싸고 진행되고 있는 상황들에 대해 침묵하는 것은 공직자로서 올바른 자세가 아니라는 데에 뜻을 함께 한 대검찰청 중간 간부들은 2020. 11. 26. 아래와 같이 의견을 모았습니다.</p> <p>○ 검찰공무원은 범죄로부터 우리 국민들을 보호하고, 온전한 법치주의 실현을 통해 자유롭고 안정된 민주사회를 구현해야 할 사명이 있습니다.</p> <p>○ 검찰총장에 대한 11. 24. 징계청구와 직무집행정지는 적법절차를 따르지 않고, 충분한 진상확인 과정도 없이 이루어진 것으로 위법, 부당합니다.</p> <p>○ 이는 검찰의 정치적 중립성은 물론이고, 검찰개혁, 나아가, 소중하게 지켜온 대한민국</p>

	<p>의 법치주의 원칙을 크게 훼손하는 것이기도 합니다.</p> <p>○ 검찰이 헌법과 법률에 따라 책임과 직무를 다 할 수 있도록 징계청구와 직무집행 정지를 재고해 주실 것을 법무부장관께 간곡하게 요청드립니다.</p> <p>○ 저희들도 국민과 함께 하는 검찰공무원으로서 본연의 임무를 충실히 수행해 나가겠습니다.</p> <p>2020. 11. 26.</p> <p>손준성 이정봉 최성국 이창수 박기동 강범구 전무곤 고필형 구승모 임승철 이만흠 반종욱 최창민 진현일 박혁수 김용자 김 우 백수진 한기식 김승연 김종현 신준호 추혜윤 장준호 손진욱 김연아 정태원</p> <p>배○○ 기자 -----@hani.co.kr</p>
<p>기사 본문으로 볼 수 없는 부가 정보 의 나열 등</p>	<p>■ 인천·경기지역 시급 현안</p> <p>‘인천·경기에서 가장 우선적으로 해결해야 할 사안이 무엇이라고 생각하느냐’는 질문에 ‘일자리 창출’이 28.0%로 가장 높았다. 이어 ‘지역간 균형발전’이 19.1%, ‘부동산 가격안정화’가 15.0%, ‘광역교통망 구축’ 13.6%, ‘미세먼지 대책마련’이 10.7%, ‘수도권 규제완화’가 3.6% 순이다. ‘기타’가 7.5%, ‘잘 모름’이 2.4%다.</p> <p>지역별로 는 계양·부평권과 남동·연수·미추홀권은 일자리 창출이 각각 32.0%와 30.0%로 가장 높은 반면, 동·서·중구·강화·옹진권은 지역간 균형발전이 22.9%로 가장 높았다.</p> <p>연령대별로 대부분은 일자리 창출을 시급한 현안으로 꼽았지만, 유일하게 40~49세에서만 지역간 균형발전이 가장 높았다.</p> <p>○○○기자</p> <p>어떻게 조사했나</p> <p>이번 조사는 경기일보의 의뢰로 조원씨앤아이가 2019년 12월28일(土)부터 30일(月)까지 사흘간, 인천광역시 거주 만19세 이상 남녀를 대상으로 ARS 여론조사(유선전화 RDD 12%+통신사 제공 휴대전화 가상번호 88% 방식, 성,연령,지역별 비례할당무작위 추출)를 실시한 결과이며,표본수는 805명(총 통화시도 17,366명, 응답률 4.6%), 표본오차는 95% 신뢰수준에 ±3.5%p임. 그 밖의 사항은 중앙선거여론조사심의위원회 홈페이지 참조</p> <p>※오차보정방법 : [립가중] 성별, 연령별, 지역별 가중값 부여(2019년 11월말 행정안전부 발표 주민등록인구기준)</p> <p>한국갤럽이 지난 27~29일 전국 만 18세 이상 1001명을 대상으로 조사(표본오차는 95% 신뢰수준에서 ±3.1% 포인트·중앙선거여론조사심의위원회 참조)한 결과 민주당 지지율은 전주 보다 5%포인트 오른 40%로 집계됐다. 국민의힘도 3%포인트 상승한 20%를 기록했다. 실제 선거가 실시되는 서울에서는 민주당(39%)이 국민의힘(16%)을 크게 따돌렸지만, 부산·울산·경남에서는 국민의힘(33%)이 민주당(31%)을 근소하게 앞섰다.</p> <p>=====</p> <p>위의 내용은 기사 본문과 이어지는 정보로 볼 수 있기에 사용한다.</p> <p>이후 2020대한민국지속가능혁신리더대상 조직위 운영 사무국으로 이메일 또는 우편(서울시 중구 청계천로 11(서린동, 청계한국빌딩 16층))을 통해 6월 30(화)(오후 6시까지 도착분에 한함)까지 제출하면 된다.</p> <p>응모 자격은 정부 상훈 관련법에 부합하는 지자체·기관·법인 및 단체·개인으로 접수된 신청서류는 반환되지 않는다. 평가는 1차 서류심사, 2차 실사를 포함한 심층심사, 3차 최종평가를 거쳐 최종 수상자를 선정한다.</p> <p>자세한 내용은 아래 개요를 참고하시기 바랍니다. 대한민국을 이끄는 혁신리더들의 많은</p>

	<p>참여 바랍니다.</p> <p>[2020 대한민국지속가능혁신리더대상 개요]</p> <p>주 최 : 2020 대한민국지속가능혁신리더대상 조직위원회</p> <p>주 관 : 머니투데이, 더리더</p> <p>접수마감 : 2020년 6월 30(화)</p> <p>접수문의 : 02-724-0952(머니투데이 더리더)</p> <p>접 수 처 : 이메일(awards@mt.co.kr)</p> <p>시상일시 : 2020년 7월 중</p> <p>시상장소 : 여의도 쉼튼호텔</p> <p>신청대상 : 정치·사회·경제·교육·체육·문화·예술·환경 등 각 분야의 지속적인 혁신 공로가 인정되는</p> <p>지자체 및 우수 기관, 단체, 개인리더 등</p> <p>신청양식 : 더리더 홈페이지 우측 상단 배너 클릭, 기사하단 신청서 다운로드에서 클릭 후 내려받기 가능</p>
	<p>특히 생체리듬으로 알려진 ‘써카디안(circadian) 리듬’은 간헐적 단식에서 아주 중요한 요소다. 써카디안 리듬과 중추시계, 말초시계가 일치돼야 건강한 일상이 가능하기 때문. 햇빛이 비출 때 일어나고 일정한 시간에 건강한 음식을 취하며 해가 지면 잠자리에 드는 ‘원시 인류’의 생활을 따라야 한다고 저자는 강조한다.</p> <p>◇호르메시스와 간헐적 단식=박용우 지음. 블루페가수스 펴냄. 276쪽/1만5000원.</p>
	<p>주목받은 신인에게 주는 '넥스트 리더'는 위클리, 크래비티, 엔하이픈에게 돌아갔다.</p> <p>{IMG:2}다음은 '2020 TMA' 수상자(작) 명단.</p> <p>▲ 대상 : 방탄소년단</p> <p>▲ 리스너스 초이스 : 방탄소년단</p> <p>▲ 월드와이드 아이콘 : 세븐틴, 방탄소년단</p> <p>▲ TMA 인기상 : 슈퍼주니어</p> <p>▲ 올해의 아티스트 : 마마무&amp;화사, 강다니엘, 방탄소년단, 갓세븐, 트와이스, 뉴이스트, 아이즈원, 몬스타엑스, 세븐틴, 슈퍼주니어</p> <p>▲ 글로벌 핫티스트 : 스트레이 키즈, (여자)아이들, 에이티즈, 더보이즈</p> <p>▲ 베스트 퍼포머 : 있지, 투모로우바이투게더, 제시</p> <p>▲ 넥스트 리더 : 위클리, 크래비티, 엔하이픈</p> <p>▲ 팬엔스타 최다 득표상(가수) : 슈퍼주니어</p> <p>▲ 팬엔스타 최다 득표상(개인) : 황치열</p> <p>▲ 팬엔스타 초이스상(가수) : 슈퍼주니어</p> <p>▲ 팬엔스타 초이스상(개인) : 황치열</p>
	<p>지난 25일 서울 성동구에서 23년째 PC방을 운영하고 있는 이모씨(47)는 올해 추석 계획을 묻는 기자의 질문에 "가게를 지키는 일"이라고 답했다. 코로나19로 적자가 너무 심해져 하루라도 가게를 비울 수 없다는 것이다.(관련 기사 <a href="#">▶ "나라도 돈 벌겠다" 중2 아들말에...PC방 사장님 3일째 집에 못갔다</a>)</p>
	<p>유족으로는 딸 이희경씨, 동생 은화(전 이화여대 교수)·효숙·성숙씨, 올케 이부자씨가 있다. 여성단체들은 여성장으로 고인을 배웅하기로 했다. 빈소는 창원경상대병원 장례식장 VIP 1호실에 마련됐다. (055)214-1910.</p> <p>김정화 기자 <a href="mailto:clean@seoul.co.kr">clean@seoul.co.kr</a></p>

	<p>경찰은 유서 내용 등을 토대로 A 소방사가 극단적 선택을 한 것으로 보고 유족 등을 상대로 정확한 사망원인을 조사하고 있다.</p> <p>※ 우울감 등 말하기 어려운 고민이 있거나 주변에 이런 어려움을 겪는 가족·지인이 있을 경우 자살 예방 핫라인 ☎1577-0199, 희망의 전화 ☎129, 생명의 전화 ☎1588-9191, 청소년 전화 ☎1388 등에서 24시간 전문가의 상담을 받을 수 있습니다.</p> <p>이보희 기자 boh2@seoul.co.kr</p>
	<p>■이부영은 누구인가</p> <p>이부영 전 열린우리당 의장은 1980년대를 대표하는 재야 민주투사이자 정치 원로다. 동아일보 해직 언론인 출신으로 민주화 투쟁을 하다 수차례 옥고를 치렀다. 1990년에 3당 합당에 반대해 만든 민주당을 통해 정계에 입문한 뒤 14~16대 서울 강동갑에서 3선을 했다. 1995년 당시 김대중 총재가 이끄는 새정치국민회의에 합류하지 않고 통합민주당에 남아 있다가 합당 후 한나라당에서 원내총무, 부총재 등을 지냈다. 2004년 17대 총선에서 과반 의석인 152석을 차지했던 열린우리당 의장을 맡았다. 2015년 정계를 은퇴했고, 지난해부터는 자유언론실천재단 이사장으로서 올바른 언론 환경 조성에 노력하고 있다. ▲1942년 서울 출생 ▲서울대 정치학과 ▲동아일보 기자 ▲14~16대 국회의원 ▲한나라당 부총재 ▲열린우리당 의장 ▲동아시아평화국제회의 조직위원장 ▲자유언론실천재단 이사장</p>
	<p>몸매관리 비법으로 밀크어트를 소개한 그녀는 자신만의 다이어트 비법인 건강음료를 공개해 화제가 되고 있다.</p> <p>손쉽게 만들 수 있는 오영주 표 밀크어트 건강음료 3가지를 소개한다.</p> <p>■ 아보카도 스무디</p> <p>&lt;재료&gt;</p> <p>우유 200ml, 아보카도 1/2개, 바나나 1개</p> <p>&lt;만드는 방법&gt;</p> <p>아보카도를 반으로 갈라 씨와 껍질을 제거하고, 우유, 아보카도, 바나나 등 모든 재료를 믹서에 넣고 갈아주면 완성이다.</p> <p>■ 고구마라떼</p> <p>&lt;재료&gt;</p> <p>우유 300ml, 삶은 고구마 1개</p> <p>&lt;만드는 방법&gt;</p> <p>삶은 고구마는 껍질을 벗긴 뒤 우유와 함께 믹서에 넣고 갈아준다. 만약 고구마라떼를 마실 때 목 넘김을 부드럽게 하고 싶다면 고구마를 잘게 잘라 믹서에 넣으면 된다. 고구마를 대신해 블루베리, 바나나, 딸기 등 과일로 대체 가능하며, 기호에 따라 꿀이나 시럽으로 당도를 조절한다.</p>
<p>기사와 상관 없는 광고 혹은 반복되는 문장, 오류의 경우</p>	<p>아래 기사는 본 기사와 상관없이 다른 기사의 내용이 오류로 잘못 들어간 경우이다. 본 기사와 상관없는 내용은 삭제한다.</p> <p>=====</p> <p>이병헌 한가인 한효주 등이 소속된 BH엔터테인먼트와 정려원 손담비 박하선 등의 소속사 키이스트, 문채원 신세경 등의 매니지먼트를 담당하는 나무엑터스도 같은 입장을 발표하며 ‘강경 대응’을 예고했다. 동방신기의 소속사 SM엔터테인먼트 또한 “현재 온라인</p>

커뮤니티 및 SNS 상에 특정 종교와 관련해 당사 아티스트가 언급되어 유포되고 있는 내용은 사실이 아니다. 이는 전혀 근거 없는 루머로, 당사 아티스트는 특정 종교와 무관함을 말씀드린다”고 입장을 밝혔다. 이들 또한 “법적 조치를 취할 것”이라고 전했다.

한편 질병관리본부 중앙방역대책본부는 4일 오전 0시 기준 코로나19 확진자가 5328명이라고 밝혔다. 전날 오전 0시와 비교하면 516명이 늘었다. 사망자는 전날 하루 사이에 4명이 추가돼 총 32명이다. 격리 해제된 확진자는 7명이 늘어 41명이다.

[출처: 서울신문에서 제공하는 기사입니다.]

<https://en.seoul.co.kr/news/newsView.php?id=20200304500092#csidx4dab120876716f7a9506745d61f1391>





### <부록3>

말뭉치 종류별 구축 예시

❖ 각 말뭉치 종류별 비교 예시(기사 내용 중 일부 발췌)

신문 기사 말뭉치	신약개발 바이오기업 메디프론은 최근 전남대 약대 조원제 교수팀과 NLRP3 저해제(뇌염증) 기전의 치매치료제 개발을 위한 공동연구계약을 체결했다고 24일 밝혔다. 회사 관계자는 "지금까지의 알츠하이머성 치매치료제는 아밀로이드베타, 타우 중심 치매 신약 개발 타겟에서 최근에는 뇌의 염증 반응에 주목했다"며 "염증조절복합체(inflammasome)를 조절함으로써 알츠하이머성 치매를 치료하려는 염증 타겟으로 확대되고 있다"고 말했다.
인용 부호 수정 말뭉치	<p>신약개발 바이오기업 메디프론은 최근 전남대 약대 조원제 교수팀과 NLRP3 저해제(뇌염증) 기전의 치매치료제 개발을 위한 공동연구계약을 체결했다고 24일 밝혔다.</p> <p>회사 관계자는 “지금까지의 알츠하이머성 치매치료제는 아밀로이드베타, 타우 중심 치매 신약 개발 타겟에서 최근에는 뇌의 염증 반응에 주목했다”며 “염증조절복합체(inflammasome)를 조절함으로써 알츠하이머성 치매를 치료하려는 염증 타겟으로 확대되고 있다”고 말했다.</p>
문장 말뭉치	<p><s>신약개발 바이오기업 메디프론은 최근 전남대 약대 조원제 교수팀과 NLRP3 저해제(뇌염증) 기전의 치매치료제 개발을 위한 공동연구계약을 체결했다고 24일 밝혔다.</s></p> <p><s>회사 관계자는 “지금까지의 알츠하이머성 치매치료제는 아밀로이드베타, 타우 중심 치매 신약 개발 타겟에서 최근에는 뇌의 염증 반응에 주목했다”며 “염증조절복합체(inflammasome)를 조절함으로써 알츠하이머성 치매를 치료하려는 염증 타겟으로 확대되고 있다”고 말했다.</s></p>
내용	인용 부호 수정, 오타 수정.

신문 기사 말뭉치	KB증권은 지난달 중개형 ISA 계좌 신규개설 고객 대상으로 공모주 이벤트 행사를 벌였다. ISA 계좌에 2000만원 이상을 납입하면 이달부터 진행되는 공모주 청약에서 청약 2배수 우대를 줬다.
인용 부호 수정 말뭉치	<p>KB증권은 지난달 중개형 ISA 계좌 신규개설 고객 대상으로 공모주 이벤트 행사를 벌였다. ISA 계좌에 2000만원 이상을 납입하면 이달부터 진행되는 공모주 청약에서 청약 2배수 우대를 줬다.</p>
문장 말뭉치	<p><s>KB증권은 지난달 중개형 ISA 계좌 신규개설 고객 대상으로 공모주 이벤트 행사를 벌였다. ISA 계좌에 2000만원 이상을 납입하면 이달부터 진행되는 공모주 청약에서 청약 2배수 우대를 줬다.</s></p>
내용	오타 수정

신문 기사 말뭉치	CJ제일제당이 식품 산업의 새로운 패러다임을 함께 만들어갈 유망 스타트업을 육성한다. CJ제일제당은 '프론티어 랩스' 프로그램을 론칭했다고 15일 밝혔다. 글로벌 엑셀러레이터 '스파크랩'과 공동으로 진행하는 이번 프로그램은 기술력과 <b>아이디어를 아이디어를</b> 보유한 스타트업을 선발해 기업당 5000만원에서 1억원을 초기 투자한다. 이를 위해 CJ제일제당은 10억원을 출자했다.
인용 부호 수정 말뭉치	<p>CJ제일제당이 식품 산업의 새로운 패러다임을 함께 만들어갈 유망 스타트업을 육성한다.</p> <p>CJ제일제당은 '프론티어 랩스' 프로그램을 론칭했다고 15일 밝혔다. 글로벌 엑셀러레이터 '스파크랩'과 공동으로 진행하는 이번 프로그램은 기술력과 <b>아이디어를</b> 보유한 스타트업을 선발해 기업당 5000만원에서 1억원을 초기 투자한다. 이를 위해 CJ제일제당은 10억원을 출자했다.</p>
문장 말뭉치	<p><s>CJ제일제당이 식품 산업의 새로운 패러다임을 함께 만들어갈 유망 스타트업을 육성한다.</s></p> <p><s>CJ제일제당은 '프론티어 랩스' 프로그램을 론칭했다고 15일 밝혔다.</s> <s>글로벌 엑셀러레이터 '스파크랩'과 공동으로 진행하는 이번 프로그램은 기술력과 <b>아이디어를</b> 보유한 스타트업을 선발해 기업당 5000만원에서 1억원을 초기 투자한다.</s> <s>이를 위해 CJ제일제당은 10억원을 출자했다.</s></p>
내용	문장 말뭉치 분할, 인용 부호 수정, 중복 문구 삭제

신문 기사 말뭉치	세계적 도시부동산 연구단체인 ULI(Urban Land Institute)와 이지스자산운용이 ESG(환경? 사회? 지배구조)와 도시 모빌리티의 트렌드를 짚고 지속가능한 도시의 미래상을 그리기 위한 글로벌 콘퍼런스를 오는 13일 개최한다고 7일 밝혔다. 공간비즈니스에 ESG 접목을 모색 중인 이지스자산운용은 이번 'ULI한국 2021 연례 콘퍼런스'에 주요 후원사와 프로그램 파트너로 참여한다.
인용 부호 수정 말뭉치	<p>세계적 도시부동산 연구단체인 ULI(Urban Land Institute)와 이지스자산운용이 ESG(환경·사회·지배구조)와 도시 모빌리티의 트렌드를 짚고 지속가능한 도시의 미래상을 그리기 위한 글로벌 콘퍼런스를 오는 13일 개최한다고 7일 밝혔다.</p> <p>공간비즈니스에 ESG 접목을 모색 중인 이지스자산운용은 이번 'ULI한국 2021 연례 콘퍼런스'에 주요 후원사와 프로그램 파트너로 참여한다.</p>
문장 말뭉치	<p><s>세계적 도시부동산 연구단체인 ULI(Urban Land Institute)와 이지스자산운용이 ESG(환경·사회·지배구조)와 도시 모빌리티의 트렌드를 짚고 지속가능한 도시의 미래상을 그리기 위한 글로벌 콘퍼런스를 오는 13일 개최한다고 7일 밝혔다.</s></p> <p><s>공간비즈니스에 ESG 접목을 모색 중인 이지스자산운용은 이번 'ULI한국 2021 연례 콘퍼런스'에 주요 후원사와 프로그램 파트너로 참여한다.</s></p>
내용	문장 부호 가운뎃점 수정

신문 기사 말뭉치	<p>휴넷은 내년 부터 주 4일제를 전면 도입하고 주 32시간 근무를 시행한다고 22일 밝혔다.</p> <p>휴넷은 지난 2019년 부터 `주 4.5일 근무`를 실시해왔다. 이번에 `주 4일`로 확대 시행으로 내년 1월 1일부터 부서별 시범 운영을 거쳐 하반기부터 전사 시행할 예정이다.</p> <p>주 4일 근무제는 직원이 일주일 중 하루를 자유롭게 선택해 쉬는 형태로 시행된다.</p>
인용 부호 수정 말뭉치	<p>&lt;p&gt;휴넷은 내년 부터 주 4일제를 전면 도입하고 주 32시간 근무를 시행한다고 22일 밝혔다.&lt;/p&gt;</p> <p>&lt;p&gt;휴넷은 지난 2019년 부터 ‘주 4.5일 근무’를 실시해왔다. 이번에 ‘주 4일’로 확대 시행으로 내년 1월 1일부터 부서별 시범 운영을 거쳐 하반기부터 전사 시행할 예정이다.&lt;/p&gt;</p> <p>&lt;p&gt;주 4일 근무제는 직원이 일주일 중 하루를 자유롭게 선택해 쉬는 형태로 시행된다.</p>
문장 말뭉치	<p>&lt;p&gt;&lt;s&gt;휴넷은 내년 부터 주 4일제를 전면 도입하고 주 32시간 근무를 시행한다고 22일 밝혔다.&lt;/s&gt;&lt;/p&gt;</p> <p>&lt;p&gt;&lt;s&gt;휴넷은 지난 2019년 부터 ‘주 4.5일 근무’를 실시해왔다.&lt;/s&gt; &lt;s&gt;이번에 ‘주 4일’로 확대 시행으로 내년 1월 1일부터 부서별 시범 운영을 거쳐 하반기부터 전사 시행할 예정이다.&lt;/s&gt;&lt;/p&gt;</p> <p>&lt;p&gt;&lt;s&gt;주 4일 근무제는 직원이 일주일 중 하루를 자유롭게 선택해 쉬는 형태로 시행된다.&lt;/s&gt;</p>
내용	<p>문장 말뭉치 분할,</p> <p>인용 부호 수정. `(U+0060) → ‘(U+2018), ’(U+2019)</p>

신문 기사 말뭉치	<p>CU는 7일 업계 최초로 벤티 사이즈 컵얼음과 델라페(delaffe) 아이스드링크를 출시한다고 밝혔다. 벤티 사이즈 컵얼음은 기존 대용량 컵얼음인 <b>‘빅컵얼음(230g)’</b>보다 두 배 가량 커진 400g 용량의 상품이다. 이에 따라 함께 출시되는 델라페도 기존 대용량 아이스드링크 용량인 335ml보다 1.5배 늘어난 500ml로 기획됐다.</p>
인용 부호 수정 말뭉치	<p>&lt;p&gt;CU는 7일 업계 최초로 벤티 사이즈 컵얼음과 델라페(delaffe) 아이스드링크를 출시한다고 밝혔다.&lt;/p&gt; &lt;p&gt;벤티 사이즈 컵얼음은 기존 대용량 컵얼음인 <b>‘빅컵얼음(230g)’</b>보다 두 배 가량 커진 400g 용량의 상품이다. 이에 따라 함께 출시되는 델라페도 기존 대용량 아이스드링크 용량인 335ml보다 1.5배 늘어난 500ml로 기획됐다.&lt;/p&gt;</p>
문장 말뭉치	<p>&lt;p&gt;&lt;s&gt;CU는 7일 업계 최초로 벤티 사이즈 컵얼음과 델라페(delaffe) 아이스드링크를 출시한다고 밝혔다.&lt;/s&gt;&lt;/p&gt; &lt;p&gt;&lt;s&gt;벤티 사이즈 컵얼음은 기존 대용량 컵얼음인 <b>‘빅컵얼음(230g)’</b>보다 두 배 가량 커진 400g 용량의 상품이다.&lt;/s&gt; &lt;s&gt;이에 따라 함께 출시되는 델라페도 기존 대용량 아이스드링크 용량인 335ml보다 1.5배 늘어난 500ml로 기획됐다.&lt;/s&gt;&lt;/p&gt;</p>
내용	<p>인용 부호 누락 수정, 문장 말뭉치 분할.</p>

<기획·연구>

국립국어원 강미영 언어정보과장

국립국어원 이선영 연구원

<사업 참여자>

사업 책임자 윤종웅(주윤즈정보개발 소장)

사업 참여자 남가윤(주윤즈정보개발 연구원)

박지영(주윤즈정보개발 연구원)

서경찬(주윤즈정보개발 책임연구원)

안소연(주윤즈정보개발 연구원)

윤종성(주윤즈정보개발 팀장)

이승철(주윤즈정보개발 수석연구원)

임순영(주윤즈정보개발 연구원)

최원수(주윤즈정보개발 연구원)

---

발행인: 국립국어원장

발행처: 국립국어원

서울시 강서구 금남화로 154

전화 02-2669-9775, 전송 02-2669-9757

인쇄일: 2022년 12월 30일

발행일: 2022년 12월 30일

인 쇄: 다큐팩토리

---

※ 이 보고서는 국립국어원의 용역비로 수행한 ‘2022년 신문 기사 원문  
자료 수집 및 정제’ 사업의 결과물을 발간한 것입니다.



NATIONAL INSTITUTE OF KOREAN LANGUAGE