

국립국어원 2023-01-22

발간등록번호
11-1371028-000953-01

## 2023년 신문 기사 원문 자료 수집 및 정제

사업책임자

윤 중 응





## 제 출 문

국립국어원장 귀하

국립국어원과 체결한 연구용역 계약에 따라 ‘2023년 신문 기사 원문 자료 수집 및 정제’에 관한 연구 보고서를 작성하여 제출합니다.

■ 사업 기간: 2023년 4월 18일 ~ 2023년 10월 18일

2023년 10월 18일

사업책임자: 윤종웅(주)윤즈정보개발)

연구 기관: (주)윤즈정보개발

사업 책임자: 윤종웅

사업 참여자: 남가운, 서경찬

안소연, 윤종성, 이승철

임순영, 임승락, 최원수



## 〈국문 요약〉

# 2023년 신문 기사 원문 자료 수집 및 정제

국립국어원은 실제 언어 사용을 반영하는 신문 기사 말뭉치를 구축하여 개인이나 기업, 학계에 필요한 말뭉치를 제공하고 있다. ‘신문 기사 원문 자료 수집 및 정제’ 사업은 올해로 5년 차를 맞는 연속 사업으로 인공 지능 산업계와 연구기관 등에서 공공재로 활용할 수 있는 대규모 한국어 학습 자료 구축 사업을 이어가고 있다.

이 사업의 수행 범위는 신문 기사 원문 자료 수집(2022년 작성 기사, 월별 1,000만 어절 이상), 저작 권리자와의 이용 허락 계약을 통한 저작권 해결, 중복 기사 제거 및 정제, 신문 기사 3종 말뭉치 구축(원시 말뭉치, 인용 부호 수정 말뭉치, 문장 말뭉치), 기사별 메타 정보 작성 및 목록 작성으로 구분되어 있다.

다양한 분야에서 활용할 수 있는 데이터 생산에 목적을 두고 신문 기사 원문 말뭉치에서 인용 부호의 오류를 수정한 ‘인용 부호 수정 말뭉치’로 정제하고, 다시 단락 단위로 구성된 인용 부호 수정 말뭉치를 문장 단위로 나눈 ‘문장 분할 말뭉치’로 정제를 진행하였다.

지난해처럼 신문 기사 원문 자료의 저작권을 확보하는 일 외에, 신문 기사 원문 일부를 국립국어원의 다른 사업(번역 말뭉치, 점역 말뭉치, 한국어-한국수어 병렬 말뭉치 구축 및 배포)에 활용할 수 있도록 번역을 포함한 2차적 저작물 작성권을 확보하는 일도 2023년의 사업 범위에 새로 포함되었다. 경향신문, 서울신문 등 전국일간지를 비롯하여 경제지, 인터넷매체 등 8개 매체의 2차적 저작물 작성권을 확보하였다.

원문 자료 수집 대상은 국립국어원과 협의하여 총 28개 매체를 선택하였고 ‘한국언론진흥재단(26개 매체의 저작권 신탁 기관)’과 ‘조선일보’, ‘경향신문’, 이렇게 세 기관과 계약서 및 부속합의서를 작성하여 저작권 해결을 진행하였다. 번역 저작권과 더불어 신문 기사 수집 및 정제 사업에 사용되는 말뭉치 이용 허락 최소 기간은 2034년 12월 31일까지로 하였고, 저작자인 언론사가 이용 허락 중지 의사를 밝히지 않으면 이용 허락이 1년 단위로 자동 갱신되도록 하였다. 28개 매체로부터 확보한 원시 자료는 총 537,880,451개의 어절로 이루어진 2,487,359건의 기사이다.

28개 매체가 제공한 문서(html 등)에는 중복 기사, 광고성 기사, 유사한 기사, 데이터 소실이 발생한 기사, 캡션과 소제목이 정확하게 분리가 되어 있지 않아 의미 전달에 문제가 있는 기사 등 학습에 사용하기 부적절한 수많은 요소가 포함되어 있다. 그리고

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. 여러 개의 문장으로 이루어진 긴 단락을 단위로 말뭉치를 학습하는 것은 효율이 떨어질 수밖에 없다. 이에 단락을 문장으로 세분한 문장 말뭉치를 구축하였다. 이 말뭉치는 문장 분할에 활용할 수 있다. 문장 분할이 중요한 이유는 문장이 인공 지능 학습의 기본 단위이기 때문이다.

보통 엄청나게 많은 데이터를 수집하면 데이터가 좀 부실해도 문제가 없을 것으로 생각하는데 데이터의 증가는 필연적으로 학습 비용으로 귀결되기 때문에 무한정으로 데이터를 확대할 수는 없다. 단순히 학습 데이터의 양을 기하급수적으로 늘리기만 해서는 인공지능 모델의 성능 향상 혹은 더 나은 의사결정으로 연결시킬 수 없다. 데이터의 양만큼이나 중요한 것이 데이터의 품질이다. “쓰레기가 들어가면 쓰레기가 나온다.(Garbage in, garbage out)”라는 말처럼 질 낮은 데이터는 아무리 양이 많아도 좋은 결과를 낼 수 없다. 인공지능 모델은 학습 데이터의 품질에 따라 성능이 좌우된다.

국립국어원에서는 기사로 볼 수 없는 불필요한 내용을 제거하고 저작권 문제가 있는 기사들을 제거한 신문 기사 말뭉치와 함께 기사 안에서 다양한 코드로 표현되거나 오류가 있는 인용 부호를 수정하여 최종적으로 225,215,915개의 어절로 이루어진 신문 기사 1,023,431개의 말뭉치를 구축하였다.

이 사업을 통해 구축한 말뭉치는 실제 언어 사용을 반영하고 있는 최신 말뭉치로서 3종의 말뭉치를 함께 이용하게 된다면 4차 산업혁명을 대비한 인공 지능 기술의 개발과 학계 연구 등 여러 분야에서 활용함으로써, 인공 지능 학습에 유용한 자료가 될 것으로 기대한다.

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

<Abstract>

## Collection and Refinement of Data from Original Newspaper Articles in 2023

The National Institute of Korean Language (NIKL) has built the Newspaper Corpus that reflects the use of actual language and provides the corpus that individuals, companies, and academia require. This project which is collection and refinement of data from original newspaper articles, now in its fifth year, was created to collect and refine data from original newspaper articles, and continues to build large-scale Korean learning data that can be utilized as public resources by the artificial intelligence (AI) industry and related research institutes.

The scope of this project is divided into: 1) collecting the original text data from newspaper articles written in 2022, amounting to more than 10 million *Eojeol* (word-spacing unit) per month, 2) addressing copyright-related issues through usage agreements with copyright holders, 3) removing and refining duplicate articles, 4) building three types of corpora (raw corpus; corpus in which quotation marks are corrected; and corpus in which paragraphs are segmented into sentences) from newspaper articles, and 5) creating and listing metadata for each article.

To produce data that can be used in a variety of fields, the corpus of original text data of newspaper articles was refined by correcting quotation marks, and then refined again by segmenting sentences.

Like last year, the copyright of the original text data of newspaper articles was secured, but the scope for the 2023 project expanded to secure the right to create derivative works, including translation, so that some of these texts can be used for other NIKL projects (such as construction and distribution of a translation corpus, a braille translation corpus, and a parallel corpus between Korean and Korean sign language). The rights to create derivative works were secured from eight media, including national daily newspapers such as the *Kyunghyang Shinmun* and *Seoul Shinmun*, economic magazines, and internet media.

In consultation with NIKL, a total of 28 media were selected for collection. The Korea Press Foundation, which is the copyright management organization for 26 media, the Chosun Ilbo, and the Kyunghyang Shinmun concluded the contracts and annexed schedules, as well as notarized the content to resolve copyright issues. In addition to copyright, the minimum period of permission to use the corpus for collecting and refining newspaper articles is until December 31, 2034, and is set to automatically renew permissions on a yearly basis unless the media as authors express their intention to suspend permissions. The raw data secured from the 28 media included 2,487,359 articles with a total of 537,880,451 *Eojeol* (*word-spacing unit*).

The documents (html, etc.) from the 28 media contain a large amount of unsuitable elements for learning, including duplicate articles, advertising articles, similar articles, articles with data loss, and articles with problems conveying meaning due to captions and subheadings not being accurately separated. 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. There are also cases of typos within the articles. Learning such unrefined data inevitably limits the efficiency of corpus utilization.

In most cases of AI learning, sentence is used as a basic unit. Learning a corpus as a long paragraph unit that consists of several sentences is ultimately inefficient. To combat these issues, a corpus in which paragraphs are subdivided into sentences was constructed. This corpus can be used for sentence segmentation, which is important as it is better for AI to learn as a single sentence than a long paragraph unit.

Usually, when a huge amount of data is collected, many believe there will be no issues even if the data is poor. However, since the increase in data inevitably results in increased learning costs, the data cannot be expanded indefinitely. Simply increasing the amount of learning data exponentially cannot lead to improved performance or better decisions of AI models. Data quality is as important as data quantity. As the performance of AI models depends on the quality of learning data, low-quality data cannot produce good results, no matter how much there is, as the saying “Garbage in, garbage out” says.



Along with the corpus of newspaper articles that removed unnecessary content that could not be viewed as articles and that removed articles with copyright issues, NIKL built the corpus of 1,023,431 newspaper articles with 225,215,915 *Eojeol* by correcting various instances of quotation mark errors in the articles.

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

**Key words:** Newspaper Corpus, artificial intelligence (AI), AI training data, refined data, data copyright, newspaper articles, corpus in which paragraphs are split into sentences, contemporary Korean, correction of quotation marks





# 차 례

## 제1장 서론

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

## 제2장 사업 수행 내용

1. 매체 선정 및 계약 .....	7
2. 데이터 수집 .....	9
3. 데이터 1차 정제 .....	26
4. 데이터 2차 정제 .....	31
5. 메타데이터 작성 .....	43
6. 인용 부호 수정 말뭉치 .....	44
7. 문장 말뭉치 구축 .....	52

## 제3장 사업 수행 결과

1. 신문 기사 정제 결과 .....	55
2. 매체별 납품 파일명 .....	59

<부록 1> 국가 언어 자원(말뭉치) 구축 및 활용 저작권 이용 허락 계약서 .....	61
<부록 2> 저작재산권 비독점적 이용허락 계약서 .....	67
<부록 3> 데이터 정제 작업 지침 .....	73
<부록 4> 말뭉치 종류별 구축 예시 .....	81
<부록 5> 신문 기사 말뭉치 오류 검색 목록 .....	84

## 표 차례

<표 1> 사업 공정표 .....	5
<표 2> 선정된 매체 구분 .....	8
<표 3> 최초 수집 기사와 어절 수 .....	10
<표 4> 원시 데이터 특징 예시(태그가 남아있는 경우) .....	10
<표 5> 원시 데이터 특징 예시(불특정 문자 포함) .....	11
<표 6> 원시 데이터 특징 예시(데이터 소실 1) .....	12
<표 7> 원시 데이터 특징 예시(데이터 소실 2) .....	12
<표 8> 데이터 특징 .....	13
<표 9> 저작권 이용 문제로 인해 사용하지 않는 기사의 특징 .....	28
<표 10> 불필요한 요소 제거 내용 .....	35
<표 11> 원시 데이터와 정제된 데이터 비교 1 .....	38
<표 12> 원시 데이터와 정제된 데이터 비교 2 .....	39
<표 13> 원시 데이터와 정제된 데이터 비교 3(기사로 볼 수 없는 정보 삭제) .....	40
<표 14> 2023년 신문 기사 주제별 통계 .....	43
<표 15> 인용 부호 치환 표 .....	45
<표 16> 인용 부호 수정 데이터 정제 전후 .....	46
<표 17> 최종 선정 기사 수 .....	47
<표 18> ‘한·중·일 호환용 한자 영역’ 한자 치환 표 .....	48
<표 19> 치환 코드 목록 .....	49
<표 20> 오타 글자 .....	51
<표 21> 문장 말뭉치 데이터 정제 .....	53
<표 22> 신문 기사 정제 총괄표 .....	56
<표 23> 구축 연도별 기사와 어절 수 .....	57

## 표 차례

<표 24> 월별 구축 어절 수 .....	58
<표 25> 주제별 기사 및 구축 어절 수 .....	58
<표 26> 말뭉치 파일명 .....	59

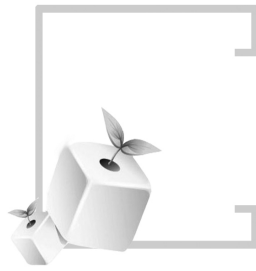
## 그림 차례

<그림 1> 구축 공정별 내용 .....	4
<그림 2> 연도별 매체 비율과 상위, 하위 기사수 매체 .....	9
<그림 3> 오류 유형 ①: 글자 깨짐 .....	14
<그림 4> 오류 유형 ②: 캡션 정보와 본문이 구분되지 않는 경우 .....	14
<그림 5> 오류 유형 ③: 중간 제목이 본문 사이에 들어간 경우 .....	15
<그림 6> 오류 유형 ④-1: 웹 페이지와는 다르게 기사 내용이 변형됨 .....	16
<그림 7> 오류 유형 ④-2: 웹 페이지와는 다르게 기사 내용이 변형됨 .....	17
<그림 8> 오류 유형 ⑤: 문장이 임의로 줄바꿈되어 있는 경우 .....	17
<그림 9> 오류 유형 ⑥: 외부 기고가 정보가 삭제되어 원시 데이터에 없는 경우 .....	18
<그림 10> 오류 유형 ⑦: 평소에 쓰이지 않는 단어가 깨지면서 ?로 치환되는 현상 ..	19
<그림 11> 오류 유형 ⑧: 오류 수정 .....	20
<그림 12> 오류 유형 ⑨: 불필요한 부분 삭제 .....	21
<그림 13> 오류 유형 ⑩: 특수기호 오류 .....	22
<그림 14> 오류 유형 ⑪: 기사 내용 반복 .....	23
<그림 15> 오류 유형 ⑫: 캡션이나 문장이 붙어 버리는 오류 .....	24
<그림 16> 오류 유형 ⑬: 용어 설명이 중간에 삽입되면서 임의로 줄바꿈됨 .....	24
<그림 17> 오류 유형 ⑭: 서명 기호(<, >)와 내용이 사라짐 .....	25
<그림 18> 원본 데이터와 정제된 데이터의 예 .....	37
<그림 19> 작업 편집 화면 .....	41
<그림 20> 작업 프로그램 화면 .....	41
<그림 21> 데이터 정제 2차 검수 공정 .....	42
<그림 22> 인공 지능을 활용한 주제 분류 .....	43
<그림 23> 연도별 기사 주제 통계 .....	44

## 그림 차례

<그림 24> 문장 말뭉치 개념 .....	52
<그림 25> 문단 내 문장 분할 수(상/하위 5개 매체) .....	53
<그림 26> 구축 공정별 내용 .....	55
<그림 27> 매체별 최종 기사 수 및 월별 구축 어절 수 .....	57





## 제 1 장

# 서 론



## 제1장 서론

### 1. 사업 목적

국립국어원은 ‘21세기 세종계획’을 통해 2억 어절의 자료를 구축하였으며, ‘모두의 말뭉치’에 매년 새로운 자료들을 추가로 구축하는 노력을 기울이고 있다. 연속 사업으로 신문 기사 원문 자료 수집 및 정제 사업이 이루어지고 있으며 매년 대량의 말뭉치 데이터를 생산하고 있다. 이 사업은 신문 기사 원문 자료를 대량으로 구축하여 언어 처리 인공지능 기술 개발에 활용될 수 있도록 하고 이용하는 데 저작권 문제가 발생하지 않도록 저작권자로부터 이용 허락을 확보하는 것이 목적이다.

신문 기사 원시 말뭉치의 구축량은 월별로 약 1,000만 어절 이상, 총 1.2억 어절 이상이다. 엄청나게 많은 데이터를 수집하면 데이터 오류가 일부 존재하더라도 일반적으로 모델의 성능을 향상시킬 수 있다고 주장하는 경우가 있으나 모델 학습에 오류는 영향을 반드시 끼치게 된다. 신문 기사는 생각보다 많은 오류를 포함하고 있다. 이 사업에서 구축될 말뭉치는 단순히 신문 기사를 수집만 한 것이 아니라, 정제를 거쳐 불필요한 정보를 제거한 말뭉치가 될 것이다.

### 2. 사업 수행 범위

이 사업의 범위는 네 부분으로 나눌 수 있다. 첫 번째는 **신문 기사의 원문 자료 수집**이다. 원문 자료 수집 대상은 2022년 1월~12월에 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. 기존 사업과 같은 방식으로 이 사업에 필요한 저작권을 확보하여 저작권 침해를 방지함으로써 사업 수행 결과물을 누구나 자유롭게 이용할 수 있도록 하는 과정이 있고 기술협상에 포함된 내용으로 국립국어원에서 시행 중인 한국어-외국어 번역 말뭉치 사업 등을 위한 저작권을 확보하는 내용도 포함이 되었다. 이는 이 사업에서 확보하는 2차적 저작물 작성권(번역 이용 허락 등)으로 번역 말뭉치 구축을 포함한 국립국어원의 일부 사업에서 저작권에 어려움 없이 활용할 수 있도록 하기 위한 것이다.

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. 불필요한 내용을 제거한 신문 기사 말뭉치, 신문 기사 내 인용부호를 수정한 인용부호 수정 말뭉치와 단락 단위를 문장 단위로 분할한 문장 말뭉치, 이렇게 총 3종의 말뭉치를 구축한다.

대부분의 인공지능 학습은 문장을 기본 단위로 하고 있으며, 형태소 분석과 기계 번역은 대부분 문장을 기본 단위로 하고 있다. 따라서 단락을 최소 단위로 하는 말뭉치가 아니라, 문

장을 최소 단위로 하는 말뭉치 구축이 필요하다. 이 사업에서는 단락을 문장으로 세분하여 문장 말뭉치를 구축하였다.

가장 중요한 것은 저작권에 위배되는 기사를 제거해야 한다. 기사 내에서 저작권에 문제가 될 수 있는 기사들은 전부 확인하여 저작권 문제가 일어나지 않도록 해당 기사는 확인 후 사용하지 않는다.

마지막으로 구축된 기사 데이터의 기자 정보, 어절 수, 주제 분류, 기사 작성일 등의 **메타데이터**를 작성하는 것이 사업의 범위이다.

## 가. 신문 기사 원문 자료 수집(2022년 작성 기사, 1.2억 어절 이상)

- ❖ 신문 기사 말뭉치 구축에 필요한 신문 기사 원문 자료를 수집.
- ❖ 대상은 2022년 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 균형성을 갖춰 1.2억 어절 이상 구축.
- ❖ 파일명과 표지의 종류 및 부착 형식 등은 국립국어원의 지침을 따름.

## 나. 신문 기사 저작 권리자와의 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.
- ❖ 계약은 법률 검토를 받은 후 주관 기관이 제공한 계약서 양식에 따라 국립국어원과 협의하여 체결.
- ❖ 저작권 이용 허락 내용은 신문 기사 원문 자료 및 신문 기사 말뭉치의 저장, 복제, 전송, 배포, 2차적 저작물 작성권을 포함함.
- ❖ 이용 허락 기간은 계약일로부터 최소 2034년 12월 31일까지로 함.
- ❖ 기술협상: 번역 활용을 허락한 매체 10% 이상 포함(전체 구축 어절의 10% 이상을 충족해야 하며, 사업 착수 2개월 이내에 이용 허락 완료할 것).

## 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 작성자의 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.
- ❖ 기사 주제(사회, 경제, 생활, ... 연예, IT/과학) 간의 비율 차(최고-최저)가 가능한 한 25% 포인트 이하가 되도록 하되, 필요시 국립국어원과 협의하여 비율 조정 가능함.

## 라. 메타 데이터 작성

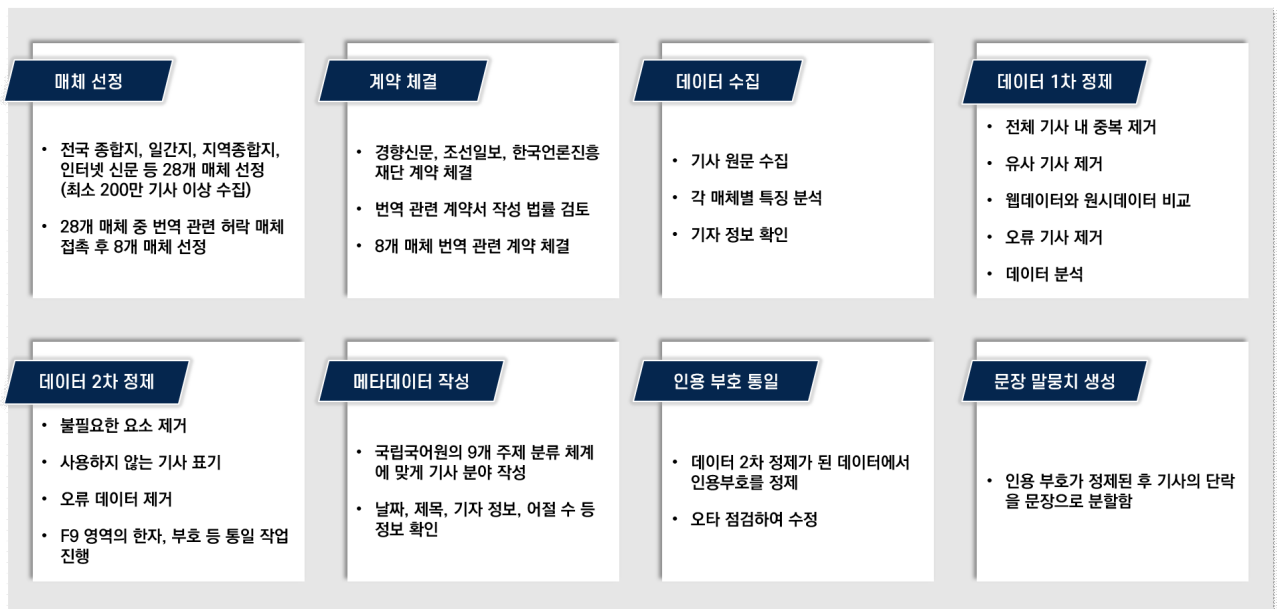
- ❖ 국립국어원이 지정하는 아홉 가지 분류 체계로 신문 기사 주제 재분류
- ❖ 신문사명, 기사 작성일, 주제 분류, 기사 제목, 어절 수 등 국립국어원이 지정하는 항목과 형식으로 기사별 메타 정보 입력 및 수집 기사 목록 작성

### 3. 사업 수행 절차

공정은 크게 8단계로 구분된다. 먼저 한국언론진흥재단(경향신문, 조선일보를 제외한 26개 매체의 저작권 계약 신탁 기관), 경향신문, 조선일보와 계약을 체결하여 기사를 확보하고 저작권을 해결하였다. 원시 데이터를 분석한 이후 해당 데이터를 가공하여 쓸 수 있는 데이터로 구분하는 1차 정제를 하였고, 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.

데이터 납품의 경우 이 사업은 세 번에 걸쳐 진행되었다. 1차 데이터 납품의 경우 국어원의 한국어-외국어 병렬 말뭉치 구축팀에서 사용해야 하기 때문에 해당 8종 매체(경향신문, 서울신문, 세계일보, 이데일리, 이투데이, 헤럴드경제, 노컷뉴스, 뉴스핌)의 구축 완료 데이터를 먼저 납품하였다. 사업 착수 후 약 3개월 만에 8종 매체의 신문 기사 말뭉치, 인용 부호 수정 말뭉치, 문장 말뭉치를 제작하여 납품하였다. 이후 사업 종료 2주, 1주 전 각각 10종 매체의 최종데이터를 각각 납품하였다.



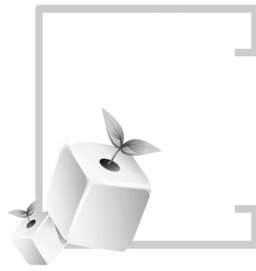
<그림 1> 구축 공정별 내용

#### 4. 사업 추진 경과

이 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							
	데이터 확보							
	번역 매체 저작권 확보 및 계약							
정 제	데이터 1차 정제							
	데이터 2차 정제							
	인용 부호 수정 말뭉치							
	문장 말뭉치							
메타데이터 생성	통계 추출							
	검수 및 반영							
납품 및 종료	중간, 완료 보고회							
	3회에 걸친 데이터 납품					8개 매체		10개 매체 10개 매체

<표 1> 사업 공정표



## 제 2 장

# 사업 수행 내용



## 제2장 사업 수행 내용

### 1. 매체 선정 및 계약

제안요청서에 나온 매체 선정 기준은 다음과 같다. 최소 25개 이상 매체를 선정하여야 하고 인터넷 매체는 전체 매체 수 대비 10% 이하로 선정해야 하며 사업비의 약 45% 이상을 저작권 계약에 사용한다는 내용이다.

수행사는 작년의 프로젝트 경험을 토대로, 계약 조건을 충족시키기 위해 최소 200만 건 이상의 기사와 5억 어절 이상의 데이터 확보가 필요하다는 판단을 하였고, 이를 위해 한국언론진흥재단에 계약 가능한 매체와 함께 기사 수를 문의하여, 최종적으로 28개의 매체를 선정하였다. 전국 종합지 8개 매체를 선정하였고, 인터넷 매체는 전체 매체 수 대비 10% 이하로 선정하였다.

저작물은 2022년 1월 1일부터 2022년 12월 31일까지의 기사였으며, 계약 대상자는 한국언론진흥재단과 경향신문, 조선일보로, 저작권 이용 허락 계약을 통해 최대한 많은 기사를 확보하였다.

이 사업의 원자료 저작권을 확보할 때 기술협상에서 결정된 다음 내용이 가장 큰 화두였다.

기술협상 내용 중 하나인 번역 활용을 위하여 사업 착수 후 2개월 이내에 번역 저작권을 확보하라는 내용이었다. 한국언론진흥재단은 2차적 저작물(번역)에 대한 계약 권한이 없기 때문에 매체를 선정하여 각 매체와 따로 번역 관련 이용 허락 계약을 하는 조건으로 진행하였다.

기사의 저작권을 확보하는 것은 큰 비용이 들어가거나, 각 매체의 동의를 얻기가 쉽지 않은 일이다. 게다가 전체 계약 매체 수의 10% 이상의 매체를 번역 매체에 포함시켜야 했으며, 전체 구축 어절의 10% 이상이 번역 기사로 활용될 수 있어야 했다. 사업 수행사는 기사를 많이 보유한 매체, 전국종합일간 등 다양한 주제를 다루고 이름이 알려져 있는 매체를 조건으로 하여 번역 관련 매체를 접촉하였다. 한정적인 예산 안에서 국립국어원의 수요와 사업의 연속성 등을 고려하여 최종 8개 매체와 계약이 성사되었고, 번역과 관련된 2차적 저작물 작성권을 확보할 수 있었다. 번역 저작물 작성권은 2022년의 기사 전체에 대한 저작권을 확보하는 것이 아닌 번역 사업에 제공될 기사에 한하여 확보하였다.

번역 저작물 작성권과 더불어 신문 기사 수집 및 정제 사업에 사용되는 말뭉치 이용 허락 최소 기간은 2034년 12월 31일까지로 하였고, 저작자인 언론사가 이용 허락 중지 의사를 밝히지 않으면 이용 허락이 1년 단위로 자동 갱신되도록 하였다.



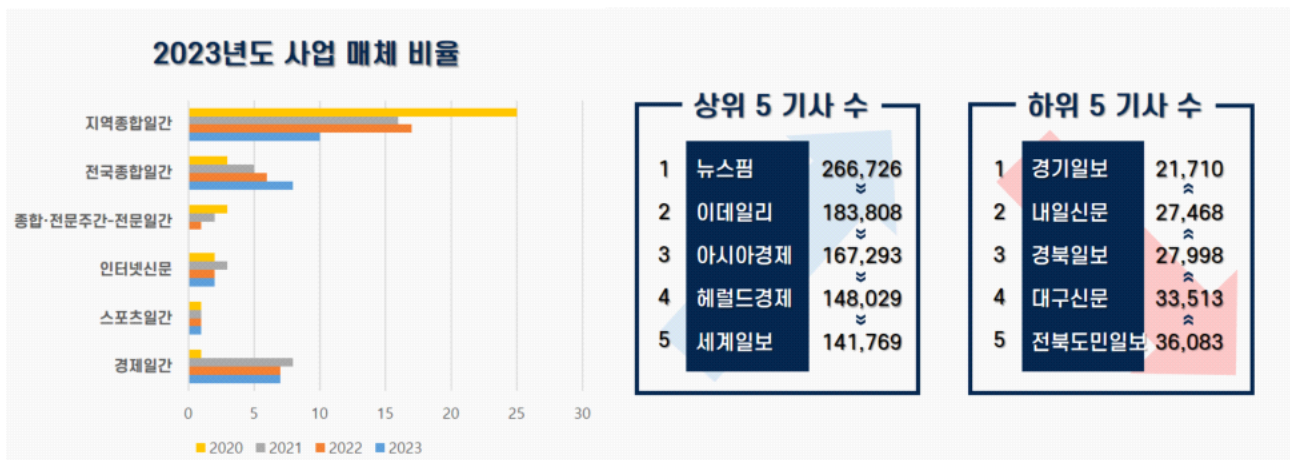
구분	2023년 신문 기사 원문 자료 수집 및 정제의 대상(총 28개 매체)
전국종합일간(8)	<ul style="list-style-type: none"> <li>경향신문(번역), 국민일보, 내일신문, 서울신문(번역), 세계일보(번역), 조선일보, 한겨레, 한국일보</li> </ul>
지역종합일간(10)	<ul style="list-style-type: none"> <li>강원일보, 경기일보, 경북일보, 남도일보, 대구신문, 부산일보, 전북도민일보, 중도일보, 중부일보, 충청일보</li> </ul>
경제일간(7)	<ul style="list-style-type: none"> <li>머니투데이, 서울경제, 아시아경제, 아주경제, 이데일리(번역), 이투데이(번역), 헤럴드경제(번역)</li> </ul>
스포츠일간(1)	<ul style="list-style-type: none"> <li>스포츠서울</li> </ul>
인터넷신문(2)	<ul style="list-style-type: none"> <li>노컷뉴스(번역), 뉴스핌(번역)</li> </ul>

<표 2> 선정된 매체 구분

## 2. 데이터 수집

데이터를 수집할 때 2백만 건 이상, 5억 어절 이상의 조건을 충족해야만 양질의 말뭉치를 무리 없이 구축할 수 있다. 한국언론진흥재단으로부터 기사 매체 명단과 매체별 기사 수 통계를 받아 28개의 매체를 선정하였고 수집된 기사와 어절 수는 각각 2,487,359건, 537,880,451개이다. 매체 수는 2022년도 사업과 비교하여 6개 매체가 줄었지만, 최초 수집된 기사는 조건을 충족하였다. 수집된 데이터의 어절 수 집계는 정제가 되기 전 불필요한 정보를 가지고 있는 데이터를 대상으로 한 것이다. 어절 수는 문장의 공백과 줄바꿈 수로 집계하였다.

매체별 기사와 어절 수에 관한 내용을 정리하면 다음과 같다.



<그림 2> 연도별 매체 비율과 상위, 하위 기사수 매체

매체명	기사 수	어절 수	매체명	기사 수	어절 수
강원일보	51,789	6,339,133	세계일보	141,769	36,728,629
경기일보	21,710	5,100,189	스포츠서울	98,657	13,835,337
경북일보	27,998	5,794,014	아시아경제	167,293	37,475,022
경향신문	74,506	14,170,331	아주경제	78,524	21,269,806
국민일보	81,025	20,900,800	이데일리	183,808	34,917,001
남도일보	37,596	8,081,081	이투데이	106,781	21,317,923
내일신문	27,468	8,562,816	전북도민일보	36,083	6,351,528
노컷뉴스	130,429	32,528,843	조선일보	49,847	11,044,217
뉴스핌	266,726	42,399,013	중도일보	65,607	12,210,837
대구신문	33,513	7,397,030	중부일보	41,774	8,779,705
머니투데이	137,207	32,131,262	충청일보	62,620	9,727,287
부산일보	72,615	17,197,990	한겨레	45,599	15,069,584
서울경제	134,620	31,818,165	한국일보	65,335	18,162,483
서울신문	98,431	21,770,907	헤럴드경제	148,029	36,799,518
계			총 합	2,487,359	537,880,451

<표 3> 최초 수집 기사와 어절 수

## 가. 원시 데이터 특징 분석

한국언론진흥재단에서는 데이터를 제공할 때, 2022년도까지는 엑스엠엘(XML) 데이터로 제공하였으나, 올해에는 제이슨(JSON) 파일로 제공하였으며, 한 기사가 하나의 제이슨(JSON) 파일로 되어 있다.

제이슨 파일로 변환하는 과정에서 원 매체의 태그가 제대로 정제되지 않고 남아 있는 경우, “??????...” 와 같이 정체를 알 수 없는 불특정 문자가 본문에 포함되거나, 데이터가 소실되는 문제 등 원시 데이터에 문제가 있는 것을 확인하였다.

```
"content": "포스코ICT는 성남지역 초등학교들을 대상으로 '인공지능(AI) 로봇 챌린지 스쿨 프로그램'을 운영한다고 21일 밝혔다. 이 프로그램은 동작으로 작동하는 AI 로봇을 제작하고 이 로봇으로 축구 게임을 하는 활동이다. \n\n 이를 위해 포스코ICT 직원들로 구성된 'AI 로봇 봉사단'은 지난 20일 포스코ICT 관교 사옥에서 성남지역 초등학교들과 함께 모션을 통해 제어되는 AI 축구 로봇을 제작했다. 프로그램을 통해 학생들이 직접 만든 AI 축구로봇으로 팀을 나눠 축구경기를 진행했다. \n\n 포스코ICT 기업시민사무국 김지연 프로는 "AI, 로봇 등 소프트웨어와 하드웨어를 결합한 IT 융합 교육의 중요성은 강조되고 있지만, 실질적으로 제대로 된 교육을 받거나 체험할 수 있는 기회는 부족한 상황"이라며 "이러한 교육의 사각지대에 놓인 아동들에게 관련 프로그램을 제공하고 로봇과 AI에 대한 꿈을 키울 수 있도록 하기 위해 프로그램을 기획했다"고 말했다. \n\n 이외에도 포스코ICT는 포스코그룹 임직원 모두가 함께 참여하는 봉사주간인 '글로벌 모범시민 위크(6월14일~6월25일)'를 맞아 다양한 봉사활동을 펼치고 있다. 화귀식물 '히어리' 보존활동, 취약계층을 위한 전기수리봉사, 취약계층 아동대상 코딩 교육, 수중 멸종 위기종 보호를 위한 스킨스쿠버 활동 등 각 지역별로 봉사활동을 진행 중이다. \n\n 박세정 기자 \n\n {\\"mean\\":\"[\" <dicwordclass style=\\\"\"user-select: text:\\\"\" >strong> AI </strong> 미국.영국[?e? ?a?] <br> artificial insensation <br> artificial intelligence <br> </dicwordclass> \",\" <dicwordclass style=\\\"\"user-select: text:\\\"\" >strong> ai </strong> 미국.영국[ái] <br> 아아 ((고통·슬픔·연민 등을 나타냄)) <br> </dicwordclass> \",\" <dicwordclass style=\\\"\"user-select: text:\\\"\" >strong> ai </strong> 미국.영국[?i:] <br> 동물 세발가락나 무늬보 ((중남미산(産))) <br> </dicwordclass> \",\" <dicwordclass style=\\\"\"user-select: text:\\\"\" >artificial intelligence <br> 컴
```



민사항을 담은 영상을 KB손해보험 공식 유튜브 채널에서 시청할 수 있다.?????? \n \n ??????? \n \n \n \n cook@heraldcorp.com",

<표 5> 원시 데이터 특징 예시(불특정 문자 포함)

"현대건설, 라틴파이낸스 선정 '올해의 딜' 구조화 금융부문 수상",

"content": "[아시아경제 조강욱 기자] 현대건설은 미국 매체 라틴파이낸스(LatinFinance)가 선정한 '2021 올해의 딜(Deal of the year)'에서 국내 건설사 최초로 구조화 금융 부문상을 수상했다고 14일 밝혔다. \n \n라틴파이낸스는 1988년 미국 뉴욕과 마이애미에서 창간된 중남미 및 카리브해 지역의 경제와 금융 시장에 대한 대표 매체다. 매년 라틴 아메리카 및 카리브해 지역 자본 시장에서 이뤄진 기념비적인 딜(Deal)을 선정해 발표하고 있다. \n \n현대건설은 파나마 메트로 3호선 사업과 관련해 지난해 7월 체결한 20억 달러 규모의 중장기 금융약정이 해당 언론사의 올해의 딜에 선정돼 이번 상을 받았다. 라틴파이낸스는 수상 선정배경으로 △파나마 인프라 사업 역사상 가장 큰 규모의 딜이었다는 부분 △아시아, 유럽, 미국 등 세계 각국의 역량 있는 금융기관들이 참여해 구조화시킨 금융이라는 점 △동 사업이 파나마에 미친 사회·환경적인 영향이 평가에 주요했다고 설명했다. \n \n파나마 메트로 3호선 건설 사업은 파나마 수도 파나마시티와 서부 아라이잔 지역 연결을 위해 총 25km의 고가철로(모노레일)와 13개 역사, 1개 차량기지를 건설하는 총 28억 달러 口瓚".

[아시아경제 조강욱 기자] 현대건설은 미국 매체 라틴파이낸스(LatinFinance)가 선정한 '2021 올해의 딜(Deal of the year)'에서 국내 건설사 최초로 구조화 금융 부문상을 수상했다고 14일 밝혔다.

라틴파이낸스는 1988년 미국 뉴욕과 마이애미에서 창간된 중남미 및 카리브해 지역의 경제와 금융 시장에 대한 대표 매체다. 매년 라틴 아메리카 및 카리브해 지역 자본 시장에서 이뤄진 기념비적인 딜(Deal)을 선정해 발표하고 있다.

현대건설은 파나마 메트로 3호선 사업과 관련해 지난해 7월 체결한 20억 달러 규모의 중장기 금융약정이 해당 언론사의 올해의 딜에 선정돼 이번 상을 받았다. 라틴파이낸스는 수상 선정배경으로 △파나마 인프라 사업 역사상 가장 큰 규모의 딜이었다는 부분 △아시아, 유럽, 미국 등 세계 각국의 역량 있는 금융기관들이 참여해 구조화시킨 금융이라는 점 △동 사업이 파나마에 미친 사회·환경적인 영향이 평가에 주요했다고 설명했다.

파나마 메트로 3호선 건설 사업은 파나마 수도 파나마시티와 서부 아라이잔 지역 연결을 위해 총 25km의 고가철로(모노레일)와 13개 역사, 1개 차량기지를 건설하는 총 28억 달러 口瓚의 파나마 정부 최대 규모 인프라 사업이다. 해당 사업으로 증진하는 지역 교통체증을 해소하고 연간 2만 톤 상당의 이산화탄소 배출 감축 효과가 있을 것으로 기대된다.

밀줄 뒤 내용 데이터 소실

현대건설은 지난 2019년 포스코건설, 현대엔지니어링과 컨소시엄을 구성해 프로젝트에 입찰해 2020년 2월 기술, 상업, 금융 전 부문에서 최고점을 획득하며 최종 수주했다. 프로젝트 수주를 위한 금융경쟁력 제고를 위해 한국수출입은행과 한국무역보험공사에서는 '중장기 수출채권 매입' 제도를 신설하고, 최초로 현대건설 컨소시엄을 지원해 '팀 코리아(Team Korea)'를 이뤘다. 중장기 수출채권 매입제도는 수출자가 발주처로부터 발급받은 수출채권을 '무소구(non-resource)' 조건으로 매입해, 수출자에게 자금을 공급하고 3~7년 후 발주처로부터 해당 자금을 회수하는 방식으로 한국 기업의 중남미 진출을 적극 지원한다. 무소구 조건은 수입자의 지급 불이행에 대해 수출채권을 매입한 금융기관이 수출자에게 상환청구 요청을 하지 않는 것을 말한다. 이어 지난해 7월 29일에는 파나마 국책은행을 비롯한 10여 개의 글로벌 은행들이 참여한 총 9년간 27억 달러 규모의 '건설대금 지급 확약서'를 매입하는 20억 달러 규모 중장기 금융약정을 체결했다.

<표 6> 원시 데이터 특징 예시(데이터 소실 1)

"근대기 대구 문화예술인들의 활동상을 알 수 있는 자료들이 대거 대구시에 기증됐다. 일제강점기 작사가이자 아동문학가로 활약한 윤복진(1907~1991년)의 유족이 소장했던 자료들이다. \n\n'가을밤 외로운 밤, 별레우는 밤~(후략)'으로 시작하는 동요 '가을밤'의 원작이 '울밑에 귀뚜라미 우는 달밤에, 기럭기럭 기러기 날아갑니다'로 시작하는 윤복진의 '기러기'(1929년)이다. '가을인가 가을인가/아 가을인가 봐~(후략)'의 동요 '아 가을인가'도 윤복진의 노랫말이다. \n\n이번 기증 자료에는 육필 노트, 필사 악보 등을 비롯해, 박태준 작곡, 윤복진 작사, 이인성 표지화로 만든\_(1939년), 윤복진이 1929년 펴낸 , (1931년), (1933년) 등 1920~40년대 악보집들이 다수 포함돼 있다. 그 중 \_은 1920년대 이름난 작사·작곡가들의 곡 35곡이 수록됐고 그간 내용이 공개되지 않았던 귀한 자료다. \n\n또 1938년 대구공회당에서 열린 제1회 신인가수선발콩쿠르 결선 프로그램(박태준, 윤복진 심사위원) 등의 공연 팸플릿과 '어린이', '아동', '음악평론' 등의 잡지, 무영당 광고지 등 당대 문화예술계 상황을 알 수 있는 다양한 자료들을 확인할 수 있다. \n\n특히 1936년 발행된 우리나라 최초의 월간 음악 평론잡지 '음악평론'4월호(윤복진 평론 게재),

이번 기증 자료에는 육필 노트, 필사 악보 등을 비롯해, 박태준 작곡, 윤복진 작사, 이인성 표지화로 만든 <물새발 자옥>(1939년), 윤복진이 1929년 펴낸 <동요곡보집>, <초등동요유희집>(1931년), <현재명작곡집>(1933년) 등 1920~40년대 악보집들이 다수 포함돼 있다. 그 중 <동요곡보집>은 1920년대 이름난 작사·작곡가들의 곡 35곡이 수록됐고 그간 내용이 공개되지 않았던 귀한 자료다.

또 1938년 대구공회당에서 열린 제1회 신인가수선발콩쿠르 결선 프로그램(박태준, 윤복진 심사위원) 등의 공연 팸플릿과 '어린이', '아동', '음악평론' 등의 잡지, 무영당 광고지 등 당대 문화예술계 상황을 알 수 있는 다양한 자료들을 확인할 수 있다.

특히 1936년 발행된 우리나라 최초의 월간 음악 평론잡지 '음악평론'4월호(윤복진 평론 게재), 1946년 창간된 아동잡지 '아동' 창간호(윤복진 동요 수록)와 최남선의 '백팔번뇌'(1926년) 등의 초판본 도서들과 대구 출신 영화감독 이규환이 해방 후 제작한 영화 <돌돌이의 모험> 시나리오도 찾아볼 수 있다. 대부분의 자료에는 윤복진의 친필 사인이 적혀 있으며 그 외 윤복진의 습작 과정을 살펴볼 수 있는 친필 노트들도 다수 포함돼 있다.

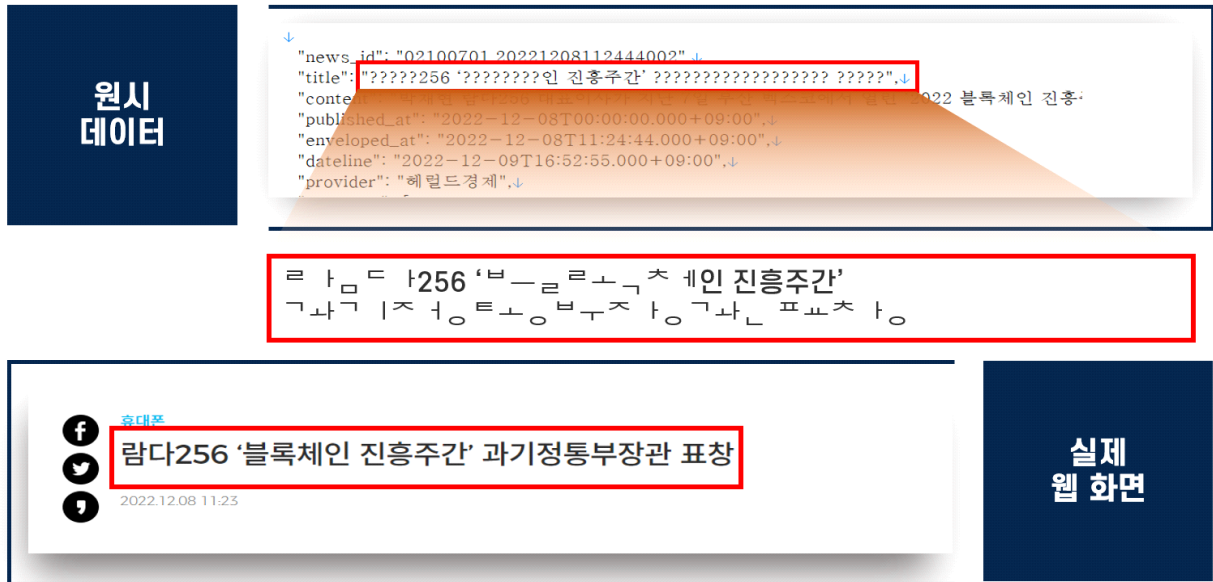
<표 7> 원시 데이터 특징 예시(데이터 소실 2)

• 하나의 기사를 한 개의 JSON 파일로 제공함.
• &lt; &gt; 등의 부호가 그대로 남아 있음.
• 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 가운데점, 마침표, 쉼표 등이 여러 가지 코드로 일관성 없이 사용됨.
• 이(李), 리(李)와 같은 한자 호환용 코드가 달리 사용되어 데이터의 공유와 유통에 문제를 일으킴.

<표 8> 데이터 특징

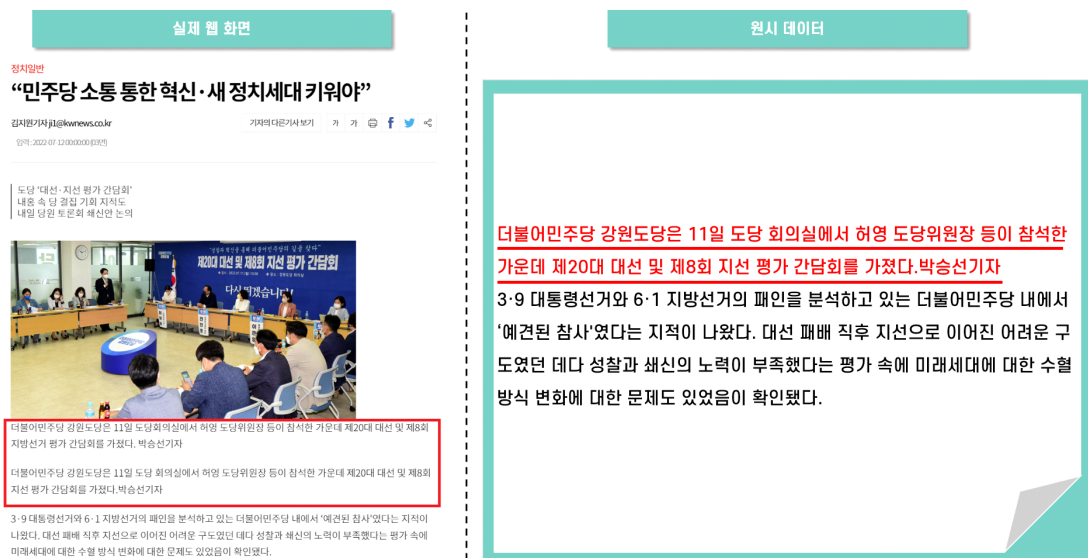
원시 데이터는 위와 같은 특징을 가지고 있다. 세심하게 데이터를 확인하지 않고 원시 데이터를 그대로 사용하게 된다면 오류를 가지고 있는 데이터를 생성할 확률이 높다. 데이터 오류의 유형과 처리되는 공정은 다음과 같다.





<그림 3> 오류 유형 ①: 글자 깨짐

원시 데이터 중 일부 자료에서는 조합형으로 작성된 경우 제목이 깨지는 경우가 발견되었다. 이러한 오류는 해당 부분의 웹 사이트를 확인 후, 원래 제목으로 수정해 주는 방식으로 처리하였다.



<그림 4> 오류 유형 ②: 캡션 정보와 본문이 구분되지 않는 경우

또한, 일부 자료에서는 캡션 정보가 본문과 구분되지 않는 경우가 존재하였다. 캡션 정보가 본문의 일부인 것처럼 나타나는데, 캡션 정보는 불필요한 요소로 데이터 정제 과정에서 삭제 대상이다. 이러한 유형은 기사 원문을 확인하면서 캡션 정보를 삭제하는 방식으로 처리하였다.

#### 실제 웹 화면



[아시아경제 세종=이동우 기자] 앞으로 고철과 폐유리 등 폐기물의 재활용 규제가 대폭 개선된다. 폐비닐과 폐플라스틱에서 추출한 열분해유를 플라스틱 원료 생산에도 사용할 수 있게 된다. 이로써 연간 2100억원의 폐기물 처리비용을 절감할 수 있을 것으로 전망된다.

환경부는 26일 대구 성서산업단지 내 아진엑스텍에서 열린 제1회 규제혁신전략회의에서 이 같은 내용이 담긴 '환경규제 혁신 방안'을 윤석열 대통령에게 보고했다. 환경 분야 규제 혁신은 기존 환경관련 규제 합리화와 탄소중립·순환경제 등 환경정책 목표 관련 규제 개선에 중점을 뒀다.

#### 달린 규제에서 열린 규제로 전환

환경부는 이번 규제혁신 방안을 크게 4가지다. 미리 정해놓은 금지행위 외 행위를 모두 허용하는 '네거티브'(열린) 규제 전환을 비롯해 위험에 비례하는 차등적 규제 전환, 쌍방향 소통 및 협의형 규제 전환, 탄소중립·순환경제 직결 규제 우선 개선 등이다.

#### 원시 데이터

환경부는 26일 대구 성서산업단지 내 아진엑스텍에서 열린 제1회 규제혁신전략회의에서 이 같은 내용이 담긴 '환경규제 혁신 방안'을 윤석열 대통령에게 보고했다. 환경 분야 규제 혁신은 기존 환경관련 규제 합리화와 탄소중립·순환경제 등 환경정책 목표 관련 규제 개선에 중점을 뒀다. **달린 규제에서 열린 규제로 전환**

환경부는 이번 규제혁신 방안을 크게 4가지다. 미리 정해놓은 금지행위 외 행위를 모두 허용하는 '네거티브'(열린) 규제 전환을 비롯해 위험에 비례하는 차등적 규제 전환, 쌍방향 소통 및 협의형 규제 전환, 탄소중립·순환경제 직결 규제 우선 개선 등이다.

#### <그림 5> 오류 유형 ③: 중간 제목이 본문 사이에 들어간 경우

기사의 중간에 있는 소제목과 그다음 단락이 붙어 버리는 오류도 존재한다. 이 경우에는 작업자가 기사를 정독하면서 확인하고, 해당 기사의 인터넷 사이트 주소(URL)를 참조하지 않으면 발견하기 어렵다.

위의 오류 유형 ①, ②, ③은 웹 페이지 데이터를 참조하여 작업을 진행하였다. 원시 데이터에는 없는 정보가 웹 페이지에서는 표기되어 있어 이를 활용하여 오류를 바로잡을 수 있다.



#### 웹에서 확인한 실제 기사 내용

기상청에 따르면 올해 **분얼기** 강수량은 지난해 293.3mm의 2/3 수준인 195.9mm에 불과했다.

**분얼기**  
요가 최근한 다음부터 영양생장기까지 분  
얼이 이루어지는 시기

부터 8월 상순까지는 강수량 부족에 일조시간 부족까지 겹쳐 1㎡당  
25개보다 1308개 적은 2만 9417개에 그쳤다.

올해 벼 재배면적은 72만 7158헥타르로, 지난해보다 0.7% 줄었다. 통계청은 벼 이외 작물 재배  
지원 등 정부의 벼 재배면적 조정 정책과 쌀값 하락 등에 따른 결과로 분석했다.

#### 한국언론진흥재단으로부터 받은 데이터 내용

이와 관련해 통계청은 "분얼기 즉, 벼가 가지를 치는 6월 상순부터 7월 상순까지 강수량이 부족해 재배면적 1㎡당 이삭  
수가 지난해 22.5개에서 올해 20.9개로 1.6개 감소했다"고 설명했다.

기상청에 따르면 올해

**분얼기요가 최근한 다음부터 영양생장기까지 분얼이 이루어지는 시기\***

강수량은 지난해 293.3mm의 2/3 수준인 195.9mm에 불과했다.

또, 낱알이 형성되는 7월 상순부터 8월 상순까지는 강수량 부족에 일조시간 부족까지 겹쳐 1㎡당 완전 낱알 수가 지난해  
3만 725개보다 1308개 적은 2만 9417개에 그쳤다.

올해 벼 재배면적은 72만 7158헥타르로, 지난해보다 0.7% 줄었다. 통계청은 벼 이외 작물 재배 지원 등 정부의 벼 재배  
면적 조정 정책과 쌀값 하락 등에 따른 결과로 분석했다.



#### 정제 데이터

이와 관련해 통계청은 "분얼기 즉, 벼가 가지를 치는 6월 상순부터 7월 상순까지 강수량이 부족해 재배면적 1㎡당 이삭  
수가 지난해 22.5개에서 올해 20.9개로 1.6개 감소했다"고 설명했다.

기상청에 따르면 올해 **분얼기** 강수량은 지난해 293.3mm의 2/3 수준인 195.9mm에 불과했다.

또, 낱알이 형성되는 7월 상순부터 8월 상순까지는 강수량 부족에 일조시간 부족까지 겹쳐 1㎡당 완전 낱알 수가 지난해  
3만 725개보다 1308개 적은 2만 9417개에 그쳤다.

올해 벼 재배면적은 72만 7158헥타르로, 지난해보다 0.7% 줄었다. 통계청은 벼 이외 작물 재배 지원 등 정부의 벼 재배  
면적 조정 정책과 쌀값 하락 등에 따른 결과로 분석했다.

<그림 6> 오류 유형 ④-1: 웹 페이지와는 다르게 기사 내용이 변형됨

웹 페이지에서는 문제가 없으나 원시 데이터에만 생기는 오류도 있다. 위의 그림에서 볼 수  
있듯, 웹 페이지에서는 '분얼기'라는 단어의 뜻을 하이퍼링크로 연결해 안내하고 있는데, 구매  
한 원시 데이터에는 단어의 설명이 문장 사이에 위치하면서 문장이 강제로 줄바꿈되는 오류가  
발생하였다. 이 경우는 데이터로 변형하면서 하이퍼링크의 내용을 본문으로 인식하여 생긴 오  
류이므로 문장 사이에 들어간 단어 설명을 삭제하고, 강제로 줄바꿈된 문장을 수정해 주는 방  
식으로 처리하였다.

기사 내용이 변형됨	
1 기사 내용이 끊김	
<p><b>실제 웹 화면</b></p> <p>소셜미디어 상에는 인근 해역에서 같은 임무를 수행하던 영국 해군 경비함 '스페이'호 역시 같은 방식으로 솔로몬제도에 접근하지 못했다는 소문이 퍼지고 있다. 이에 영국 해군 대변인은 보안상 이유를 들며 소문에 대해 구체적인 언급을 하지 않았다고 외신은 전했다.</p> <p>솔로몬제도의 머내시 소가바레 총리는 2019년 대만과 단교하고 중국과 국교를 수립하는 등 친중국 행보를 지속하고 있다. 올해 4월에는 중국이 솔로몬제도에 군 병력과 군함을 파견할 수 있도록 하는 내용이 있는 안보협정을 맺었다.</p>	<p><b>원시 데이터</b></p> <p>소셜미디어 상에는 인근 해역에서 같은 임무를 수행하던 영국 해군 경비함 '스페이'호 역시 같은 방식으로 솔로몬제도에 접근하지 못했다는 소문이 퍼지고 있다. 이에 영국 해군 대변인은 보안상 이유를 들며 소문에 대해 구체적인 언급을 하지 않았다고 외신은 전했다.</p> <p>솔로몬제도의 머내시 소가바레 총리는 2019년 대만과 단교하고 중국과 국교를 수립하는 등 친중국 행보를 지속하고 있다. <b>올해 4월에는 중국이 솔로</b> <b>몬</b></p>
2 기사 정보가 기사 중간에 삽입됨	
<p><b>실제 웹 화면</b></p> <p>강릉시는 유전자증폭(PCR) 검사와 신속항원검사 장소를 이원화하는 방안 등을 놓고 세부 내용을 조율 중인 것으로 알려졌다. 속초시는 다중 이용 시설에 대한 방역 상황 특별 점검을 진행 중이다. 도내 추모공원들도 성묘객들의 방문을 막기 위해 추모시간을 제한하는 등 온라인 성묘를 독려하고 있다.</p> <p>한편 오미크론 감염 사례가 지속적으로 증가하는 가운데 설 연휴를 앞두고 높은 수준의 확진자 발생이 이어졌</p>	<p><b>원시 데이터</b></p> <p>강릉시는 유전자증폭(PCR) 검사와 신속항원검사 장소를 이원화하는 방안 등을 놓고 세부 내용을 조율 중인 것으로 알려졌다. 속초시는 다중 이용 시설에 대한 방역 상황 특별 점검을 진행 중이다. 도내 추모공원들도 성묘객들의 방문을 막기 위해 추모시간을 제한하는 등 온라인 성묘를 독려하고 있다.</p> <p><b>김도균·권순찬</b> <b>(한편)</b> 오미크론 감염 사례가 지속적으로 증가하는 가운데 설 연휴를 앞두고 높은 수준의 확진자 발생이 이어졌다.</p>

<그림 7> 오류 유형 ④-2: 웹 페이지와는 다르게 기사 내용이 변형됨

위의 오류는 웹 사이트에서는 기사가 정상적으로 이어지고 있으나 원시 데이터에서는 기사가 중간에 끊긴 경우이다. 이때 어떠한 오류로 인해 변형이 발생했는지 확인이 어렵고, 작업자가 모든 기사를 웹 사이트와 비교하여 수정하기가 쉽지 않으므로 데이터 오류를 최소한으로 줄이기 위해 해당 기사를 사용하지 않았다.

## 기사 사용 X

삼성물산은 지난 2020년 10월 '탈석탄 선언'의 연장선에서 이사회를 중심으로 탄소중립 필요성에 대한 공감대를 형성하고 체계적 이행 방안을 추진할 것을

문장의 뒷 부분이 삭제됨

삼성물산은 2030년까지 전 사업장 재생에너지 사용을 달성하기 위해 재생에너지 공급 여건이 양호한 해외 사업장에서 이를 우선 추진하고 국내 사업장에서도 재생에너지 사용을 점진적으로 확대해 나갈 계획이다.

## 데이터 수정

한편 청해부대에서 확진자 1명이 추가되면서 총 확진자가 271명으로

늘었다. 22일 국방부에 따르면 지난 20일 입국 후 유전자증폭(PCR) 검사에서 음성 판정을 받은 뒤 1인 격리 중이던 청해부대 병사 1명이 증상 발현으로 다시 진단검사를 받아 확진됐다.

서울=이무현기자


<그림 8> 오류 유형 ⑤: 문장이 임의로 줄바꿈되어 있는 경우

원시 데이터에는 문장이 이유 없이 잘라내어져 있는 기사들이 존재한다. 이러한 기사는 데이터를 직접 보면서 확인하고 수정한다. 문장이 끊겨 뒤의 내용이 사라진 기사는 사용하지 않으며, 단순히 잘라내어져 있는 기사는 수정하여 이어 붙여 준다.

HOME > 오피니언 > 의료칼럼

## 2022년, '슈뢰딩거의 백신'의 현실을 호시(호랑이의 눈)로

© 승인 2022.01.02 19:50



**캡션 정보**

A씨는 어머니가 코로나19에 감염되어 재택 치료 중 증상 악화로 며칠씩 병실을 구하지 못하여 발을 동동 굴렀다. 어머니가 어렵사리 입원하였으나 사흘 만에 증상 악화로 중환자실로 옮겼다. 일주일 후에 병원에서 전담 중환자실에서 일반 중환자실로 전원해야 한다는 연락이 왔다. 증상이 발생하고 20일이 지난 환자는 강제로 전원한다는 내용이다. 옮기지 않으면 중환자실 입원비는 A씨가 내야 하며 과태료도 내야 한다는 것이다. 교수님은 "병실을 옮길 때 작동 중인 예크모나 고정형 인공호흡기 등 장비를 떼야 하는데, 그 과정에서 증상 악화나 사망의 위험성이 있다"라는 설명을 했다. 코

자영업자, 백신을 맞았으나 스마트 폰이 없는 어르신, 청소년 코로나 백신 패스 등, 이런 상황을 만든 정부는 국민들의 신뢰를 잃는 것이 당연하다. 국민의 신뢰를 잃은 정부라면 이런 난감한 상황을 풀기 어렵다. Wn 2022년 검은 호랑이해를 맞아서 우리들 스스로 호랑이의 눈(호시탐탐이 여기에서 나왔다)으로 무능한 정부의 정책을 살펴야 할 것이다. 언제든지 어디든지 보고 싶은 사람을 함께 만나서 마스크를 벗고 정다운 이야기를 나눌 수 있는 올해를 맞이하기 위하여...

"published\_at": "2022-01-02T00:00:00.000+09:00",  
"enveloped\_at": "2022-01-02T19:53:35.000+09:00",  
"dateline": "2022-01-02T19:50:07.000+09:00",  
"provider": "대구신문",  
"category": [  
"사회>의료 건강"  
"사회>여성"  
],  
"category\_incident": [],

```

<doc>↓
<newsitemid>01501001.20220102195335001</newsitemid>↓
<HeadLine>2022년, '슈뢰딩거의 백신'의 현실을 호시(호랑이의 눈)로</HeadLine>↓
<url>https://www.idaegu.co.kr/news/articleView.html?idxno=368472</url>↓
<used></used>↓
<byline>윤덕우</byline>↓
<content>↓
A씨는 어머니가 코로나19에 감염되어 재택 치료 중 증상 악화로 며칠씩 병실을 구하지 못하여 발을 동동 굴렀다. 어머니가 어렵사리 입원하였으나 사흘 만에 증상 악화로 중환자실로 옮겼다. 일주일 후에 병원에서 전담 중환자실에서 일반 중환자실로 전원해야 한다는 연락이 왔다. 증상이 발생하고 20일이 지난 환자는 강제로 전원한다는 내용이다. 옮기지 않으면 중환자실 입원비는 A씨가 내야 하며 과태료도 내야 한다는 것이다. 교수님은 "병실을 옮길 때 작동 중인 예크모나 고정형 인공호흡기 등 장비를 떼야 하는데, 그 과정에서 증상 악화나 사망의 위험성이 있다"라는 설명을 했다. 코로나로 경제적으로 힘든 A씨는 어머니를 코로나 전담 중환자실에서 옮기자니 어머니의 상태가 걱정되고 계속 있자니 얼마마한 치료비가 걱정되어 이리저리 저러지도 못하고 있다. 이런 지침을 내린 정부가 야속하고 '나는 대한민국 국민인가.' 하는 의심이 들었다.↓

```

**외부 기고가 작성한 기사이나  
기자 이름으로 기입됨**

<그림 9> 오류 유형 ⑥: 외부 기고가 정보가 삭제되어 원시 데이터에 없는 경우

몇몇 신문의 경우 기자 정보만으로 외부 기고문 여부를 알 수 없다. 외부인이 언론사의 기사 입력 시스템에 직접 접속할 수 없으므로 기고문을 기자가 받아서 올렸기 때문이다. 보통 이런 경우 외부 기고문의 기사 하단부에 존재하는 외부 기고가 정보로 기사를 걸러낼 수 있었다. 하지만 대구신문의 경우 원시 데이터에서도 정보가 기입되어 있지 않아 해당 내용을 알 수 없었으며 캡션 정보마저 사라져 해당 데이터만 보고서 이 글이 외부 기고가의 글인지 구별할 수 없었다. 따라서 이를 해결하기 위해 전체 기사의 사이트 주소(url)를 추출하여 엑스엠엘(xml) 파일에서 카테고리 정보를 얻어 외부 기고가의 글을 제거하였다.

웹  
데이터

## 펭수, 신임 EBS 사장에 '꼰대짓' "잘 할 수 있습니<sup>ㅈ</sup>?"



▲ 사진=EBS '자이언트 펭tv' 캡처

EBS 인기 캐릭터 겸 크리에이터 펭수가 김명중 전 EBS 사장 퇴임식에 참석한 가운데 신임 EBS 사장에 '꼰수짓'을 펼쳤다.



평소에 쓰이지 않는 단어가 깨짐

원시  
데이터

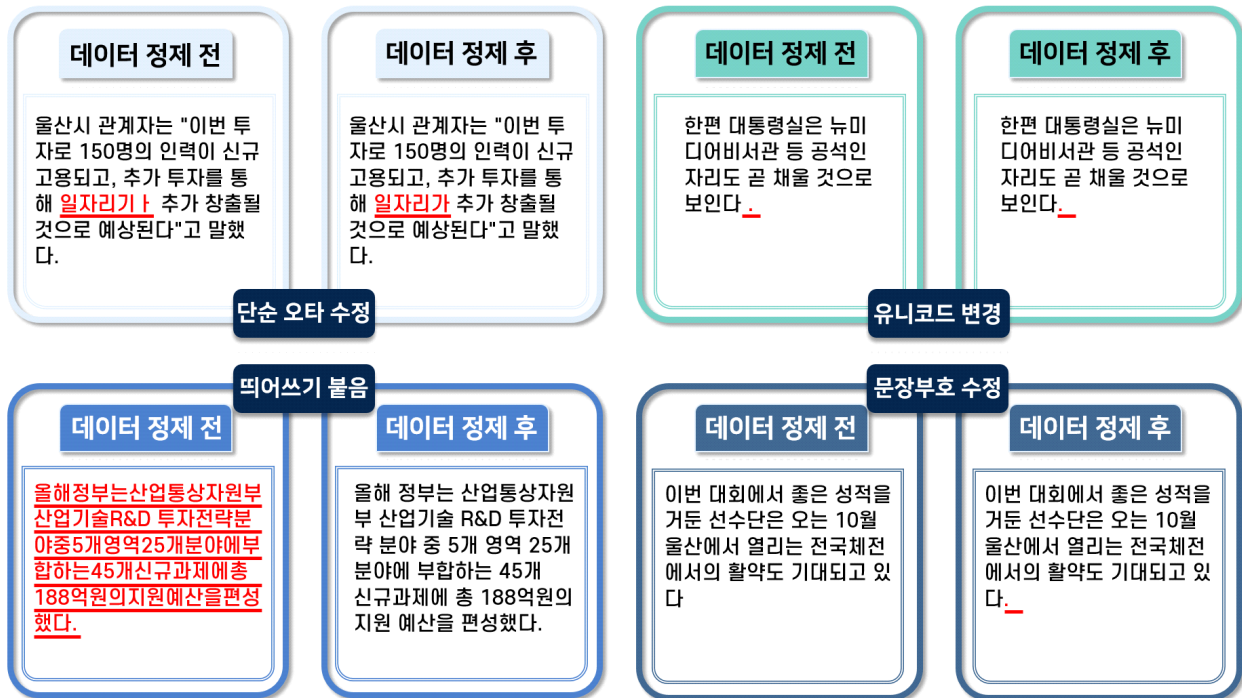
"title": "펭수, 신임 EBS 사장에 '꼰대짓' "잘 할 수 있습니<sup>ㅈ</sup>?",  
"content": "사진=EBS '자이언트 펭tv' 캡처 Wn EBS 인기 캐릭터 겸 크리에이터 펭수가 김명중 전 EBS 사장 퇴임식에 참석한 가운데 신임 EBS 사장에 '꼰수짓'을 펼쳤다. Wn Wn 22일 '자이언트 펭tv' 유튜브 채널에는 '김명중과 헤어졌습니다..."

정제  
데이터

"title": "펭수, 신임 EBS 사장에 '꼰대짓' "잘 할 수 있습니<sup>ㅈ</sup>?",  
"content": "사진=EBS '자이언트 펭tv' 캡처 Wn EBS 인기 캐릭터 겸 크리에이터 펭수가 김명중 전 EBS 사장 퇴임식에 참석한 가운데 신임 EBS 사장에 '꼰수짓'을 펼쳤다. Wn Wn 22일 '자이언트 펭tv' 유튜브 채널에는 '김명중과 헤어졌습니다..."

<그림 10> 오류 유형 ⑦: 평소에 쓰이지 않는 음절이 깨지면서 ?로 치환되는 현상


평소에 쓰이지 않는 단어가 제이슨(JSON)으로 변경되면서 단어가 깨지는 현상이 발견되었다. 위 기사의 제목에서 사용된 'ㅈ'은 의도적인 표기이지만, 원시 데이터에서는 글자가 깨지면서 ?로 치환된다. 이러한 오류는 정규식을 이용해 해당 오류를 검색 후, 웹 데이터와 비교해 보고 본래의 글자로 수정한다.



<그림 11> 오류 유형 ⑧: 오류 수정

작업자가 작업을 진행하면서 발견한 오류는 수정하거나 삭제하는 방향으로 진행하게 된다. 오류 수정의 예를 들자면 기자가 작성하면서 생긴 단순 오타는 올바른 단어로 수정해 준다. 띄어쓰기가 없이 전부 붙은 문장은 띄어쓰기를 넣어 주며, 문장 부호가 없는 문장의 끝에는 적절한 문장 부호를 넣어 준다. 또한, 원시 데이터에는 가운뎃점, 쉼표, 마침표 등이 여러 가지 코드로 일관성 없이 사용된 걸 볼 수 있는데, 표준 코드로 통일해 준다. 위와 같은 오류들은 수정하지 않고 삭제하거나 놔두는 방향으로 진행한다면 기사의 내용이 변질되어 기자의 의도와는 다른 잘못된 방향으로 기사를 인식할 수 있으므로 해당 오류를 찾아 수정해야 한다.



<p><b>불필요한 부호 사용 삭제(용어 설명 없음)</b></p> <p>수소경제 활성화를 위해 제정된 '수소경제 육성 및 수소안전관리에 관한 법률'에 따라 올해 2월부터 안전관리분야 시행됨에 따라 수소용품*에 대해 제조허가, 등록제도 및 안전검사를 실시하고 있다. (기사 내 *는 있으나 *에 관한 용어 설명이 없음.)</p> <p>수소용품 검사지원센터가 본격 운영되면 전북도에 수소용품 관련 기업이 집적화되고, 유동 인구 증가로 지역경제 활성화에 기여할 것으로 기대된다.</p>	<p><b>전문 인용 부분 삭제</b></p> <p>이어 "지난 12일 뷔와 다른 멤버들 간 접촉이 있었으나, 모두 마스크를 착용한 상태였고 밀접한 수준의 접촉은 없었다"라면서 "뷔를 제외한 방탄소년단 멤버들은 현재 특별한 증세는 없으며, 선제적으로 자가진단키트 검사 결과 모두 음성 판정을 받았다"라고 설명했다.</p> <p>방송가에서는 전현무, 김성주 등도 코로나에 확진됐다.</p>  <p>BTS 팬 커뮤니티에 올라온 글 전문</p> <p>안녕하세요. 빅히트 뮤직입니다.</p> <p>방탄소년단 멤버 뷔의 코로나19 확진 관련에 안내드립니다.</p>
<p>울진·익산산단 중심으로 이차전지 생산 중심지 조성 소상공인 코로나 피해 지원 지속...출산장려금 대폭 확대 망덕포구~배알도~근린공원 잇는 해상로드 연대 완성 목표</p> <p>▶ 3부 전남 동부권 6개 시·군 지자체 구상</p> <p>① 순천시, "한 뱀도 정원" 30만 정원도시 초석의 해로 ② 여수시, 여순사건 특별법 시행 첫해...섬박람회 준비도 ③ 광양시, 일상회복 넘어 미래산업 육성의 원년으로 (계속)</p> <p>전남 광양시는 임인년 새해를 일상으로의 복귀를 넘어 미래 신산업 육성의 원년으로 삼고 새로운 성장 시대를 열어가길 방침입니다.</p> <p><b>불필요한 글목록 삭제</b></p>	<p>박희영 용산구청장은 "위드 코로나로 조금씩 경기가 회복되는 상황이지만 하나 긴장의 끈을 놓을 수가 없다"면서 "지금까지 고통을 감내한 소상공인, 중소기업을 위해 구에서 할 수 있는 최대한의 지원을 아끼아겠다"고 말했다.</p> <p>2022년 9월 현재 용산구 중소기업육성기금 총액은 396억원이며, 423개 업체에 155억원을 지원 중이다.</p> <p>◆용산구 인사 &lt;4급 승진&gt; ▲문화환경국장 신동기 &lt;5급 전보&gt; ▲주차관리과장 김재두 &lt;10월1일자&gt;</p> <p><b>인사 명단 삭제</b></p>

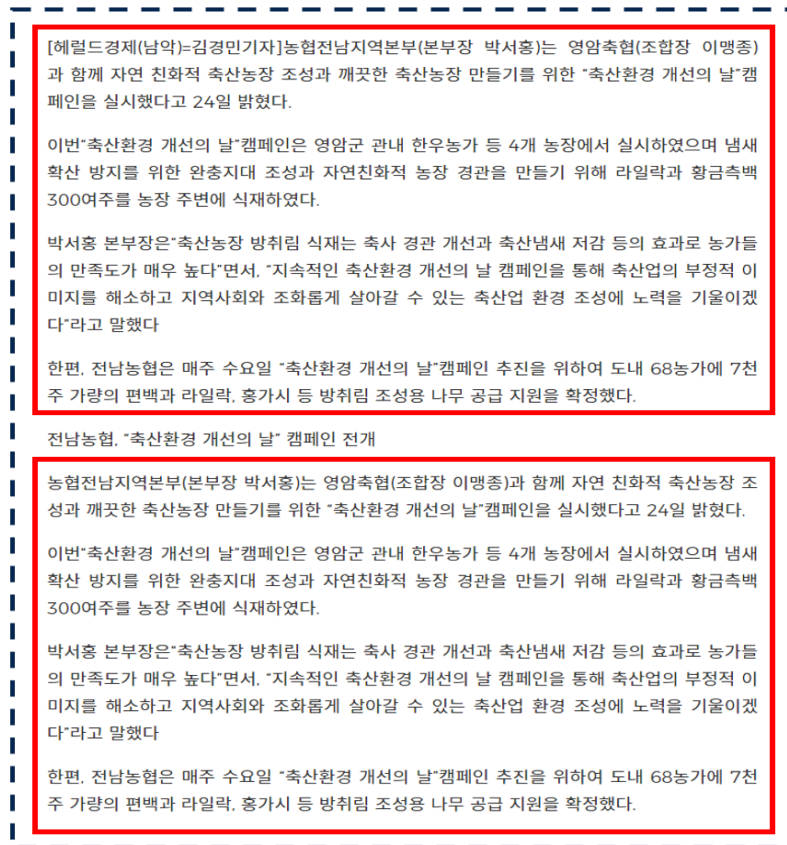
<그림 12> 오류 유형 ⑨: 불필요한 부분 삭제

반면 오류를 삭제하는 방향으로 진행해야 하는 때도 있다. 기사에 불필요한 부호나 글 목록, 중간에 들어가 있는 전문이나 인사 명단 등은 놔두면 기사를 이해하는 데 불필요한 요소가 되어 기사를 읽는 것을 방해하므로 데이터 정제 시 해당 오류는 삭제하게 되었다.



<그림 13> 오류 유형 ⑩: 특수기호 오류

특수기호가 ?로 치환된 오류도 발견되었다. 여기서 ?는 문장 부호 용도로 사용된 것이 아니다. 이는 웹 사이트 자체에서 기자가 기사를 올릴 때 기호가 깨지거나 웹 데이터 정보를 원시 데이터로 가공할 때 생기는 오류로 인해 원래 형태에서 변형된 것이다. 따라서 원래 형태에 맞는 특수기호로 수정해야 한다. 또한, 특수기호를 코드 번호로 적은 매체도 존재했다. 해당 코드 번호 검색 후, 그에 맞는 특수기호로 변경하는 방향으로 진행하였다.



<그림 14> 오류 유형 ⑪: 기사 내용 반복

기사 전체 내용이 한 번 더 사용된 기사도 존재했다. 이런 경우 기사의 아랫부분이 중복 내용으로 인해 삭제된 것인지, 마지막 문장으로 마무리된 것인지 구분하기 어렵다. 전자라면 해당 기사를 사용하지 않고, 후자라면 중복된 부분만 삭제하는 방향으로 진행하면 되지만, 작업자의 주관으로 판단하기엔 확실하지 않으므로 해당 유형의 기사는 사용하지 않았다.





원시 데이터에서 데이터가 소실된 유형도 발견하였다. ‘&lt;’와 같이 서명 기호가 사용된 경우 실제 데이터 자체에서 해당 기호 안 글자가 누락된 경우가 있었다. 유형에 대한 예시는 다음과 같다.

<> 괄호와 내용이 사라짐	
웹 데이터	<p>지역 음악 단체 ‘담소’에서는 조 시인의 시에 곡을 붙여 공연을 선보인다.</p> <p>&lt;오 솔레미오&gt;, &lt;축배의 노래&gt;, &lt;우정의 노래&gt; 등 바리톤 황성철, 소프라노 임현진 성악가의 축하 무대도 준비되어 있다.</p> <p>조태일시문학기념관 일대에서는 추모 시화전이 열린다.</p>
원시 데이터	<p>지역 음악 단체 ‘담소’에서는 조 시인의 시에 곡을 붙여 공연을 선보인다.</p> <p>등 바리톤 황성철, 소프라노 임현진 성악가의 축하 무대도 준비되어 있다.</p> <p>조태일시문학기념관 일대에서는 추모 시화전이 열린다.</p>

<그림 17> 오류 유형 ⑭: 서명 기호(<,>)와 내용이 사라짐

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

### 3. 데이터 1차 정제

#### 가. 중복 기사, 유사한 데이터 제거

중복 기사와 유사 데이터를 제거하는 방법은 다음과 같다. 원시 데이터를 수령한 뒤에 전체 매체를 대상으로 본문 내용이 같은 기사는 제외하게 된다. 이때 먼저 생성된 날짜의 데이터를 사용하게 된다. 유사 기사의 경우 같은 매체 기사 전후 14일 내의 기사들을 대상으로 하여 유사도 비교를 진행하게 되고 85% 이상 유사한 기사들은 사용하지 않게 된다.

○○○○○매체 기사 중 유사도 비교를 통해 사용하지 않는 기사의 예(유사도 85%)	
<p>제목: “2030 골퍼 잡자” 현대백, 역대 최대 규모 골프페어 연다</p> <p><b>[○○○○○=○○○ 기사]</b> 현대백화점은 리오프닝(경제활동 재개)에 대한 기대감 확산과 골프 성수기를 맞아 역대 최대 규모의 골프 페어를 진행한다고 7일 밝혔다.</p> <p>현대백화점이 전사 차원에서 골프 테마 행사를 진행하는 것은 이번이 처음으로 오는 8일부터 24일까지 압구정본점 등 전국 16개 점포에서 ‘현대백화점 그린 마스터’를 테마로 진행된다. 이번 행사는 최근 사회적 거리두기 등으로 해외여행 수요가 골프로 흡수되며 2030세대 고객들이 골프 시장에서 주요 고객층으로 떠오르면서 기획됐다.</p> <p>실제로 현대백화점의 지난 1~3월 골프 부문 매출은 전년동기 대비 70.3% 신장하며 지난해(65.5%)에 이어 꾸준히 증가하고 있다. 특히, 2030 고객의 골프 수요가 급격하게 늘고 있다. 같은 기간 매출은 2배 이상(103.2%) 신장했으며, 전체 골프 매출에서 2030 고객이 차지하는 비중 또한 사상 처음으로 20%를 넘기며 골프 부문 주요 고객층으로 자리잡고 있다.</p> <p>이에 현대백화점은 2030 고객을 타겟으로 한 SNS 이벤트 ‘골프 패셔니스타’를 진행하</p>	<p>제목: 현대백, 역대 최대 규모 골프페어 연다</p> <p>현대백화점은 리오프닝(경제활동 재개)에 대한 기대감 확산과 골프 성수기를 맞아 역대 최대 규모의 골프 페어를 진행한다고 7일 밝혔다.</p> <p>현대백화점이 전사 차원에서 골프 테마 행사를 진행하는 것은 이번이 처음으로 오는 8일부터 24일까지 압구정본점 등 전국 16개 점포에서 ‘현대백화점 그린 마스터’<b>(사진)</b>를 테마로 진행된다. 이번 행사는 최근 사회적 거리두기 등으로 해외여행 수요가 골프로 흡수되며 2030세대 고객들이 골프 시장에서 주요 고객층으로 떠오르면서 기획됐다.</p> <p>실제로 현대백화점의 지난 1~3월 골프 부문 매출은 전년동기 대비 70.3% 신장하며 지난해(65.5%)에 이어 꾸준히 증가하고 있다. 특히, 2030 고객의 골프 수요가 급격하게 늘고 있다. 같은 기간 매출은 2배 이상(103.2%) 신장했으며, 전체 골프 매출에서 2030 고객이 차지하는 비중 또한 사상 처음으로 20%를 넘기며 골프 부문 주요 고객층으로 자리잡고 있다.</p> <p>이에 현대백화점은 2030 고객을 타겟으로</p>

며 하이엔드 골프웨어 상품권, 프리미엄 퍼터, 프로암(아마추어와 프로 선수가 팀을 이뤄 치루는 대회) 대회 초대권 등을 증정한다.

또한, 영글퍼를 타깃으로 현대백화점 자체 캐릭터 ‘흰디’를 활용해 제작한 다양한 볼마커, 골프공, 드라이버·아이언 커버 세트 등 다양한 골프 굿즈들도 선보인다. 행사 기간 현대백화점 전점포에서 20만원 이상 구매 시 흰디 볼마커(2개)를 한정수량으로 증정하는 프로모션도 진행한다.

아울러 오는 8일부터 24일까지 전국 16개 점포 골프 브랜드에서 30만원 이상 구매 시 현금처럼 사용 가능한 플러스포인트 2만포인트를 증정하며, 오는 8일부터 10일까지는 현대백화점카드로 50만·100만원 이상 구매 시 현대백화점 상품권을 최대 7% 증정한다. 오는 8일부터 14일까지 무역센터점 지하1층 대행사장에서 왁·던롭·쉐르보·헤라디아 등 골프 브랜드의 이월 상품을 최대 70% 할인 판매하는 ‘봄맞이 골프대전’을 진행하는 등 점포별로 할인 행사와 팝업 매장도 다양하게 운영한다. 무역센터점, 판교점 등 전국 6개 점포 행사장에는 퍼팅존도 구성한다.

현대백화점 관계자는 “고객들에게 다양한 혜택과 즐거움을 줄 수 있는 행사가 되길 바란다”며 “앞으로도 트렌드에 맞는 다양한 고객 맞춤형 행사를 선보일 계획”이라고 말했다.

한편, 현대백화점은 지난해 거래하는 골프 브랜드 수를 62개로 전년 대비 2배 늘리며 골프 부문 강화에 나서고 있다. 또한, 골프 시타 체험 등이 가능한 골프 용품 매장을 지난해 8개점에서 연내 백화점 16개 전 점으로 확대하는 등 골프 부문을 지속 강화해 나갈 계획이다.

○○@○○○○○○○○.com

한 SNS 이벤트 ‘골프 패셔니스타’를 진행하며 하이엔드 골프웨어 상품권, 프리미엄 퍼터, 프로암(아마추어와 프로 선수가 팀을 이뤄 치루는 대회) 대회 초대권 등을 증정한다.

또한, 영글퍼를 타깃으로 현대백화점 자체 캐릭터 ‘흰디’를 활용해 제작한 다양한 볼마커, 골프공, 드라이버·아이언 커버 세트 등 다양한 골프 굿즈들도 선보인다. 행사 기간 현대백화점 전점포에서 20만원 이상 구매 시 흰디 볼마커(2개)를 한정수량으로 증정하는 프로모션도 진행한다.

아울러 오는 8일부터 24일까지 전국 16개 점포 골프 브랜드에서 30만원 이상 구매 시 현금처럼 사용 가능한 플러스포인트 2만포인트를 증정하며, 오는 8일부터 10일까지는 현대백화점카드로 50만·100만원 이상 구매 시 현대백화점 상품권을 최대 7% 증정한다. 오는 8일부터 14일까지 무역센터점 지하1층 대행사장에서 왁·던롭·쉐르보·헤라디아 등 골프 브랜드의 이월 상품을 최대 70% 할인 판매하는 ‘봄맞이 골프대전’을 진행하는 등 점포별로 할인 행사와 팝업 매장도 다양하게 운영한다. 무역센터점, 판교점 등 전국 6개 점포 행사장에는 퍼팅존도 구성한다.

한편, 현대백화점은 지난해 거래하는 골프 브랜드 수를 62개로 전년 대비 2배 늘이며 골프 부문 강화에 나서고 있다.

○○○ 기자

○○@○○○○○○○○.com

## 나. 기사 선택

오류가 발생했거나 중복 기사, 그리고 유사도 비교를 통해 기사를 제거함으로써 이후 단계에서 작업의 효율성을 높일 수 있었다. 이 단계에서는 유사도 비교를 통해 한 차례 걸러낸 기사 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(정제를 하기 전에 이미 100어절에 미치지 못한 기사가 1차 제외되었고, 정제 후 100어절 이하, 1,000어절 이상인 기사가 2차 제외되었음).
- ❖ 단순 광고, 떠벌 오늘의 운세, 퀴즈 등 기사로 보기 어려운 것은 제외함.
- ❖ 승진자 명단이나 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 공동취재단인 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	명예 기자	삭제	해당 언론사 소속 기자 이외의 작성자가 쓴 기사 (명예 기자, 대학생 기자, 시민기자, 학생 기자, 어린이 기자 등)
각 매체	연합뉴스	삭제	연합뉴스가 출처인 기사, 또는 제공받은 기사
각 매체	공동취재단	삭제	해당 매체와 계약 등을 일일이 확인 불가 (국방부 공동취재단, 올림픽 공동취재단, 대선공동취재단 등)
각 매체	특별취재팀	삭제	해당 매체와 계약 등을 일일이 확인 불가 (대선특별취재팀 등)
각 매체	전국종합	삭제	해당 매체와 계약 등을 일일이 확인 불가 (전국종합, 지역종합, 지방종합 등)

매체명	기자명 정보	사용여부	내용
각 매체	대담	삭제	인터뷰가 아니라 대담임을 밝히고 있는 경우
각 매체	전문기자	삭제	분야별 전문가 작성 (에디터, 이코노미스트 등)
각 매체	아나운서	삭제	라디오 방송 또는 유튜브 영상을 그대로 옮겨 적음 (아나운서, PD, 프로듀서, 진행 등)
각 매체	리포터	삭제	모집 프리랜서 기자
각 매체	객원기자	삭제	모집 프리랜서 기자
OO매체	○○비즈 기자	삭제	해당 매체 미디어 그룹에 속한 별도의 법인
OO매체	디지털 ○○ 기자	삭제	해당 매체 미디어 그룹에 속한 별도의 법인
OO매체	어린이 ○○ 기자	삭제	해당 매체 미디어 그룹에 속한 별도의 법인
OO매체	월간 ○○	삭제	해당 매체 미디어 그룹에 속한 별도의 법인
OO매체	주간 ○○	삭제	해당 매체 미디어 그룹에 속한 별도의 법인
OO매체	메디컬 리포트	삭제	저작권자 확인 불가
OO매체	영상미디어	삭제	저작권자 확인 불가

<표 9> 저작권 이용 문제로 인해 사용하지 않는 기사의 특징

인공 지능(AI) 챗봇 ‘챗지피티(ChatGPT)’의 개발사인 오픈에이아이(OpenAI)가 개인 정보를 무단으로 수집했다는 이유로 미국 캘리포니아에서 집단 소송에 휘말렸다.<sup>1)</sup> 이 소송은 2023년 6월 28일 클락슨 법률사무소에 의해 미국 북부 캘리포니아 연방지방법원에 제기되었다. 소송은 오픈에이아이가 ‘챗지피티’를 훈련시키기 위해 수백만 개의 소셜 미디어 댓글, 블로그 글, 위키백과 글, 가족 레시피 등에서 데이터를 수집했지만, 이를 위한 사용자들의 동의를 받지 않았다고 주장하고 있다. 따라서 오픈에이아이가 수백만 인터넷 사용자들의 저작권을 침해했고 개인 정보 보호의 의무를 다하지 않았다는 것이다.

최근에 뉴욕타임스(NYT)는 챗지피티 개발사인 오픈에이아이사에게 저작권 위반으로 소송을 준비하고 있고, 오픈에이아이사는 세계 최대 뉴스 통신사 에이피(AP)통신과 2년간의 뉴스 기사 라이선스 계약을 체결했다고 한다. 저작권 문제가 본격적으로 수면 위에 떠오르게 되었다. 인공 지능 분야의 4대 석학으로 인정받는 앤드류 응(Andrew Ng) 스탠포드 대학교 겸임 교수는 “좋은 데이터를 수집하고 가공하는 것이 인공 지능을 만드는 과정의 80%를 차지하는데, 이는 데이터가 인공 지능의 핵심적인 부분임을 의미한다.”<sup>2)</sup>라고 말했다. 그는 데이터 중심 인공 지능(Data-centric AI) 개발의 중요성을 강조하면서, 인공 지능 개발자들이 코드 수정을 통한 모델 하이퍼파라미터 변경에 매몰되지 않고 좋은 데이터를 확보하고 유지하려는 노력을

1) <https://kr.cointelegraph.com/news/open-ai-hit-with-class-action-lawsuit-over-chatgpt-data-theft>

2) A Chat with Andrew on MLOps: From Model-centric to Data-centric AI, <https://www.youtube.com/watch?v=06-AZXmwHjo&t=0s>, DeepLearningAI, 7:00

해야 한다고 덧붙였다. 그리고 데이터로 성능 개선을 이루어 내는 것을 실제 사례를 통해 보여 주었다.

보통 엄청나게 많은 데이터를 수집하면 데이터가 좀 부실해도 문제가 없을 것으로 생각하는데 데이터의 증가는 필연적으로 학습 비용으로 귀결되기 때문에 무한정으로 데이터를 확대할 수는 없다. 단순히 학습 데이터의 양을 기하급수적으로 늘리기만 해서는 인공 지능 모델의 성능 향상 혹은 더 나은 의사결정으로 연결시킬 수 없다. 데이터의 양만큼이나 중요한 것이 데이터의 품질이다. “쓰레기가 들어가면 쓰레기가 나온다.(Garbage in, garbage out)”라는 말처럼 질 낮은 데이터는 아무리 양이 많아도 좋은 결과를 낼 수 없다.

인공 지능 모델은 학습 데이터의 품질에 따라 성능이 좌우된다. 인공지능이 생성한 콘텐츠가 다음 인공 지능 모델의 학습 데이터에 포함되면, 모델 붕괴 현상이 발생할 수 있다. 모델 붕괴란 인공 지능 모델이 학습 데이터의 오류나 편향성을 그대로 반복하여, 정확도와 유용성이 저하되는 현상을 말한다. 영국 옥스포드 대학 교수 일리아 슈마일로프가 주축으로 구성된 연구팀이 발표한 논문인 ‘재귀의 저주(The curse of Recursion)’에 따르면, 인공 지능이 생성한 콘텐츠가 차기 생성형 인공 지능의 훈련 데이터에 조금이라도 포함되면 결국 해당 모델에게 악영향을 주게 되는 ‘모델 붕괴(Model Collapse)’가 발생한다고 주장했다. 모델 붕괴를 실제로 확인하기 위해, 해당 논문에서 연구팀은 먼저 사람이 만들어 낸 데이터로 학습한 챗지피티와 같은 대규모 언어 모델로 일부 텍스트를 출력했다. 이렇게 출력된 텍스트들을 새로운 모델의 학습 데이터로 활용한 뒤, 학습된 모델로부터 텍스트를 다시 뽑아낸다. 이 같은 과정을 세 번째, 네 번째에 걸쳐 계속 반복하면 회차마다 오류가 쌓이게 되는데, 이에 따라 열 번째 모델에게 영국 건축에 대해 글을 쓰도록 요청했을 때 모델이 질문과 관련 없는 터무니 없는 대답을 내렸다고 연구팀은 설명했다. 또한, 최근 월스트리트저널(WSJ)은 인공 지능이 생성한 쓰레기 정보들이 인터넷을 오염시키기 시작했다고 보도했다. 이 모두가 데이터의 품질이 중요하다는 것을 지적하고 있다.

국립국어원의 신문 기사 원문 자료 수집 및 정제 사업은 신문 기사를 대상으로 말뭉치를 수집하고 정제하고 저작권에 문제가 되는 기사를 사용하지 않는 것에 집중하고 있다. 신문 기사에는 외부 기고가 또는 공동취재단 같이 저작권에 대해 이해를 다르게 하는 기사들이 상당수 포함되어 있기 때문이다. 이에 전체 기사 정보를 추출하여 위의 <표 9>와 같이 저작권 이용에 위험이 있는 기사를 걸러냈다.



## 4. 데이터 2차 정제

데이터 1차 정제를 마친 기사는 데이터 총괄 관리자가 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

원시 데이터만으로는 데이터를 구축할 수 없다. 수많은 오류가 포함되어 있는 다양한 매체의 원시 데이터를 그대로 사용할 경우, 추후에 수정하는 공정이 많이 필요하다. 그렇기 때문에 최초 기사 선정 단계에서 많은 데이터를 확인하고 오류를 제거하거나 수정해 주어야 한다. 이 방법을 활용하기 위해서는 웹 페이지의 정보를 참조하여 작업을 진행해야 한다. 데이터의 소실 문제, 기사 중간의 소제목이 다음 단락과 붙어 버리는 문제 등 이러한 문제는 작업자가 기사를 정독하지 않는다면 발견하기가 상당히 어렵다.

하지만 방대한 양의 말뭉치를 구축하면서 모든 데이터를 정독하기란 쉽지 않다. 수십만의 기사를 정독하더라도 오류를 놓칠 수 있다. 또한 캡션이 특별한 정보 없이 평이한 문장으로 들어간 경우라면 캡션임을 구분해내기가 어렵다. 이런 오류 등은 작업자가 기사를 웹에서 직접 확인하고 해결해야 한다. 웹 페이지 데이터에는 원시 데이터에 없는 정보가 표시되어 있어 이를 활용하여 오류를 바로잡을 수 있다.



❖ 중간 제목이 본문 사이에 들어간 경우

웹에서 확인한 실제 기사 내용
<p>환경부는 26일 대구 성서산업단지 내 아진엑스텍에서 열린 제1회 규제혁신전략회의에서 이 같은 내용이 담긴 '환경 규제 혁신 방안'을 윤석열 대통령에게 보고했다. 환경 분야 규제 혁신은 기존 환경관련 규제 합리화와 탄소중립·순환 경제 등 환경정책 목표 관련 규제 개선에 중점을 뒀다.</p> <hr/> <p><b>닫힌 규제에서 열린 규제로 전환</b></p> <hr/> <p>환경부는 이번 규제혁신 방안을 크게 4가지다. 미리 정해놓은 금지행위 외 행위를 모두 허용하는 '네거티브'(열린) 규제 전환을 비롯해 위험에 비례하는 차등적 규제 전환, 쌍방향 소통 및 협의형 규제 전환, 탄소중립·순환경제 직접 규제 우선 개선 등이다.</p>
원시 데이터 내용
<p>환경부는 26일 대구 성서산업단지 내 아진엑스텍에서 열린 제1회 규제혁신전략회의에서 이 같은 내용이 담긴 '환경규제 혁신 방안'을 윤석열 대통령에게 보고했다. 환경 분야 규제 혁신은 기존 환경관련 규제 합리화와 탄소중립·순환경제 등 환경정책 목표 관련 규제 개선에 중점을 뒀다. <b>닫힌 규제에서 열린 규제로 전환</b></p> <p>환경부는 이번 규제혁신방안을 크게 4가지다. 미리 정해놓은 금지행위 외 행위를 모두 허용하는 '네거티브'(열린) 규제 전환을 비롯해 위험에 비례하는 차등적 규제 전환, 쌍방향 소통 및 협의형규제 전환, 탄소중립·순환경제직결 규제 우선 개선 등이다.</p>

❖ 캡션 정보가 본문과 구분되지 않아 본문처럼 보이는 내용

## 웹에서 확인한 실제 기사 내용

정치일반

# “민주당 소통 통한 혁신·새 정치세대 키워야”

김지원기자ji1@kwnews.co.kr

기자의 다른기사 보기

가 가    

입력: 2022-07-12 00:00:00 (03면)

도당 '대선·지선 평가 간담회'  
내용 속 당 결집 기회 지적도  
내일 당원 토론회 채신안 논의



더불어민주당 강원도당은 11일 도당 회의실에서 허영 도당위원장 등이 참석한 가운데 제20대 대선 및 제8회 지방선거 평가 간담회를 가졌다. 박승선기자

더불어민주당 강원도당은 11일 도당 회의실에서 허영 도당위원장 등이 참석한 가운데 제20대 대선 및 제8회 지선 평가 간담회를 가졌다. 박승선기자

3·9 대통령선거와 6·1 지방선거의 패인을 분석하고 있는 더불어민주당 내에서 '예견된 참사'였다는 지적이 나왔다. 대선 패배 직후 지선으로 이어진 어려운 구도였던 데다 성찰과 채신의 노력이 부족했다는 평가 속에 미래세대에 대한 수혈 방식 변화에 대한 문제도 있었음이 확인됐다.

## 한국언론진흥재단으로부터 받은 데이터 내용

더불어민주당 강원도당은 11일 도당 회의실에서 허영 도당위원장 등이 참석한 가운데 제20대 대선 및 제8회 지선 평가 간담회를 가졌다. 박승선기자

3·9 대통령선거와 6·1 지방선거의 패인을 분석하고 있는 더불어민주당 내에서 '예견된 참사'였다는 지적이 나왔다. 대선 패배 직후 지선으로 이어진 어려운 구도였던 데다 성찰과 채신의 노력이 부족했다는 평가 속에 미래세대에 대한 수혈 방식 변화에 대한 문제도 있었음이 확인됐다.

## 나. 불필요한 요소 제거

불필요한 요소 제거 공정에서는 기사를 읽어 가며 저작권에 위배되는 기사를 다시 한번 확인하고, 기사 내용에 불필요한 요소를 삭제하는 공정이다. 전체 작업 중에서 가장 많은 인력이 투입되고 핵심적인 작업이라고 할 수 있다.

### 1) 제외 대상 기사

- ❖ 저작권에 위배되는 기사
- ❖ 문장이 도중에 잘렸거나 오류가 많은 기사
- ❖ 기사의 제목을 스크랩한 기사
- ❖ 기사로 볼 수 없는 문장이 나열되는 기사, 방송을 그대로 옮겨 적은 기사
- ❖ 불필요한 요소를 제거한 뒤 기사 내용이 극히 적은 기사

사용하지 않는 기사 예
<p>■ 방송 : CBS 라디오 &lt;김현정의 뉴스쇼&gt; FM 98.1 (07:20~09:00)</p> <p>■ 진행 : 김현정 앵커</p> <p>■ 대담 : 고등학생 000군 (익명)</p> <p>잠시 후에 러시아, 우크라이나전 이야기 할 텐데요. 다시 좀 심각한 이야기 들어가기 전에 우리의 마음을 정화하고 가겠습니다. 짧은 인터뷰를 하나 준비했는데요. 요즘 참 불의를 봐도 내 일처럼 나서는 사람이 점점 줄어들고 있는 세상에 세 명의 고교생이 용감하고 정의롭게 다른 시민을 구했습니다. 어떤 일이라면 등곳길에 누군가가 불법 촬영을 하고 있는 장면을 목격하고 용감하게 그를 제압해서 경찰에 넘긴 고등학생들이 지금 화제예요. 심지어 이 범인이 휴대폰을 부수기 시작하니까 그 장면을 동영상으로, 증거로 찍어놓는 기지까지 발휘했다고 하는데요. 오늘 화제의 인터뷰, 용기 있는 친구들 가운데 한 명, 연결해보죠. 불러보겠습니다. 용감한 학생 나와 주세요?</p> <p>◆ 익명&gt; 네, 안녕하세요.</p> <p>◇ 김현정&gt; 지금 학교십니까?</p> <p>◆ 익명&gt; 학교, 그냥 기다리고 있었어요.</p> <p>◇ 김현정&gt; 전화 기다리면서 지금 학교, 수업 못 들어가고 계셨어요?</p> <p>◆ 익명&gt; 괜찮습니다. 아직 시작을 안 해서.</p> <p>◇ 김현정&gt; (웃음)그래요. 학교에서 지금 친구들이나 선생님들이 엄청 자랑스러워 하시겠는데요.</p> <p>◆ 익명&gt; 맞습니다.</p>

방송을 그대로 옮겨 적어 문어체 문장의 집합이 아닌 기사도 사용하지 않는 기사로 표기하고 제외하였다.

## 2) 불필요한 본문 내용 삭제

기사와 무관한 정보들을 삭제하는 과정이다. 이 정보들은 전체 맥락을 해치므로 제거한다. 또한 연설문, 입장문, 누리소통망서비스(SNS) 게시물 등 기사와는 다른 외부 전문이 실린 경우, 이러한 전문은 기사가 작성한 기사로 보기 어렵다. 이를 그대로 사용하는 것은 신문 기사 말뭉치를 구축하는 이 사업의 목적에 맞지 않으며, 저작권 문제가 발생할 수 있으므로 제거한다. 그리고 이를 제거하면서 다음에 전문이 존재함을 알리는 문장도 기사 문맥의 완결성에 유의하며 제거한다.

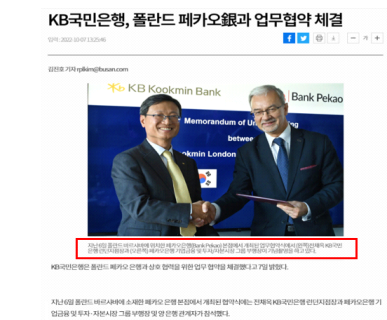
아래 예시의 ‘전문’, ‘문장으로 볼 수 없는 정보’, ‘문장의 오류’ 항목에서 굵은 붉은색 글꼴이 삭제해야 할 대상이다.

삭제 정보	예시
표, 그림, 그래프 등의 캡션 정보	<div> <div>사진제공=화성시</div> <div>[사진 제공= 로이터]</div> <div>사진제공   카카오톡</div> <div>사진/이○○ 숲 해설가</div> <div>[영상=시너지영상팀]</div> </div> <div> <div>사진=‘쇼! 음악중심’ 방송 캡처</div> <div>사진·자료=경기관광공사</div> <div>출처  송가인 SNS</div> <div>(사진설명)</div> <div>(조감도)</div> </div>
기자의 이름, ID 등의 정보	<div>의정부=김○○기자</div> <div>[세종=○○○○제 주○○ 기자]</div> <div>이○○기자/사진=강화군 제공</div>
‘Copyright©’ 등 저작권 관련 내용	<div>랄프 깁슨 'Salon Litteraire'. ©Ralph Gibson</div> <div>-----@---.co.kr/2022-10-16 10:03:05/&lt;저작권자 ©</div> <div>1980-2022 ○○일보. 무단 전재 재배포 금지.&gt;</div>
전문	<div><b>[○○○○○   남○○기자]</b>NC 엔터테인먼트가 AOA 출신 지민과의 전속계약 만료 소식을 전했다.</div> <div>13일 FNC는 공식 홈페이지에 “소속 아티스트 지민과의 전속 계약 기간이 2022년 1월 12일로 종료되어 안내드립니다”고 밝혔다.</div> <div>이어 소속사는 “지난 9년간 당사 소속 가수로서 활발한 활동을 이어온 지민에게 감사의 마음을 전한다. 비록 당사와 함께하는 인연은 마무리되었지만, 지민의 앞날과 향후 행보에 따뜻한 격려와 응원 부탁 드립니다”고 덧붙였다.</div> <div><b>이하 FNC 글 전문.</b></div> <div><b>안녕하세요,</b></div> <div><b>FNC엔터테인먼트입니다.</b></div> <div><b>소속 아티스트 지민과의 전속 계약 기간이 2022년 1월 12일로 종료되어 안내드립니다.</b></div> <div><b>지난 9년간 당사 소속 가수로서 활발한 활동을 이어온 지민에게 감사의 마음을 전합니다.</b></div> <div><b>비록 당사와 함께하는 인연은 마무리되었지만, 지민의 앞날과 향후 행보에 따뜻한 격려와 응원 부탁 드립니다.</b></div>

삭제 정보	예시
	<p><a href="#">감사합니다. ---@sportsseoul.com</a></p> <p><a href="#">사진출처   지민 인스타그램</a></p>
문장으로 볼 수 없는 정보	<p>◆MBC 여론 조사 결과...윤석열 41.1%, 이재명 32.9%, 안철수 10.5%, 심상정 3.1%</p> <p>설 연휴를 앞두고 MBC가 코리아리서치에 의뢰, 26~27일 여론 조사를 한 결과 4자 가상대결에선 국민의힘 윤석열 후보가 41.1%, 더불어민주당 이재명 후보는 32.9%를 얻어 윤 후보가 이 후보를 오차범위 밖에서 앞서는 것으로 나타났다.</p> <p>이번 대선이 이재명 윤석열 심상정 안철수 후보의 4자대결로 치러지면 누구에게 투표할 지 물은 결과 국민의힘 윤석열 41.1, 더불어민주당 이재명 32.9, 국민의당 안철수 10.5, 정의당 심상정 3.1%로, 윤 후보가 이 후보를 8.2%p 차이로 앞섰다. 2주 전 조사에선 두 후보의 격차가 오차범위 내인 6% 포인트였지만, 이번 조사에선 조금 더 벌어져 오차범위 밖으로 나타났다.</p> <p><a href="#">조사의뢰 : MBC 조사기관 : (주)코리아리서치인터내셔널 조사대상 : 전국 거주 만 18세 이상 남녀 1,002명 조사기간 : 2022년 1월 26일 ~ 27일(2일간) 조사방법 : (국내 통신 3사 제공) 휴대전화 가상번호 100% 이용 무선전화면접 피조사자 선정방법 : 성/연령/지역별 할당 응답률: 17% (5,910명 통화 1,002명 응답) 가중값 산출·적용방법 : 성/연령/지역별 가중치 부여 (셀 가중, 2021년 12월 말 행정안전부 주민등록인구통계 기준) 표본오차: 95% 신뢰 수준 ±3.1% 포인트 질문내용: 중앙선거여론조사심의위원회 홈페이지 참조</a></p> <p>10일 통계청에 따르면 7월 외식물가지수는 전년 동월 대비 8.4% 상승했다. 지난 1992년 10월(8.8%) 이후 오름폭이 가장 컸다. 통계청이 집계하는 전체 외식 품목 39개의 물가도 모두 상승했다. 결식아동이 주로 찾는 분식집의 김밥, 라면 등 메뉴 가격이 두 자릿수 상승률을 기록한 것으로 나타났다.</p> <p>이렇다 보니 현장에서는 "김밥 한 줄도 4000원인데 7000원 가지고 라면하고 김밥도 못 먹는다" 는 등 목소리가 나왔다. 전문가와 아동보호기관도 물가상승에 따라 꿈나무카드 급식지원 단가를 인상해야 한다고 제언하기도 했다.</p> <p><a href="#">(관련기사: "7000원으로 배고픔 사라질까요?" 한도 빠듯한 '꿈나무카드' [결식아동 배부르게①])</a></p> <p>○ 후보자의 발탁을 두고 노동계는 전혀 예상하지 못한 깜짝 발탁이라는 반응이 주를 이뤘다. 그동안 장관 후보자로 이름이 거론된 적이 없는 데다 하마평에 올랐던 인물들이 국회의원과 노동학자, 고용부 전직 관료가 대부분이었기 때문이다. 노동계 출신 고용부 장관도 김○○ 전 장관, 방○○ 전 장관 등 손에 꼽을 만큼 드물다. 이 때문에 노동계에서는 윤석열 정부가 예고한 친기업 정책에 대한 우려를 반영한 인사라는 분석이 나온다. 한 노동계 인사는 이 후보자에 대해 “문○○ 경사노위 위원장과 함께 노사정 전문가로 평가된다”며 “새 정부가 사회적 대화의 동력을 얻었다고 기대할 수 있다”고 말했다.</p> <p><a href="#">■이○○ 고용노동부 장관 후보자 프로필</a></p>

삭제 정보	예시
	△19--년 충북 제천 △대전 대전고 △서울대 경제학과 △한국노총 정책기획국장 △최저임금심의위원회 연구위원 △노사관계개혁위원회 전문위원 △노사정위원회 전문위원 △21세기노사관계연구회 회장 △서울디지털대 e-경영학부 전임교수 △건설근로자공제회 비상임이사 △한국노총 정책본부장 △고용부 최저임금위원회 근로자위원 △한국노총 사무처장 △노사발전재단 사무총장 △삼성전자 자문위원
문장의 오류	<p>농심이 매출액의 <b>매출액의</b> 2.15%를 소아암 환아에게 기부하는 백산수 한정판을 출시했다고 29일 밝혔다.</p> <p>앞서 핀란드와 <b>핀란드와</b> 스웨덴이 오는 6월 스페인에서 개최하는 나토 정례회의에서 가입 신청서를 제출할 수 있다는 관측이 현지 매체 등을 통해 제기됐다.</p>

<표 10> 불필요한 요소 제거 내용



원본 데이터

지난 6일 폴란드 바르샤바에 위치한 페카오은행(Bank Pekao) 본점에서 개최된 업무협약식에서 (왼쪽)전채옥 KB국민은행 런던지점장과 (오른쪽) 페카오은행 기업금융 및 투자/자본시장 그룹 부행장이 기념촬영을 하고 있다. KB국민은행은 폴란드 페카오 은행과 상호 협력을 위한 업무 협약을 체결했다고 7일 밝혔다.

정제 데이터

KB국민은행은 폴란드 페카오 은행과 상호 협력을 위한 업무 협약을 체결했다고 7일 밝혔다. 지난 6일 폴란드 바르샤바에 소재한 페카오 은행 본점에서 개최된 협약식에는



원본 데이터

1화성 송산그린시티 동측지구인 새솔동 도로 곳곳에서 균열이 발생, 주민들이 불안해하고 있다. 새솔동의 한 도로 옆에 긴급보수에 사용된 것으로 추정되는 아스팔트 프라이머(Asphalt Primer)가 놓여져 있다. 아스팔트 프라이머는 콘크리트와 시멘트 등 표면에 도포해 하층에 피막을 형성, 시공성과 접착력을 강화시키는 방수·방습용 제품이다. 이런 가운데, 입주 4년여 밖에 안 된 새솔동 곳곳에서 심각한 도로균열 현상이 발생, 주민들이 대책을 호소하고 있다.

정제 데이터

이런 가운데, 입주 4년여 밖에 안 된 새솔동 곳곳에서 심각한 도로균열 현상이 발생, 주민들이 대책을 호소하고 있다.

<그림 18> 원본 데이터와 정제된 데이터의 예



데이터 정제 전	정제 데이터
<p><u>코로나로 중단됐다 3년만에 열리는 2022 한마음 건강 걷기대회가 9일 시민 4천여명이 참가한 가운데 열렸다. 시민들이 중랑천 변을 따라 걷고 있다. 김○○기자</u> 코로나로 중단됐다 3년만에 열리는 2022 한마음 건강 걷기대회가 9일 비가 내리는데도 시민 4천여명이 참가한 가운데 중랑천 동막교 인라인 스के이트장에서 펼쳐졌다. (중략)</p> <p><u>코로나로 중단됐다 3년만에 열리는 2022 한마음 건강 걷기대회가 9일 시민 4천여명이 참가한 가운데 열렸다. 시민들이 기념촬영을 하고 있다. 김○○기자</u> 해마다 대회에 참가했다는 호원동 거주 한 어르신(74)은 "사위와 딸 모두 나오려고 했는데 비가 내려 혼자 왔다. 이런 정도는 견딜만하다. 모처럼 많은 시민과 함께 걸으니 기분이 좋다"고 말했다. (중략)</p> <p><u>코로나로 중단됐다 3년만에 열리는 2022 한마음 건강 걷기대회가 9일 시민 4천여명이 참가한 가운데 열렸다. 김동근 시장이 시민들과 악수를 하고 있다. 김○○기자</u> 김동근 시장은 "많은 지자체가 걷기 좋은 도시 만들기에 나서고 있다. 걷기좋은 도시는 대중교통 접근성, 생태환경, 안전한 환경 등을 잘 갖춰 시민이 살기 좋은 도시를 말한다. 의정부도 걷기 좋은 도시 만들기에 최선을 다하겠다"고 말했다.</p> <p><u>의정부=김○○기자</u></p>	<p>코로나로 중단됐다 3년만에 열리는 2022 한마음 건강 걷기대회가 9일 비가 내리는데도 시민 4천여명이 참가한 가운데 중랑천 동막교 인라인 스के이트장에서 펼쳐졌다. (중략)</p> <p>해마다 대회에 참가했다는 호원동 거주 한 어르신(74)은 "사위와 딸 모두 나오려고 했는데 비가 내려 혼자 왔다. 이런 정도는 견딜만하다. 모처럼 많은 시민과 함께 걸으니 기분이 좋다"고 말했다. (중략)</p> <p>김동근 시장은 "많은 지자체가 걷기 좋은 도시 만들기에 나서고 있다. 걷기좋은 도시는 대중교통 접근성, 생태환경, 안전한 환경 등을 잘 갖춰 시민이 살기 좋은 도시를 말한다. 의정부도 걷기 좋은 도시 만들기에 최선을 다하겠다"고 말했다.</p>

<표 11> 원시 데이터와 정제된 데이터 비교 1

데이터 정제 전	정제 데이터
<p>①이정립- <u>A Space for inner thoughts, 34.5x34.5cm, Incense flame on the Korean paper</u></p> <p>②이영훈-어울림 <u>Acrylic on Canvas 2020</u></p> <p>③전성규 -<u>Hidden Passage-Flapping</u></p> <p>④윤수보-<u>image forest456 oil on canvas 72.5x60.5cm 2021</u> 정치인에서 서양화가로, 또 갤러리 관장으로 변신한 이수진 베카갤러리(BEKA Gallery) 관장이 개관 초대전을 선보인다. (중략)</p> <p><u>(원)이근택 - 72.7X53.0cm Acrylic on Canvas 2017</u></p> <p><u>(오)김일중-베르사이유 거울의 방 112cm × 162cm / 자개, 아크릴릭, 바니쉬 2017</u></p> <p>윤수보 작가의 ‘이미지 포레스트 image forest 456’은 시대 정신과 구조가 잘 녹아든 색채, 형태의 전개가 있다. 화려하고 선명한 색감의 줄기들이 화면 전체를 휘감고 여기에서 확산하는 빛의 약동은 무수한 신비의 공간을 엿보게 한다. 이영훈 작가의 어울림은 만물이 존재하는 데 필수적인 생명의 물질과 물질을 감싸고 있는 공간, 공간이 변화하고 확장하는 그 변화를 표현한다. 이근택 작가의 이스탄불은 시선에 대한 시리즈의 연작이다.</p> <p>(중략)</p> <p>이수진 관장은 “베카갤러리는 현대미술시장의 중심에서 동시대 작가들을 발굴하고 신진작가 및 중견작가들의 전시를 기획하기 위해 개관했다”며, “앞으로 갤러리가 당면하고 있는 현대미술시장에서의 상생 방안을 모색하는데 역점을 두고 갤러리를 운영해 나가겠다”라고 말했다.</p> <p><u>과천=김○○기자</u></p>	<p>정치인에서 서양화가로, 또 갤러리 관장으로 변신한 이수진 베카갤러리(BEKA Gallery) 관장이 개관 초대전을 선보인다. (중략)</p> <p>윤수보 작가의 ‘이미지 포레스트 image forest 456’은 시대 정신과 구조가 잘 녹아든 색채, 형태의 전개가 있다. 화려하고 선명한 색감의 줄기들이 화면 전체를 휘감고 여기에서 확산하는 빛의 약동은 무수한 신비의 공간을 엿보게 한다. 이영훈 작가의 어울림은 만물이 존재하는 데 필수적인 생명의 물질과 물질을 감싸고 있는 공간, 공간이 변화하고 확장하는 그 변화를 표현한다. 이근택 작가의 이스탄불은 시선에 대한 시리즈의 연작이다.</p> <p>(중략)</p> <p>이수진 관장은 “베카갤러리는 현대미술시장의 중심에서 동시대 작가들을 발굴하고 신진작가 및 중견작가들의 전시를 기획하기 위해 개관했다”며, “앞으로 갤러리가 당면하고 있는 현대미술시장에서의 상생 방안을 모색하는데 역점을 두고 갤러리를 운영해 나가겠다”라고 말했다.</p>

<표 12> 원시 데이터와 정제된 데이터 비교 2



데이터 정제 전	정제 데이터
<p><u>[스포츠서울   윤○○기자]</u> 이듬해 1월과 5월 상무에 입대하는 프로야구선수 27명이 확정됐다.</p> <p>국군체육부대는 1일 상무 최종합격자 명단을 각 구단에 공지했다. 10구단으로부터 합격자를 취합한 결과 KT 유격수 심○○, SSG 필승조 김○○ 등이 합격자 명단에 이름을 올렸다.</p> <p>10구단 중 가장 많은 합격자를 배출한 구단은 SSG와 삼성이었다. SSG와 삼성은 각각 5명이 상무에 입대한다. LG와 롯데는 4명이 최종합격자 명단에 포함됐다. KIA는 상무 지원자가 없었던 만큼 합격자도 나오지 않았다.</p> <p>이번 합격자들은 1월 16일과 5월 8일로 나눠서 입대한다. 전역 시기 또한 입대 시기에 맞춰서 다르다. 1월 입대자는 2024년 7월, 5월 입대자는 2024년 11월에 전역한다. <u>다음은 각 구단 발표를 취합한 합격자 명단.</u></p> <p><u>2023년 1월 16일 입대: 김○○, 김○○, 장○○, 조○○(이상 SSG), 이○○, 임○○(이상 LG), 심○○, 권○○(이상 KT), 김○○, 박○○(이상 NC), 김○○, 박○○(이상 삼성), 추○○(롯데), 허○○(한화)</u></p> <p><u>2023년 5월 8일 입대: 전○○(SSG), 박○○, 이○○(이상 키움), 허○○, 송○○(이상 LG), 이○○, 박○○, 이○○(이상 삼성), 이○○, 조○○, 한○○(이상 롯데), 박○○(두산), 정○○(한화)</u></p> <p><u>---@-----.com</u></p>	<p>이듬해 1월과 5월 상무에 입대하는 프로야구선수 27명이 확정됐다.</p> <p>국군체육부대는 1일 상무 최종합격자 명단을 각 구단에 공지했다. 10구단으로부터 합격자를 취합한 결과 KT 유격수 심○○, SSG 필승조 김○○ 등이 합격자 명단에 이름을 올렸다.</p> <p>10구단 중 가장 많은 합격자를 배출한 구단은 SSG와 삼성이었다. SSG와 삼성은 각각 5명이 상무에 입대한다. LG와 롯데는 4명이 최종합격자 명단에 포함됐다. KIA는 상무 지원자가 없었던 만큼 합격자도 나오지 않았다.</p> <p>이번 합격자들은 1월 16일과 5월 8일로 나눠서 입대한다. 전역 시기 또한 입대 시기에 맞춰서 다르다. 1월 입대자는 2024년 7월, 5월 입대자는 2024년 11월에 전역한다.</p>

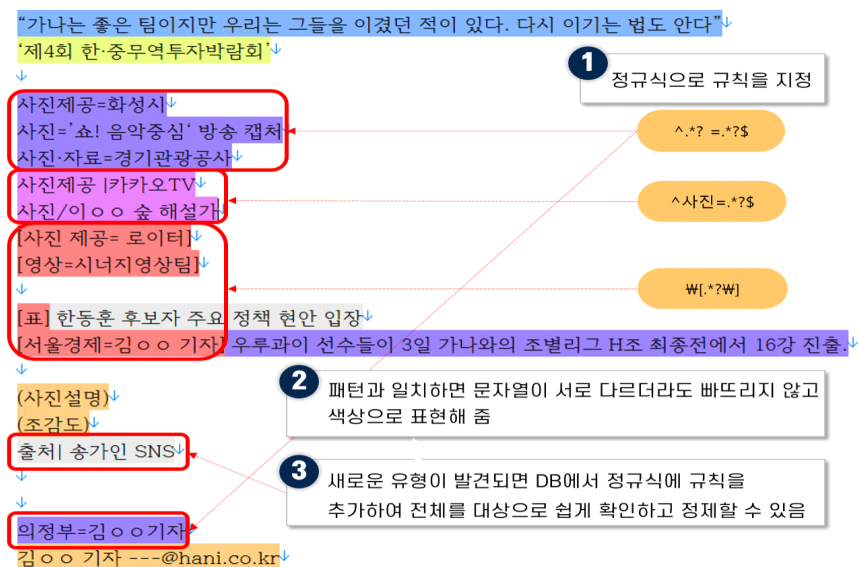
<표 13> 원시 데이터와 정제된 데이터 비교 3(기사로 볼 수 없는 정보 삭제)

<url>https://news.kmib.co.kr/article/view.asp?arcid=0924276972&amp;code=11121100</url>  
 <used>Y</used>  
 <byline>박○○○</byline>  
 <content>  
 <d>사진=이○○ 기자</d>  
 한동훈<d>{사진}</d> 법무부 장관이 직접 자신을 향한 '국민의힘 당대표 차출설' 진화에 나섰다. 그러나 친윤(친윤석열)계 핵심 장제원 의원과 정진석 비상대책위원장이 말씨름을 벌이는 등 '한동훈 차출설'의 여진은 계속됐다.  
 한 장관은 7일 자신을 둘러싼 국민의힘 당대표 차출설에 대해 "중요한 할 일이 많기에 장관의 역할에 최선을 다하겠다고 분명히, 단호하게 말씀드린다"고 강조했다. 한 장관은 이날 국회 법제사법위원회에 참석하기 전 기자들과 만나 "제가 아직 많이 부족하다고 생각하지만, 장관으로서 최선을 다해왔고 앞으로도 그 생각밖에 없다"고 말했다. 한 장관은 '정계에서 당대표 제안이 있었느냐'는 질문에는 '저에게 그런 얘기를 한 사람은 아무도 없다'고 답했다.  
 윤석열 대통령이 '한동훈 차출설'에 대해 강한 불쾌감을 표출했다는 보도가 나온 하루 뒤에 한 장관이 직접 나서 자신의 차출설을 일축한 것이다.  
 이런 상황에서 장 의원은 이날 친윤계가 이끄는 공부모임 '국민공감' 첫 모임에서 기자들과 만나 "비대위원장은 (전당대회) 심판인데, 선거에 기준을 제시하는 건 어른의 자세가 아니다"며 "부적절하다"고 비판했다. 정 비대위원장이 지난 5일 "MZ·미래세대의 새로운 물결에 공감하는 지도부가 탄생하기를 바란다"고 말하면서 '한동훈 차출설'에 힘이 실렸던 것을 비판한 것이다.  
 장 의원은 그러면서 "그런 얘기를 자주 하니까 한 장관 차출론이 나오지 않나"라며 "대통령도 한 장관 차출론을 결코 원하지 않을 것"이라고 강조했다.  
 정 위원장은 곧장 장 의원의 발언을 받아쳤다. 정 위원장은 "심판이기에 당연히 해야 하는 이야기이지 심판으로서 해선 안 될 이야기인가"라고 반박했다. 정 위원장은 경기도 용인 '용인 반도체 클러스터 조성 사업' 현장방문 후 기자들과 만나 "새로운 물결을 구축하기 위해서는 국민의힘은 MZ세대·미래세대와 늘 공감하는 지도부를 구성하고 그런 자세로 임해야 한다"고 강조했다. 정 위원장은 또 "내가 이야기한 것은 집권여당의 자세에 대한 이야기이지 인물에 대한 이야기가 아니다"며 "누구누구 차출론이나 이런 건 아무 상관이 없는 것"이라고 주장했다.  
 친윤계 핵심인 권성동 의원도 이날 '국민공감' 참석 후 "(한 장관 차출론은) 아주 극히 일부에서 주장하는 것 아닌가, 이렇게 보고 있다"고 말했다. 권 의원은 이어 "한 장관이 스스로 판단을 내릴 것"이라고 지적했다.  
 <d>박○○○ 기자 pmj@kmib.co.kr</d>  
 <d>GoodNews paper © 국민일보(www.kmib.co.kr), 무단전재 및 수정, 재배포금지</d>  
 </content>

<그림 19> 작업 편집 화면

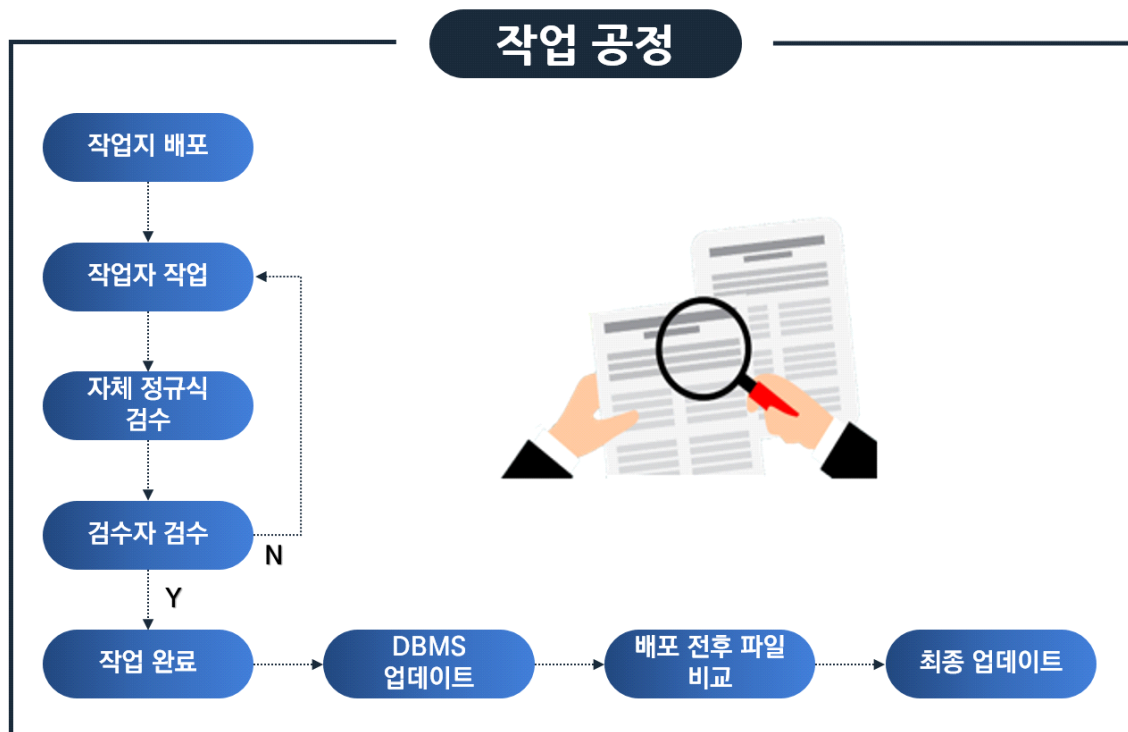
불필요한 요소는 정규식 목록을 활용하여 처리함으로써 확인 요소임을 분명히 하였다. 작업은 수행사가 가지고 있는 프로그램을 사용하였으며 모든 데이터는 기사 단위로 데이터 베이스 관리 시스템(DBMS)에서 처리하였다. 작업자들은 해당 기사를 엑스엠엘(XML) 데이터로 받아 정규식 목록을 이용하여 직접 삭제가 아닌 마크업을 부여하는 방식으로 작업하였다. 이때 사용하지 않는 기사는 기사 사용 여부를 나타내는 속성인 '유즈드(used)' 항목에 사용하지 않음을 표기하여 작업하였다. 각각의 요소들은 색깔로 구분하여 놓치지 않고 작업할 수 있도록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

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.



<그림 20> 작업 프로그램 화면

예를 들어 ‘~기자’로 끝나는 문장을 검색하는 정규식을 지정하였는데 기사 입력 시의 오타로 ‘기지’, ‘가지’ 등으로 입력이 된 경우가 발견된다면 이러한 오류 사항을 작업자들에게 공유하여 확인토록 하며, 오류가 빈번하다면 정규식으로 규칙을 추가함으로써 빠르고 정확하게 처리하는 것이 가능하다.



<그림 21> 데이터 정제 2차 검수 공정

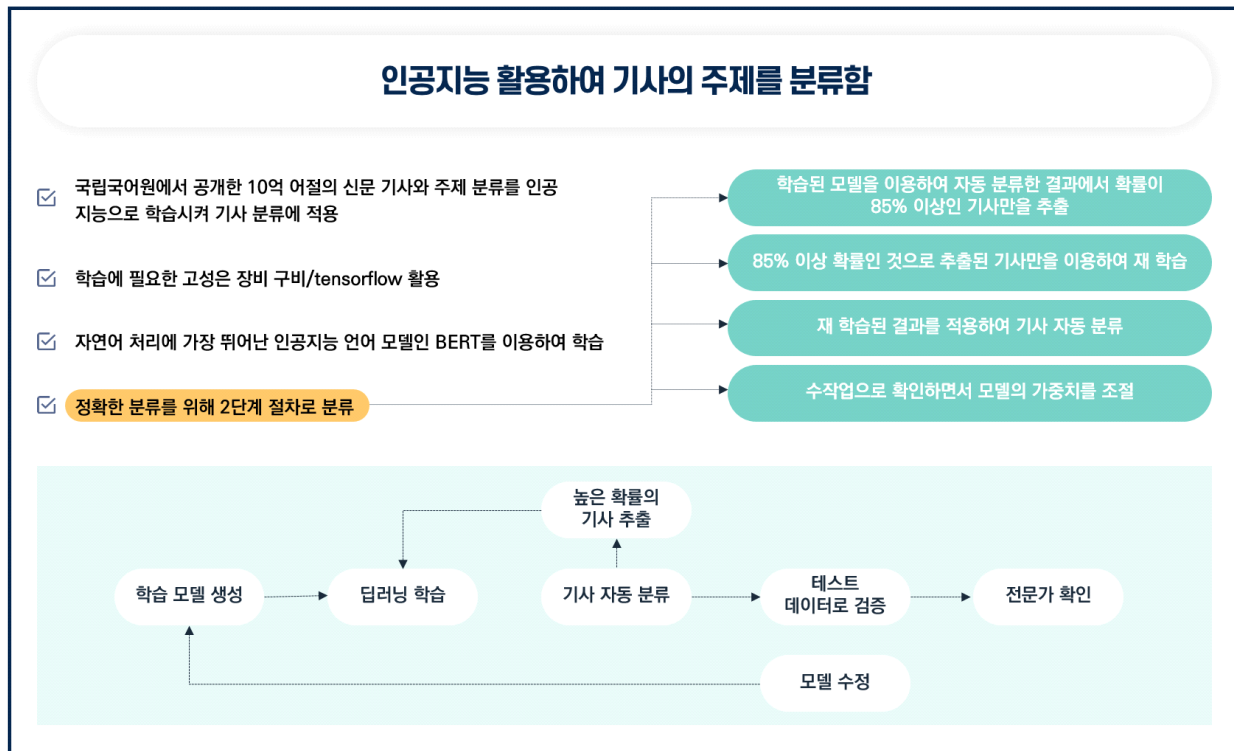
데이터 검수는 할당된 작업을 완료한 후 검수자가 만들어 놓은 오류 유형을 활용하여 1차로 자체 검수를 실시하였다. 사진, 출처, 전문, 전자 우편(이메일)으로 끝나는 문장 등 작업 완료된 내용을 작업자 스스로 1차 검수를 진행한 후, 검수 폴더에 올리면 검수자가 2차로 해당 파일을 전수 검수하였다.

오류 유형은 계속 갱신하여 작업자들에게 공유하였으며, 검수 도중 작업자의 오류가 많이 발견된 경우에는 파일을 반려한 뒤, 오류 유형에 대해 피드백하며 교육을 실시하였다.

## 5. 메타데이터 작성

메타데이터 작성은 신문 매체의 구분, 기사 제목, 기자 이름, 매체명, 어절 수, 원 주제, 국립국어원에서 제시한 아홉 가지의 주제 등을 작성하는 공정이다. 신문사별로 주제 범주를 명명하는 이름이 다양하다. 그렇기 때문에 원 주제명을 넣고 국립국어원이 지정한 주제 분류로 기사를 분류하여 해당 주제의 정보도 삽입하게 된다.(정치, 경제, 사회, 생활, IT/과학, 연예, 스포츠, 문화, 미용/건강의 통합 분류 체계)

기사의 주제 분류는 수행사가 가지고 있는 인공 지능 모델을 신문 기사 학습에 최적화시켜 진행하였으며 기존에 공개된 신문 기사와 주제를 활용하였으며, 이 중 정확도가 85% 이상인 데이터만을 선별하여 재학습한 모델을 사용하였다.

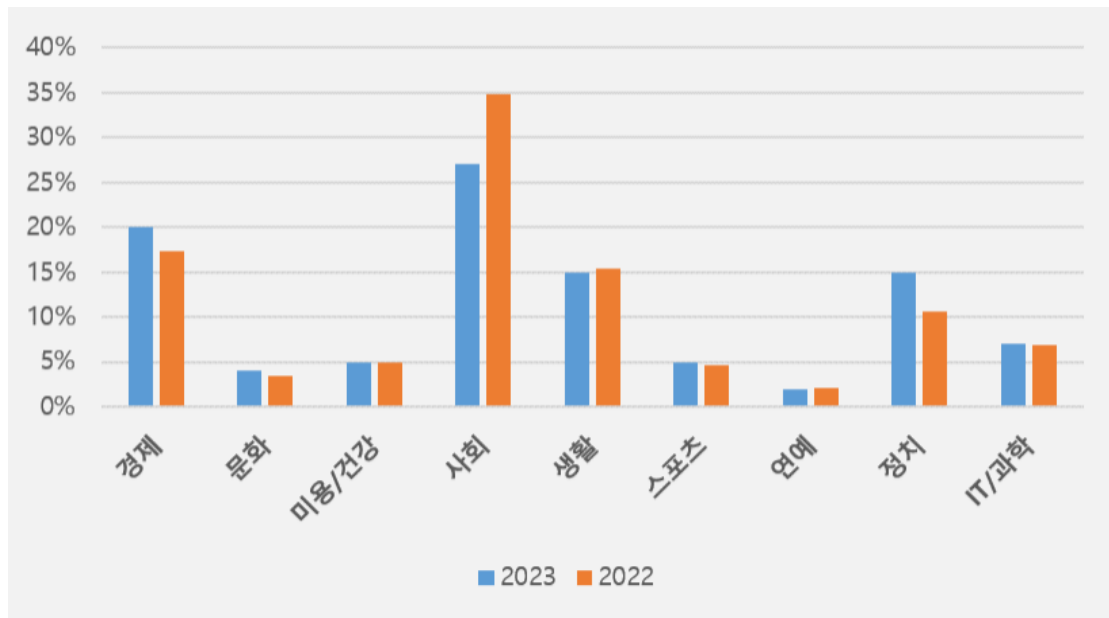


<그림 22> 인공 지능을 활용한 주제 분류

사회	경제	생활	정치	IT/과학	미용/건강	스포츠	문화	연예
26.8%	19.6%	15.3%	15.1%	7.3%	5%	5.3%	3.5%	2.1%

<표 14> 2023년 신문 기사 주제별 통계

기사 수 1,023,431



<그림 23> 연도별 기사 주제 통계

## 6. 인용 부호 수정 말뭉치

2차 정제가 끝난 데이터는 ‘신문 말뭉치’ 1종으로 구축된다. 이 사업 제안요청에 포함된 ‘인용 부호 수정 말뭉치’는 신문 말뭉치에서 인용 부호를 수정한 말뭉치를 의미하고 본 말뭉치가 모두의 말뭉치 공개 대상이다.

### 가. 인용 부호의 통일

같은 매체 안에서도 인용 부호의 표현은 다양하다. 표준 인용 부호를 사용하지 않고 키보드 엔터 키의 옆에 있는 작은따옴표'(0027)와 큰따옴표"(0022)를 표준 인용 부호 대신에 사용한 매체가 많았다.

인용 부호의 통일 작업에 해당하는 작업은 다음과 같다.

- ❖ 인용 부호가 열리고 닫히지 않거나, 열리지 않고 닫히는 등 짝이 맞지 않는 내용
- ❖ 인용 부호의 순서가 다르거나, 표준 인용 부호 대신에 다른 부호가 사용된 내용

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, 영어의 ‘아포스트로피(Apostrophe)’의 경우에는 그대로 살려 주었다.

다른 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 15> 인용 부호 치환 표

데이터 정제 전	데이터 정제 후
<p>‘ _발전과 안보 균형_’문구는 지난해 11월 열린 19기 공산당 중앙위원회 6차 전체회의(19기 6중전회)에서 채택된 공산당의 제3차 역사결의전문에도 들어갔다.</p>	<p>‘ _발전과 안보 균형_’문구는 지난해 11월 열린 19기 공산당 중앙위원회 6차 전체회의(19기 6중전회)에서 채택된 공산당의 제3차 역사결의전문에도 들어갔다.</p>
<p>A씨는경찰에서 “ _별다른 수입이 없는 상태에서 고액 아르바이트 자리가 있다는 말에 가담하게 됐다 ” _는취지의 진술을 했다.</p>	<p>A씨는경찰에서 “ _별다른 수입이 없는 상태에서 고액 아르바이트 자리가 있다는 말에 가담하게 됐다 ” _는취지의 진술을 했다.</p>
<p>‘ _Makeup_’, ‘ _Natural &amp; Organic_’, ‘ _Skin Care_’,  Tools &amp; Devices_’, ‘ _Hair_’등 5개 카테고리 부문에서 파이널(Final)을 선정해 어워즈를수여한다.</p>	<p>‘ _Makeup_’, ‘ _Natural &amp; Organic_’, ‘ _Skin Care_’,  ‘ _Tools &amp; Devices_’, ‘ _Hair_’등 5개 카테고리 부문에서 파이널(Final)을 선정해 어워즈를수여한다.</p>

<표 16> 인용 부호 수정 데이터 정제 전후

매체명	기사 수	어절 수	매체명	기사 수	어절 수
강원일보	16,202	2,942,252	세계일보	54,186	13,093,998
경기일보	9,116	1,881,120	스포츠서울	26,121	5,004,838
경북일보	15,815	3,185,012	아시아경제	83,062	18,209,492
경향신문	41,607	10,525,866	아주경제	15,431	3,860,139
국민일보	43,640	10,195,636	이데일리	86,516	20,193,756
남도일보	21,082	3,892,504	이투데이	52,610	11,635,613
내일신문	15,417	3,831,382	전북도민일보	15,334	2,635,454
노컷뉴스	59,278	12,256,628	조선일보	14,006	3,742,750
뉴스핌	57,801	11,498,298	중도일보	27,894	5,033,979
대구신문	12,211	2,358,123	중부일보	22,837	4,182,469
머니투데이	62,779	14,653,959	충청일보	16,195	2,618,936
부산일보	22,908	5,010,756	한겨레	22,785	6,018,200
서울경제	63,249	14,020,678	한국일보	32,549	7,894,278
서울신문	47,777	10,804,340	헤럴드경제	65,023	14,035,104
총 합				1,023,431	225,215,915

<표 17> 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 학습 및 데이터 유통에서 통일되지 않은 한자 코드는 기술적, 그리고 운영적 문제를 일으킬 소지가 있다. 데이터의 일관성과 신뢰성을 보장하기 위해 ‘한·중·일 호환용 한자 영역’ 내의 한자 문자 코드를 표준화하는 작업을 진행하였다.

데이터의 통일성을 확보하기 위해 같은 글자이면서 문자 코드가 다른 한자 문자들을 표준 유니코드로 일치시켜 주었다.



❖ 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	壟	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 18> ‘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.

전각 알파벳(A, B, C, a 등), 전각 부호( [ , ? , @ , ; , ( , ' , & 등), 전각 숫자( 0 , 1 , 2 등)는 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	~	
FF65	·	00B7	·	
FFE5	₩	00A5	₩	

대상 코드	대상 문자	대상 코드	치환 문자	비고
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 19> 치환 코드 목록

## 라. 오타 후보 문자 수정

작업자들이 신문 기사 전체를 읽어가는 과정에서 발견된 오타는 바로 수정을 진행하였다. 이후 최종적으로 오류 후보 글자들을 확인하여 오류 후보 글자들이 있는지 확인하는 과정을 거쳐 기사를 최종 선정하였다.

오류 후보 글자	해당 내용
겹	이집트 대표팀에서 해임된 이합 갈탈 감독이 지난 14일 서울 상암월드 <u>겹</u> 경기장에서 열린 한국과 평가전에서 작전을 지시하고 있다.
판	전라남도는 의사 1명과 전담 간호사 2명이 관리할 경우 환자 75명까지 대응할 수 있을 것으로 <u>판</u> 단하고 있다.
손	또 전투기가 항모 활주로에 충돌한 뒤 화염을 내 뿜은 채 활주로를 빙글빙글 도는 장면과 바다 속으로 추락 후 바다 속에서 엄청난 물기둥이 <u>손</u> 구치는 장면, 또 항모에 탑승한 해군 병사들이 놀라 우왕좌왕하는 모습 등도 고스란히 담고 있다.
졌	“사건의 진행 절차를 고민해 봐야 <u>졌</u> 지만, 새로운 정부 출범 이전에 마무리 될 사건도 있고, 계속 이어질 사건도 있을 것”이라며 “다만 마무리 될 수 있는 사건은 마무리 해야 한다고 본다”고 말했다.
복	또, “그건 중앙정치나 지방정치나 이치가 똑 같다. 정부 출범 초기에 정부조직법이나 급속을 요하는 기관 통폐합의 경우 <u>복</u> 잡한 정부절차를 거치기 보다는 통상 당정회의를 거쳐 의원입법 형태로 발의 되기도 한다”고 설명했다.
한	“내년 국고예산으로 반영된 <u>한</u> 안사업에 대해선 신속한 예산집행과 행정 절차 이행관리 등을 통해 사업성과를 극대화 하겠다”며 “2024년도 국고 확보도 선제적으로 대응해 신규사업을 집중 발굴하고 전남 행복시대를 열겠다”고 말했다.
를	이 작품은 인생의 가장 절망적인 순간에도 다시 일어서 나아가는 메시지가 담긴 그의 문장을 독창적인 스토리 <u>를</u> 통해 뮤지컬 형식으로 재구성했다.
벚	선정된 골목상권은 △동천역상가 △합지공원 먹골촌 △고성동 <u>벚</u> 꽃테마거리 △들안길 먹거리타운 △신평화골목이다.
휠르	큰 관심을 가져주시고 아낌없는 응원을 보내준 덕분에 지난해 대 <u>휠르</u> 성황리에 마칠 수 있었다”고 돌아봤다.
반	“임시주거시설은 피해 주민의 의견을 수렴해 상하수도 시설이 <u>반</u> 비된 임시주택, 원룸, 펜션, 리조트, 카라반 등 원하는 곳에서 거주할 수 있도록 지원하고, 친인척 집에서 거주하는 이재민에 대한 생계비 지원 기준도 마

오류 후보 글자	해당 내용
	련하라”고 지시했다.
ㅃ	NSC 상임위 회의는 이날 오전 10시40 <del>ㅃ</del> 부터 12시까지 80분간 진행됐다.

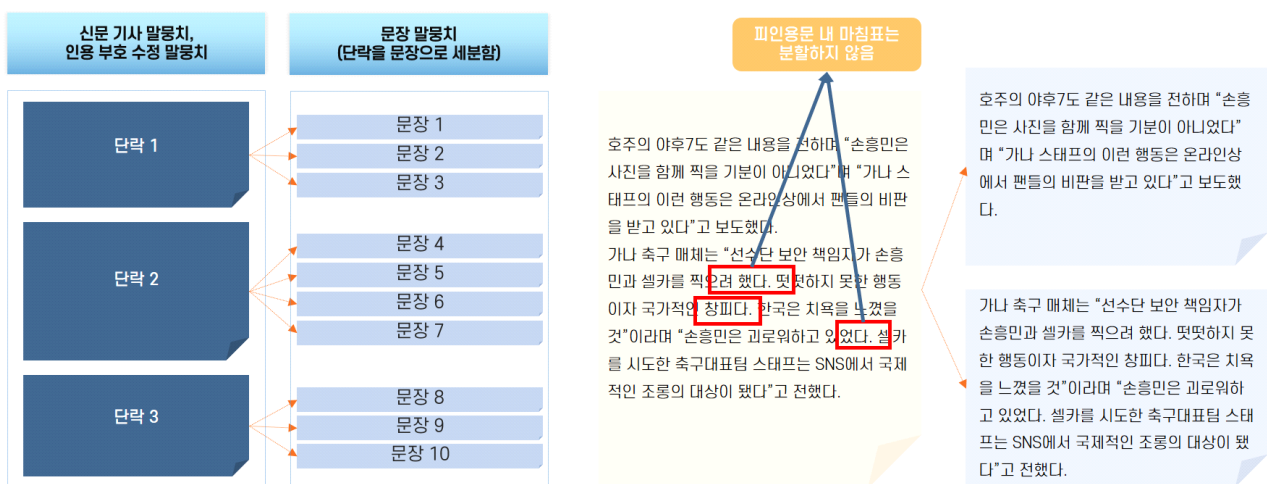
<표 20> 오타 글자

## 7. 문장 말뭉치 구축

문장 말뭉치는 인용 부호 수정 말뭉치와 함께 활용하기 위해 단락 구분을 표시하는 '<p>' 태그와, 문장 구분을 표시하는 '<s>' 태그를 삽입하여 문장을 구분하였다. 복수의 인용문을 한꺼번에 인용하는 경우에는 개개의 인용문을 문장 종결 부호 단위로 분할하지 않았다. 이렇게 구성된 데이터는 보다 정확한 문장 분할 작업을 지원하고, 자연어 처리 모델의 성능 향상에 기여할 것으로 기대된다.

### 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.
- ❖ 한꺼번에 인용되는 복수의 인용문은 마침표(.), 느낌표(!), 물음표(?) 등의 문장 부호에서 분할하지 않음.



<그림 24> 문장 말뭉치 개념

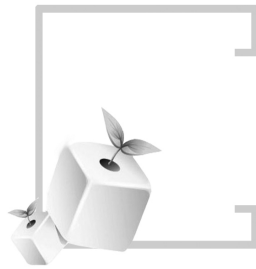
인용 부호 수정 말뭉치	문장 말뭉치
<p>&lt;p&gt;김 예비후보는 특히 윤석열 당선인과의 깊은 신뢰를 장점으로 내세웠다. 윤 당선인에게 강력한 지지를 보낸 대구시민의 기대를 염두에 둔 것으로 보인다.&lt;/p&gt;</p>	<p>&lt;p&gt;&lt;s&gt;김 예비후보는 특히 윤석열 당선인과의 깊은 신뢰를 장점으로 내세웠다.&lt;/s&gt; &lt;s&gt;윤 당선인에게 강력한 지지를 보낸 대구시민의 기대를 염두에 둔 것으로 보인다.&lt;/s&gt;&lt;/p&gt;</p>
<p>&lt;p&gt;그는 “대선 당시 클린선거전략본부장을 맡아 당선인에 대한 상대의 네거티브 공격을 방어했다. 서로 신뢰가 깊다”며 “당선인은 16개의 대구지역 대선 공약을 약속했는데, 이 공약들이 국정과제에 반영되어야만 정책 추진력이나 예산확보에 수월하다”고 설명했다. 그러면서 “당선인과 대구시장 간에 깊은 신뢰가 없다면 다른 지역공약에 밀려 후순위로 밀려날 것이 분명하다”며 “저는 누구보다 윤 당선인과 호흡을 잘 맞출 수 있는 책임자”라고 덧붙였다.&lt;/p&gt;</p>	<p>&lt;p&gt;&lt;s&gt;그는 “대선 당시 클린선거전략본부장을 맡아 당선인에 대한 상대의 네거티브 공격을 방어했다. 서로 신뢰가 깊다”며 “당선인은 16개의 대구지역 대선 공약을 약속했는데, 이 공약들이 국정과제에 반영되어야만 정책 추진력이나 예산확보에 수월하다”고 설명했다.&lt;/s&gt; &lt;s&gt;그러면서 “당선인과 대구시장 간에 깊은 신뢰가 없다면 다른 지역공약에 밀려 후순위로 밀려날 것이 분명하다”며 “저는 누구보다 윤 당선인과 호흡을 잘 맞출 수 있는 책임자”라고 덧붙였다.&lt;/s&gt;&lt;/p&gt;</p>

<표 21> 문장 말뭉치 데이터 정제

상위 5 문장분할수 ↑			↓ 하위 5 문장분할수		
1	조선일보	2.99	1	충청일보	1.06
2	한겨레	2.54	2	경북일보	1.17
3	경향신문	2.22	3	전북도민일보	1.17
4	한국일보	2.16	4	중부일보	1.18
5	서울경제	2.03	5	중도일보	1.27

<그림 25> 문단 내 문장 분할 수(상/하위 5개 매체)

<그림 25>는 한 문단(매체가 나눈 단락) 내에서 수행사가 문장으로 나눈 평균 분할 수이다. 조선일보의 경우 문장 분할 평균이 한 단락당 2.99개로 가장 높았으며, 충청일보의 경우 1.06개로 가장 낮은 수치를 보였다. 평균 문장 분할 수는 1.67개이다.



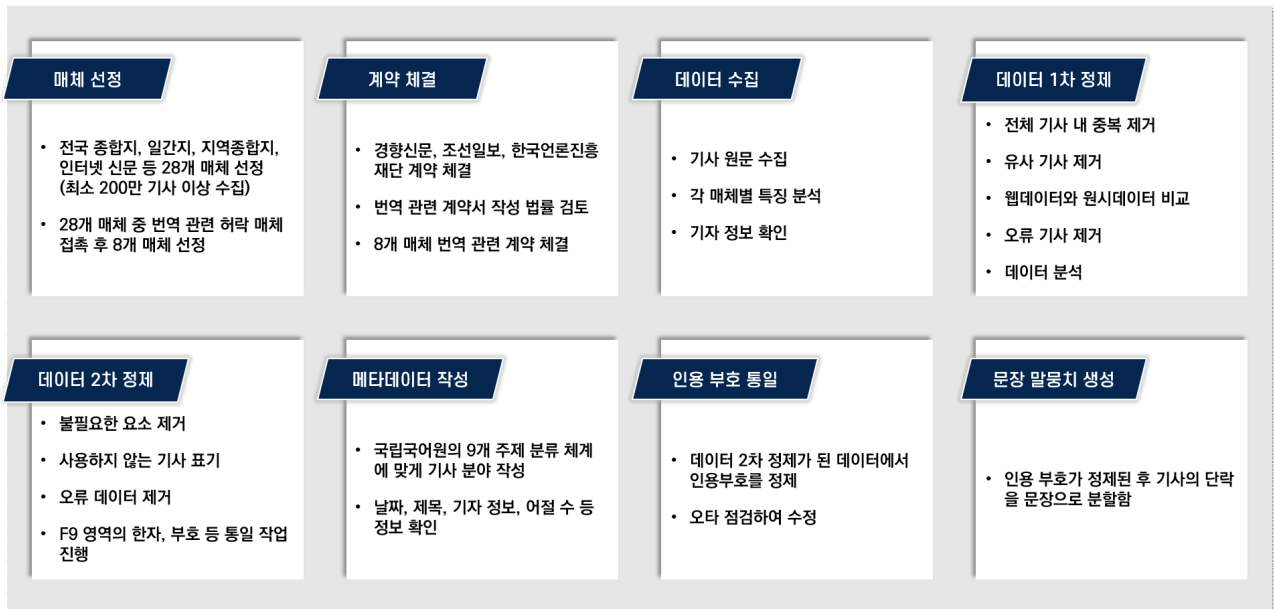
## 제 3 장

# 사업 수행 결과



## 제3장 사업 수행 결과

### 1. 신문 기사 정제 결과



<그림 26> 구축 공정별 내용

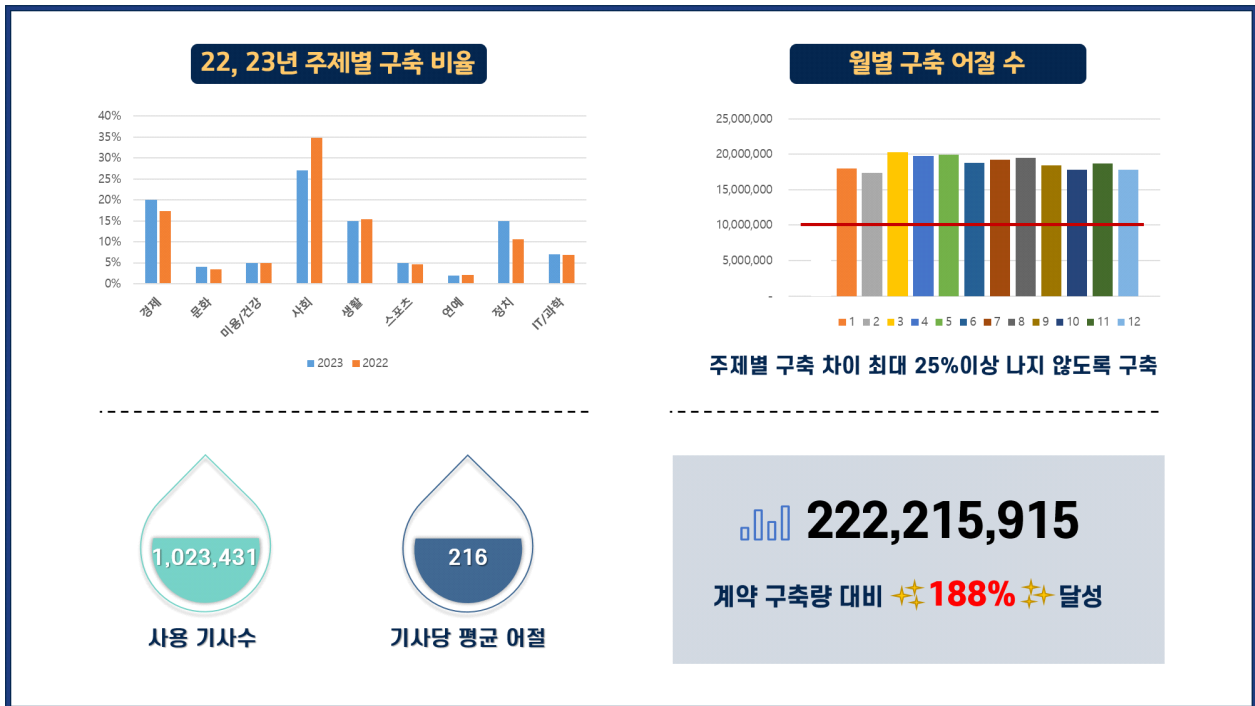
신문 기사의 정제 결과는 다음과 같다. 말뭉치 구축 기사 건수는 1,023,431건이며, 총 구축 어절 수는 계약 구축량 대비 188% 이상 달성한 222,215,915개의 어절을 구축하였다. 납품 말뭉치는 총 3종으로 구성되어 있으며 저작권 문제가 없고 불필요한 요소와 오류가 없는 신문 기사 말뭉치, 신문 기사 말뭉치에서 인용 부호를 수정한 인용 부호 수정 말뭉치, 인용 부호 수정 말뭉치에서 단락을 문장으로 분할한 문장 말뭉치를 구축하여 납품하였다.



매체명	최초 수집 기사 수	최초 수집 어절 수	정제 수집 기사 수	정제 수집 어절 수
강원일보	51,789	6,339,133	16,202	2,942,252
경기일보	21,710	5,100,189	9,116	1,881,120
경북일보	27,998	5,794,014	15,815	3,185,012
경향신문	74,506	14,170,331	41,607	10,525,866
국민일보	81,025	20,900,800	43,640	10,195,636
남도일보	37,596	8,081,081	21,082	3,892,504
내일신문	27,468	8,562,816	15,417	3,831,382
노컷뉴스	130,429	32,528,843	59,278	12,256,628
뉴스핌	266,726	42,399,013	57,801	11,498,298
대구신문	33,513	7,397,030	12,211	2,358,123
머니투데이	137,207	32,131,262	62,779	14,653,959
부산일보	72,615	17,197,990	22,908	5,010,756
서울경제	134,620	31,818,165	63,249	14,020,678
서울신문	98,431	21,770,907	47,777	10,804,340
세계일보	141,769	36,728,629	54,186	13,093,998
스포츠서울	98,657	13,835,337	26,121	5,004,838
아시아경제	167,293	37,475,022	83,062	18,209,492
아주경제	78,524	21,269,806	15,431	3,860,139
이데일리	183,808	34,917,001	86,516	20,193,756
이투데이	106,781	21,317,923	52,610	11,635,613
전북도민일보	36,083	6,351,528	15,334	2,635,454
조선일보	49,847	11,044,217	14,006	3,742,750
중도일보	65,607	12,210,837	27,894	5,033,979
중부일보	41,774	8,779,705	22,837	4,182,469
충청일보	62,620	9,727,287	16,195	2,618,936
한겨레	45,599	15,069,584	22,785	6,018,200
한국일보	65,335	18,162,483	32,549	7,894,278
헤럴드경제	148,029	36,799,518	65,023	14,035,104
<b>총 합</b>	<b>2,487,359</b>	<b>537,880,451</b>	<b>1,023,431</b>	<b>225,215,915</b>

<표 22> 신문 기사 정제 총괄표

2023년 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 216개이다. 올해 구축 분이 최대 어절 수를 기록했음을 아래 <그림 27>과 <표 23>으로 알 수 있다.



<그림 27> 매체별 최종 기사 수 및 월별 구축 어절 수

내용/연도	2020년도	2021년도	2022년도	2023년도
기사 기간	1년치 기사	1년치 기사	1년치 기사	1년치 기사
매체 수	35	35	34	28
기사 수	630,095	730,017	978,344	1,023,431
어절 수	150,669,174	203,585,743	208,320,912	222,215,915

<표 23> 구축 연도별 기사와 어절 수

월	어절 수
1월	17,992,302
2월	17,368,552
3월	20,239,605
4월	19,735,085
5월	19,906,578
6월	18,732,930
7월	19,177,211
8월	19,427,073
9월	18,414,237
10월	17,767,353
11월	18,677,278
12월	17,777,711
합계	225,215,915

<표 24> 월별 구축 어절 수

주제별로 차이가 25% 이상이 되지 않도록 해야 하는 기술협상의 내용도 충족하였다. 기사 단위 주제별 분포는 다음과 같다.

주제별	기사 수	어절 수
경제	200,739	19.6%
문화	36,208	3.5%
미용/건강	51,050	5.0%
사회	274,416	<u>26.8%</u>
생활	156,393	15.3%
스포츠	54,662	5.3%
연예	21,120	<u>2.1%</u>
정치	154,183	15.1%
IT/과학	74,660	7.3%
합	1,023,431	100%

<표 25> 주제별 기사 및 구축 어절 수

## 2. 매체별 납품 파일명

말뭉치 유형 구분의 N은 신문 기사 말뭉치를, 분석 층위 구분의 RW는 원시를 의미하며, 매체 장르 및 구분 정보는 다음과 같다. (W: 전국 종합지, L: 지역 종합지, P: 전문지, I: 인터넷 기반 신문, Z: 기타)

말뭉치 유형 구분	매체 및 장르 분류	분석 층위 구분	구축 연도	매체일련번호	매체명
N	I	RW	23	00000001	노컷뉴스
N	I	RW	23	00000002	뉴스핌
N	L	RW	23	00000001	강원일보
N	L	RW	23	00000002	경기일보
N	L	RW	23	00000003	경북일보
N	L	RW	23	00000004	남도일보
N	L	RW	23	00000005	대구신문
N	L	RW	23	00000006	부산일보
N	L	RW	23	00000007	전북도민일보
N	L	RW	23	00000008	충도일보
N	L	RW	23	00000009	충부일보
N	L	RW	23	00000010	충청일보
N	P	RW	23	00000001	머니투데이
N	P	RW	23	00000002	서울경제
N	P	RW	23	00000003	아시아경제
N	P	RW	23	00000004	아주경제
N	P	RW	23	00000005	이데일리
N	P	RW	23	00000006	이투데이
N	P	RW	23	00000007	스포츠서울
N	P	RW	23	00000008	헤럴드경제
N	W	RW	23	00000001	경향신문
N	W	RW	23	00000002	국민일보
N	W	RW	23	00000003	내일신문
N	W	RW	23	00000004	서울신문
N	W	RW	23	00000005	세계일보
N	W	RW	23	00000006	조선일보
N	W	RW	23	00000007	한겨레
N	W	RW	23	00000008	한국일보

<표 26> 말뭉치 파일명

<부록 1>

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 2022년 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 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 2023년 12월 31일까지로 한다.

(2) ‘대상저작물’ 및 ‘복제·변형물’의 이용허락 최소 기간은 계약체결일부터 2034년 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 이용할 때 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 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 권리와 이용자가 상호 원만한 합의에 이르도록 노력하여야 하며, 분쟁이 원만히 해결되지 않는 경우에는 소 제기에 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.

2023년      월      일

권리자 :

성명

주소

이용자 :

성명    국립국어원장 (인)

주소 서울특별시 강서구 금남화로 154

제안요청서에 삽입된 계약서 형태

## <부록 2>

저작재산권 비독점적 이용허락 계약서(2차적 저작물 작성권[번역 이용 허락] 관련)

## 저작재산권 비독점적 이용허락 계약서

저작자 또는 저작재산권 이용허락자 \_\_\_\_\_(이하 “권리자” 이라 함)와 저작재산권 이용자 국립국어원(이하 “이용자” 이라 함)는 아래 저작물 신문기사에 관한 저작재산권 이용허락과 관련하여 다음과 같이 계약을 체결한다.

### 다 음

#### 제1조 (계약의 목적)

이 계약은 저작재산권 이용허락과 관련하여 권리자와 이용자 사이의 권리관계를 명확히 하는 것을 목적으로 한다.

#### 제2조(정의)

이 계약에서 사용하는 용어의 뜻은 다음과 같다.

1. ‘전체 기사’란 권리자가 제공하는 2022년 1년 동안 생산된 신문 기사 원문 자료를 말한다.
2. ‘수집 기사’란 국립국어원 및 국립국어원이 발주한 용역 사업의 수행자(이하 “과업수행자”라 함)가 ‘전체 기사’에서 수집한 기사를 말한다.
3. ‘대상저작물’이란 ‘수집 기사’ 중 국립국어원 및 과업수행자가 말뭉치 구축 대상으로 선정한 기사 원문을 말한다.
4. ‘복제·변형물’이란 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 및 범위 등)

① 이 계약에 따라 권리자가 이용자에게 비독점적 이용을 허락하는 저작물 및 그에 관한 저작권의 대상과 범위는 아래와 같다.

1. 원저작권자 : \_\_\_\_\_
2. 저작물 종류 : ☐ 어문저작물, ☐ 음악저작물, ☐ 연극저작물, ☐ 미술저작물,  
☐ 건축저작물, ☐ 사진저작물, ☐ 영상저작물, ☐ 도형저작물,  
☐ 컴퓨터프로그램저작물, ☐ 기타( )
3. 이용허락 대상 저작물 : \_\_\_\_\_ 매체의 2022년 신문 기사 중 ‘대상저작물’,

‘복제 변형물’

4. 이용 허락 저작권 : ☐ 복제권, ☐ 공연권, ☐ 공중송신권(☐ 방송권, ☐ 전송권, ☐ 디지털음성송신권), ☐ 전시권, ☐ 배포권, ☐ 대여권, ☐ 2차적저작물작성권 중 번역저작물의 생성·작성·복제·전시·배포에 관한 권리
5. 2차적저작물작성권의 범위 : 영어, 베트남어, 인도네시아어, 태국어, 인도 힌디어, 캄보디아 크메르어, 필리핀 타갈로그어, 러시아어, 우즈베크어, 아랍어를 포함하여 국립국어원이 한국어의 홍보와 교류를 위하여 필요한 국가의 언어로 번역하여 생성되거나 작성된 번역물의 생성 및 작성에 관한 권리. 이때 위 10개의 언어 이외의 다른 언어로 번역될 경우, 이용자는 권리자에게 번역되는 언어를 통지하기로 한다.
6. 저작권 이용료: 과업수행자’는 권리자에게 저작권 이용료로 \_\_\_\_\_을 본 계약 일로부터 \_\_\_\_\_ 이내에 일시금으로 권리자가 지정하는 계좌에 지급한다.
- ② 이용자는 권리자의 ‘전체기사’를 \_\_\_\_\_의 계약을 통해 데이터를 제공받는다.
- ③ 이용자는 제1항의 ‘대상저작물’, ‘복제 변형물’에 관하여 저작권법에 따라 아래 각 호의 행위를 할 수 있다.
1. 저작권법 제33조에 따라 시각장애인 등을 위하여 점자로 변환·복제·배포하는 일
  2. 저작권법 제33조의2에 따라 청각장애인을 위하여 한국수어로 변환·복제·배포·공연 또는 공중송수신하는 일
- ④ 이용자가 제1항에서 정한 저작물 및 저작권의 범위 외에, 그 범위에 해당하지 않는 저작물이나 저작권을 비독점적으로 이용하고자 할 때에는, 권리자와 별도로 대상 저작물 및 그에 관한 저작권의 범위에 관하여 사전에 서면 합의를 해야 한다.

제4조 (이용허락 기간)

‘대상저작물’ 및 ‘복제·변형물’의 이용허락 최소 기간은 계약체결일부터 2034년 12월 31일까지로 한다. 최소 기간 만료 후 권리자 또는 저작자인 언론사가 이용허락 중지 의사를 밝히지 아니하면 이용허락이 1년 단위로 자동 갱신되고, 갱신 기간 중 언론사도 권리자 또는 저작자인 언론사가 이용허락 중지 의사를 밝히면 그 의사 내용에 따라 이용허락이 중지된다.

### 제5조 (권리자의 의무)

- ① 권리자는 이용자에게 대상저작물에 관하여 이 계약서 제3조에 따른 저작권재산권을 이용할 권리를 제4조의 기간 동안 비독점적으로 허락한다.
- ② 권리자는 대상저작물에 제3자의 이용허락권, 질권 등이 존재하는 경우, 이용자에게 그 사실을 사전에 알려야 한다.
- ③ 권리자는 대상저작물의 저작권재산권 전부 또는 일부를 제3자에게 양도하거나 이에 대하여 질권을 설정하고자 하는 경우, 사전에 이용자에게 이 사실을 통보하여야 한다.

### 제6조 (이용자의 권리 및 의무)

- ① 이용자는 제4조의 기간 동안 제3조에서 정한 대상과 범위 내에서 이를 비독점적으로 자유롭게 이용할 수 있다.
- ② 이용자는 제3조에서 정한 권리자가 이용을 허락한 대상 및 범위에 관한 이용자의 권리에 관하여, 권리자의 서면에 의한 사전 동의가 없이는, 제3자에게 양도하거나 이에 대하여 질권을 설정할 수 없다.
- ③ 이용자는 저작권법에서 정한 바와 일반적인 관례에서 행하는 바에 따라, 이 계약에서 정한 비독점적 이용이 가능한 대상을 이용하는 경우 및 이 계약에서 정한 비독점적 이용이 가능한 권리를 행사하는 경우, 그 원저작자 및 저작권재산권자의 성명 등을 표시하여야 한다.
- ④ 이용자는 제3조에서 정한 대상과 범위를 비독점적으로 이용하고 행사할 때 원저작자의 저작인격권을 침해하지 아니한다. 다만, 비독점적 이용권을 행사할 때 저작인격권을 침해하지 않는 범위 내에서 필수불가결적으로 사소한 수정이나 편집이 요구된다는 사실을 원저작권자에게 사전에 고지하면, 비독점적 이용권을 행사할 때 필수불가결적으로 요구되는 바에 따라 최소한으로 사소한 수정이나 편집을 할 수 있다.

### 제7조 (확인 및 보증)

- ① 권리자는 이용자에게 다음 각 호의 사항을 확인하고 보증한다.
  - 1. 이 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. 제3조에서 정한 대상 및 범위 내에서만 비독점적으로 자유롭게 이용한다는 것
  - 2. 제1호의 이용에서 제3자의 명예 등의 인격적 권리를 침해하지 아니할 것이라는 점

3. 제1호의 이용에서 그 이용에 따른 결과물의 제공·배포 시 이용허락 조건 및 재배포 금지, 목적 외 사용 금지 등 주의사항을 고지할 것

#### 제8조 (계약내용의 변경)

이 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 효력이 발생한다.

#### 제9조 (계약의 해지)

- ① 당사자는 천재지변 또는 기타 불가항력으로 계약을 유지할 수 없는 경우에 이 계약을 해지할 수 있다.
- ② 당사자는 상대방이 정당한 이유 없이 이 계약을 위반하는 경우에 적절한 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 이 계약에 대한 해지권의 행사는 상대방에 대한 손해배상청구권 행사에 영향을 미치지 아니한다.

#### 제10조 (손해배상)

당사자가 정당한 이유 없이 이 계약을 위반하는 경우, 그로 인하여 상대방에게 발생한 모든 손해를 배상할 책임이 있다. 다만, 제9조 제1항의 사유로 이 계약을 이행하지 못한 경우에는 손해배상책임을 면한다.

#### 제11조 (비용의 부담)

계약 체결에 따른 비용은 권리자와 이용자가 동등하게 부담한다.

#### 제12조 (분쟁해결)

- ① 이 계약에서 발생하는 모든 분쟁은 권리자와 이용자가 상호 원만한 합의에 이르도록 노력하여야 하며, 분쟁이 원만히 해결되지 않는 경우에는 소 제기에 앞서 한국저작권위원회에 조정을 신청할 수 있다.
- ② 제1항에 따라 분쟁이 해결되지 아니할 때에는 1심 소송의 관할법원은 서울남부지방법원으로 한다.

### 제13조 (비밀유지)

양 당사자는 이 계약의 체결 및 이행과정에서 알게 된 상대방에 관한 정보, 이 계약의 내용 및 대상저작물의 내용을, 상대방의 서면에 의한 승낙 없이 제3자에게 공개하여서는 아니 된다.

### 제14조 (계약의 해석 및 보완)

이 계약서에서 명시되어 있지 아니하거나 해석상 이견이 있을 경우에는 저작권법, 민법 등을 준용하고 사회 통념과 조리에 맞게 해결한다.

### 제15조 (계약 효력 발생일)

이 계약의 효력은 계약 체결일부터 발생한다.

2023년 00월 00일

권리자

이용자

국립국어원장

서울특별시 강서구 금남화로 154

원장 (인)



### <부록 3>

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>외부 기고가가 작성한 기사               <ul style="list-style-type: none"> <li>교수, 박사, 의원, 소장, 대표, 변호사 등</li> <li>해당 언론 소속 기자 이외의 모든 직업 및 직책</li> </ul> </li> <li>명예기자, 객원기자, 시민기자, 도민기자, 학생기자가 작성한 기사               <ul style="list-style-type: none"> <li>명예기자나 객원기자라고 표기되지 않고, 이름만 나오는 경우는 그대로 사용</li> </ul> </li> <li>외부 기고문임을 의미하는 단어를 포함하는 기사               <ul style="list-style-type: none"> <li>[기고], [특별기고], [발언대], [확대경], [아침의 창] 등</li> <li>매체별로 다양하게 쓰이고 있으므로 반드시 확인</li> </ul> </li> <li>외국 기사를 번역한 기사</li> <li>다른 매체의 헤드라인이나 단신, 일정을 모아 놓은 기사               <ul style="list-style-type: none"> <li>00월 00일 언론동정, 00일 국회 일정 등</li> </ul> </li> <li>공동취재단이 작성한 기사               <ul style="list-style-type: none"> <li>공동취재단, 공동취재팀, 공동취재반, 특별취재단 등</li> </ul> </li> <li>기자명이 비어있는 기사</li> <li>방송을 그대로 옮겨 적은 기사</li> <li>동일 미디어 그룹 내의 다른 매체 소속 기자가 작성한 기사</li> <li>연합뉴스발 기사</li> </ul>
구어체 기사	<ul style="list-style-type: none"> <li>대부분이 구어체로 이루어진 기사는 사용하지 않음               <ul style="list-style-type: none"> <li>친근하게 전달하기 위해 구어체로 쓴 기사, 인터뷰를 구어체 그대로 옮긴 기사 등</li> <li>구어체의 경우 ‘~입니다’ 등과 같은 경우 사용함</li> </ul> </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 않음</li> </ul>

불완전하게 종료되는 기사	<ul style="list-style-type: none"> <li>• ‘관계자는...라고 말’, ‘정부는...예정인’ 처럼 ‘했다.’, ‘다.’가 빠진 것으로 유추할 수 있는 경우에는 사용</li> <li>• 누락된 글자를 유추하여 기사의 마지막 문장을 완성을 시킬 수 있는 경우에는 사용하지만, 문장의 대부분, 또는 주요성분이 누락되어 유추가 어려운 수준으로 불완전한 기사는 사용하지 않음</li> </ul>
한글이 모아쓰기 되지 않은 기사	<ul style="list-style-type: none"> <li>• 한글이 모아쓰기 되지 않은 경우, 원래의 단어를 확정할 수 있으면 사용하고, 수정 후보가 많으면 사용하지 않음</li> </ul>
명확한 광고 기사	<ul style="list-style-type: none"> <li>• 기사에서 명확히 광고라고 표기하는 경우에는 사용하지 않음 <ul style="list-style-type: none"> <li>– 별도의 협찬, 제공을 받았음을 밝힌 기사도 사용하지 않음</li> </ul> </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> </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 정보	<p>(사진제공=건국대학교)      표&gt; 공정위 망 이용대가 불공정 조사 쟁점</p> <p>(사진제공=SBA)      &lt;표&gt; 한○○ 후보자 주요 ICT 정책 현안 입장</p> <p>[사진 제공= 로이터]      [표]</p> <p>사진제공=tvN      ▲영상제공=</p> <p>사진=CJ엔터테인먼트 제공      사진=FNC엔터테인먼트 제공</p> <p>사진제공   카카오TV      출처: 라디오타임스 / 굿모닝브리튼, 사진=스타쉽 제공</p> <p>사진/이○○ 숲 해설가      출처  송○○ SNS</p> <p>[그래픽]      &lt;그래픽&gt;      [영상=시너지영상팀]</p> <p>일러스트       (사진설명)</p> <p>화면 캡처      (조감도)</p>
기자의 이름, ID 등	<p>[세종= 주○○ 기자]</p> <p>이○○기자/사진=강화군 제공</p> <p>글·사진=양○○ 기자 -----@---co.kr</p> <p>김○○ -----@---.co.kr, 사진=인터파크 제공</p>
‘Copyright©’ 등 저작권 관련 내용	<p>랄프 김슨 'Salon Litteraire'. ©Ralph Gibson</p> <p>Copyright © ○○○○. All rights reserved. 무단 전재 및 재배포 금지.</p> <p>&lt;저작권자 © 1980-2022 ○○일보. 무단 전재 재배포 금지.&gt;</p>
전문	<p>[○○○○○   남○○기자] NC 엔터테인먼트가 ○○과의 전속계약 만료 소식을 전했다.</p> <p>13일 FNC는 공식 홈페이지에 “소속 아티스트의 전속 계약 기간이 2022년 1월 12일로 종료되어 안내드린다”고 밝혔다.</p> <p>이어 소속사는 “지난 9년간 당사 소속 가수로서 활발한 활동을 이어온 아티스트에게 감사의 마음을 전한다. 비록 당사와 함께하는 인연은 마무리되었지만, 지민의 앞날과 향후 행보에 따뜻한 격려와 응원 부탁 드린다”고 덧붙였다.</p> <p>이하 FNC 글 전문.</p> <p>안녕하세요.</p> <p>FNC엔터테인먼트입니다.</p> <p>소속 아티스트의 전속 계약 기간이 2022년 1월 12일로 종료되어 안내드립니다.</p> <p>지난 9년간 당사 소속 가수로서 활발한 활동을 이어온 아티스트에게 감사의 마음을 전합니다.</p> <p>비록 당사와 함께하는 인연은 마무리되었지만, 지민의 앞날과 향후 행보에 따뜻한 격려와 응원 부탁 드립니다.</p> <p>감사합니다. ---@sportsseoul.com</p> <p>사진출처  ○○ 인스타그램</p>
기사 본문으로 볼 수 없는 부가 정보 의 나열 등	<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 가격 안정화’가 15.0%, ‘광역교통망 구축’ 13.6%, ‘미세먼지 대책마련’이 10.7%, ‘수도권 규제 완화’가 3.6% 순이다. ‘기타’가 7.5%, ‘잘 모름’이 2.4%다.</p>

	<p>지역별로는 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><u>○○○기자</u>  <u>어떻게 조사했나</u>          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><u>※오차보정방법 : [립가중] 성별, 연령별, 지역별 가중값 부여(2019년 11월말 행정안전부 발표 주민등록인구기준)</u></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>통계 정보를 설명하고 있으나 위와 같은 내용은 기사 본문과 이어지는 것으로 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><u>자세한 내용은 아래 개요를 참고하시기 바랍니다. 대한민국을 이끄는 혁신리더들의 많은 참여 바랍니다.</u></p> <p><u>[2020 대한민국지속가능혁신리더대상 개요]</u>  <u>주 최 : 2020 대한민국지속가능혁신리더대상 조직위원회</u>  <u>주 관 : 머니투데이, 더리더</u>  <u>접수마감 : 2020년 6월 30(화)</u>  <u>접수문의 : 02-724-0952(머니투데이 더리더)</u>  <u>접 수 처 : 이메일(awards@mt.co.kr)</u>  <u>시상일시 : 2020년 7월 중</u>  <u>시상장소 : 여의도 쉼튼호텔</u>  <u>신청대상 : 정치·사회·경제·교육·체육·문화·예술·환경 등 각 분야의 지속적인 혁신 공로가 인정되는</u>  <u>지자체 및 우수 기관, 단체, 개인리더 등</u>  <u>신청양식 : 더리더 홈페이지 우측 상단 배너 클릭, 기사하단 신청서 다운로드에서 클릭 후 내려받기 가능</u></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>10일 통계청에 따르면 7월 외식물가지수는 전년 동월 대비 8.4% 상승했다. 지난 1992년 10월(8.8%) 이후 오름폭이 가장 컸다. 통계청이 집계하는 전체 외식 품목 39개의 물가도 모두 상승했다. 결식아동이 주로 찾는 분식집의 김밥, 라면 등 메뉴 가격이 두 자릿수 상승률을 기록한 것으로 나타났다.</p> <p>이렇다 보니 현장에서는 "김밥 한 줄도 4000원인데 7000원 가지고 라면하고 김밥도 못 먹는다" 는 등 목소리가 나왔다. 전문가와 아동보호기관도 물가상승에 따라 꿈나무카드 급식지원 단가를 인상해야 한다고 제언하기도 했다. (관련기사: "7000원으로 배고픔 사라질까요?" 한도 빠듯한 '꿈나무카드' [결식아동 배부르게①])</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, 갯세븐, 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>▲ 팬앤스타 최다 득표상(가수) : 슈퍼주니어</p> <p>▲ 팬앤스타 최다 득표상(개인) : 황치열</p> <p>▲ 팬앤스타 초이스상(가수) : 슈퍼주니어</p> <p>▲ 팬앤스타 초이스상(개인) : 황치열</p> <p>유족으로는 딸 이희경씨, 동생 은화(전 이화여대 교수)·효숙·성숙씨, 율케 이부자씨가 있다. 여성단체들은 여성장으로 고인을 배웅하기로 했다. 빈소는 창원경상대병원 장례식장 VIP 1호실에 마련됐다. (055)214-1910.</p> <p>김○○ 기자 ○○○○@○○○○.co.kr</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>이○○ 기자 b○○○○@○○○○.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</p>
--	---

	<p>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. <b>▲1942년 서울 출생 ▲서울대 정치학과 ▲동아일보 기자 ▲14~16대 국회의원 ▲한나라당 부총재 ▲열린우리당 의장 ▲동아시아평화국제회의 조직위원장 ▲자유언론실천재단 이사장</b></p> <hr/> <p>몸매관리 비법으로 밀크어트를 소개한 그녀는 자신만의 다이어트 비법인 건강음료를 공개해 화제가 되고 있다.  손쉽게 만들 수 있는 오영주 표 밀크어트 건강음료 3가지를 소개한다.</p> <p><b>■ 아보카도 스무디</b></p> <p><b>&lt;재료&gt;</b>  우유 200ml, 아보카도 1/2개, 바나나 1개</p> <p><b>&lt;만드는 방법&gt;</b>  아보카도를 반으로 갈라 씨와 껍질을 제거하고, 우유, 아보카도, 바나나 등 모든 재료를 믹서에 넣고 갈아주면 완성이다.</p> <p><b>■ 고구마라떼</b></p> <p><b>&lt;재료&gt;</b>  우유 300ml, 삶은 고구마 1개</p> <p><b>&lt;만드는 방법&gt;</b>  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 상관없는 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 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.</p> <p><b>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.</b></p> <p><b>[출처: ○○○○에서 제공하는 기사입니다.]</b></p> <p><b><a href="https://○○○○.co.kr/news/newsView.php?id=○○○○○○○#csidx4dab120876716f7a9506745d61f1391">https://○○○○.co.kr/news/newsView.php?id=○○○○○○○#csidx4dab120876716f7a9506745d61f1391</a></b></p>

문장의 오류	농심이 매출액의 <u>매출액의</u> 2.15%를 소아암 환아에게 기부하는 백산수 한정판을 출시했다고 29일 밝혔다.
	<p>앞서 핀란드와 <u>핀란드와</u> 스웨덴이 오는 6월 스페인에서 개최하는 나토 정례 회의에서 가입 신청서를 제출할 수 있다는 관측이 현지 매체 등을 통해 제기됐다.</p> <p>=====</p> <p>‘3년이 지난 지난해 10월’, ‘비정규직 정규직화’ 같은 문장들을 단어 반복 오류로 보고 삭제하지 않도록 유의한다.</p>



#### <부록 4>

말뭉치 종류별 구축 예시

원시데이터	<p><b>강화군, 'DMZ 평화의 길' 테마노선 운영</b> 강화군이 16일부터 평화와 통일로 가는 'DMZ 평화의 길 테마노선' 참가자 신청을 받는다.</p> <p>'디엠지(DMZ) 평화의 길' 테마노선은 강화 전쟁박물관을 시작으로 연미정과 고려천도공원을 거쳐 평화전망대까지 연결된다. 이어 의두분초와 의두돈대를 찍고 교동대교 건너 대룡시장까지 61.1km를 한강하구 너머 북녘과 마주하며 강화도 북부지역을 걷는 노선이다. 북녘 땅을 내려다볼 수 있는 의두분초와 의두돈대는 민간에 개방되지 않는 군사시설과 야생동물 서식지로 참가자들의 안전을 위해 군부대의 협조를 받아 이동하게 된다. 모든 코스를 둘러보는데 6시간 정도 소요된다.</p> <p>참가신청은 '디엠지(DMZ) 평화의 길' 누리집을 통해 사전 예약하면 된다. 프로그램은 매주 금, 토, 일에 진행되며, 12월 18일까지 운영된다.</p> <p>한편, 강화군은 방문객들이 쉬어갈 수 있도록 평화전망대에 인근에 남북 1.8센터를 지난 해 준공했다. 교동도에는 화개정원 및 전망대 조성사업으로 380억 원을 투자해 새로운 볼거리를 마련하고 있다. 또한, 산이포 민속마을 조성사업과 강후초 문화재생 및 별자리 관측소 건립사업도 속도를 내고 있다.</p> <p>유천호 군수는 "DMZ 관광자원을 활용한 관광 상품 개발, 운영을 통해 북부지역의 관광을 활성화하겠다"며 "북부지역에 부족한 관광·문화 기반시설을 체계적으로 조성해 남부 지역과 균형을 이루며, 미래로 도약할 수 있도록 하겠다"고 말했다.</p> <p>00=000 기자 0000@</p>
신문기사말뭉치	<p>강화군이 16일부터 평화와 통일로 가는 'DMZ 평화의 길 테마노선' 참가자 신청을 받는다.</p> <p>'디엠지(DMZ) 평화의 길' 테마노선은 강화 전쟁박물관을 시작으로 연미정과 고려천도공원을 거쳐 평화전망대까지 연결된다. 이어 의두분초와 의두돈대를 찍고 교동대교 건너 대룡시장까지 61.1km를 한강하구 너머 북녘과 마주하며 강화도 북부지역을 걷는 노선이다. 북녘 땅을 내려다볼 수 있는 의두분초와 의두돈대는 민간에 개방되지 않는 군사시설과 야생동물 서식지로 참가자들의 안전을 위해 군부대의 협조를 받아 이동하게 된다. 모든 코스를 둘러보는데 6시간 정도 소요된다.</p> <p>참가신청은 '디엠지(DMZ) 평화의 길' 누리집을 통해 사전 예약하면 된다. 프로그램은 매주 금, 토, 일에 진행되며, 12월 18일까지 운영된다.</p> <p>한편, 강화군은 방문객들이 쉬어갈 수 있도록 평화전망대에 인근에 남북 1.8센터를 지난 해 준공했다. 교동도에는 화개정원 및 전망대 조성사업으로 380억 원을 투자해 새로운 볼거리를 마련하고 있다. 또한, 산이포 민속마을 조성사업과 강후초 문화재생 및 별자리 관측소 건립사업도 속도를 내고 있다.</p> <p>유천호 군수는 "DMZ 관광자원을 활용한 관광 상품 개발, 운영을 통해 북부지역의 관광을 활성화하겠다"며 "북부지역에 부족한 관광·문화 기반시설을 체계적으로 조성해 남부 지역과 균형을 이루며, 미래로 도약할 수 있도록 하겠다"고 말했다.</p>
인용부호	<p>&lt;p&gt;강화군이 16일부터 평화와 통일로 가는 'DMZ 평화의 길 테마노선' 참가자 신청을 받는다.&lt;/p&gt;</p>

수정 말뭉치	<p>&lt;p&gt;‘디엠지(DMZ) 평화의 길’ 테마노선은 강화 전쟁박물관을 시작으로 연미정과 고려천도공원을 거쳐 평화전망대까지 연결된다. 이어 의두분초와 의두돈대를 찍고 교동대교 건너 대룡시장까지 61.1km를 한강하구 너머 북녘과 마주하며 강화도 북부지역을 걷는 노선이다.&lt;/p&gt;</p> <p>&lt;p&gt;북녘 땅을 내려다볼 수 있는 의두분초와 의두돈대는 민간에 개방되지 않는 군사시설과 야생동물 서식지로 참가자들의 안전을 위해 군부대의 협조를 받아 이동하게 된다. 모든 코스를 둘러보는데 6시간 정도 소요된다.&lt;/p&gt;</p> <p>&lt;p&gt;참가신청은 ‘디엠지(DMZ) 평화의 길’ 누리집을 통해 사전 예약하면 된다. 프로그램은 매주 금, 토, 일에 진행되며, 12월 18일까지 운영된다.&lt;/p&gt;</p> <p>&lt;p&gt;한편, 강화군은 방문객들이 쉬어갈 수 있도록 평화전망대에 인근에 남북 1.8센터를 지난해 준공했다. 교동도에는 화개정원 및 전망대 조성사업으로 380억 원을 투자해 새로운 볼거리를 마련하고 있다. 또한, 산이포 민속마을 조성사업과 강후초 문화재생 및 별자리 관측소 건립사업도 속도를 내고 있다.&lt;/p&gt;</p> <p>&lt;p&gt;유천호 군수는 “DMZ 관광자원을 활용한 관광 상품 개발, 운영을 통해 북부지역의 관광을 활성화하겠다”며 “북부지역에 부족한 관광·문화 기반시설을 체계적으로 조성해 남부지역과 균형을 이루며, 미래로 도약할 수 있도록 하겠다”고 말했다.&lt;/p&gt;</p>
문장 말뭉치	<p>&lt;p&gt;&lt;s&gt;강화군이 16일부터 평화와 통일로 가는 &lt;l&gt;‘DMZ 평화의 길 테마노선’&lt;/l&gt; 참가자 신청을 받는다.&lt;/s&gt;&lt;/p&gt;</p> <p>&lt;p&gt;&lt;s&gt;‘디엠지(DMZ) 평화의 길’ 테마노선은 강화 전쟁박물관을 시작으로 연미정과 고려천도공원을 거쳐 평화전망대까지 연결된다.&lt;/s&gt; &lt;s&gt;이어 의두분초와 의두돈대를 찍고 교동대교 건너 대룡시장까지 61.1km를 한강하구 너머 북녘과 마주하며 강화도 북부지역을 걷는 노선이다.&lt;/s&gt;&lt;/p&gt;</p> <p>&lt;p&gt;&lt;s&gt;북녘 땅을 내려다볼 수 있는 의두분초와 의두돈대는 민간에 개방되지 않는 군사시설과 야생동물 서식지로 참가자들의 안전을 위해 군부대의 협조를 받아 이동하게 된다.&lt;/s&gt; &lt;s&gt;모든 코스를 둘러보는데 6시간 정도 소요된다.&lt;/s&gt;&lt;/p&gt;</p> <p>&lt;p&gt;&lt;s&gt;참가신청은 ‘디엠지(DMZ) 평화의 길’ 누리집을 통해 사전 예약하면 된다.&lt;/s&gt; &lt;s&gt;프로그램은 매주 금, 토, 일에 진행되며, 12월 18일까지 운영된다.&lt;/s&gt;&lt;/p&gt;</p> <p>&lt;p&gt;&lt;s&gt;한편, 강화군은 방문객들이 쉬어갈 수 있도록 평화전망대에 인근에 남북 1.8센터를 지난해 준공했다.&lt;/s&gt; &lt;s&gt;교동도에는 화개정원 및 전망대 조성사업으로 380억 원을 투자해 새로운 볼거리를 마련하고 있다.&lt;/s&gt; &lt;s&gt;또한, 산이포 민속마을 조성사업과 강후초 문화재생 및 별자리 관측소 건립사업도 속도를 내고 있다.&lt;/s&gt;&lt;/p&gt;</p> <p>&lt;p&gt;&lt;s&gt;유천호 군수는 “DMZ 관광자원을 활용한 관광 상품 개발, 운영을 통해 북부지역의 관광을 활성화하겠다”며 “북부지역에 부족한 관광·문화 기반시설을 체계적으로 조성해 남부지역과 균형을 이루며, 미래로 도약할 수 있도록 하겠다”고 말했다.&lt;/s&gt;&lt;/p&gt;</p>
<p>- 신문 기사 말뭉치 : 원시데이터에서 캡션 정보 등 불필요한 요소를 제거</p> <p>- 인용 부호 수정 말뭉치 : 신문 기사 말뭉치에서 인용 부호를 수정</p> <p>- 문장 말뭉치 : 인용 부호 수정 말뭉치에서 문장 단위로 &lt;s&gt;태그를 부착</p>	

## <부록 5>

신문기사 말뭉치 오류 검색 목록

## 신문기사 말뭉치 오류 검색 목록

아래 정규식은 제출된 json 파일을 대상으로 했을 때의 오류 검색 방법이다.

### □ 정제 말뭉치

정규식	[가-힣]+[^\.]"\n(_____)\\n(_____)\\n(_____)\\n(_____)\\n(_____)id")						
설명	<ul style="list-style-type: none"> <li>기사의 마지막 문장이 마침표(.)로 끝나지 않은 기사를 검색합니다.</li> <li>- 처리되지 않은 기자 정보, 사진 캡션 또는 마지막에 끊긴 기사, 마침표 이외의 기호로 끝난 기사들이 검색됩니다.</li> </ul> <table border="1"> <tr> <td>[가-힣]</td><td>‘가’ ~ ‘힣’ 까지 한글 검색</td></tr> <tr> <td>\\n</td><td>엔터키에 의한 줄바꿈</td></tr> <tr> <td>[^\.]</td><td>마침표(.)가 아님(대괄호 안에서 ^는 not을 의미)</td></tr> </table>	[가-힣]	‘가’ ~ ‘힣’ 까지 한글 검색	\\n	엔터키에 의한 줄바꿈	[^\.]	마침표(.)가 아님(대괄호 안에서 ^는 not을 의미)
[가-힣]	‘가’ ~ ‘힣’ 까지 한글 검색						
\\n	엔터키에 의한 줄바꿈						
[^\.]	마침표(.)가 아님(대괄호 안에서 ^는 not을 의미)						
예시	이를 “악순환”이라 분석했다. 온라인뉴스 <u>부</u> <u>_____}</u> <u>_____]</u> <u>_____},</u> <u>_____ {</u> <u>_____ "id"</u>						

정규식	\\([2-9] [1-4]{1}[0-9]{1})",\\n_____"(form").*?[가-힣][^\.]				
설명	<ul style="list-style-type: none"> <li>기사 내에서 마침표(.) 없이 줄바꿈된 문장을 검색합니다.</li> <li>- 불필요하게 줄바꿈된 문장, 마침표를 찍지 않은 문장, 소제목이 검색됩니다.</li> </ul> <table border="1"> <tr> <td>([2-9] [1-4]{1}[0-9]{1})</td><td>숫자 2부터 49까지 검색(1번 문장은 기사 제목이므로 제외)</td></tr> <tr> <td>{1}</td><td>바로 앞의 대괄호 [1-4], [0-9]가 지정하는 범위에서 하나를 선택하여 검색 [1-4]{1}[0-9]{1}는 1과 4 사이의 숫자 하나를, 0과 9 사이의 숫자 하나를 검색하게 되며, 각각 십의 자리, 일의 자리가 되어 10~49까지를 검색</td></tr> </table>	([2-9] [1-4]{1}[0-9]{1})	숫자 2부터 49까지 검색(1번 문장은 기사 제목이므로 제외)	{1}	바로 앞의 대괄호 [1-4], [0-9]가 지정하는 범위에서 하나를 선택하여 검색 [1-4]{1}[0-9]{1}는 1과 4 사이의 숫자 하나를, 0과 9 사이의 숫자 하나를 검색하게 되며, 각각 십의 자리, 일의 자리가 되어 10~49까지를 검색
([2-9] [1-4]{1}[0-9]{1})	숫자 2부터 49까지 검색(1번 문장은 기사 제목이므로 제외)				
{1}	바로 앞의 대괄호 [1-4], [0-9]가 지정하는 범위에서 하나를 선택하여 검색 [1-4]{1}[0-9]{1}는 1과 4 사이의 숫자 하나를, 0과 9 사이의 숫자 하나를 검색하게 되며, 각각 십의 자리, 일의 자리가 되어 10~49까지를 검색				
예시	001.26216. <u>9</u> , <u>_____ "form": "■군, 온실가스 배출 관리 ‘사각지대’"</u>				



정규식	ㅇ ㅏ   ㄱ ㄴ ㄷ ㄹ ㅁ ㅂ ㅅ ㅇ ㅈ ㅊ ㅋ ㅌ ㅍ ㅎ ㄲ ㅃ ㅆ ㅉ ㅑ ㅕ ㅗ ㅛ ㅜ ㅠ ㅡ ㅣ ㅞ ㅟ ㅠ ㅡ 과 내 더 게 나 거 니
설명	• 텍스트가 깨진 문장, 문장에 포함된 오타 등을 검색합니다. - 기자 정보에도 오타가 있을 수 있으므로 체크해야 합니다. - ‘제보자 ㄱ씨’, ‘ㅁ자형 건물’ 처럼 문맥상 필요한 자음과 모음 사용은 수정하지 않습니다.
예시	<div style="margin-bottom: 10px;"><span style="color: red; font-weight: bold;">ㅇ</span> <span style="color: red; font-weight: bold;">ㅏ</span>  <span style="font-size: 1.2em;">ㅓ</span>ㅓ</div> <p>[헤럴드경제=이원율 기<span style="color: red; font-weight: bold;">ㅏ</span>]</p> <p>공연을 오는 내다다<span style="color: red; font-weight: bold;">ㅏ</span><span style="color: red; font-weight: bold;">르</span> 7일부터 10일까지</p>

정규식	이 기사는      본 기획물은      위 텍스트는      대담      정리=
설명	<ul style="list-style-type: none"> <li>• 인공지능이 작성한 기사, 협찬을 받아 작성된 기사를 검색합니다.</li> <li>- 인공지능 작성, 협찬/지원 여부를 명백히 밝히고 있으므로 기사를 삭제 처리합니다.</li> <li>• 대담을 나눈 것을 옮긴 기사를 검색하여 삭제 처리합니다.</li> </ul>
예시	<p>※ <b>이 기사는</b> 한국언론진흥재단-세명대 기획탐사 디플로마 교육 과정의 일환으로 작성됐습니다.</p> <p>[박세환 헤럴드경제 전국부장 <b>대담, 정리</b>=이영기 기자]</p>

정규식	<div> <div>"사진</div> <div>이미지=</div> <div>자료 :</div> <div>포스터</div> <div>제공"</div> </div> <div> <div>&gt;사진</div> <div>출처</div> <div>자료=,</div> <div>조감도</div> <div>캡처"</div> </div> <div> <div>자료:</div> <div>그림</div> <div>투시도</div> <div>장면"</div> </div>
설명	<ul style="list-style-type: none"> <li>• 사진 캡션을 검색합니다.</li> <li>- json 형식에서는 " 기호 이후 모든 문장이 구분되어 있어 위와 같이 검색이 가능하나, 문장이 구분되어 있지 않은 문서라면 ‘사진’이라는 단어로 시작하는 문장을 검색하는 ^(사진.*?) 또는 제공, 캡처, 장면 이후 줄바꿈을 의미하는 \n 을 붙여 제공\n 으로 검색합니다.</li> </ul>
예시	<p>‘동양의 올리브유’라고 불린다.</p> <p><b>사진</b> 한국관광공사</p> <p>큰 역할을 해오고 있다.</p> <p>[<b>사진제공</b>=모코이엔티, 티모넷]</p> <p>획득하겠다”라며 각오를 전했다.</p> <p>스포츠서울 <b>제공</b></p>

정규식	" < > " _ <
설명	<ul style="list-style-type: none"> <li>· 사진 캡션, 인사/승진 정보 등을 검색합니다.</li> </ul> <div>   정규식에서 or를 의미 (키보드에서 엔터 위 원화기호(W)+Shift 키를 눌러 입력) </div>
예시	<p>"&lt;승진&gt; ◇전무이사 ▷동원산업 순례를 다녀온 사람들은 오래 머문다. &lt;계속&gt;" ▶아시아나항공 &lt;전무 승진&gt; ▷원유석 ▷ 두성국</p>

정규식	<div> 유니코드 2008 ‘ ’      유니코드 2028 ‘ ’      유니코드 3164 ‘ ’  유니코드 202F ‘ ’      유니코드 3000 ‘ ’ </div>
설명	<ul style="list-style-type: none"> <li>· 여러가지 공백을 일반적인 공백(space bar, 유니코드 0020)으로 수정합니다.</li> <li>- 따옴표 사이의 공백을 복사+붙여넣기 하여 검색합니다.</li> </ul>
예시	<p>물결이 넘실댄다._대한축구협회는 물결이 넘실댄다. 대한축구협회는</p>

정규식	<div> 유니코드 FF0E ‘ . ’      유니코드 FF0C ‘ , ’ </div>
설명	<ul style="list-style-type: none"> <li>· 코드가 다른 마침표와 쉼표를 검색하여 일반적인 마침표와 쉼표로 수정합니다.</li> <li>- 따옴표 사이의 기호를 복사+붙여넣기 하여 검색합니다.</li> </ul>
예시	<p>2020년 34%로 감소했다._ 2020년 34%로 감소했다. 국민연금기금,_사학연금기금,_ 국민연금기금, 사학연금기금,</p>

정규식	^.\n ^[.] ".
설명	<ul style="list-style-type: none"> <li>· 마침표를 잘못 찍은 경우를 검색합니다.</li> </ul>
예시	<p>"_1882년 창단된 베를린 필은</p>



정규식	다," \\w,"
설명	<ul style="list-style-type: none"> <li>마침표가 와야 할 자리에 쉼표를 잘못 사용한 것을 검색합니다.</li> <li>- 잘못 나뉜 문장이라면 붙여주고, 마침표가 와야 할 곳에는 부호를 마침표로 바꿔 줍니다.</li> </ul>
예시	경영책임자에게 징역형도 부과될 수 있다,"

정규식	([가-힣]{2,4})_1		
설명	<ul style="list-style-type: none"> <li>반복되는 어절을 검색합니다.</li> <li>- 문맥에 유의하며 반복되는 어절을 삭제합니다.</li> <li>- ‘3년이 지난 지난해 8월’, ‘비정규직 정규직화’ 같은 경우를 삭제하지 않도록 유의합니다.</li> </ul>		
	<table border="1"> <tr> <td>\\1</td><td>앞에서 검색한 소괄호 ( )의 내용이 동일하게 반복됨을 의미</td></tr> </table>	\\1	앞에서 검색한 소괄호 ( )의 내용이 동일하게 반복됨을 의미
\\1	앞에서 검색한 소괄호 ( )의 내용이 동일하게 반복됨을 의미		
예시	여론조사 <u>응답률은 응답률은</u> 노원구 6.2%, 도봉구·광진구 6.4%		

정규식	다\ 다\ 따\ 쏟\ 날\.
설명	<ul style="list-style-type: none"> <li>이후 공정에 지장을 줄 수 있으므로 문장 끝에서 자주 발생하는 오타를 검색하여 수정합니다.</li> </ul>
예시	김해시에 BSS 30기를 설치하 <u>날</u> . 압력으로 작용하고 있 <u>따</u> .

정규식	[?][가-힣]      [?][a-z]      [?][0-9]      ??      \\\
설명	<ul style="list-style-type: none"> <li>특수기호, 한자 등 글자가 깨진 부분을 검색하여 수정합니다.</li> </ul>
예시	<u>초?중?고</u> 교생을 대상으로 한 이번 초·중·고교생을 대상으로 한 이번

정규식	<div>■ 진행                    영상취재                    [영상]                    PD                    [앵커]</div> <div>■ 진 행                    영상편집                    촬영:  촬영                    프로듀서                    앵커</div> <div>:</div>
설명	<div>· 영상/라디오 뉴스를 그대로 옮긴 기사를 검색합니다. 해당 기사는 사용하지 않습니다.</div>
예시	<div>■ <b>진행</b> : 김현정 앵커 (노컷뉴스 CBS라디오&lt;김현정의 뉴스쇼&gt;)</div>

□ 문장부호 수정 말뭉치

정규식	$\wedge[\wedge\text{"}]*\text{"}$ $\wedge[\wedge\text{'}]*'$
설명	<ul style="list-style-type: none"> <li>여는 따옴표(“, ’)없이 닫는 따옴표만 있는 문장을 검색합니다. (~”), (~’)</li> <li>- 빠진 따옴표를 넣어주거나 잘못 쓰인 따옴표를 바꿔 줍니다.</li> </ul>
예시	<p>이날 논평을 내고 <u>“</u>이명박 부패 세력 40회 금산인삼축제<u>”</u>도 같은 날</p>

정규식	‘[ ^ ]’*?\$ “[ ^ ”]*?\$
설명	· 여는 따옴표만으로 끝난 문장을 검색합니다. (‘~), (“~)
예시	‘2022 세계유산축전 - 제주 화산섬과 용암동굴’을 진행한다.

정규식	“_” ‘_’
설명	<ul style="list-style-type: none"> <li>· 따옴표 앞뒤로 띄어쓰기가 잘못된 문장을 검색합니다.</li> <li>- 띄어쓰기가 잘못된 경우 또는 따옴표의 짝이 맞더라도 뒤집혀 있는 경우가 검색됩니다.</li> </ul>
예시	‘과학 기술과 지식의 의식’ 등 그의 연구를

[illegible]



<기획·연구>

국립국어원 언어정보과장 강미영

국립국어원 학예연구관 김문오

국립국어원 연구원 이선영

<사업 참여자>

사업 책임자 윤종웅(주윌즈정보개발 소장)

사업 참여자 남가윤(주윌즈정보개발 연구원)

서경찬(주윌즈정보개발 책임연구원)

안소연(주윌즈정보개발 연구원)

윤종성(주윌즈정보개발 팀장)

이승철(주윌즈정보개발 수석연구원)

임순영(주윌즈정보개발 연구원)

임승락(주윌즈정보개발 연구원)

최원수(주윌즈정보개발 연구원)

---

발행인: 국립국어원장

발행처: 국립국어원

서울시 강서구 금남화로 154

전화 02-2669-9636, 전송 02-2669-9757

인쇄일: 2023년 10월 17일

발행일: 2023년 10월 17일

인 쇄: 다큐팩토리

---

※ 이 보고서는 국립국어원의 용역비로 수행한 ‘2023년 신문 기사 원문  
자료 수집 및 정제’ 사업의 결과물을 발간한 것입니다.