

국립국어원 2024-01-19

발간등록번호
--------

11-1371028-001016-01
----------------------

## 2024년 신문 기사 원문 자료 수집 및 정제

사업책임자

윤 중 응



국립국어원



## 제 출 문

국립국어원장 귀하

국립국어원과 체결한 연구용역 계약에 따라 ‘2024년 신문 기사 원문 자료 수집 및 정제’에 관한 연구 보고서를 작성하여 제출합니다.

■ 사업 기간: 2024년 4월 24일 ~ 2024년 10월 31일

2024년 10월 31일

사업책임자: 윤종웅(주)윤즈정보개발)

연구 기관: (주)윤즈정보개발

사업 책임자: 윤종웅

사업 참여자: 남가운, 서경찬

안소연, 윤종성, 이승철

임승락, 최원수



## 〈국문 요약〉

# 2024년 신문 기사 원문 자료 수집 및 정제

국립국어원은 현대 한국어의 실제 사용 양상을 반영하는 신문 기사 말뭉치를 구축하여, 개인, 기업, 학계가 폭넓게 활용할 수 있도록 제공하고 있다. ‘신문 기사 원문 자료 수집 및 정제’ 사업은 대규모 한국어 학습 자료 구축의 일환으로, 올해로 6년째 이어지고 있으며, 특히 인공 지능 산업계나 연구 기관 등이 정제된 고품질 데이터를 공공재로 활용할 수 있도록 지원하고 있다.

이 사업의 수행 범위는 신문 기사 원문 자료 수집(2023년 작성 기사, 월별 1,600만 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.

이 사업은 다양한 분야에서 활용 가능한 신뢰성 높은 데이터의 구축을 목표로, 수집된 신문 기사 원문에서 발생할 수 있는 오류를 체계적으로 정제하였다. 우선, 기사 내 그림 캡션, 비문, 데이터의 오류 등 불필요하거나 비정형적인 요소를 제거한 ‘신문 기사 말뭉치’, 언어 데이터의 정확성을 높이기 위해 인용 부호의 오류를 수정한 ‘인용 부호 수정 말뭉치’를 생성했다. 대부분의 자연어 처리 모델이 문장을 기본 단위로 한다는 점을 고려하여, 단락 단위의 데이터를 문장 단위로 나눈 ‘문장 분할 말뭉치’를 구축하여 형태소 분석 및 기계 번역의 정확도를 향상시켰으며, 언어 모델 학습 시 세분화된 데이터 사용으로 더 나은 성능을 보장할 수 있도록 구축하였다.

이 사업의 핵심은 참여 매체와의 계약을 통해 기사 저작권을 확보하는 일이다. 이 사업에서는 통신사, 종합 일간지, 지방 일간지, 경제지 등 다양한 매체와 협의 후 계약을 체결하여 필요한 저작권을 확보하였다. 저작권과 관련한 매체 담당자들의 문의 사항은 전문 법무법인의 자문을 통해 상세히 설명하여 매체 담당자들의 궁금증을 해소할 수 있도록 하였다.

원문 자료 수집 대상은 국립국어원과 협의하여 총 24개 매체를 선택하였고 신문 기사 수집 및 정제 사업에 사용되는 말뭉치 이용 허락 최소 기간은 2035년 12월 31일까지로 하였고, 저작자인 언론사가 이용 허락 중지 의사를 밝히지 않으면 이용 허락이 1년 단위로 자동 갱신되도록 하였다. 24개 매체로부터 확보한 원시 자료는 총 418,010,722개의 어절로 이루어진 2,069,751건의 기사이다.

이 사업에서는 엑스엠엘(XML) 형식으로 기사를 제공하도록 24개 매체에 요청하였고, 각 매체는 이를 준수하여 데이터를 제출하였다. 그러나 매체마다 데이터 특성과 구조에 차이가 있어, 학습에 사용하기 부적절한 요소들이 포함된 경우가 많았다. 예를 들어, 중복 기사와 광고성 기사, 유사 기사, 데이터 일부가 소실된 기사 등이 있었으며, 캡션이 명확히 분리되지 않아 의미 전달에 문제가 있는 기사나, 기자 정보가 본문에만 있고 별도 필드에는 없는 기사도 발견되었다.

또한, 동일 매체 내에서도 인용 부호 사용이 일관되지 않아 열고 닫는 부호가 맞지 않는 경우가 많았고, 많은 기사에서 오타도 확인되었다. 이러한 비정제 데이터가 그대로 학습에 활용될 경우, 말뭉치의 활용 효율이 크게 떨어질 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. 문장 분할이 중요한 이유는 문장이 인공 지능 학습의 기본 단위이기 때문이다.

대량의 데이터를 수집하는 것은 인공 지능 모델의 성능 향상에 긍정적인 영향을 미칠 것이라는 일반적인 믿음이 있지만, 실제로는 데이터의 양보다 품질이 더욱 중요한 요소로 작용한다. 단순히 데이터를 기하급수적으로 늘린다고 해서 자동적으로 성능 향상으로 이어지지는 않는다. 학습 데이터의 양이 많을수록 비용이 증가하므로, 무한정으로 데이터를 확대할 수는 없다. 결국 인공 지능 모델의 성공은 고품질 데이터를 얼마나 효과적으로 활용하는지에 달려 있다.

이 사업에서는 데이터의 양과 함께 품질을 고려하여 수집 및 정제를 진행하였다. 먼저, 다양한 매체의 데이터를 수집하였으며, 수집된 데이터는 정제 과정을 거쳐 불필요한 정보를 제거하고, 정확한 데이터로 가공하였다. 또한, 데이터의 중복을 최소화하고, 데이터의 분포를 고려하여 균형 잡힌 말뭉치를 구축하였다.

최종적으로 265,724,419개의 어절로 이루어진 신문 기사 1,119,753건의 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 인공 지능 학습에 유용한 자료가 될 것으로 기대한다.

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

<Abstract>

## Collection and Refinement of Data from Original Newspaper Articles in 2024

The National Institute of Korean Language(NIKL) has developed a newspaper corpus that reflects the actual use of modern Korean, making it widely available to individuals, businesses, and academia. The “Newspaper Article Collection and Refinement” project, now in its sixth year, is part of a larger initiative to build a massive repository of Korean learning data. This project is aimed at providing refined, high-quality data as a public resource for the AI industry and research institutions.

The project scope includes: 1) collecting original newspaper articles from 2023, amounting to over 16 million Eojeols(word-spacing units) per month; 2) resolving copyright issues through usage agreements with copyright holders; 3) removing duplicate articles and refining content; 4) building three types of corpora(raw corpus, corrected quotation mark corpus, and sentence corpus); and 5) creating metadata and article lists.

The project aims to create a reliable dataset that can be utilized in various fields by systematically refining errors found in collected original newspaper articles. First, a “Newspaper Article Corpus” was created by removing unnecessary or irregular elements such as image captions, ungrammatical sentences, and data errors. Then, to improve linguistic accuracy, a “Corrected Quotation Mark Corpus” was developed. Since most natural language processing(NLP) models operate at the sentence level, a “Segmenting Sentence Corpus” was built by segmenting paragraph-level data into sentences. This refinement improves the accuracy of morphological analysis and machine translation, ensuring better model performance by leveraging segmented data during language model training.

A key component of this project is securing copyright permissions from

participating media through agreements. This year, agreements were concluded with a variety of media outlets, including news agencies, comprehensive dailies, local dailies, and economic newspapers. Legal inquiries from media representatives regarding copyright were addressed through consultations with a professional law firm, providing detailed explanations to alleviate concerns.

In collaboration with NIKL, 24 media outlets were selected as data sources. The minimum usage period for the collected and refined corpus is set until December 31, 2035, with an automatic annual renewal unless media organizations explicitly withdraw permissions. The raw data from these 24 media outlets included 2,069,751 articles, comprising a total of 418,010,722 Eojeols.

The project required media outlets to submit articles in XML format, and all 24 complied. However, the diversity in data structure and characteristics among the media often led to the inclusion of elements unsuitable for training. These included duplicate, advertising, and similar articles, articles with partial data loss, as well as those where captions were not clearly separated, affecting readability. Some articles contained reporter information embedded in the main text rather than in designated fields.

Moreover, even within the same media outlet, the usage of quotation marks was inconsistent, resulting in mismatched opening and closing marks. Numerous typos were also detected across articles. Training models with such unrefined data would severely hinder the efficient utilization of the corpus.

Most AI learning processes rely on sentences as their fundamental unit. Training a corpus at the paragraph level, which consists of multiple sentences, is inherently less efficient. Hence, a sentence corpus was developed by segmenting paragraphs into sentences. Sentence segmentation is crucial since sentences serve as the foundational learning units for AI models.

While there is a common belief that collecting large amounts of data positively impacts AI model performance, quality is often a more decisive factor than quantity. Simply increasing data volume exponentially does not guarantee performance improvement. As data volume increases, so do associated costs, limiting the scope for unrestrained expansion. Ultimately, the success of an AI model hinges on how effectively high-quality data is utilized.



In this project, both data quantity and quality were prioritized during collection and refinement. Diverse media sources were tapped, and collected data underwent rigorous refinement to eliminate extraneous information and transform it into accurate datasets. Efforts were made to minimize data duplication and maintain balanced corpora by considering data distribution.

Ultimately, a corpus comprising 1,119,753 newspaper articles and 265,724,419 Eojeols was constructed. The corpus, reflecting contemporary language use, is expected to serve as valuable training data for AI learning in various fields, including the development of AI technologies and academic research in preparation for the Fourth Industrial Revolution.

**Key words:** Newspaper Corpus, Contemporary Korean, Artificial Intelligence (AI), AI Training Data, Refined Data, Data Copyright, Newspaper Articles, Segmenting Sentence Corpus, Correction of Quotation Marks



# 차 례

## 제1장 서론

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

## 제2장 사업 수행 내용

1. 매체 선정 및 계약 .....	6
2. 데이터 수집 .....	7
3. 데이터 1차 정제 .....	16
4. 데이터 2차 정제 .....	20
5. 메타데이터 작성 .....	34
6. 인용 부호 수정 말뭉치 .....	35
7. 문장 말뭉치 구축 .....	42

## 제3장 사업 수행 결과

1. 신문 기사 정제 결과 .....	46
2. 매체별 납품 파일명 .....	50

<부록 1> 국가 언어 자원(말뭉치) 구축 및 활용 저작권 이용 허락 계약서(양식안) .....	51
<부록 2> 데이터 정제 작업 지침 .....	58
<부록 3> 말뭉치 종류별 구축 예시 .....	66
<부록 4> 신문 기사 말뭉치 오류 검색 목록 .....	69

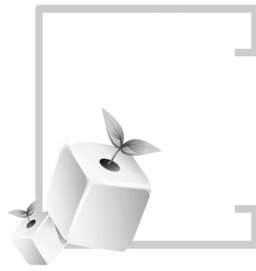
## 표 차례

<표 1> 사업 공정표 .....	4
<표 2> 선정된 매체 구분 .....	6
<표 3> 최초 수집 기사와 어절 수 .....	7
<표 4> 원시 데이터 특징 예시 .....	8
<표 5> 기사의 메타데이터의 누락(기자 정보가 본문에만 존재하는 경우) .....	10
<표 6> 원시 데이터 특징 예시(데이터 소실 1) .....	10
<표 7> 원시 데이터 특징 예시(데이터 소실 2) .....	11
<표 8> 데이터 특징 .....	15
<표 9> 유사도 비교를 통해 사용하지 않는 기사의 예 .....	16
<표 10> 저작권 이용 문제로 인해 사용하지 않는 기사의 특징 .....	18
<표 11> 불필요한 요소 제거 내용 .....	23
<표 12> 원시 데이터와 정제된 데이터 비교 1 .....	27
<표 13> 원시 데이터와 정제된 데이터 비교 2 .....	28
<표 14> 2024년 신문 기사 주제별 통계 .....	34
<표 15> 인용 부호 치환 표 .....	36
<표 16> 인용 부호 수정 데이터 정제 전후 .....	37
<표 17> ‘한·중·일 호환용 한자 영역’ 한자 치환 표 .....	39
<표 18> 치환 코드 목록 .....	40
<표 19> 오타 글자 .....	42
<표 20> 문장 말뭉치 데이터 정제 .....	44
<표 21> 신문 기사 정제 총괄표 .....	47
<표 22> 월별 구축 어절 수 .....	49
<표 23> 주제별 기사 및 구축 어절 수 .....	49
<표 24> 말뭉치 파일명 .....	50

## 그림 차례

<그림 1> 구축 공정별 내용 .....	3
<그림 2> 상·하위 기사 수 매체, 성격별 매체 수 .....	7
<그림 3> 오류 유형 ①: 언론사 내부 설명이 있는 경우 .....	13
<그림 4> 오류 유형 ②: 문자 코드가 조합형으로 나타난 경우 .....	14
<그림 5> 오류 유형 ③: 서명 기호 안 텍스트가 사라진 경우 .....	15
<그림 6> 사용하지 않는 기사의 예 (기고문) .....	21
<그림 7> 오류 유형 ④: 원저작자의 의도가 훼손될 우려가 있는 경우 .....	22
<그림 8> 오류 유형 ⑤: 기사 내용이 끊긴 경우 .....	22
<그림 9> 기사 수정 예 ① .....	30
<그림 10> 기사 수정 예 ② .....	30
<그림 11> 기사 수정 예 ③ .....	31
<그림 12> 기사 수정 예 ④ .....	31
<그림 13> 작업 편집 화면 .....	32
<그림 14> 작업 프로그램 화면 .....	32
<그림 15> 데이터 작업 및 검수 공정 .....	33
<그림 16> 연도별 주제별 통계 .....	35
<그림 17> 문장 말뭉치 개념 .....	43
<그림 18> 매체별 평균 문장 분할 수 .....	44
<그림 19> 구축 공정별 내용 .....	46
<그림 20> 최근 5년 신문 기사 구축 어절 수, 올해 구축 어절 수 .....	48
<그림 21> 월별 구축 어절 수 .....	48





## 제 1 장

# 서 론



# 제1장 서론

## 1. 사업 목적

이 사업은 국립국어원의 ‘모두의 말뭉치’ 구축 사업의 일환으로, 인공 지능 언어 처리 기술 개발에 활용될 수 있는 방대한 신문 기사 원문 자료를 구축하고, 저작권 문제 없이 이용할 수 있도록 저작권자로부터 이용 허락을 확보하는 데 목적이 있다.

올해는 월 1,600만 어절 이상, 총 2억 어절 이상의 대규모 데이터 구축을 목표로 하며, 이는 작년 대비 약 0.8억 어절이 증가한 수치이다. 이처럼 방대한 데이터는 양적 확장뿐만 아니라 정제 과정을 통해 오류를 수정함으로써, 인공 지능 학습에 활용 가능한 고품질 자료로 구축된다. 단순 수집이 아닌 데이터 정제 작업을 통해 언어 모델의 성능을 높이고자 하였으며, 이는 학습 과정에서 오류가 최소화된 데이터의 중요성을 반영한 것이다.

결국, 저작권이 확보된 대규모 신문 기사 데이터를 기반으로 신뢰성 있는 인공 지능 언어 모델 개발을 지원하고, 정확하고 유용한 말뭉치 자료를 제공하는 것이 이 사업의 주된 목표이다.

## 2. 사업 수행 범위

이 사업의 범위는 다음과 같이 네 가지로 나눌 수 있다.

첫 번째는 신문 기사 원문 자료 수집이다. 2023년 1월부터 12월까지 작성된 기사를 대상으로, 매월 1,600만 어절 이상의 구축을 목표로 한다. 매체는 최소 20개 이상 선정하며, 인터넷 기반 매체는 전체의 30% 이상을 구성하도록 하였다.<sup>1)</sup>

두 번째는 기사 저작권 확보이다. 저작권 확보는 사업 결과물을 자유롭게 사용할 수 있도록 하기 위한 필수 절차로, 저작권자로부터 2차적 저작물 작성권을 포함한 원문 이용 허락을 얻어 구축한 국립국어원의 신문 기사 말뭉치를 허락 조건에 부합하는 한 누구나 이용할 수 있어야 한다. 이를 통해, 이 사업에서 수집된 자료가 저작권 문제 없이 다양한 연구나 사업에 사용될 수 있도록 한다.

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 관련 학계 및 인공 지능 학습 분야에서 활용할 수 있는 데이터를 생성해야 한다. 불필요한 내용을 제거한 신문 기사 말뭉치, 신문 기사 내 인용 부호를 수정한 인용 부호 수정 말뭉치와 단락 단위를 문장 단위로 분할한 문장 말뭉치, 이렇게 총 3종의 말뭉치를 구축한다.

네 번째는 구축된 데이터의 메타데이터 작성이다. 수집된 데이터에 기자 정보, 어절 수, 주제 분류, 기사 작성일 등의 메타데이터를 추가하여 사용자들이 손쉽게 활용할 수 있도록 한다.

---

1) 약 2억 어절의 말뭉치를 구축하기 위해서는 대량의 기사가 필요하며, 제안요청서상에 명시된 인터넷 기반 매체 비율을 10%로 제한할 경우 2개 정도의 매체로는 목표 기사 수를 충족하기 어려울 것으로 판단하였고, 각 매체와의 협의 및 계약이 지연되는 상황에서 빠르게 인터넷 기반 매체를 확보하여 목표치를 달성하기 위해, 사업 계획 변경 신청 및 승인을 거쳐, 인터넷 기반 매체 비율을 30% 이하로 수정하게 되었다.



## 가. 신문 기사 원문 자료 수집(2023년 작성 기사, 2억 어절 이상)

- 신문 기사 말뭉치 구축에 필요한 신문 기사 원문 자료를 수집함.
- 대상은 2023년 기사로 월별 1,600만 어절 이상을 구축함.
- 전국 종합지는 3개 이상 포함하고, 인터넷 기반 매체는 전체 매체 수의 30% 이내로 한정함.
- 신문 기사 원시 말뭉치는 매체별, 월별, 기사 주제별로 균형을 갖춰 2억 어절 이상을 구축함.
- 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 확보함.
- 수집한 기사 원문 자료 중 국립국어원에서 말뭉치 구축 대상으로 선정하는 매체의 기사 원문에 대해서 저작권자와 저작물 이용 허락 계약을 체결함.
- 계약은 법률 검토를 받은 후 주관 기관이 제공한 계약서 양식에 따라 국립국어원과 협의하여 체결함.
- 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 2035년 12월 31일까지로 함.

## 다. 기사 데이터의 정제

- 수집된 기사 중에 동일 매체 내에서 기사 내용이 동일한 기사는 제거함.
- 신문 기사 내에 삽입된 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, 기고문 등은 제거함.
- 중복 기사, 길이가 너무 짧은 기사 등 말뭉치로 구축하기에 부적절한 기사 원문은 대상에서 제외하고, 정제된 신문 기사 원문은 3종(신문 기사 말뭉치, 인용 부호 수정 말뭉치, 문장 말뭉치)으로 가공해야 함.
- 기사 주제(사회, 경제, 생활, ... 연예, 정보통신/과학) 간의 비율 차(최고-최저)가 가능한 한 25% 포인트 이하가 되도록 하되, 필요시 국립국어원과 협의하여 비율 조정을 할 수 있음.

## 라. 메타데이터 작성

- 국립국어원이 지정하는 아홉 가지 분류 체계로 신문 기사 주제를 재분류함.
- 신문사명, 기사 작성일, 주제 분류, 기사 제목, 어절 수 등 국립국어원이 지정하는 항목과 형식으로 기사별 메타 정보를 입력하고 수집 기사 목록을 작성함.

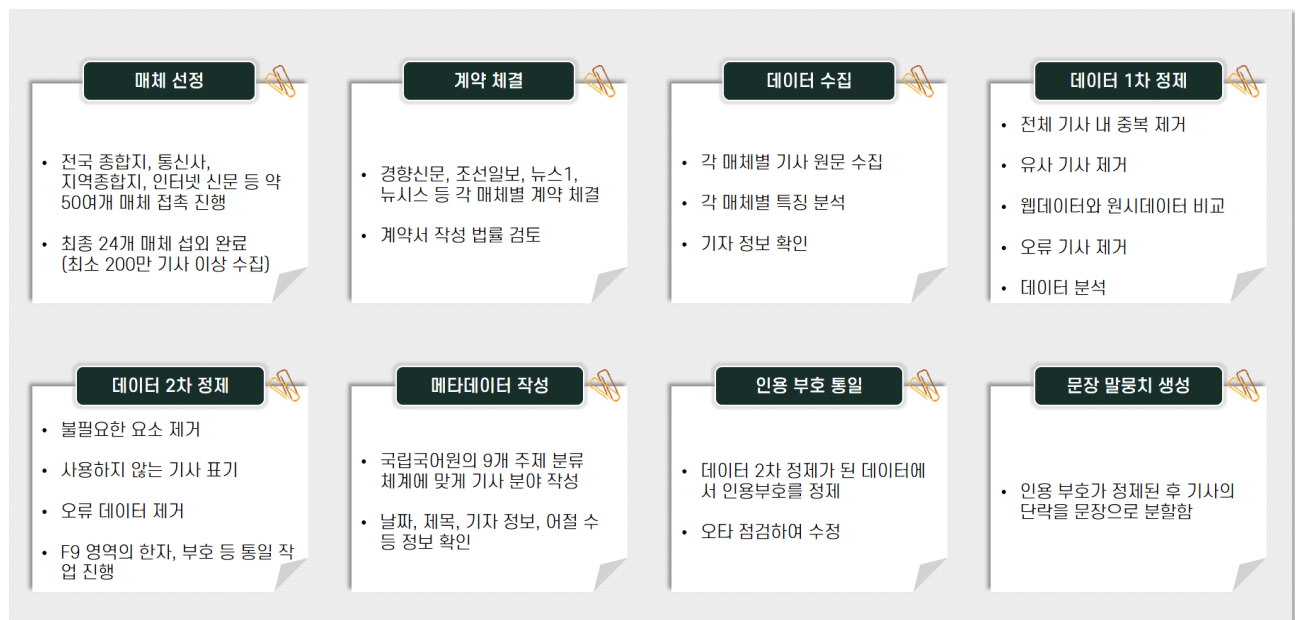
### 3. 사업 수행 절차

사업 수행 절차는 아래 <그림 1>과 같이 총 8단계로 구분된다. 우선, 사업 계약 후 가장 먼저 진행해야 할 매체 선정 및 계약 단계에서는 전체 매체 기사 수를 웹을 통해 파악하였다. 이를 바탕으로 약 50여 개 매체와 접촉하였고, 최종적으로 24개 매체와 계약을 체결하여 저작권 문제를 해결하였다.

그 다음으로, 각 매체별 원시 데이터를 분석하여 해당 매체의 특징을 파악하고, 활용 가능한 데이터로 구분하는 1차 정제를 진행하였다. 이후에는 1차 정제된 데이터를 바탕으로 불필요한 요소를 제거하고, 외부 기고가가 쓴 데이터 등 저작권에 문제가 있는 기사를 선별하는 2차 정제를 실시하여 신문 기사 말뭉치 데이터를 생성하였다.

메타데이터는 최종 선정된 기사를 바탕으로 작성하였으며, 데이터 2차 정제까지 완료된 기사는 1차로 신문 기사 말뭉치를 구축하였고, 인용 부호 통일 공정이 완료된 데이터는 2차로 인용 부호 수정 말뭉치를 구축하였고, 신문 기사의 각 단락을 문장 단위로 분할하는 공정이 완료된 데이터는 3차로 문장 단위로 분할된 말뭉치(이하 ‘문장 말뭉치’라 함)를 구축하였다. 이로써 하나의 기사에 총 3종의 데이터가 구축되었다.

데이터 납품은 착수 후 4개월 이내에 중간 납품으로 10개 매체와 최종 납품으로 14개 매체를 선정하여 2회에 걸쳐 진행하였다.



<그림 1> 구축 공정별 내용

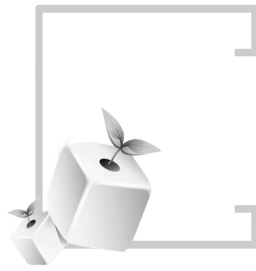
#### 4. 사업 추진 경과

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							
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							
	데이터 납품							

<표 1> 사업 공정표





## 제 2 장

# 사업 수행 내용



## 제2장 사업 수행 내용

### 1. 매체 선정 및 계약

국내 주요 신문 매체는 대부분 한국언론진흥재단에 신탁하여 저작권을 관리하고 있다. 그러나 올해는 한국언론진흥재단이 이 사업에 참여하지 않기로 결정함에 따라, 저작권 계약을 독자적으로 진행할 수 있는 매체 섭외가 필수적이었다. 이번 사업에서는 이전 사업 대비 약 두 배에 달하는 2억 어절 이상을 구축해야 하는 만큼, 충분한 기사 확보가 중요한 과제가 되었다.

다음과 같은 두 가지 조건을 충족하는 언론 매체를 선정하고 섭외를 진행하였다.

- 신문 매체 독자적으로 저작권 계약이 가능할 것
- 충분한 기사 수를 제공할 수 있는 매체일 것

이 조건을 충족하기 위해 국내 매체를 목록화하고 기사 수를 전체적으로 파악하였다. 이후 약 50개 매체와 접촉을 시도하였으며, 그 결과 24개 매체와의 계약을 체결하는 성과를 거두었다. 사업 초기에는 이 사업에 대한 이해가 부족한 상황이었으나 지속적인 소통과 협의를 통하여 통신사, 종합 일간지, 지역 일간지, 전문지 등 다양한 매체들로 구성하여, 다양한 기사와 관점을 포함할 수 있도록 하였다.

신문 기사 수집 및 정제 사업에 사용되는 말뭉치 이용 허락 최소 기간은 2035년 12월 31일까지로 하였고, 저작자인 언론사가 이용 허락 중지 의사를 밝히지 않으면 이용 허락이 1년 단위로 자동 갱신되도록 하였다.

구분	2024년 신문 기사 원문 자료 수집 및 정제의 대상(총 24개 매체)
경제지(2)	디지털타임스, 매일경제
인터넷(5)	뉴스투데이, 마이데일리, 아이뉴스24, 오마이뉴스, 오에스이엔
전문지(2)	스포츠경향, 스포츠투데이
종합지(3)	경향신문, 일간투데이, 조선일보
주간지(1)	굿모닝충청
지역종합지(8)	경남매일, 경인매일, 대경일보, 대전투데이, 울산제일일보 전남매일, 충남일보, 충청매일
통신사(3)	뉴스1, 뉴스웍스, 뉴시스

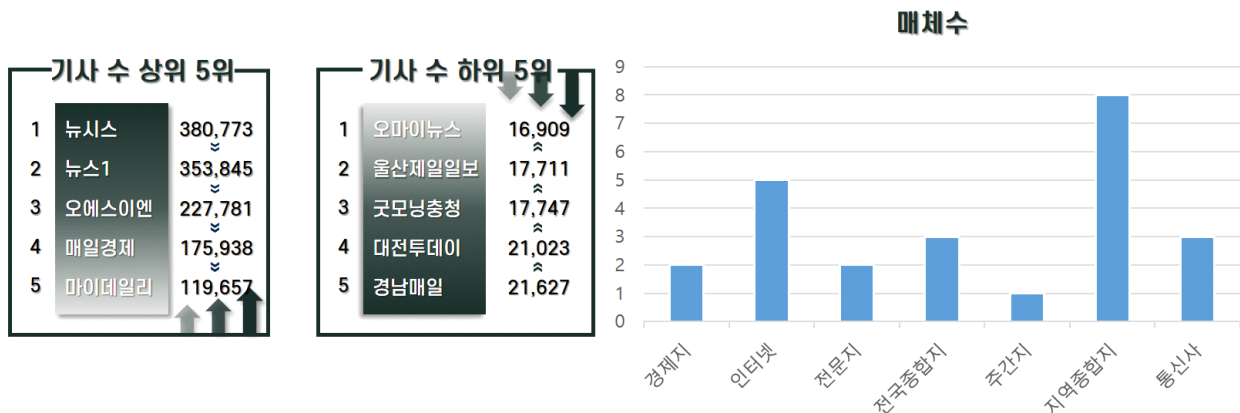
<표 2> 선정된 매체 구분

## 2. 데이터 수집

데이터 최종 목표인 2억 어절 이상을 구축하기 위해서는 기사 수 2백만 건 이상의 조건이 충족되어야 양질의 말뭉치를 무리 없이 구축할 수 있다.

24개 매체의 최초 수집된 기사와 어절 수(본문+제목)는 각각 2,069,751건, 418,010,722어절이다. 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> 상·하위 기사 수 매체, 성격별 매체 수

매체명	기사 수	어절 수	매체명	기사 수	어절 수
경남매일	21,627	3,309,024	매일경제	175,938	43,174,410
경인매일	42,368	7,207,076	스포츠경향	76,202	18,190,894
경향신문	74,116	22,020,899	스포츠투데이	57,464	8,223,681
굿모닝충청	17,747	4,301,532	아이뉴스24	80,869	16,346,113
뉴스1	353,845	69,606,852	오마이뉴스	16,909	5,332,432
뉴스웍스	27,415	6,916,277	오에스이엔	227,781	34,709,111
뉴스투데이	24,267	7,936,536	울산제일일보	17,711	2,816,252
뉴스스	380,773	81,220,617	일간투데이	49,148	9,322,305
대경일보	36,534	5,953,026	전남매일	24,795	4,485,644
대전투데이	21,023	3,872,669	조선일보	46,670	11,475,428
디지털타임스	70,197	16,532,859	충남일보	48,170	7,548,770
마이데일리	119,657	19,748,979	충청매일	58,525	7,759,336
총 합				2,069,751	418,010,722

<표 3> 최초 수집 기사와 어절 수

## 가. 원시 데이터 특징 분석

원시 데이터는 24개 매체에서 수집되어 다양한 형식을 내포하고 있기 때문에 일관된 형식을 유지하는 것이 중요하였다. 이를 위해 각 매체에 엑스앰엘(XML) 파일 형식을 요구하고, 제목, 날짜, 기사 정보, 카테고리 등의 메타 정보와 본문을 포함하도록 요청하였다. 대부분의 매체가 정해진 형식을 준수하여 데이터를 제공했으나, 매체마다 고유한 형식적 차이가 존재하였으며 이는 데이터 검증과 오류 수정을 통해 확인 및 조정이 필요했다.

```
<content><![CDATA[<b><b>[기사요약]<BR>차별화된 에코생태계 구축이 디지털 플랫폼 경쟁에 있어 핵심 전략<BR>화물맨, 전
국 '개인(개별)화물자동차운송사업협회'와 업무협약 체결<BR>AI 추천 화물운임정보, H-pass, 쾌적운송경로추천 등 에코생태계 니드
반영한 독자적인 지능체계 구축 중</b>
</b><br/><p style="text-align: justify;">'알파고'의 바둑대결로 AI가 주목받게 되었듯이 2021년 3월 쿠팡의 뉴욕증권거래소 입성
(86조원 시가총액 인정)은 일반 국민들의 물류에 대한 관심을 고조시켰다. 더욱이 의아했던 점은 당시 쿠팡의 적자 규모가 4조원에
달했다는 점이다. 한편 쿠팡 상장 1년 전 '우아한형제들'의 배민을 독일계 DH(딜리버리 히어로)가 4조7500억원에 인수하는 사건도 있
었다. 창고와 트럭으로 대변되던 3D업종 물류가 핫한 주목을 받게 된 다이내믹스(Dynamics, 역동성)는 과연 무엇이고, 그렇다면 미래
에도 물류는 계속 주목받는 산업으로 남게 될까? 역동적인 물류의 미래를 들여다본다. <strong><편집자 주></strong><br>
<p style="text-align: justify;">
<br>
<p style="text-align: justify;">
<br>
<table align="center" border="0"
class="class_div_main image" width="500"><tbody><tr><td>

</td></tr><tr><td>
[출처=freepik]
</td></tr></tbody></table>
<p style="text-align: justify;">
<br>
<p style="text-align: justify;">[뉴스투데이=김승한
(주)화물맨 부사장/경기대 겸직교수] 화물운송시장 내 미들마일 플랫폼 경쟁이 뜨겁다.<br>
<p style="text-align: justify;">
<br>
<p style="text-align: justify;">지난해 카카오톡비리터의 화물마당 인수 소식에 이어 최근에는 LG유플러스도 화물운송중계
플랫폼을 신규 사업으로 추진한다고 알려지면서, 이전 SK텔레콤, KT 등의 사업 런칭과 함께 통신 3사 모두 미들마일 플랫폼 시장 진
출을 하는 모양새이다.<br>
<p style="text-align: justify;">
<br>
<p style="text-align: justify;">그렇다면 기존 자체적 토종
미들마일 플랫폼, 화물정보망 시장의 강자인 '화물맨'의 미래 플랫폼 성장을 위한 노력은 어떠한가? 무엇을 준비 중인지 살펴보기로 하
자.<br>
<p style="text-align: justify;">
<br>
<table align="center" border="0" class="class_div_main image"
width="500"><tbody><tr><td>

</td></tr><tr><td>
[출처=각사 홈페이지]
</td></tr></tbody></table>
<p style="text-align: justify;">
<br>
<hr>
<p style="text-align: justify;"><strong>● 디지털
플랫폼에서 에코생태계의 중요성, 고객 니드 정확히 파악하고 빠르게 반영해야.</strong><br>
<p style="text-align: justify;">
<br>
<p style="text-align: justify;">가치창출의 관점에서 전통적 가치사슬 모델과 디지털 플랫폼 모델의 차이를 구체
적으로 설명하고 있는 저서 '플랫폼 비즈니스의 미래'(저자 이성렬 등, 2022년)를 보면 디지털 플랫폼 모델의 특징으로 에코생태계의
중요성을 강조하고 있다.<br>
<p style="text-align: justify;">
<br>
<p style="text-align: justify;">여기서 디지털 혁신 전략
의 수립은 고객 및 파트너들과 함께 창조적인 혁신을 해나가는 과정이라 설명하고 있는데, 고객, 파트너들과 에코생태계를 만들어가면
서 고객을 중심으로 서로 소통하고, 고객의 의견을 빠르게 반영하는 디지털 플랫폼을 수립하는 것이 디지털 혁신 전략의 핵심이라고
강조하고 있다.<br>
<p style="text-align: justify;">
<br>
<p style="text-align: justify;">디지털 플랫폼 모델을 실제 운영 중
인 필자의 의견도 동일하며, 에코생태계의 니드(need)를 얼마나 정확하게 이해하고 신속하게 구현해서 생태계 구성원들의 피드백을
지속적으로 주고받을 수 있을지가 성공의 핵심요소라는 점을 경험적으로 인지하고 있다.<br>
<p style="text-align: justify;">
<br>
<p style="text-align: justify;">특히 화물운송망을 다루는 플랫폼의 경우는 화주와 차주를 구성원으로 갖고 있는 다면 네트
워크 플랫폼 형태로서 각각의 니드를 충족하면서 이들 상호간의 이해충돌을 적절하게 방지 혹은 해결해 나가는 방안이 기능적으로 구
현되어야 하는 요소들이다.<br>
<p style="text-align: justify;">
<br>
<p style="text-align: justify;">(※플랫폼 운영자가 직
```



접 서비스를 제공하는 단면 모델과는 달리 복수의 생산자와 소비자가 참여하는 다면 네트워크 모델에서는 파트너들 간 교환 과정에서  
의 이해충돌 조정이 필연적으로 발생한다.)<br> <p style="text-align: justify;"> <br> <table align="center" border="0" class="class\_div\_main image" width="500"><tr><td>  </td></tr><tr><td> [출처=cenit] </td></tr></tbody></table>

<p style="text-align: justify;"> <br> <p style="text-align: justify;">이전 기고문에서도 잠시 언급했지만 기존 대기업들의 화물 운송 정보망 진입의 실패는 사실 에코생태계의 이해 부족이 근본 원인이라 할 수 있다.(필자의 본 시리즈 27편(2022.12.23) https://www.news2day.co.kr/article/20221222500162)<br> <p style="text-align: justify;"> <br> <br> <p style="text-align: justify;"><strong>● 화물맨, 전국 ‘개인(개별)화물자동차운송사업협회’ 디지털 전환 지원</strong><br> <p style="text-align: justify;"> <br> <p style="text-align: justify;">1998년 설립된 화물맨(창업주 임재득 회장, 임영목 대표)은 초기 ‘무전기(TRS)’ 시대 때부터 화물정보 중개 서비스를 시작한 1세대 ‘원조’ 화물 중개 플랫폼 기업이다.<br> <p style="text-align: justify;"> <br> <p style="text-align: justify;">2010년대 초반 휴대폰 앱 개발이라는 자체 디지털 혁신을 통해 당시 치열했던 ‘무전기’ 시대의 경쟁에서 성공하여 현재 전국 규모의 화물운송 네트워크를 운영 중인 강소기업이다.<br> <p style="text-align: justify;"> <br> <p style="text-align: justify;">지난해에는 경기대와의 산학 협력을 통해 화물맨 내부적으로 축적해 왔던 200만건의 빅데이터를 활용해서 요일과 운송시간, 날씨 등 조건을 반영, 적절한 화물운임을 산출하는 AI 모델을 개발하였고, 현재 화물맨의 화주 파트너에게 운임정보를 제공하고 있다.<br> <p style="text-align: justify;"> <br> <p style="text-align: justify;">(연합뉴스, “요일·시간·날씨 따라 화물운임 책정하는 AI기술 개발)<br> <p style="text-align: justify;"> <br> <table align="center" border="0" class="class\_div\_main image" width="500"><tbody><tr><td>  </td></tr><tr><td> [출처=‘화물맨’ 화주전용 프로그램 화면 캡처] </td></tr></tbody></table> <p style="text-align: justify;"> <br> <p style="text-align: justify;">에코생태계 내의 화주 파트너들 입장에서 가장 필요한 니즈가 원하는 경로에 적합한 운임을 예상하는 것인데 택시요금 같은 B2C 운임예측과는 달리 화물종류, 무게, 특성 그리고 차량의 종류, 방문지역 등 여러 요인을 고려해야하기 때문에 기존 축적된 빅데이터가 없으면 예측 자체가 어려운 과제이다.<br> <p style="text-align: justify;"> <br> <br> <p style="text-align: justify;"><strong>● 차주 니즈도 적극 반영, 차주 파트너의 디지털 복지 지원을 지향</strong><br> <p style="text-align: justify;"> <br> <p style="text-align: justify;">그럼 에코생태계의 다른 쪽인 차주의 니즈는 어떨까? 질 좋은 화물을 많이 제공받는 것과 운임 회수에 대한 리스크를 덜어주는 것이 가장 큰 바람일 것이다.<br> <p style="text-align: justify;"> <br> <p style="text-align: justify;">화물맨은 차별화 포인트로 단순 화물 매칭 서비스를 넘어서 회사를 최적화한 ‘책적경로’ 정보 제공을 준비하고 있다. 즉, 공차를 최소화한 복합 경로를 계산하는 알고리즘을 자체 개발하였고, 차주 파트너가 이들 정보에 편하게 접근할 수 있게 할 계획이다.<br> <p style="text-align: justify;"> <br> <p style="text-align: justify;">또 하나의 고민인 운임 관련 리스크를 해결하기 위해서 ‘H-pass’ 라는 화물맨 자체의 지불(payment) 조건을 차주 앱에서 제공 중인데, H-pass를 이용하는 차주에게 화물맨이 화주를 대신해서 운임을 책임지고 지불하는 제도이다.<br> <p style="text-align: justify;"> <br> <p style="text-align: justify;">현재 초기 운영과정에서 차주들의 반응이 매우 좋은 상태이고, 궁극적으로는 개인사업자인 차주들도 안정적인 월급 노동자와 같은 혜택을 받도록 하는 것이 화물맨의 최종 목표이다.<br> <p style="text-align: justify;"> <br> <table align="center" border="0" class="class\_div\_main image" width="500"><tbody><tr><td>  </td></tr><tr><td> [출처=‘화물맨’ 차주전용 프로그램 화면 캡처] </td></tr></tbody></table> <p style="text-align: justify;"> <br> <p style="text-align: justify;">화물맨은 최근 3월에 전국 16개의 시/도 ‘개인(개별)화물자동차운송사업협회’의 디지털 전환을 위한 업무 협약을 체결하였다. 이 또한 에코생태계 연계 전략의 일환으로 추진되었고, 궁극적으로는 차주 파트너들의 디지털 복지를 지원하겠다는 것이 화물맨이 지향하는 목적이다.<br> <p style="text-align: justify;"> <br> <p style="text-align: justify;">우선 협회의 홈페이지 제작 및 전용 앱을 제공해 협회와 회원들의 효율적인 소통공간을 제공한다.<br> <p style="text-align: justify;"> <br> <p style="text-align: justify;">홈페이지를 통해 화물맨은 협회회원들에게 협회 진행사업, 공지사항 등 협회 정보는 물론 경력증명서 발급안내, 대폐차, 직무보수교육, 여러 검사 및 교육의 온라인 예약 및 접수안내 및 날씨정보, 도로교통정보, 고속도로상황 등 다양한 편의 기능을 제공한다

다.<br> <p style="text-align: justify;"> <br> <p style="text-align: justify;">(<strong>※대폐차</strong> : 화물자동차 운송사업 및 화물자동차 가맹사업에 사용되는 차량을 다른 차량을 교체하는 것)<br> <p style="text-align: justify;"> <br> <p style="text-align: justify;">또한 홈페이지와 같이 제공되는 협회 전용 앱은 오는 4월 중순부터 출시될 예정이다. 해당 앱은 별도의 가입없이 화물맨의 실시간 화물정보를 비롯하여, 지도기반 화물정보, 무료전자세금계산서 서비스 등 간편 기능을 제공한다.<br> <p style="text-align: justify;"> <br> <p style="text-align: justify;">아울러 차주, 협력사, 민원 등 파트를 세분화하여 전문적인 상담이 가능한 365일 24시간 콜센터 서비스를 운영할 예정이다.<br> <p style="text-align: justify;"> <br> <table align="center" border="0" class="class\_div\_main image" width="500"><tbody><tr><td> </td></tr><tr><td> [출처='화물맨'에서 제작한 '서울 개인(개별)화물자동차운송사업협회' 홈페이지 화면 캡처] </td></tr></tbody></table> <p style="text-align: justify;"> <br> <p style="text-align: justify;">화물맨은 이외에도 화주 및 차주에 제공하는 화물 마일리지 이벤트를 확대해 지속적으로 에코생태계 내의 파트너들에게 다양한 혜택과 편리성, 그리고 안정성을 제공하는 화물 플랫폼으로 성장하기 위한 노력을 떠나갈 계획인 것으로 알려져 있다.<br> <p style="text-align: justify;"> <br> <p style="text-align: justify;">[정리=최봉 산업경제 전문기자]<br> <p style="text-align: justify;"> <br> <p style="text-align: justify;"> <br> <hr> <p style="text-align: justify;"> <br> <table align="center" border="0" class="class\_div\_main image" width="500"><tbody><tr><td> </td></tr><tr><td> </td></tr></tbody></table> /<br> / <br>]]></content>

<표 4> 원시 데이터 특징 예시

,원시데이터 기사 예시 1  
<author><![CDATA[○○○○]]></author>  
(본문 내용 일부 생략....)  
새로운 시대적 사고 발상으로 '두 마리 토끼' 의미가 위기에서 기회로 바뀌었다고 할 수 있다. 다산과 풍요, 지혜와 민첩함을 지닌 토끼처럼 2023년은 위기를 현명하게 돌파해 건강과 풍요 두 가지 목표를 모두 달성하기를 기원해 본다.  
/○○○ 기자 ]]></content>

원시데이터 기사 예시2  
<author><![CDATA[#2023011001000317400009961#]]></author>  
(본문 내용 일부 생략....)  
.....클러스터를 본격 육성해야 하는 전남에게 큰 시사점을 준다"고 말했다.  
/○○○ 기자 #2023011001000317400009961# ]]></content>

<표 5> 기사의 메타데이터의 누락(기사 정보가 본문에만 존재하는 경우)

(기사 초반 내용 생략....)  
JYP엔터테인먼트와 리퍼블릭 레코드의 협업은 K팝 역사에 오래 남을 이정표를 세웠다. 2020년 2월, 리퍼블릭 레코드와 가장 먼저 손잡고 글로벌 시장 공략에 나선 트와이스는 그해 6월 미니 9집 'MORE & MORE'(모어 앤드 모어)로 '빌보드 200' 차트 200위에 첫 입성했다. 이후 2020년 12월 정규 2집 'Eyes wide open'(아이즈 와이드 오픈) 72위, 2021년 6월 미니 10집 'Taste of Love'(테이스트 오브 러브) 6위, 2021년 11월 정규 3집 'Formula of Love: O+T=]]></content>

#### 실제 웹 화면

JYP엔터테인먼트와 리퍼블릭 레코드의 협업은 K팝 역사에 오래 남을 이정표를 세웠다. 2020년 2월, 리퍼블릭 레코드와 가장 먼저 손잡고 글로벌 시장 공략에 나선 트와이스는 그해 6월 미니 9집 'MORE & MORE'(모어 앤드 모어)로 '빌보드 200' 차트 200위에 첫 입성했다. 이후 2020년 12월 정규 2집 'Eyes wide open'(아이즈 와이드 오픈) 72위, 2021년 6월 미니 10집 'Taste of Love'(테이스트 오브 러브) 6위, 2021년 11월 정규 3집 'Formula of Love: O+T=ㄷ'(포뮬러 오브 러브: O+T=ㄷ) 3위, 2022년 8월 미니 11집 'BETWEEN 1&2'(비트윈 원앤투) 3위, 2023년 3월 미니 12집 'READY TO BE'(레디 투 비) 2위까지 계단식 성장을 일구고 K팝 걸그룹 중 최다인 총 4장의 앨범을 '빌보드 200' 톱 10 반열에 올렸다. 또 2021년 10월 첫 오리지널 영어 싱글 'The Feels'(더 필즈)를 통해 빌보드 '핫 100'에 첫 입성한 데 이어 2023년 1월 발매한 영어 싱글 'MOONLIGHT SUNRISE'(문라이트 선라이즈)로 해당 차트에 통산 두 번째 이름을 올리는 저력을 과시했다. 게다가 다섯 번째 월드투어의 일환으로 6월 로스앤젤레스 소파이, 7월 뉴욕 메트라이프 등 K팝 걸그룹 최초 미국 현지 스타디움 입성 및 매진까지 달성했다.

#### 해당 부문 이후 데이터 소실

<표 6> 원시 데이터 특징 예시(데이터 소실 1)

<content><![CDATA[- 포항문화예술지원사업 1년 간의 활동 성과를 공유하는 소통의 장 마련<br/>- 포항의 예술인을 위한 안정적인 지원과 지역 예술생태계 강화<br/>- 지난 13일부터 16일까지 4일간 문화예술팩토리 4층 아트갤러리에서 진행한 재단법인 포항문화재단의 '2023 포항문화예술지원사업 성과공유회 '당신의 예술적 삶은 반짝거려요'가 성황리에 마무리됐다.<br/>- 지역 문화예술생태계 강화 및 예술인의 안정적인 성장 도모를 위한 목적으로 운영되고 있는 포항문화예술지원사업은 법정문화도시 조성사업과 연계하여 4년 차를 맞았다.<br/>- 이번 성과공유회에서는 여러 프로그램이 선보였는데 그중 예술가-시민 참여 프로그램을 통해 코믹 작명가에게 받는 유쾌한 예명(예술활동명) 만들기가 시민들에게 큰 인기를 얻으며 1, 2회차로 진행되었다. 또한 공공프로젝트 '예술항구' 참여 작가인 박세호 서예가가 이라는 주제로 '교' 필법 퍼포먼스와 더불어 참여 예술가들이 포항과 관련된 단어들이 적힌 한지를 물에 담그는 오프닝 퍼포먼스를 연출했다.<br/>- 부대 프로그램으로는 문학 분야 도서 출간기념 북콘서트가 열려 최미경(인문 분야) 모더레이터와 문학 분야 8명의 작가(이원만, 이순영, 전은주, 조종래, 박형철, 김귀현, 박한규, 이경옥)가 함께 출간물에 대한 주요 내용과 비하인드 스토리 등 전반적인 작품 소개를 나누는 시간을 가졌다.<br/>- 마지막 날은 작가별 참여 소감을 나누고 향후 예술지원사업의 진행 방향을 재점검하며, 사업 진행과 관련된 일련의 과정들이 유쾌한 미담으로 남을 수 있는 위트있는 시상식 형태로 마무리됐다.<br/>- 올해 포항문화예술지원사업의 참여 작가는 "그간 개인 작업에만 집중한다고 같은 사업에 지원한 작가들의 작품을 잘 알지 못했었는데 성과공유회를 통해 서로 나눌 수 있는 자리가 있어 좋았다."며 소감을 밝혔다.]></content>

#### 특정 매체 서명 기호 안 글자 삭제

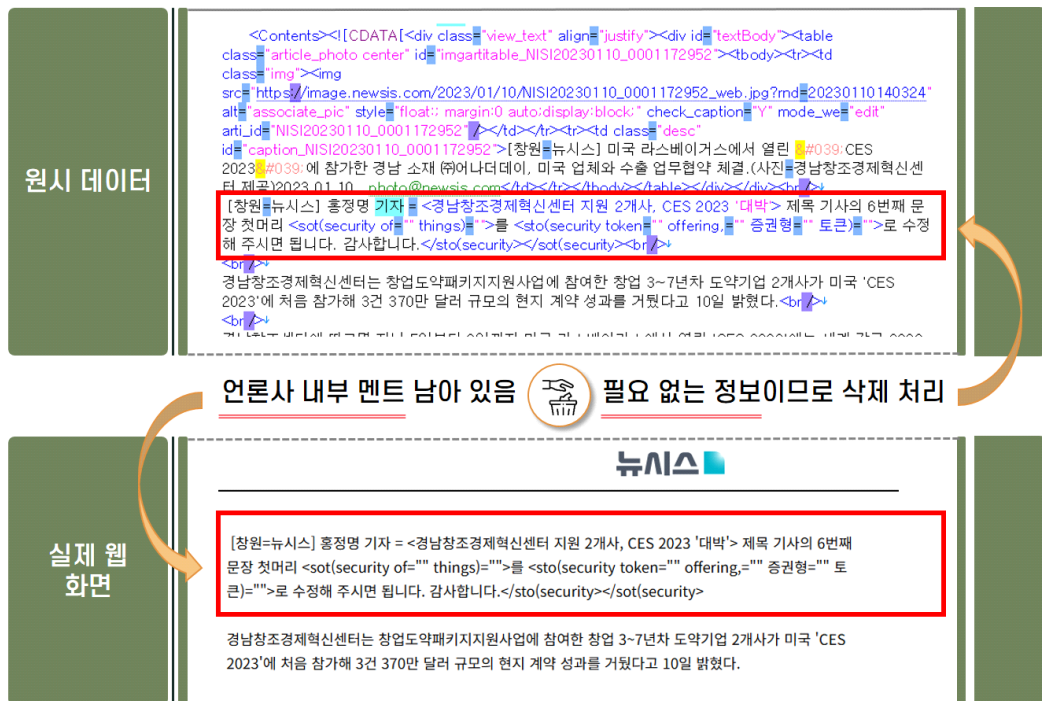
그중 예술가-시민 참여 프로그램 <나똥 이름은>을 통해 코믹 작명가에게 받는 유쾌한 예명(예술활동명) 만들기가 시민들에게 큰 인기를 얻으며 1, 2회차로 진행되었다. 또한 공공프로젝트 '예술항구' 참여 작가인 박세호 서예가가 <나똥 포항>이라는 주제로 '교' 필법 퍼포먼스와 더불어 참여 예술가들이 포항과 관련된 단어들이 적힌 한지를 물에 담그는 오프닝 퍼포먼스를 연출했다.

부대 프로그램으로는 문학 분야 도서 출간기념 <나똥 책> 북콘서트가 열려 최미경(인문 분야) 모더레이터와 문학 분야 8명의 작가(이원만, 이순영, 전은주, 조종래, 박형철, 김귀현, 박한규, 이경옥)가 함께 출간물에 대한 주요 내용과 비하인드 스토리 등 전반적인 작품 소개를 나누는 시간을 가졌다.

마지막 날은 작가별 참여 소감을 나누고 향후 예술지원사업의 진행 방향을 재점검하며, 사업 진행과 관련된 일련의 과정들이 유쾌한 미담으로 남을 수 있는 위트있는 시상식 형태로 마무리됐다.

올해 포항문화예술지원사업의 참여 작가는 "그간 개인 작업에만 집중한다고 같은 사업에 지원한 작가들의 작품을 잘 알지 못했었는데 성과공유회를 통해 서로 나눌 수 있는 자리가 있어 좋았다."며 소감을 밝혔다.

<표 7> 원시 데이터 특징 예시(데이터 소실 2)



<그림 3> 오류 유형 ①: 언론사 내부 설명이 있는 경우



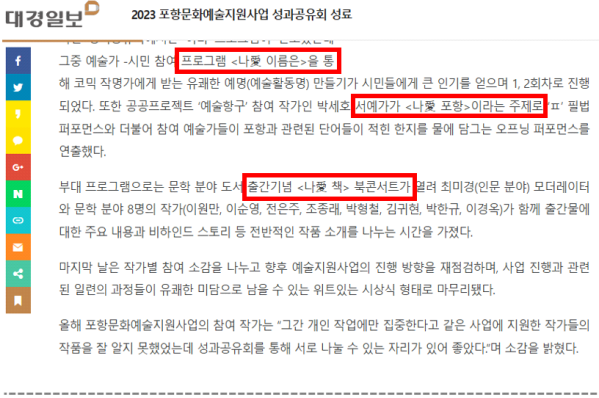


<그림 4> 오류 유형 ②: 문자 코드가 조합형으로 나타난 경우

각 매체로부터 받은 원시 데이터는 고유한 특징을 가지고 있다. 이러한 특징을 파악하지 않고 데이터를 처리하면 오류가 발생할 가능성이 높다. 예를 들어, 특정 요소 이름이 어떤 매체에서는 본문 내용으로 필요하지만 다른 매체에서는 삭제해야 할 정보일 수 있다.

이러한 오류는 기사를 꼼꼼히 검토해도 발견하기 어렵기 때문에 작업 초기 단계에서 각 매체의 특징을 파악하고 이를 고려하여 데이터를 처리하는 것이 중요하다. 이를 통해 보다 정확하고 신뢰성 높은 데이터를 생성할 수 있었다.

특정 매체 서명 기호 안 글자 삭제 : ◇ 기호 안 내용 데이터 소실

실제 웹 화면	원시 데이터
	<pre> &lt;content&gt;&lt;![CDATA[~ 포항문화예술지원사업 1년 간의 활동 성과를 공유하는 소통의 장 마련&lt;br&gt;~ 포항의 예술인을 위한 안정적인 지원과 지역 예술생태계 강화&lt;br&gt;~ 지난 13일부터 16일까지 4일간 문화예술팩토리 4층 아트갤러리에서 진행한 재단법인 포항문화재단의 '2023 포항문화예술지원사업 성과공유회' 당신의 예술적 삶은 반짝거려요'가 성황리에 마무리됐다.&lt;br&gt;~ 지역 문화예술생태계 강화 및 예술인의 안정적인 성장 도모를 위한 목적으로 운영되고 있는 포항문화예술지원사업은 법정문화도시 조성사업과 연계하여 4년 차를 맞았다.&lt;br&gt;~ 이번 성과공유회에서는 여러 프로그램이 선보였는데 그중 예술가-시민 참여 프로그램인 '포항 문화예술지원사업'을 통해 지역 예술가에게 받는 유쾌한 예명(예술활동명) 만들기가 시민들에게 큰 인기를 얻으며 1, 2회자로 진행되었다. 또한 공공프로젝트 '예술항구' 참여 작가인 박세호 서예가가 이라는 주제로 '표' 필법 퍼포먼스와 더불어 참여 예술가들이 포항과 관련된 단어들에 적힌 한지를 붙여 만드는 오프닝 퍼포먼스를 연출했다.  부대 프로그램으로는 문학 분야 도서 출간기념 '나눔' 책 북콘서트' 열려 최미경(인문 분야) 모더레이터와 문학 분야 8명의 작가(이원만, 이순영, 전은주, 조종래, 박형철, 김귀현, 박한규, 이경욱)가 함께 출간물에 대한 주요 내용과 비하인드 스토리 등 전반적인 작품 소개를 나누는 시간을 가졌다.  마지막 날은 작가별 참여 소감을 나누고 향후 예술지원사업의 진행 방향을 재점검하며, 사업 진행과 관련된 일련의 과정들이 유쾌한 미담으로 남을 수 있는 위트있는 시상식 형태로 마무리됐다.  올해 포항문화예술지원사업의 참여 작가는 "그간 개인 작업에만 집중한다고 같은 사업에 지원한 작가들의 작품을 잘 알지 못했는데 성과공유회를 통해 서로 나눌 수 있는 자리가 있어 좋았다"며 소감을 밝혔다.  ]]&gt;&lt;/content&gt; </pre>

<그림 5> 오류 유형 ③: 서명 기호 안 텍스트가 사라진 경우

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.

- 하나의 기사를 한 개의 엑스엠엘(XML) 파일로 제공함.
- &lt; &gt;, html 언어 등 그대로 남아 있음.
- 기사가 불완전하게 종결되는 경우가 있음.
- 원시 데이터에서 서명 기호 안의 글자가 누락되는 등 데이터 소실이 발견됨.
- 기자 정보 등 메타데이터의 누락이 있음.
- 인용 부호로 ', ', ", " 등을 사용해 표준에 맞지 않음. 인용 부호의 열고 닫는 짝이 맞지 않음.
- 같은 의미로 사용되는 가운뎃점, 마침표, 쉼표 등이 여러 가지 코드로 일관성 없이 사용됨.
- 이(李), 리(李)와 같은 한자 호환용 코드가 달리 사용되어 데이터의 공유와 유통에 문제를 일으킴.

<표 8> 데이터 특징

### 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.

※ 좌우 비교를 위해 줄바꿈 임의 삽입

<p>제목: 영주 서천 배고개둔치서 34개 의료서비스 누리요</p> <p><u>18일까지 시민건강 체험마당</u> <u>갑상선·구강·약물 오남용 등</u> <u>건강상담부터 검사·관리까지</u> <u>영주 시민건강 체험마당이 16일 오후 7시</u> <u>개막해 18일까지 서천 배고개둔치에서 펼쳐</u> <u>진다.</u> <u>코로나19로 4년 만에 재개된 시민건강 체험</u> <u>마당은 시민들의 편의와 접근성을 높이기</u> <u>위해 오후 7시부터 진행된다.</u></p> <p>올해는 영주시의사회, 한의사회, 치과의사회, 약사회, 안경사회, 간호사회 등 지역보건 의료단체와 대학·병원·국민건강보험공단 영주봉화지사·영주시건강기협회 등 34개 기관 단체가 참여해 다양한 체험 부스를 운영한다.</p> <p>14회째를 맞는 이번 행사에서는 진료상담, 건강검사, 건강생활실천 등 30여 개 체험관을 운영하고 걷기 활성화를 위한 야간 걷기 행사도 진행된다.</p> <p>주요 프로그램으로는 갑상선초음파, 한방진료, 구강진료, 약물 오남용 상담 등 진료상담과 혈압 및 당뇨, 빈혈, 체성분검사, 시력 검사 등 평소 소홀할 수 있는 기본적인 건</p>	<p>제목: 영주시 시민건강 체험마당 실시</p> <p><u>영주시가 오는 16일부터 18일까지 3일간의</u> <u>일정으로 서천 배고개둔치 주차장에서 '제</u> <u>14회 시민건강 체험마당'을 개최 한다고 11</u> <u>일 밝혔다.</u> <u>2005년부터 매년 진행되던 시민건강 체험</u> <u>마당은 코로나19로 인해 4년 만에 재개된</u> <u>다.</u></p> <p>올해는 영주시의사회, 한의사회, 치과의사회, 약사회, 안경사회, 간호사회 등 지역보건 의료단체와 대학·병원·국민건강보험공단 영주봉화지사·영주시건강기협회 등 34개 기관 단체가 참여해 다양한 체험 부스를 운영한다.</p> <p>14회째를 맞는 이번 행사에서는 진료상담, 건강검사, 건강생활실천 등 30여 개 체험관을 운영하고 걷기 활성화를 위한 야간 걷기 행사도 진행된다.</p> <p>주요 프로그램으로는 갑상선초음파, 한방진료, 구강진료, 약물 오남용 상담 등 진료상담과 혈압 및 당뇨, 빈혈, 체성분검사, 시력 검사 등 평소 소홀할 수 있는 기본적인 건</p>
--	---



<p>강검사, 올바른 걷기체험, 금연·절주, 치매검진, 감염병예방 등 건강생활실천 체험, 심폐소생술, 구강관리, 메이크업 등 생활 속에서 실천할 수 있는 건강관리체험 등이 진행된다.</p> <p>행사기간 중 3개 어린이집(보현·아트·풍기) 원아들의 율동, 드럼북, 댄스 시연과 걷기체조 및 의용소방대원들의 심폐소생술 시연 등 공연, 야간 서천걷기행사 등도 진행된다. 특히 야간 서천걷기행사는 17일 <b>오후</b> 8시 영주시걷기협회와 걷기동호회의 협조로 참여 <b>행사장에서 제2가흥교까지 코스를</b> 시민들과 함께 <b>걷는다. 시는 이번 행사로</b> 걷기의 중요성을 홍보하고 시민들이 일상생활에서 걷기를 실천할 수 있는 계기를 마련하<b>고자 한다.</b></p> <p>권경희 보건소장은 "가족과 <b>함께</b> 서천도 걸으시고, 시민건강체험마당에서 건강도 챙기시고 코로나19로 지친 심신에 조금이나마 충전의 기회가 되기를 바란다"고 <b>밝혔다.</b></p>	<p>강검사, 올바른 걷기체험, 금연·절주, 치매검진, 감염병예방 등 건강생활실천 체험<b>부스</b>, 심폐소생술, 구강관리, 메이크업 등 생활 속에서 실천할 수 있는 건강관리체험 등이 진행된다.</p> <p>행사기간 중 3개 어린이집 원아들의 율동, 드럼북, 댄스 시연과 걷기체조 및 의용소방대원들의 심폐소생술 시연 등 공연, 야간 서천걷기행사 등도 진행된다. 특히 야간 서천걷기행사는 <b>오는</b> 17일 <b>저녁</b> 8시 영주시걷기협회와 걷기동호회의 협조로 참여 시민들과 함께 <b>진행된</b>다. 걷기의 중요성을 홍보하고 시민들이 일상생활에서 걷기를 실천할 수 있는 계기를 마련하<b>기 위해서</b>다.</p> <p>권경희 보건소장은 "<b>가정의달 5월에는 가족과</b> 서천도 걸으시고, 시민건강체험마당에서 건강도 챙기시고 코로나19로 지친 심신에 조금이나마 충전의 기회가 되기를 바란다"고 <b>말했다.</b></p>
---	--

<표 9> 유사도 비교를 통해 사용하지 않는 기사의 예

## 나. 기사 선택

작업 대상을 선정할 때 최초 원시데이터에서 1차적으로 선별 작업을 통해 기사를 선별하게 된다. 중복 기사, 어절 수가 부족한 기사, 기사로 볼 수 없는 기사 등을 제거함으로써 이후 단계에서 작업의 효율성을 높일 수 있었다. 기준은 아래와 같다.

- 기사 길이가 너무 적은 100어절 이하 기사는 제외함(정제를 하기 전에 이미 100어절에 미치지 못할 것으로 예상되는 기사가 1차 제외되었고, 정제 후 어절 수 계산으로 100어절 미만 기사는 2차로 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, 타 기관 소장, 부장, 의사 등 매체에 속하지 않은 외부 기고가 및 전문가가 작성한 기사 등은 제외함.
- 기자 정보가 매체 소속이 아닌 같은 계열사인 경우는 제외함.
- 기자 정보가 없는 데이터는 제외함(사설 제외).
- 기자 정보가 들어가 있으나 본문 내용에서 외부 기고가 확실한 경우 제외함.
- 기자 정보가 공동 취재단인 경우 해당 기사는 제외함.
- 시민 기자의 글은 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, 원장 등	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, ~제공	삭제	연합뉴스가 출처인 기사, 또는 통신사로부터 제공받은 기사
각 매체	공동 취재단	삭제	해당 매체와 계약 등을 일일이 확인 불가 (국방부 공동 취재단, 올림픽 공동 취재단, 대선 공동 취재단 등)
각 매체	특별취재팀	삭제	해당 매체와 계약 등을 일일이 확인 불가 (대선평취재팀 등)
각 매체	전국종합	삭제	해당 매체와 계약 등을 일일이 확인 불가 (전국종합, 지역종합, 지방종합 등)
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 적은 경우 (아나운서, 앵커, 피디(PD), 프로듀서, 진행 등)
각 매체	리포터	삭제	모집 프리랜서 기자
각 매체	객원기자	삭제	모집 프리랜서 기자
OO매체	디지털 ○○ 기자	삭제	해당 매체 미디어 그룹에 속한 별도의 법인 (헬스 ○○, 월간 ○○, 주간 ○○ 등)
OO매체	어린이 ○○ 기자	삭제	해당 매체 미디어 그룹에 속한 별도의 법인

<표 10> 저작권 이용 문제로 인해 사용하지 않는 기사의 특징

매체에 속한 기자의 기사 저작권은 각 매체에 귀속되지만, 매체에 속하지 않은 외부 기고가 작성한 글 등의 경우 매체와의 계약 조건에 따라 저작권 소유가 달라질 수 있다. 따라서 특정 기사에 대해 매체가 저작권을 보유하지 않는 경우도 있을 수 있다. 이러한 이유로 저작권 문제가 조금이라도 발생할 가능성이 있는 기사는 모두 사용하지 않는다.

예를 들어, 오마이뉴스는 시민기자가 기사 작성에 참여하는 매체로, 연간 기사 중 일정 부분은 시민기자가 작성한 것이다. 이에 따라 오마이뉴스와 시민기자의 계약 형태를 검토하고 매체로부터 문제없다는 확인을 받았지만, 오마이뉴스의 시민기자 작성 기사는 모두 제외하였다. 공공 말뭉치로 구축되는 기사에서는 조금의 저작권 분쟁 요소도 발생해서는 안 되기 때문이다.

또한, 일부 기사의 경우 기자 이름이 표기되어 있지만 실제 내용은 외부 기고가가 작성한 경우도 있다. 이러한 기사는 불필요 요소를 제거하는 과정에서 기사를 일일이 확인하면서 모두 사용하지 않는 기사로 처리하였다.

최근 인공 지능 새싹기업(창업초기기업)인 퍼플렉시티(Perplexity)가 저작권 문제로 미국 월스트리트저널(WSJ)과 뉴욕포스트로부터 소송을 당하면서<sup>2)</sup>, 인공 지능과 저작권 간의 갈등이 부각되고 있다. 월스트리트저널과 뉴욕포스트는 퍼플렉시티가 인공 지능 훈련에 자사 콘텐츠를 무단으로 사용해 독자와 수익을 빼앗고 브랜드 가치를 훼손시켰다고 주장하며, 저작물에 대한 데이터베이스 파괴와 건당 최대 15만 달러의 손해배상을 요구하고 있다.

퍼플렉시티는 최신 정보와 출처 링크를 제공하는 인공 지능 기반 검색 엔진으로 급성장 중이지만, 저작권 침해 논란이 지속되고 있다. 이는 최근 인공 지능 기업들이 저작권 문제로 여러 소송에 직면하고 있는 가운데 발생한 사건으로, 인공 지능 모델의 데이터 수집과 저작권 준수의 중요성을 보여 준다.

국립국어원의 신문 기사 원문 자료 수집 및 정제 사업은 신문 기사를 대상으로 말뭉치를 수집하고 정제하고 저작권에 문제가 되는 기사를 사용하지 않는 것에 특별히 유의하고 있다. 신문 기사에는 외부 기고가 또는 공동 취재단같이 저작권에 대해 이해를 달리하는 기사들이 상당수 포함되어 있기 때문이다. 이에 전체 기사 정보를 추출하여 위의 <표 10>와 같이 저작권 이용에 위험이 있는 기사를 걸러 냈다.

---

2) <https://www.news1.kr/world/usa-canada/5575340>

## 4. 데이터 2차 정제

데이터 1차 정제를 마친 기사는 데이터 총괄 관리자가 매체별로 오류 등을 1차로 수정하고 정제하였다. 이는 모든 오류를 수정한 것이 아니고 기계적으로 검증하여 유형별 오류를 1차적으로 처리한 것으로 아직까지 기사에는 불필요한 요소와 더불어 많은 정보가 포함되어 있다. 최종적으로 작업자가 직접 기사를 읽으며 불필요한 요소를 제거하고, 사용하지 않는 기사들은 불용 표시를 하여 작업을 진행하였다.

### 가. 불필요한 요소 제거

불필요한 요소 제거 과정에서는 기사 내 불필요한 정보를 세밀하게 검토하여 삭제하는 작업이 이루어지며, 전체 공정 중에서 가장 많은 인력이 투입되는 핵심 단계이다. 기사마다 캡션 정보, 기자 이름, 전문 또는 비문 등의 요소를 제거하고, 기사의 저작권을 재검토하여 사용 여부를 결정한다. 이 과정을 통해 최종적으로 사용할 기사와 사용하지 않을 기사를 구분하고 데이터의 정확성을 높이는 작업이다.

#### 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

## (6) 뇌경색의 골든타임

입력 : 2023.08.04 06:59  기자

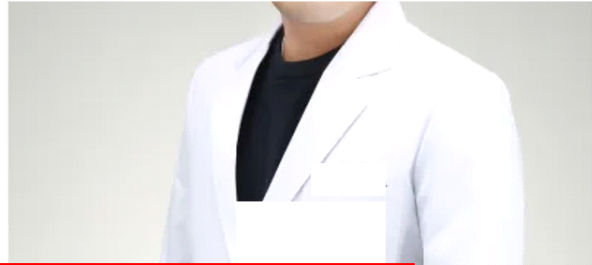
■ 뇌의 정맥·동맥 '혈전제거술' 시행

■ 4시간 이내, 빠를수록 후유증 감소

심장질환, 고혈압, 당뇨, 고지혈증 같은 만성질환 혹은 암과 같은 다양한 원인에 의해 뇌혈관이 갑자기 막혀버릴 경우 반신마비, 언어장애, 감각 이상, 어지러움, 의식 저하와 같이 다양한 신경학적 이상이 발생하게 된다.

뇌혈관 폐색으로 뇌가 혈액을 공급받지 못하면 뇌세포들은 수 분 내에 죽기 시작한다. 다만 환자 혈관 상태에 따라서는 주변 혈관의 도움을 통해 더 오랜 시간을 버티게 되는 경우도 있다. 하지만 이런 경우라도 일반적으로 수 시간을 버텨내는 것은 어려우며, 골든타임을 지나게 된다면 돌이킬 수 없는 뇌의 후유증을 남기게 된다. 뇌경색에 있어서 골든타임이 중요한 이유이다.

사용을 하게 되는 경우에는 출혈 위험성이 증가하기 때문에 모든 환자에서 가능한 치료는 아니다. 환자의 피검사 결과, 과거력, CT, MRI 같은 뇌 영상 사진 결과를 종합적으로 판단하여 시행하게 된다.



뇌졸중센터 교수

동맥 내 혈전 제거술의 경우 가느다란 도관을 다리의 동맥에 삽입한 이후 도관을 이용하여 직접 막힌 대뇌 혈관까지 접근하여 폐색된 혈관을 막고 있는 혈전을 제거한다. 수년 전까지만 하더라도 6시간 이내의 환자에게서만 시술이 가능하였으나 최근 기구의 발전, 최신 연구결

<그림 6> 사용하지 않는 기사의 예 (기고문)

기자 정보에는 기자 이름이 들어가 있으나 실제 기사를 확인해 보면 외부 기고가가 작성한 글로 확인이 된다. 단순 자문을 받고 전문가가 답변을 한 내용이 아니라, 글 자체를 외부 기고가가 작성한 글로 파악이 되는 기사이다. 위와 같은 경우, 오류가 있는 기사는 아니지만 저작권에 문제가 될 수 있기 때문에 선택하지 않는다.

원 저작자(기자 또는 매체)의 의도가 훼손되지 않은 선에서 최소한의 수정 진행

: 내용을 알 수 없는 경우 사용하지 않음

원시 데이터

```

<content><[[CDATA[[스포츠투데이 김영훈 기자] 세계 랭킹 1위 손영사의 벽은 높았다. 신유빈은 손영사를 상대로 완패해 동메달을 획득했다.
↓
한국 탁구 간판 신유빈은 1일(한국시각) 중국 저장성에 위치한 공수 캐널 스포츠파크 체육관에서 열린 제19회 2022 항저우 아시안게임 탁구 여자 단식 4강전에서 손영사에게 0-4(7-11 8-11 12-14 10-12)로 완패했다.
↓
신유빈은 전날(9월 30일) 8강에서 대만의 저즈영우를 꺾고 4강에 진출하며 동메달을 확보했다.
↓
탁구 중족은 4강전 패자 2명에게 모두 동메달을 수여한다.
↓
이번 상대는 여자 단식 랭킹 1위 손영사였다. 신유빈은 손영사와 총 4번의 맞대결을 펼친 바 있는데 모두 패했다.
↓
신유빈은 준결승전 손영사를 상대로 2세트를 연달아 내주며 골라간 가운데 3세트에서는 먼저 10점을 따내 승리를 앞섰지만
        
```

---

신유빈은 준결승전 **손영사를 상대로 불전을 펼쳤다.** 2세트 연달아 내주며 골라간 가운데 3세트에서는 먼저 10점을 따내 승리를 앞섰지만 손영사의 반격에 흔들렸고 듀스 끝에 12-14로 패했다.

이어 4세트까지 손영사를 꺾지 못하며 결승 진출에 실패했다.

그러나 분명 수확은 있다. 이전까지 신유빈은 국제 종합 대회 단식 메달이 없었는데 이번 항저우 대회에서 동메달을 따내는 쾌거를 이뤘다.

이제 신유빈은 내일(2일) 있을 여자 복식 준결승에서 일본을 꺾고 결승 진출을 노린다.

실제 웹 화면

<그림 7> 오류 유형 ④: 원저작자의 의도가 훼손될 우려가 있는 경우

**뉴스웍스** 광명시, 군 입대자 임명지원금 10만원 지원

역이며, 신청일 기준 광명시에 1년 이상 연속하여 주소를 두고 거주한 광명시민 가운데 9월13일 이후 입영자이다.

신청은 10일부터이며, 입영통지서 수령일로부터 입영일 전까지 신분증과 입영통지서를 가지고 관할 주소지 행정복지센터를 방문해 신청하면 된다. 부득이한 사유로 입영 전에 신청하지 못한 경우에는 입영 후 3개월 이내에 신청할 수 있으며, 이 경우에는 직계존비속 등 조건에 부합하는 대리인이 대신 신청할 수 있다.

지원금은 광명시 지역화폐인 광명사랑화폐로 지급되므로 신청 전에 지역화폐에 가입해야 한다. 신청일로부터 8일 이내 지급되며 지급일로부터 3년 내에 지역화폐 사용처에서 사용할 수 있다. 입영지원금에 관한 자세한 사항은 광명시 안전총괄과 또는 주소지 행정복지센터에 문의하면 안내받을 수 있다.

박승원 광명시장은 "입영지원금을 통해 광명시 청년들의 병역의무 이행을 격려하고, 광명시민으로서의 자부심을 높일 수 있을 것으로 기대한다"고 말했다.

**한편,**

**경인매일** 경기도의회 김미리 의원, 열악한 학교급식 환경 개선 촉구

[24차 완만팀] 김미리 의원도 연 내로는 48% 합선임크 곧 종료>

이와 함께 김 의원은 얼마 전 서울시에서 발생한 사고로 이슈가 된 '교실배식 안전 문제'에 대해서도 개선책 마련을 촉구했다.

김 의원은 "중장기적으로는 급식시설을 별도로 설치해야 하지만, 단기적으로는 교실배식 학교에 대한 이물질 혼입 방지 대책, 식중독 방지 대책을 철저히 수립해야 한다"라고 강조하였다.

이에 대해 경기도교육청의 정수호 대외협력국장은 "경기도 내 전 학교 급식시설을 대상으로 2027년까지 환기시설을 개선하겠다"고 밝히고 "교실 배식이 이루어지는 학교의 안전한 급식을 위해서 급식용 엘리베이터 부근에 CCTV를 설치하는 등 **적극적으로 대책을 마련하겠다**"고

저작권자 © 경인매일 - 세력에 타협하지 않는 신문 무단 전재 및 재배포 금지

<그림 8> 오류 유형 ⑤: 기사 내용이 짧은 경우

## 2) 불필요한 본문 내용 삭제

기사와 무관한 정보들을 삭제하는 과정이다. 이 정보들은 전체 맥락을 해치므로 제거한다. 또한 연설문, 입장문, 누리 소통망 서비스(SNS) 게시물 등 기사와는 다른 외부 전문이 실린 경우, 이러한 전문은 기자가 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 뒤이어 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	[사진 및 참고자료 서울○○○○ / 사진 ©○○○] 사진 제공 몰디브 ○○○ ○○○ © 사진 설명 : 표창 수상을 기념하는 모습
기자의 이름, 계정(ID) 등의 정보	[인천=○○○기자] ○○○○=○○○기자] 【서울=○○○】○○○ 기자 = 취재협조 = ○○○○사업단
저작권 관련 내용 (‘Copyright©’ 등)	랄프 깁슨 'Salon Litteraire'. ©Ralph Gibson -----@--.co.kr/2022-10-16 10:03:05/<저작권자 © 1980-2022 ○○일보. 무단 전재 재배포 금지.>
전문	○○○와 자신의 딸 결혼소식이 과거 주가조작 사건이 대중들로부터 언급되자 □□□도 반박에 나선 것이다. □□□는 이날 공개된 인터뷰에서 “오해의 소지가 있었던 건 인정하지만 사실 이 왜곡돼 있는 부분이 많다”며 “납편이 허위공시에 의한 부당한 이익을 취했 다는 것인데 그 돈은 개인명의로 쓸 수 없는 회삿돈이고 개인이나 가족에게 쓴 일이 없다”고 주장했다. 예비 사위 ○○○에 대해서는 “반듯하고 건실한 남성을 사윗감으로 맞다는 것만으로 너무나 고마운 일”이라며 “매우 용기있고 배려와 아량, 희망 에너지 가 넘친다. ○○○가 식구로 합류하면서 집안 분위기도 많이 밝아졌다”고 말했 다. <b>□□□ 법률대리인 입장문 전문</b> <b>□□□ 씨에 대한 허위 사실 관련 공식입장입니다.</b> <b>안녕하십니까. □□□ 씨와 소속사 ○○○의 법률대리인을 맡고 있는 법무법인</b> <b>○○입니다.</b> <b>먼저 □□□ 씨는 이와 같은 입장을 전할 수밖에 없게 되어 무척 죄송한 마음</b> <b>을 가지고 있다는 점을 말씀 드립니다.</b>

삭제 정보	예시
	<p>다만 <u>□□□ 씨와 가족들을 둘러싼 회복할 수 없을 만큼 확대 재생산되는 뉴스들이 더는 묵과할 수 없는 지경에 이르러, 이를 올바르게 바로 잡기 위해 입장을 전달하여 드립니다.</u></p> <p><u>다시 한 번 이와 같은 입장을 전하게 되어 □□□ 씨는 송구스러운 마음을 가지고 있는 점을 말씀 드리며, 허위 사실이 급속도로 무분별하게 유포되고 어느덧 기정사실화 되는 현 상황은, □□□씨 가족과 새롭게 가족이 되는 분들을 위해서라도 더는 묵과하지 않을 것이며, 엄중한 대응으로 사실과 다른 부분을 끝까지 바로 잡겠습니다.</u></p> <p><u>○○○ 기자</u></p>
문장으로 볼 수 없는 정보	<p>프로야구 LG 트윈스가 2023시즌 코칭스태프 구성을 마무리했다.</p> <p>LG는 4일 "코칭스태프 구성을 마쳤다"고 밝혔다.</p> <p>LG는 2022년 구단 창단 이래 한 시즌 역대 최다승인 87승(2무 55패)을 거뒀지만, 플레이오프(5전 3선승제)에서 키움 히어로즈에 가로막히며 최종 3위에 그쳤다. 그러자 지난해 11월 류지현 전 감독과의 동행을 끝내고 염경엽 신임 감독에게 지휘봉을 맡겼다.</p> <p>염 감독이 부임하며 코칭스태프에도 큰 변화가 생겼다. 이날 LG의 발표에 따르면 김정준 수석 코치와 김일경 수비코치, 박경완 배터리 코치, 배요한 컨디셔닝 코치가 새로 LG 유니폼을 입는다. 이 밖에도 퓨처스(2군)리그에 있던 이종범 감독과 김경태 코치, 스티브홍 코치도 각각 1군 주루·외야 코치, 1군 투수 코치, 1군 컨디셔닝 코치를 맡는다.</p> <p>한편 1군 코칭스태프 구성이 완료되며 퓨처스리그 감독 및 코치들도 다수 변경됐다. LG 2군 사령탑은 지난시즌 류지현 전 감독을 보좌했던 황병일 수석코치가 맡으며 김광삼 투수코치도 퓨처스리그에서 유망주 육성에 힘을 보탠다. 1군 배터리 코치를 역임했던 조인성 코치는 잔류군 코치로 이동한다.</p> <p><b>▼ 2023시즌 LG 트윈스 코칭스태프 구성</b></p> <p><b>▲ 1군</b></p> <p><u>감독 : 염경엽(신임)</u></p> <p><u>수석 : 김정준(신임)</u></p> <p><u>수석 트레이너 : 김용일</u></p> <p><u>타격 : 이호준, 모창민</u></p> <p><u>투수 : 경현호, 김경태(퓨처스 → 1군)</u></p> <p><u>수비 : 김일경(신임)</u></p> <p><u>작전 : 김민호</u></p> <p><u>주루/외야수비 : 이종범(퓨처스 감독 → 1군)</u></p> <p><u>배터리 : 박경완(신임)</u></p> <p><u>컨디셔닝 : 박종곤, 안영태, 이권엽, 고정환, 스티브홍(퓨처스 → 1군)</u></p> <p><b>▲ 퓨처스</b></p> <p><u>감독 : 황병일(1군 수석 → 퓨처스 감독)</u></p>



삭제 정보	예시
	<p><u>타격 : 임훈(잔류군 → 퓨처스)</u>  <u>투수: 김광삼(1군 → 퓨처스), 장진용</u>  <u>수비 : 윤진호</u>  <u>작전 : 박용근</u>  <u>주루/외야수비 : 양영동</u>  <u>배터리 : 윤요섭(잔류군 → 퓨처스)</u>  <u>컨디셔닝 : 배요한(신입), 김종욱, 유현원, 최재훈</u>  <u>▲ 잔류군</u>  <u>타격(잔류군 총괄/배터리) : 조인성(1군 배터리 → 잔류군)</u>  <u>투수 : 신재웅</u>  <u>수비 : 양원혁</u></p> <p>&lt;배드 시스터즈&gt;는 지난해 8월 공개 직후 현지 매체의 호평과 함께 시청자들의 폭발적인 반응을 얻었습니다. 높은 인기에 종영 직후인 지난 11월 시즌 2 제작이 확정됐고요. 지난 연말에는 워싱턴 포스트 등 다수의 영미권 매체가 ‘올해의 최고 시리즈’로 선정했습니다.</p> <p>여성의 고통을 깊이 이해하고, 자매(여성)들의 연대를 다룬다는 점에서 미국 드라마 &lt;와이 우먼 킬&gt; 시즌 1,2 나 &lt;빅 리틀 라이즈&gt;를 떠올리게 합니다. 다만 &lt;배드 시스터즈&gt;의 유머러스한 분위기는 &lt;빅 리틀 라이즈&gt;보다 &lt;와이 우먼 킬&gt;에 가깝습니다.</p> <p><b>‘풋’ 지수 ★★★★★ 타율 높은 아이리쉬 블랙 유머. ‘뽕’보단 ‘풋’ 터진다.</b>  <b>자매애 폭발 지수 ★★★★★ ‘아버지는 지킨다?’ 아니, 자매는 지킨다!</b></p> <p>올트먼은 정치에도 관심이 많다. 2014년 야후 최고경영자였던 머리사 메이와 함께 버락 오바마 대통령을 위한 모금 행사를 열었다. 2017년 무렵에는 캘리포니아 주지사로 출마하는 방안을 진지하게 고려했다고 윌리 브라운 전 샌프란시스코 시장이 공개한 적이 있다. 주지사 출마 뜻은 접었지만 올트먼은 주택·의료 정책을 바꿔야 한다는 정치 운동인 '유나이티드 슬레이트'를 주도했다.</p> <p>올트먼은 채식주의자다. 그의 옷장에는 티셔츠와 청바지가 가득하다고 한다. 일찌감치 고등학생 시절 커밍아웃한 동성애자다. 루프트를 공동 창업한 닉 시보와 9년간 사귀다가 헤어졌다.</p> <p>앞으로 올트먼이 넘어야 할 산은 여럿이다. 챗GPT가 인터넷상 정보를 가져오는 행위가 절도라는 저작권 논란을 해결해야 한다. 대화형 AI 산업의 수익 모델이 생각보다 신통치 않다는 반응도 있다. AI가 인류를 위협할 것이라는 근원적 두려움도 장애물이다.</p> <p>올트먼이 과대평가된 인물이라는 냉소적 시각도 있다. 오랜 에치오니 전 앨런 AI연구소장은 "오픈AI가 마이크로소프트와 파트너십을 늘려가는 걸 보면 마이크로소프트를 넘어서거나 거대한 독립 기업이 되려는 신호가 아닌 것 같다"며 "올트먼은 빌 게이츠를 넘어설 수 없으며, 단순한 억만장자에 그칠 것"이라</p>

삭제 정보	예시
	<p>고 했다.</p> <p><u>샘 올트먼은</u></p> <p><u>1985년 시카고 태생</u></p> <p><u>2005년 스탠퍼드대 컴퓨터과학과 중퇴</u></p> <p><u>2005년 스타트업 루프트 공동 창업</u></p> <p><u>2012년 루프트 4340만달러에 매각</u></p> <p><u>2014년 벤처캐피털 Y콤비네이터 대표</u></p> <p><u>2015년 일론 머스크와 오픈AI 창업</u></p> <p><u>2022년 오픈AI, 챗GPT 출시</u></p>
기사와 관련 없는 내용	<p>지난 3월 1일 세종시의 한 아파트에는 일장기가 나부꼈다. 정치인들은 “한국과 일본이 피해자와 가해자라는 사실관계는 결코 바뀌지 않는다”고 입버릇처럼 말한다. 그렇지 않다. 증거와 증언이 사라지고 사람들의 기억에서 잊혀지면 역사도 바뀐다. 3·1절에 한국인이 일장기를 게양할 수 있다는 것을 상상한 사람이 몇이나 될까. 이는 ‘표현의 자유’가 아닌, ‘한국의 역사교육이 무엇인가 잘못된 것 아닌가’를 상징적으로 보여준다.</p> <p>“일제가 한반도를 수탈하고, 당시 조선인들을 강제동원한 증거가 있다. 왜 지금 일본에게 사죄하라고 하나”고 따지는 상황은 먼 미래의 일이 아니다. 백마디 말보다 확실한 것은 당시의 참혹함이 남은 증거를 보여주는 것이다. 인천시 부평구 산곡동 449는 그 증거가 될 뻔 했다. “진짜 있었다. 그런데 2023년에 아무도 지시를 안했는데 그게 없어졌더라”는 변명이 아닌 ‘있는 그대로’를 보여줄 기회가 우리에게 틀림없이 있었다.</p> <p><u>*사람을 찾습니다.</u></p> <p><u>1946~1948년 사이 인천 부평구 에스컴 시티(캠프마켓) 내에 있던 미군 382 위수병원에서 간호장교 교육을 받은 분들의 연락을 기다립니다. 특히, 아래 첫 번째 사진은 1948년 7월 2일 인천 캠프마켓 내에 있던 382 위수병원 학위과정을 졸업한 8명의 간호사들입니다. 앞줄 왼쪽부터 엄금례, 신영숙, 김선(순)태, 정정(청)화, 이충실, 김감음, 뒷줄 왼쪽부터 이운(은)산, 이해자. 아래 두 번째 사진 속 인물이 본인이거나 가족이신 분들의 연락도 기다립니다.</u></p> <p><u>아울러 건축가 김중업, 화가 이중섭의 부평 미군기지에서 활동 내용을 아시는 분들의 연락도 함께 기다립니다.</u></p>
문장의 오류	<p>을 1월13일까지 매주 금요일 2회씩 공개됐다. <del>됐다.</del></p>

<표 11> 불필요한 요소 제거 내용

작업 파일 데이터	정제 데이터
<p><u>(○○=○○○) ○○○ 기자 =</u> 경기도해양수산물연구소는 3월17일까지 ‘2023년 경기도 귀어학교’ 1기 교육생을 모집한다고 밝혔다.</p> <p>경기도 귀어학교는 귀어를 희망하거나 어촌에 살지만, 어업에 종사하지 않는 주민을 대상으로 어촌생활에 필요한 실습·실무 위주의 교육을 하는 기관이다. 신청 대상은 만 18세~65세 귀어 희망자 혹은 어촌에 거주하는 비어업인으로 서류와 면접심사를 거쳐 총 16명을 선발한다.</p> <p>희망자는 필요한 서류를 작성한 뒤 도 해양수산물연구소에 방문하거나 우편, 전자우편 또는 팩스로 제출하면 된다. 자세한 정보는 도 누리집 또는 도 해양수산물연구소 누리집에서 확인할 수 있다.</p> <p>교육생은 오는 4월 3일부터 4월 28일까지 4주간 숙식을 제공받으며 귀어정책, 어업·양식 기술교육 등 수산업 창업 및 어촌생활에 필요한 이론과 현장실습 교육을 받게 된다. <u>교육 수료자에게 해양레저 관련자격 취득 과정의 교육수수료 일부를 지원받을 수 있는 혜택도 제공된(사진은 기사 내용과 무관함) /</u></p>	<p>경기도해양수산물연구소는 3월17일까지 ‘2023년 경기도 귀어학교’ 1기 교육생을 모집한다고 밝혔다.</p> <p>경기도 귀어학교는 귀어를 희망하거나 어촌에 살지만, 어업에 종사하지 않는 주민을 대상으로 어촌생활에 필요한 실습·실무 위주의 교육을 하는 기관이다. 신청 대상은 만 18세~65세 귀어 희망자 혹은 어촌에 거주하는 비어업인으로 서류와 면접심사를 거쳐 총 16명을 선발한다.</p> <p>희망자는 필요한 서류를 작성한 뒤 도 해양수산물연구소에 방문하거나 우편, 전자우편 또는 팩스로 제출하면 된다. 자세한 정보는 도 누리집 또는 도 해양수산물연구소 누리집에서 확인할 수 있다.</p> <p>교육생은 오는 4월 3일부터 4월 28일까지 4주간 숙식을 제공받으며 귀어정책, 어업·양식 기술교육 등 수산업 창업 및 어촌생활에 필요한 이론과 현장실습 교육을 받게 된다.</p>

<표 12> 원시 데이터와 정제된 데이터 비교 1

데이터 정제 전	정제 데이터
<p><u>[○○○○=○○○ 기자]</u> 유튜버 아미가 지칠 줄 모르는 강력한 먹방으로 '토밥'을 뒤집어 놓는다.</p> <p>오는 11일(토) 방송되는 티캐스트 E채널 '토요일은 밥이 좋아(연출 이영식)'에서는 지난주에 이어 히밥과 함께 강남구 맛집을 찾아 여행을 떠난 아미의 모습이 그려질 예정이다.</p> <p>(중략)</p> <p>이어 "먹방 하기 전에는 라면 5봉지 그 정도 밖에 못 먹었다. 그런데 방송하다 보니까 점점 늘어나더라. 그래서 결국 17봉지까지 늘어나더라. 요즘은 조금 더 준 것 같다"라며 히밥도 놀랄 대식가 면모를 뽐낸다.</p> <p>감탄사를 연발하던 히밥은 "먹으면 배가 부르긴 불러?"라고 재차 질문했고 아미는 "적당히 먹고 끝낸다. 배 터질 때까지 먹지는 않는다"라며 도무지 믿을 수 없는 이야기를 꺼내 놀라움을 자아낸다.</p> <p>마지막 식사를 모두 마친 아미와 히밥은X 세대팀 김숙,현주엽과 합류해 돼지,소,오리 즉석 대패 고기3종으로 대미를 장식하기 위해 이동한다. (끝)기 전에는 라면 5봉지 그 정도 밖에 못 먹었다. 그런데 방송하다 보니까 점점 늘어나더라. 그래서 결국 17봉지까지 늘어나더라. 요즘은 조금 더 준 것 같다"라며 히밥도 놀랄 대식가 면모를 뽐낸다.</p> <p><u>감탄사를 연발하던 히밥은 "먹으면 배가 부르긴 불러?"라고 재차 질문했고 아미는 "적당히 먹고 끝낸다. 배 터질 때까지 먹지는 않는다"라며 도무지 믿을 수 없는 이야기를 꺼내 놀라움을 자아낸다.</u></p>	<p>유튜버 아미가 지칠 줄 모르는 강력한 먹방으로 '토밥'을 뒤집어 놓는다.</p> <p>오는 11일(토) 방송되는 티캐스트 E채널 '토요일은 밥이 좋아(연출 이영식)'에서는 지난주에 이어 히밥과 함께 강남구 맛집을 찾아 여행을 떠난 아미의 모습이 그려질 예정이다.</p> <p>(중략)</p> <p>이어 "먹방 하기 전에는 라면 5봉지 그 정도 밖에 못 먹었다. 그런데 방송하다 보니까 점점 늘어나더라. 그래서 결국 17봉지까지 늘어나더라. 요즘은 조금 더 준 것 같다"라며 히밥도 놀랄 대식가 면모를 뽐낸다.</p> <p>감탄사를 연발하던 히밥은 "먹으면 배가 부르긴 불러?"라고 재차 질문했고 아미는 "적당히 먹고 끝낸다. 배 터질 때까지 먹지는 않는다"라며 도무지 믿을 수 없는 이야기를 꺼내 놀라움을 자아낸다.</p> <p>마지막 식사를 모두 마친 아미와 히밥은X 세대팀 김숙,현주엽과 합류해 돼지,소,오리 즉석 대패 고기3종으로 대미를 장식하기 위해 이동한다. (끝)기 전에는 라면 5봉지 그 정도 밖에 못 먹었다. 그런데 방송하다 보니까 점점 늘어나더라. 그래서 결국 17봉지까지 늘어나더라. 요즘은 조금 더 준 것 같다"라며 히밥도 놀랄 대식가 면모를 뽐낸다.</p>

<p>마지막 식사를 모두 마친 아미와 히밥은 X 세대팀 김숙, 현주엽과 합류해 돼지, 소, 오리 즉석 대패 고기 3종으로 대미를 장식 하기 위해 이동한다</p> <p>/○○○○○○○○○○@○○○○.co.kr</p> <p>[사진] 제공</p> <p>]]&gt;&gt;[○○○○=○○○ 기자] 유튜버 아미가 지칠 줄 모르는 강력한 먹방으로'토밥'을 뒤집어 놓는다.</p> <p>오는 11일(토) 방송되는 티캐스트 E채널 '토요일은 밥이 좋아(연출 이영식)'에서는 지난주에 이어 히밥과 함께 강남구 맛집을 찾아 여행을 떠난 아미의 모습이 그려질 예정이다.</p> <p>(내용 반복 중략..)</p> <p>마지막 식사를 모두 마친 아미와 히밥은 X 세대팀 김숙, 현주엽과 합류해 돼지, 소, 오리 즉석 대패 고기 3종으로 대미를 장식 하기 위해 이동한다</p> <p>/○○○○○○○○○○@○○○○.co.kr</p> <p>[사진] 제공</p> <p>]]&gt;</p>	
--	--

<표 13> 원시 데이터와 정제된 데이터 비교 2

원 저작자(기사 또는 매체)의 의도가 훼손되지 않은 선에서 최소한의 수정 진행

: 명백한 오류는 수정

원시 데이터

발의했다.</p><p>국회 국토교통위원회 위원인 더불어민주당 장철민 의원(대전 동구)은 계획도시 등 원도 지원에 관한 특별법안(이하 '특별법')을 대표 발의했다고 밝혔다.</p><p>장 의원이 대표 발의한 특별법은 건축, 교육 등에 관한 특례를 부여하고 사업시행자 등에 조세 기본계획 수립과 활성화 비율 감소한 지역 ▲택지 권역을 주요 골자로 하고 있다. 세부적으로 '노후'에 더해 ▲최근 30년간 인구 밀집도가 높고 ▲노후도시 내 ▲택지 지역과 동일한 생활권을 구성하는 연접 노후도시 지역 ▲역세권개발 지역 등 정비 대상의 범위가 더욱 넓어졌다.

정지

입력 2023.03.29 13:14

댓글 0

키워드

#장철민

#노후신도시정비

#대표발의

#지역불균형

#포괄적 도시재생

업 완료 후 2년 이상 경과한 100만㎡이상의 택지 지역에 대해 ▲최근 30년간 인구가 일정 비율 감소한 지역 ▲택지 지역과 동일한 생활권을 구성하는 연접 노후도시 지역 ▲역세권개발 지역 등 정비 대상의 범위가 더욱 넓어졌다.

이번 노후도시 특별법에는 유일하게 교육 및 보육에 관한 특례도 추가됐다. 노후도시 내 교육 경비 등 교육 여건은 물론 교육과정 운영과 교육시설 지원 등을 담았으며 보육기반시설 확충이나 노후·유치시설을 정비하는 조문도 신설됐다.

실제 웹 화면

## 원 저작자(가) 또는 매체)의 의도가 훼손되지 않은 선에서 최소한의 수정 진행

## : 오타 수정

김태환 총독에게도 법학전문대학원에서는 '저작권법'에 제재를 받아야 할 사법서비스 이용자인 당사자와 관계인이 가장법인이 아닌 지방법원에 '재판을 받을 수밖에' 있다는 것은 의료서비스 이용자인 환자나 내과 진료가 필요한 데도 내과 전문의의 진료를 받지 못하고 일반과에서 진료를 받을 수밖에 있는 것과 마찬가지로 상황'이라며 '사법 접근성 확대를 위해 가장법인의 진'이라고, 성직 중

이두영 구강발달지장애분과 총독은 '재판을 받을 수밖에' >>> '재판을 받을 수밖에' 사법

지 못하고 있고, 일반법원에서 계속 사건을 담당하게 되면 사건 처리의 신속성과 전문성이 결여돼 지역민들의 어려움이 누적될 것이다'며 '국민의 통통한 재판관과 도민들의 신속·공정한 재판받을 권리를 보장해야 한다'고 주장했다.

한영숙 청주YWCA 여성통합상담소장은 '청주가정법원 설치'는 도민의 전문적 사법서비스 접근을 위한 기반과 보장

2 지역을 넘어 총독에게로 우회된다

4 코로나 재확산에도 정부 노력





### 뉴스웍스

성장을 꺾인 배달 시장...배달앱 3사 '무료배달' 카드까지 꺼냈다

도한 지난 4월 오픈서비스의 '배달 서비스 트렌드 리포트'에 따르면, 전국 만 20~59세 남녀 2000명 설문조사에서 '배달 서비스 이용이 줄어들었다'고 답한 응답자가 28.8%에 달했다. 이들은 '비싼 배달비(83.9%)', '배달 음식 가격이 비싸져서(56.9%)', '외식비 자체를 줄이려고(54.4%)' 등의 이유로 배달 서비스 이용을 줄였다고 응답했다.

통계청의 '2022년 온라인쇼핑 동향' 자료에서도 배달 시장의 성장세가 꺾이고 있음을 확인할 수 있다. 코로나 사태가 극심했던 2020~2021년에 배달 음식 온라인 거래액 증가율은 각각 78.1%, 48.1% 비중을 보였으나, 지난해는 1.4%로 성장세가 사실상 멈춰섰다.

업계 한 관계자는 "배달앱 3사 모두 올해 1분기 실적이 기대 이하여서 충격적이라는 말이 나올 정도"라며 "그동안 배달앱 서비스가 외식사업업자를 대상으로 폭리를 취한다는 부정적 인식이 팽배했던 만큼 이를 분식시킬 수 있는 새로운 서비스와 수익 모델을 제시해야 할 것"이라고 진단했다.

했다,

했다,

했다,

다자탈퇴임스

### 1분기 경제성장률 0.3%...민진소비

한국 경제가 두 분기 연속 역성장을 피했다. 코로나19 사회적 거리두기 종료 이후 민간 소비가 증가한 데 따른 것이다.

증거한 데 따른 것이다,

증가한 데 따른 것이다.



원 저작자(기자 또는 매체)의 의도가 훼손되지 않은 선에서 최소한의 수정 진행

: 반복되는 단어 수정

최강볼펜 연예 야구 축구 스포츠종합 라이프 만화 갤러리

인드레프트에서 2차 3라운드는 29라운드 역전(한 기금)의 유너움을 담은 김아성은 2020시즌을 마치고 미국 진출을 선언했다.

포스팅시스템으로 샌디에이고와 계약한 김하성은 김하성은 2021년~2022년 타율 0.235, 출루율 0.306, 장타율 0.372에 그치는 등 적응 과정을 거쳤다. MLB닷컴 역시 "KBO리그에서는 확실한 주전이었던 김하성은 빅리그에서는 벤치에서 경기를 지켜볼 때가 많았다"고 설명했다.

지상파 방송에서 '슈퍼푸드'라고 소개한 식품이 첫가루 법적인 것으로 드러난 것과 관련해 해당 식품을 유통·판매한 업자가 집행유예를 받고 석방한다.

제주지방법원 형사1단독(오지에 판사)는 14일 식품위생법 위반 등의 혐의로 구속기소된 A씨(63)에게 징역 6월에 집행유예 3년을 선고했다.

가능해요

자비스에 따르면 이번 공급 계약은 입찰 전 한 달여 기간 동안의 기술검증 평가를 통과해 수주했다. 이에 글로벌 2차전지 시장에서 자비스의 X-ray 검사장비 기술력을 인정 받았다고 평가했다. 또 배터리 국간 검사와 '이물 검사 알고리즘'의 기술력과 노하우가 이번 수주를 견인했다고 강조했다.

정 의원은 "이 대표 방탄을 위해 방탄국회 소집에 이어 방탄 장외 집회, 방탄 탄핵, 검수완박(검찰 수사권 완전탈) 시즌2까지 속셈이 여실이 드러난다"며 "민주당이 수단과 방법을 안 가리는 입법 독주를 계속하면 국민의 엄중한 심판을 면치 못할 것"이라고 경고했다.

<그림 11> 기사 수정 예 ③

기사 내 같은 내용 반복

▶ 삭제

제목이 기사 마지막에 반복

▶ 해당 부분 삭제

**기사 내 같은 내용 반복**

이 부장은 "기업의 이익 전망이 하향 조정됐음에도 실제로 상반기 성과는 매우 긍정적이었는데, 올해 지수 상승세가 개별 기업의 이익 요인보다는 밸류에이션의 요인이 컸다"며 "올해 S&P 지수에 기여한 상위 10대 종목에 비해 나머지 490개 종목은 상대적으로 저렴한 밸류에이션을 보이고 있다"고 말했다.

이어 "연초에 비해 밸류에이션이 비싸 보일 수는 있으나, 사실 올해 성장세에 대부분 이끌었던 소수의 종목을 제외한 나머지의 종목 자원에서 보았을 때 밸류에이션이 여전히 적절하거나 상대적으로 여전히 매력적이다"고 강조했다.

그러면서 "거시 경제적 환경에 상대적으로 영향을 덜 받는 주식이나 기업들을 선호하고 있고, 그리고 펀더멘털 부분에서 보았을 때는 이 우량성과 성장성이 동반된 주식, 그리고 상대적으로 밸류에이션이 적정하거나 매력적인 이런 주식들을 선별하여 투자하는 것이 매우 중요하다"고 강조했다.

**삭제**

이 부장은 ""전체 지수로 보면 밸류에이션이 비싸 보이지만 소수의 종목이 미국 증시 주식을 대부분을 기여한 것으로 고려하면 이들을 제외한 종목들의 밸류에이션은 적정 혹은 매력적인 상황"이라고 말했다.

이 부장은 "기업의 이익 전망이 하향 조정됐음에도 실제로 상반기 성과는 매우 긍정적이었는데, 올해 지수 상승세가 개별 기업의 이익 요인보다는 밸류에이션의 요인이 컸다"며 "올해 S&P 지수에 기여한 상위 10대 종목에 비해 나머지 490개 종목은 상대적으로 저렴한 밸류에이션을 보인다"고 말했다.

이어 "연초에 비해 밸류에이션이 비싸 보일 수는 있으나, 사실 올해 성장세에 대부분 이끌었던 소수의 종목을 제외한 나머지의 종목 자원에서 보았을 때 밸류에이션이 여전히 적절하거나 상대적으로 여전히 매력적이다"고 강조했다.

그러면서 "거시 경제적 환경에 상대적으로 영향을 덜 받는 주식이나 기업들을 선호하고 있고, 그리고 펀더멘털 부분에서 보았을 때는 이 우량성과 성장성이 동반된 주식, 그리고 상대적으로 밸류에이션이 적정하거나 매력적인 이런 주식들을 선별하여 투자하는 것이 매우 중요하다"고 강조했다.

이 부장은 ""전체 지수로 보면 밸류에이션이 비싸 보이지만 소수의 종목이 미국 증시 주식을 대부분을 기여한 것으로 고려하면 이들을 제외한 종목들의 밸류에이션은 적정 혹은 매력적인 상황"이라고 말했다.

**제목이 기사 마지막에 반복**

**인순이, 찢어진 청바지에 워커 신고 아이돌 변신...모니카도 감탄**

'골든걸스' 17일 방송

김민지 기자

2023.11.17 오후 1:42

**제목**

한편 무대를 본 이은미가 춤에 대한 욕심을 드러냈다고 해 궁금증이 높아진다. 이은미는 인순이, 신효범의 시원한 무대에 "흥분을 감출 수가 없었다, 배워보고 싶다"라고 밝히, '골든걸스' 합류 전까지 댄스를 극구 거부해온 '댄스 베이비' 이은미의 의욕을 불태운 인순이, 신효범의 '터치 마이 바디'는 어떨지 본 방송에 관심이 집중된다.

한편 '골든걸스'는 매주 금요일 오후 10시 방송된다. **인순이, 찢어진 청바지에 워커 신고 아이돌 변신...모니카도 감탄**

breeze52@news1.kr

**본문**

<그림 12> 기사 수정 예 ④



```

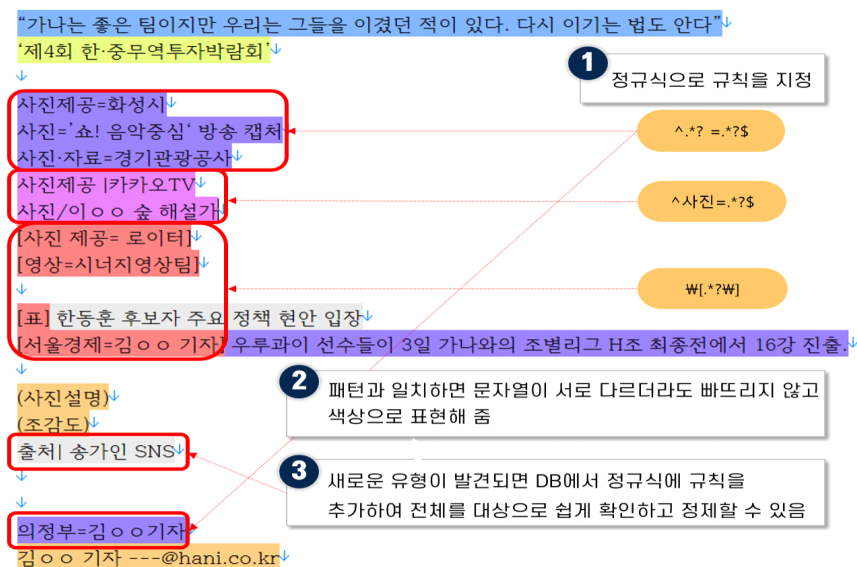
<url>https://news.kmib.co.kr/article/view.asp?arcid=0924276972&amp;code=11121100</url>
<used>Y</used>
<byline>박○○○</byline>
<content>
<d>사진=이○○ 기자</d>
한동훈<d>{사진}</d> 법무부 장관이 직접 자신을 향한 '국민의힘 당대표 차출설' 진화에 나섰다. 그러나 친윤(친윤석열)계 핵심 장제원 의원과 정진석 비상대책위원장이 말씨름을 벌이는 등 '한동훈 차출설'의 여진은 계속됐다.
한 장관은 7일 자신을 둘러싼 국민의힘 당대표 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 상관없는 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 13> 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.

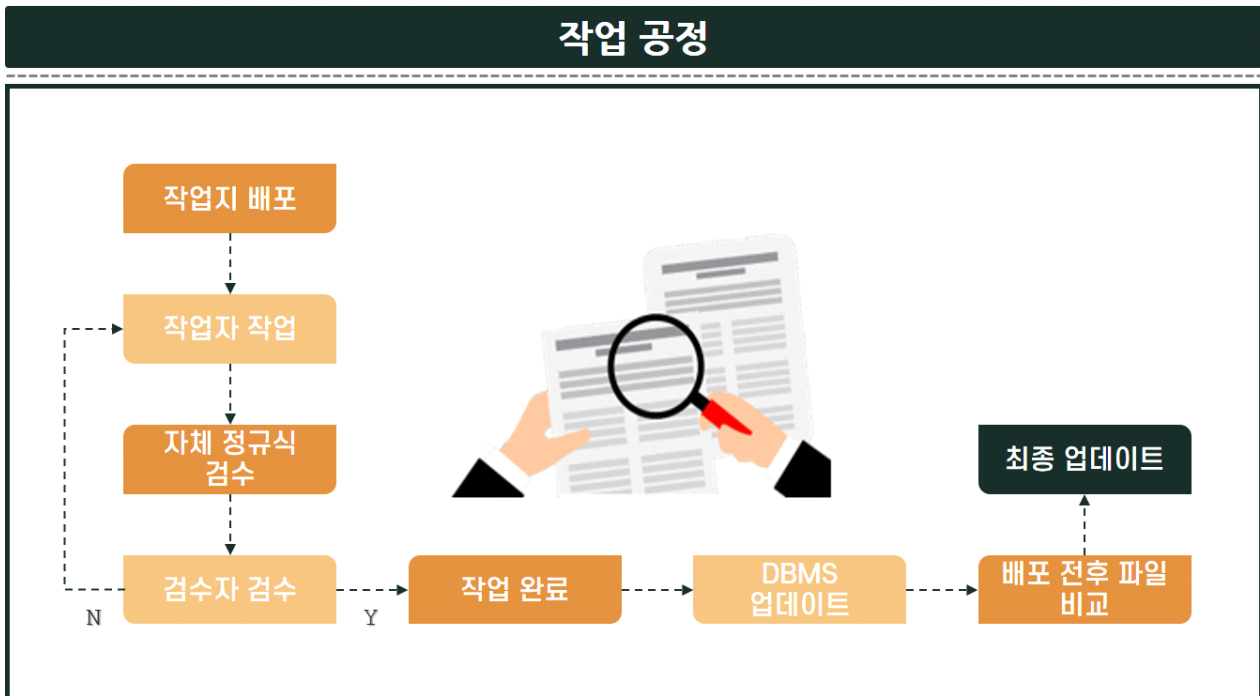
새로운 유형이 나오는 경우 작업자들이 구글 시트를 활용하여 해당 유형을 공유, 축적하였고 불필요한 요소 제거 작업을 마친 데이터는 소실 비교를 통해 문자 데이터의 누락 등을 검수하였다.



<그림 14> 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 15> 데이터 작업 및 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, 비문 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.

작업 전후의 글자만을 비교하여 소실된 데이터가 있는지 확인하는 과정을 거쳐 데이터의 소실을 예방할 수 있었다.

## 5. 메타데이터 작성

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

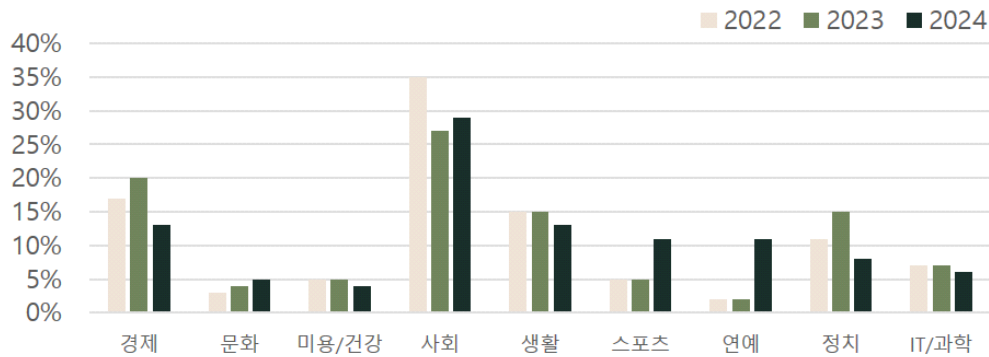
기사의 주제 분류는 수행사가 가지고 있는 인공 지능 모델을 신문 기사 학습에 최적화시켜 진행하였으며 기존에 공개된 신문 기사와 주제를 활용하여 삽입하였다.

기자 정보의 경우에는 일부 매체는 메타데이터에 삽입되어 있지 않고 본문 중에 해당 정보가 기입된 경우도 있었다. 이러한 경우에는 본문에서 기자 정보를 확인하여 메타데이터에 삽입하였다.

사회	경제	생활	정치	IT/과학	미용/건강	스포츠	문화	연예
29%	13%	13%	8%	6%	4%	11%	5%	11%

<표 14> 2024년 신문 기사 주제별 통계

## 최근 3년간 주제별 구축 비율



2022년 주제별 통계			2023년 주제별 통계			2024년 주제별 통계		
경제	169,640	17%	경제	200,739	20%	경제	156,889	13%
문화	32,530	3%	문화	36,208	4%	문화	57,253	5%
미용/건강	48,806	5%	미용/건강	51,050	5%	미용/건강	45,427	4%
사회	340,222	35%	사회	274,416	27%	사회	344,290	29%
생활	150,109	15%	생활	156,393	15%	생활	161,058	13%
스포츠	45,317	5%	스포츠	54,662	5%	스포츠	131,581	11%
연예	21,021	2%	연예	21,120	2%	연예	125,102	11%
정치	103,927	11%	정치	154,183	15%	정치	102,431	8%
IT/과학	66,772	7%	IT/과학	74,660	7%	IT/과학	75,722	6%

<그림 16> 연도별 주제별 통계

각 주제별로 구축량의 차이가 최대 25%p 이상 나지 않도록 하라는 국립국어원 제안 요청서의 내용(이 내용은 2023년부터 계속 적용되어 옴)에 따라, 이 사업에서도 주제별 구축량에서 최대와 최소의 차이는 25%p 이내가 되도록 구축하였다.

## 6. 인용 부호 수정 말뭉치

2차 정제가 끝난 데이터는 ‘신문 말뭉치’ 1종으로 구축된다. 이 사업 제안 요청에 포함된 ‘인용 부호 수정 말뭉치’는 신문 말뭉치에서 인용 부호를 수정한 말뭉치를 의미하고 이 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 경우
- 인용 부호의 순서가 다르거나, 표준 인용 부호 대신에 다른 부호가 사용된 경우

아래 예시는 부호가 짝이 맞지 않은 경우이다. 이와 같이 짝이 맞지 않은 사례가 많이 발견되었으며, 예를 들어 큰따옴표로 시작하여 작은따옴표로 끝나는 경우, 큰따옴표로 시작했으나 닫히지 않은 경우, 작은따옴표로 시작하여 닫히지 않은 경우, 또는 닫는 부호만 있는 경우 등이 있었다. 수정은 인용 부호에 한정하였으며, 영어의 아포스트로피(Apostrophe)는 원래 형태를 유지했다. 또한, 다른 비표준 기호들은 표준 기호로 수정하여 통일성을 확보하였다.

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
이종민 백학면장은 __이웃을 향한 학생들의 따뜻한 마음에 감동을 받았다.__며 미래를 향한 우리 학생들의 꿈을 응원한다__고 말했다.	이종민 백학면장은 __이웃을 향한 학생들의 따뜻한 마음에 감동을 받았다.__며 __미래를 향한 우리 학생들의 꿈을 응원한다__고 말했다.
김영현 한국코치협회 회장은 __이번 코칭컨퍼스티벌을 계기로 더더욱 코치들의 결속 강화를 통해 함께 성장하고 우리의 코칭이 대한민국에 긍정적인 영향을 끼칠 수 있도록 최선을 다하겠다.__고 말했다.	김영현 한국코치협회 회장은 __이번 코칭컨퍼스티벌을 계기로 더더욱 코치들의 결속 강화를 통해 함께 성장하고 우리의 코칭이 대한민국에 긍정적인 영향을 끼칠 수 있도록 최선을 다하겠다.__고 말했다.
이른바 __게이라고 말하지 마__(Don_t say gay) 법을 제정한 것이 발단이 됐다.	이른바 __게이라고 말하지 마__(Don_t say gay) 법을 제정한 것이 발단이 됐다.

<표 16> 인용 부호 수정 데이터 정제 전후

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

인공 지능 학습 및 데이터 유통에서 통일되지 않은 한자 코드는 검색 누락과 같은 기술적 문제나 데이터 관리상의 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 17> ‘한·중·일 호환용 한자 영역’ 한자 치환 표

## 다. 문장 부호 등 통일

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

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 18> 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
꺾	“아지 <del>꺾</del> 지 구체적인 시행지침이 없어 증권사 입장에서는 혼란스러운 상황”이라고 설명했다.
꺾	신승훈이라는 교체 멤버 활용폭을 이날 만큼은 좁게 가져 <del>꺾</del> 다.
꺾	감리분야에서는 감리자 시공도서 <del>꺾</del> 토 여부 등 감리업무 수행실태(자재 승인, 점수·검측 관련 서류)의 적정 여부도 확인한다.
꺾	21일 기준 ‘정이’는 넷플릭스 영화 부문 세 <del>꺾</del> 1위를 차지했다.
넌	데이트 코스 돌아다니는거 그 <del>꺾</del> 사부작 사부작 해보려고 했느 <del>넌</del> 공교롭게도 5월 5일 어린이날에 비가 내려서
꺾	“이 <del>꺾</del> 게 불공정한 상황에서 경기를 치른 적이 없다. 정말 미친 판정이었다”며 판정에 대한 강한 불만을 표출했다.
꺾	생전 처음 보는 박위라는 사람을 딱 <del>꺾</del> 는데 진짜 너무 웃기게 호감의 문이 확 열려버렸다.
꺾	역대 LCK에서 왕조로 <del>꺾</del> 린 팀은 T1과 디플러스 기아.
꺾	아쉽게도 참여하지 못하는 고객들을 위해 서 <del>꺾</del> 숲 매장을 중심으로 구성된 ‘피크닉 서비스’를 컨셉화한 다채로운 포토 스팟을 마련했고
꺾	스타쉽엔터테인먼트의 연기자 레이블인 킹콩 by스타쉽과 전속계약 <del>꺾</del> 맺고 본격적으로 배우 활동을 시작한다.

<표 19> 오타 글자

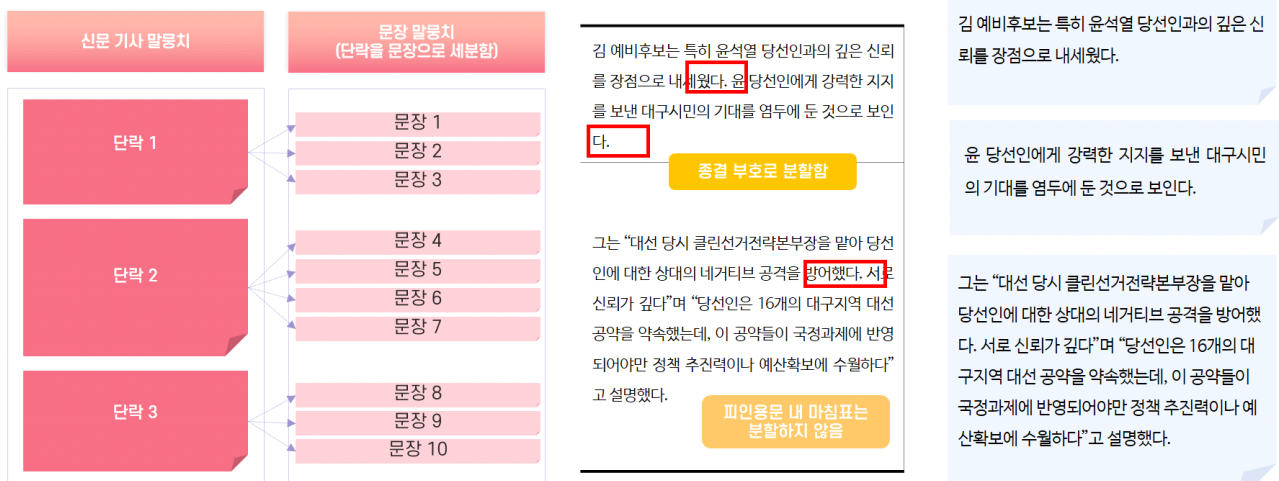
## 7. 문장 말뭉치 구축

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 학습의 속도, 정확도, 문장 구조 파악 등 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.

⊙인공지능 학습에 반드시 필요한 문장 분할 등에 활용할 수 있음



<그림 17> 문장 말뭉치 개념

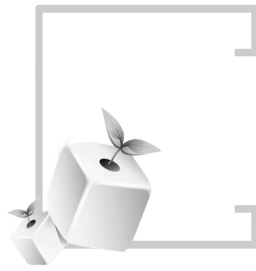
인용 부호 수정 말뭉치	문장 말뭉치
<p>&lt;p&gt;김 예비후보는 특히 윤석열 당선인과의 깊은 신뢰를 장점으로 내세웠다. 윤 당선인에게 강력한 지지를 보낸 대구시민의 기대를 염두에 둔 것으로 보인다.&lt;/p&gt;</p>	<p>&lt;p&gt;&lt;s&gt;김 예비후보는 특히 윤석열 당선인과의 깊은 신뢰를 장점으로 내세웠다.&lt;/s&gt; &lt;s&gt;윤 당선인에게 강력한 지지를 보낸 대구시민의 기대를 염두에 둔 것으로 보인다.&lt;/s&gt;&lt;/p&gt;</p>
<p>&lt;p&gt;그는 “대선 당시 클린선거전략본부장을 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 20> 문장 말뭉치 데이터 정제



<그림 18> 매체별 평균 문장 분할 수

<그림 17>은 한 문단(매체가 나눈 단락) 내의 문장 분할 수 평균을 기준으로 매체별 상위 5위와 하위 5위를 제시한 것이다. 조선일보의 경우 문장 분할 수 평균이 한 단락당 3.07개로 가장 높았으며, 충청매일의 경우 1.09개로 가장 낮았다. 평균 문장 분할 수는 1.5개이다.



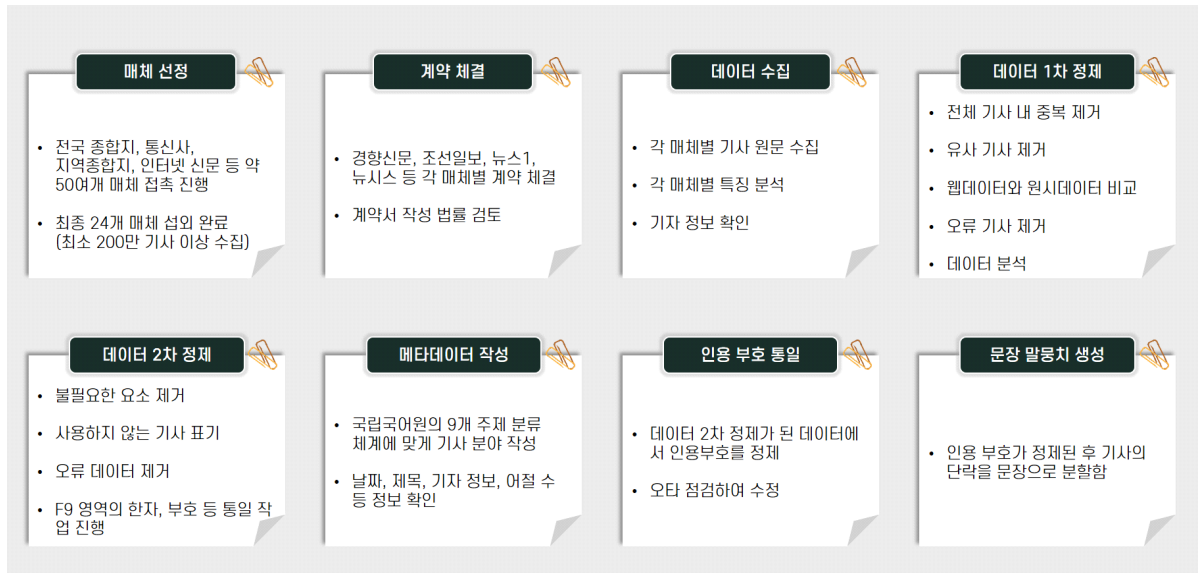
## 제 3 장

# 사업 수행 결과



## 제3장 사업 수행 결과

### 1. 신문 기사 정제 결과



<그림 19> 구축 공정별 내용

신문 기사의 정제 결과는 다음과 같다.

24개의 매체를 통해 다양한 데이터 구성을 확보했다. 참여 매체는 통신사, 종합지, 전문지, 지역 종합지, 인터넷 매체 등으로 구성되어, 여러 관점과 주제를 포함할 수 있도록 했다.

올해 총 구축 어절 수는 265,724,419어절(본문+제목)로, 당초 목표였던 2억 어절을 크게 초과하여 목표 대비 133%를 달성했다. 이는 월별 목표 어절 수인 1,600만 어절을 넘어 매월 평균 2,000만 어절 이상의 데이터를 구축함으로써 이루어졌다. 총 구축 기사 수는 1,199,753건으로, 계약을 초과하는, 방대한 데이터를 수집하고 정제했다.

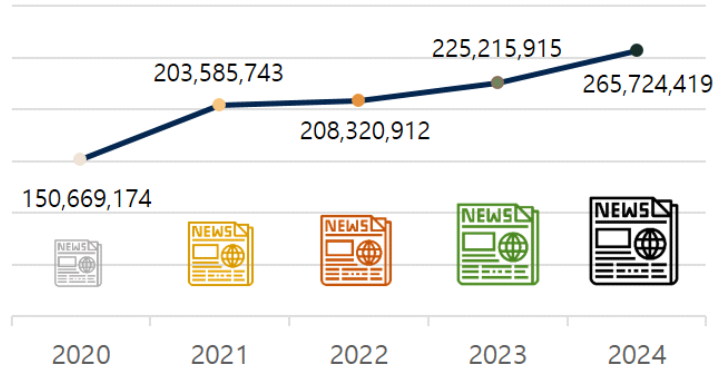
올해 구축한 말뭉치는 3종(신문 기사 말뭉치, 인용 부호 수정 말뭉치, 문장 말뭉치)으로 구분되며, 해당 말뭉치는 저작권 문제를 면밀히 검토하여 저작권이 확보된 기사만을 사용했다.

올해 총 구축 어절 수는 전년도 대비 증가했으며, 작년 2억 2천만 어절에서 올해 2억 6천만 어절 이상으로 확대되었다. 이를 통해 더 방대한 고품질의 데이터 확보가 가능해졌고, 구축 데이터의 양적·질적 성장을 동시에 이뤄 냈다.

매체명	최초 수집 기사 수	최초 수집 어절 수	정제 수집 기사 수	정제 수집 어절 수
경남매일	21,627	3,309,024	10,468	1,875,264
경인매일	42,368	7,207,076	29,314	5,050,012
경향신문	74,116	22,020,899	47,998	12,732,706
굿모닝충청	17,747	4,301,532	9,290	2,167,788
뉴스1	353,845	69,606,852	234,475	50,063,645
뉴스웍스	27,415	6,916,277	20,292	4,924,016
뉴스투데이	24,267	7,936,536	15,498	3,887,149
뉴시스	380,773	81,220,617	273,076	57,642,617
대경일보	36,534	5,953,026	18,542	3,418,126
대전투데이	21,023	3,872,669	14,044	2,658,894
디지털타임스	70,197	16,532,859	47,725	11,645,561
마이데일리	119,657	19,748,979	53,502	12,400,284
매일경제	175,938	43,174,410	88,540	21,991,599
스포츠경향	76,202	18,190,894	48,049	10,754,240
스포츠투데이	57,464	8,223,681	23,071	4,661,110
아이뉴스24	80,869	16,346,113	43,829	9,779,533
오마이뉴스	16,909	5,332,432	8,174	2,379,433
오에스이엔	227,781	34,709,111	80,586	21,124,520
울산제일일보	17,711	2,816,252	8,649	1,725,790
일간투데이	49,148	9,322,305	35,392	6,835,315
전남매일	24,795	4,485,644	13,491	2,595,703
조선일보	46,670	11,475,428	19,146	5,770,813
충남일보	48,170	7,548,770	28,072	4,867,065
충청매일	58,525	7,759,336	28,530	4,773,236
총 합	2,069,751	418,010,722	1,199,753	265,724,419

<표 21> 신문 기사 정제 총괄표

## 최근 5년, 신문 기사 구축 어절 수

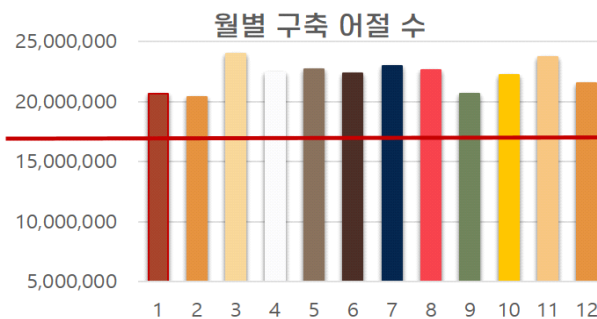


구분	2020년	2021년	2022년	2023년	2024
매체 수	35	35	34	28	24
최종 구축 어절	150,669,174	203,585,743	208,320,912	225,215,915	265,724,419
최종 구축 기사 수	630,095	730,017	978,344	1,023,431	1,199,753

<그림 20> 최근 5년 신문 기사 구축 어절 수, 올해 구축 어절 수

**265,724,419**

계약 구축량 대비 **133%** 달성



<그림 21> 월별 구축 어절 수



월	목표 어절 수	구축 어절 수
1월	16,000,000어절 이상	20,615,537
2월	16,000,000어절 이상	20,362,796
3월	16,000,000어절 이상	23,974,422
4월	16,000,000어절 이상	22,377,749
5월	16,000,000어절 이상	22,621,787
6월	16,000,000어절 이상	22,300,221
7월	16,000,000어절 이상	22,946,103
8월	16,000,000어절 이상	22,597,287
9월	16,000,000어절 이상	20,634,830
10월	16,000,000어절 이상	22,149,269
11월	16,000,000어절 이상	23,678,705
12월	16,000,000어절 이상	21,465,713
합계	2억 어절 이상	265,724,419

<표 22> 월별 구축 어절 수

주제별 구축량에서 최대와 최소의 차이는 25%p 이내(주제별 기사 수 최대 최소 간의 차이는 25%p, 주제별 기사 어절 수 최대 최소 간의 차이는 23%p)가 되도록 구축하였다. 주제별 분포는 다음과 같다.

주제별	기사 수	비율	어절 수	비율
사회	344,290	29%	72,425,941	27%
생활	161,058	13%	32,637,315	12%
경제	156,889	13%	36,356,170	14%
스포츠	131,581	11%	31,498,721	12%
연예	125,102	10%	27,366,986	10%
정치	102,431	9%	25,686,444	10%
정보통신/과학	75,722	6%	18,022,269	7%
문화	57,253	5%	11,931,370	4%
미용/건강	45,427	4%	9,799,203	4%
계	1,199,753	100%	265,724,419	100%

<표 23> 주제별 기사 및 구축 어절 수

## 2. 매체별 납품 파일명

말뭉치 유형 구분의 N은 신문 기사 말뭉치를, 분석 층위 구분의 RW는 원시 말뭉치(raw corpus)를 의미하며, 매체 장르 및 구분 정보는 다음과 같다.(매체 분류 칸의 로마자 알파벳순으로 정렬하였고, 각 매체 분류 안에서는 매체명 가나다순으로 정렬하였다. I: 인터넷 기반 신문, L: 지역 종합지, P: 전문지, W: 전국 종합지, Z: 기타)

말뭉치 유형 구분	매체 분류	분석 층위 구분	구축 연도	매체 일련번호	매체명
N	I	RW	24	00000001	뉴스투데이
N	I	RW	24	00000002	마이데일리
N	I	RW	24	00000003	아이뉴스24
N	I	RW	24	00000004	오마이뉴스
N	I	RW	24	00000005	오에스이엔
N	L	RW	24	00000001	경남매일
N	L	RW	24	00000002	경인매일
N	L	RW	24	00000003	대경일보
N	L	RW	24	00000004	대전투데이
N	L	RW	24	00000005	울산제일일보
N	L	RW	24	00000006	전남매일
N	L	RW	24	00000007	충남일보
N	L	RW	24	00000008	충청매일
N	P	RW	24	00000001	디지털타임스
N	P	RW	24	00000002	매일경제
N	P	RW	24	00000003	스포츠경향
N	P	RW	24	00000004	스포츠투데이
N	W	RW	24	00000001	경향신문
N	W	RW	24	00000002	일간투데이
N	W	RW	24	00000003	조선일보
N	Z	RW	24	00000001	굿모닝충청
N	Z	RW	24	00000002	뉴스1
N	Z	RW	24	00000003	뉴스웍스
N	Z	RW	24	00000004	뉴스시스

<표 24> 말뭉치 파일명

<부록 1>

국가 언어 자원(말뭉치) 구축 및 활용 저작권 이용 허락 계약서(양식안)

# 국가 언어 자원(말뭉치) 구축 및 활용 저작권 이용허락 계약서(양식안)

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)

이 계약은 국가 언어 자원(말뭉치) 구축 및 활용을 위한 저작권 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 \_\_\_\_년 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 신문 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’을 분석·처리·가공한 결과물인 원시 및 분석 말뭉치를 말한다.

### 제3조 (계약의 대상)

이 계약의 이용허락 대상이 되는 권리는 아래의 저작물에 대한 저작권 중 이 조에 명시한 이용허락 범위에 필요한 권리로 한다.

저작물: \_\_\_\_년 \_\_월 \_\_일 ~ \_\_\_\_년 \_\_월 \_\_일까지의 기사 중 권리자가 저작권을 이용허락할 권리를 보유한 기사

#### 저작권재산권 이용허락 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 인공 지능 기술 개발 등 언어 정보 처리 분야에 응용하기 위해 ‘대상저작물’을 복제·변형하여 원시 및 분석 말뭉치로 구축하는 일
3. 국립국어원이 ‘복제·변형물’을 국어 연구와 인공 지능 기술 개발 등 언어 정보 처리 분야 응용을 위하여 학계·연구기관·산업체 등이 이용할 수 있도록 제공하는 일
4. 국립국어원이 ‘복제·변형물’을 제공받은 학계·연구기관·산업체 등에서 국어 연구와 인공 지능 기술 개발 등 언어 정보 처리 분야 응용을 위하여 ‘복제·변형물’을 분석·처리·가공하여 사용할 수 있도록 허락하는 일

#### 제4조 (이용허락 기간)

- ① ‘전체 기사’ 및 ‘수집 기사’의 이용허락 기간은 계약체결일부터 \_\_\_\_년 \_\_\_\_월 \_\_\_\_일까지로 한다.
- ② ‘대상저작물’ 및 ‘복제·변형물’의 이용허락 최소 기간은 계약체결일부터 \_\_\_\_년 \_\_\_\_월 \_\_\_\_일까지로 한다. 최소 기간 만료 후 권리자가 이용허락 중지 의사를 밝히지 않으면 이용허락이 1년 단위로 자동 갱신되며, 권리자가 이용허락 중지 의사를 밝히면 그 의사 내용에 따라 이용허락이 중지된다.

#### 제5조 (권리자의 의무)

- ① 권리자는 이용자에게 이 계약서 제3조에 따른 저작권재산권을 이용할 권리를 제4조의 기간 동안 비독점적으로 허락한다.
- ② 권리자는 이용자에게 계약 체결일로부터 10일 이내에 ‘대상저작물’의 이용을 위해 필요한 상당한 자료를 인도하여야 한다. 이때 자료를 인도하는 형식과 방법은 부속합의서에 따른다.
- ③ 권리자는 ‘대상저작물’에 이 계약 이행에 지장을 주는 제3자의 이용허락권, 질권 등

이 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 않는다. 다만, 이 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. 이 저작재산권 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 않는다는 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 않을 것
  3. ‘대상저작물’ 및 ‘복제·변형물’의 제공 시 이용허락 조건 및 재양도 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 효력을 가진다.

### **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 상당한 기간을 정하여 상대방에게 그 시정을 최고하고, 상대방이 그 기간이 지나도록 이행하지 않는 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 않는다.

### **제10조 (손해배상)**

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

### **제11조 (분쟁해결)**

- ① 이 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 않을 때에는 대한민국의 민사소송법 등에 따른 관할법원의 소송에 의해 해결토록 한다.

### **제12조 (비밀유지)**

양 당사자는 이 계약의 체결 및 이행과정에서 알게 된 상대방에 관한 정보, 이 계약의 내용을 상대방의 서면에 의한 승낙 없이 제3자에게 공개해서는 안 된다. 다만, 계약의 내용을 저작자에게 알리는 경우는 예외로 한다.

### **제13조 (기타부속합의)**

- ① 권리와 이용자는 이 계약의 내용을 보충하거나, 이 계약에서 정하지 않은 사항을 규정하기 위하여 부속합의서를 작성할 수 있다.

② 제1항에 따른 부속 합의는 이 계약의 내용과 배치되거나 위반하지 않는 범위 내에서 유효하다.

#### 제14조 (계약의 해석 및 보완)

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

#### 제15조 (계약 효력 발생일)

본 계약의 효력은 계약 체결일로부터 발생한다.

2024년      월      일

권리자 :

성명

주소

이용자 :

성명    국립국어원장 (인)

주소 서울특별시 강서구 금남화로 154

제안요청서에 삽입된 계약서 형태



## <부록 2>

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>명예 기자, 객원 기자, 시민 기자, 도민 기자, 학생 기자가 작성한 기사               <ul style="list-style-type: none"> <li>명예 기자나 객원 기자라고 표기되지 않고, 이름만 나오는 기사는 기고문 형식을 확인한 후 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 머리기사나 단신, 일정을 모아 놓은 기사               <ul style="list-style-type: none"> <li>00월 00일 언론 동정, 00일 국회 일정 등</li> </ul> </li> <li>공동 취재단이 작성한 기사               <ul style="list-style-type: none"> <li>공동 취재단, 공동 취재팀, 공동 취재반, 특별 취재단 등</li> <li>‘공동 000, 공동000’과 같이 띄어쓰기를 달리한 경우가 있으므로 확인 필수</li> </ul> </li> <li>기자명이 비어 있는 기사</li> <li>방송, 라디오 방송, 유튜브 라이브를 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> </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...예정’ 처럼 ‘했다.’, ‘다.’가 빠진 것으로 유추할 수 있는 경우에는 사용</li> <li>• 누락된 글자를 유추하여 기사의 마지막 문장을 완성할 수 있는 경우에는 사용하지만, 문장의 대부분 또는 주요 성분이 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, 제공, 지원, 협조를 받아 작성하였음을 밝힌 기사도 사용하지 않음</li> </ul> </li> </ul>
인공 지능이 작성한 기사	<ul style="list-style-type: none"> <li>• 인공 지능이 작성하였음을 밝힌 기사는 사용하지 않음 <ul style="list-style-type: none"> <li>– “본 기사의 제목은 챗GPT가 작성한 것임을 밝힙니다.” 등의 문구로 인공 지능 작성 여부를 밝힌 기사</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>[사진 제공= 로이터]                      표&gt; 공정위 망 이용대가 불공정 조사 쟁점</p> <p>사진제공=tvN                              ▲영상제공=</p> <p>사진=CJ엔터테인먼트 제공              [사진 및 참고자료 서울○○○○ / 사진 ©○○○]</p> <p>사진제공  카카오TV                      사진 제공 몰디브 ○○○ ○○○</p> <p>사진/이○○ 숲 해설가                      © 사진 설명 : 표창 수상을 기념하는 모습</p> <p>[그래픽]                      &lt;그래픽&gt;                      [표] 2017~2021년 귀속 기부금 세액공제 현황</p> <p>일러스트                                       (사진설명)</p> <p>화면 캡처                                      (조감도)</p>
기자의 이름, ID 등	<p>[인천=○○○기자]</p> <p>○○○○=○○○기자]</p> <p>【서울=○○○】 ○○○ 기자 =</p> <p>취재협조 = ○○○○사업단</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>[서울=○○○]○○○ 기자 = KIA 타이거즈가 '뒤틀린 요구' 의혹을 받는 ○○○ 단장을 해임했다.</p> <p>KIA는 29일 "품위손상 행위로 물의를 일으킨 ○○○ 단장에 대해 징계위원회를 개최하고 해임을 결의했다"고 발표했다.</p> <p>○ 단장은 지난해 말 프리에이전트(FA) ○○○과 협상에서 뒤틀린을 요구했다는 의혹으로 파문을 일으켰다.</p> <p>구단은 "지난 주 제보를 받은 후 사실관계 등을 파악했다"며 "사실관계를 떠나 그 어떤 이유에서라도 소속 선수와 협상 과정에서 금품 요구라는 그릇된 처신은 용납할 수 없다는 판단에 ○○○ 단장을 징계위원회에 회부했고 최종 해임 조치했다"고 밝혔다. (중략)</p> <p><u>다음은 KIA 사과문 전문.</u></p> <p><u>KIA 타이거즈는 최근 불거진 ○○○ 단장의 품위 손상 행위에 대해 KIA 타이거즈 팬 여러분은 물론, 프로야구를 사랑해 주시는 모든 팬 여러분들께 머리 숙여 사과 드립니다.</u></p> <p><u>또한 개막을 앞두고 있는 KBO리그 전체에 누를 끼치게 돼 리그 모든 구성원분들에게도 사과의 말씀을 드립니다.</u></p> <p><u>KIA 타이거즈는 즉시 사실 관계를 파악하였으며 어떠한 이유에서라도 금품 요구는 정당화되지 않는다고 판단해 징계위원회를 개최, 곧바로 ○○○ 단장을 해임 조치했습니다.</u></p> <p><u>구단은 이번 사안에 대해 무거운 책임감을 느끼며 다시는 이러한 일이 재발되지 않도록 모든 구단 임직원 및 선수단의 준법 교육에 더욱 힘쓰고, 끊임없이 노력하겠습니다.</u></p> <p><u>프로야구를 사랑해 주시고 KIA 타이거즈를 응원해 주시는 팬 여러분께 심려를 끼쳐 다시 한 번 사과의 말씀을 올립니다.</u></p> <p><u>○○○○○ ○○○ ---@---.com</u></p>

	<p>기업별 조사결과를 보면, LG전자는 김치냉장고(디오스), 노트북(Gram), 무선스틱 청소기(오브제컬렉션), 드럼세탁기(트롬), 공기청정기(LG퓨리케어), TV(OLED) 부문에서 1위 브랜드를 보유하며 6관왕의 영예를 안았고, LG생활건강은 홈쇼핑화장품(수려한), 바디로션(비온드), 샴푸(엘라스틴), 칫솔(페리오), 주방세제(풍풍) 부문에서 1위 브랜드를 보유하여 5관왕을 거머쥐었다. 그 뒤를 이어 CJ제일제당(비비고, 해찬들, 백설), 매일유업(애플루트, 매일우유, 맘마밀), 삼성전자(비스포크, 무풍에어컨, 그랑테), 소노인터내셔널(비발디파크, 오션월드), 유한킴벌리(하گی스, 그린펄거), 한국피앤지(다우니, 페브리즈)가 3개 부문에서 1위 브랜드를 보유한 것으로 나타났다.</p> <p><b>어떻게 조사했나</b></p> <p><b>조사명 = 2023 제9회 브랜드고객만족도</b></p> <p><b>영문명 = 2023 9th BCSI( Brand Customer Satisfaction Index)</b></p> <p><b>조사 대상 = 국내 소비생활을 하고있는 18세 이상 남녀</b></p> <p><b>표본수 = 3,000명 조사 방법 = 온라인 설문조사</b></p> <p><b>조사 기간 = 2023년8월1일 ~ 9월26일 주최 = 소비자평가 / 한국마케팅협회</b></p> <p><b>[○○○ ○○○○ 기자]</b></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>톰 크루즈의 출연작 4편 중 3편엔 전력질주하는 장면이 나온다. 이 배우가 오래 달릴수록 그 영화의 평점이 좋다는 분석도 나왔다. 아니나 다를까, 예고편을 보니 이번에도 그는 달리기의 정석처럼 허리를 꼿꼿이 세우고 전력질주한다.</p> <p><b>* QR코드에 휴대폰을 갖다 대거나, 인터넷 주소창에 https://---.--- 을 넣으면 구독창이 열립니다. “이메일 주소”와 “존함”을 적고 “구독하기”를 누르면 이메일로 뉴스레터가 날아갑니다.</b></p>
기사 본문으로 볼 수 없는 부가 정보 의 나열 등	<p>그의 이런 출중한 재기(才器)와 훗날 구름처럼 사라진 기묘한 행적 때문에 예인이자 ‘신선(神仙)’으로 보는 이가 적지 않다. 예인으로서 노비, 거지, 평민, 양반은 물론 도도하기 그지없는 한양 기생들까지 단숨에 매료시킨 광대 달문의 ‘필살기’는 바로 ‘팔뚝무’(송어 뛰거나 자반 뒤집기가 연상되는 남사당의 살판, 즉 땅재주와 같은 춤), ‘철괴무’(쇠지팡이를 짚고 호리병을 든 모습인 중국 신화 속의 늙은 신선 이철괴가 추었다는 탈춤), ‘만석중놀이’(산대 위에 인형을 놓고 돌리는 인형극), ‘입에 주먹 집어넣기’ 등의 춤과 묘기이다.</p> <p>&amp;lt;!-- textbox_start --&amp;gt;□일시: 2023년 12월 15일(금) 오후 7시, 16일(토) 오후 2시.    □장소: 부평아트센터 달누리극장(인천광역시 부평구 아트센터로 166)    □기획/제작: 극단 집현(集賢), Ritual &amp; Play, 코티(KOTTI)   □후원: 인천광역시, 인천문화재단 [2023 문화예술지원사업 예술창작집중지원사업 선정작]     □관람료/예매: 2만 원, 인터파크&amp;lt;!-- textbox_end --&amp;gt;</p> <p>○○○ 기자 --@-----.com</p>

	<p>신인상은 TNX, 뉴진스, 르세라핌에게 각각 돌아갔으며, 인기상은 임영웅이 차지했다.</p> <p><u>&amp;lt;다음은 '제32회 서울가요대상' 수상자 리스트&amp;gt;</u></p> <p>▲ 대상 : <u>NCT 드림</u></p> <p>▲ 최고 음원상 : <u>아이브</u></p> <p>▲ 최고 앨범상 : <u>방탄소년단</u></p> <p>▲ 월드베스트아티스트상 : <u>싸이</u></p> <p>▲ 본상 : <u>에스파·김호중·싸이·강다니엘·(여자)아이들·스트레이키즈·블랙핑크·지코·NCT 드림·레드벨벳·세븐틴·갯 더 비트·아이브·태연·방탄소년단·임영웅</u></p> <p>▲ 신인상 : <u>TNX·뉴진스·르세라핌</u></p> <p>▲ 인기상 : <u>임영웅</u></p> <p>▲ 한류 대상 : <u>수호</u></p> <p>▲ OST 상 : <u>멜로망스</u></p> <p>▲ R&amp;B 힙합상 : <u>비오·빅나티</u></p> <p>▲ 밴드상 : <u>잔나비</u></p> <p>▲ 베스트 퍼포먼스상 : <u>(여자)아이들</u></p> <p>▲ 올해의 발견상 : <u>이승윤</u></p> <p>▲ 아이돌플러스 베스트 아티스트상 : <u>방탄소년단</u></p> <p>▲ 아이돌플러스 뉴스타상 : <u>템페스트</u></p> <p>▲ K팝 특별상 : <u>카라</u></p> <p>▲ 레전드 아티스트상 : <u>보아</u></p> <p>▲ 뉴웨이브 스타상 : <u>라필루스·TAN·케플러</u></p> <p><u>[○○○ ○○○○ 기자]</u></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><u>이○○ 기자 ○○○○@○○○○.co.kr</u></p> <p>덧붙여 “한국에서도 백년이 넘게 지속되는 브랜드가 나와야 한다”면서 “한국은 패션에 대한 역사가 짧아 디자이너에 대한 지원이 열악하다. 브랜드가 지속적으로 성장하기 위해서는 이를 뒷받침 하기 위한 제도적 지원이 구축돼야 한다”고 강조했다.</p> <p>■ She is… △1958년 서울 출생 △서울여고 △숙명여대 미술대 산업공예과 △1989년 ‘손정완’ 브랜드 설립 △1990년 백화점 입점 △1994년 (주)손정완 설립 △1994년 세계 패션 그룹 회원 △1997년 SFAA(Seoul Fashion Artist Association) 가입 △2006년 ‘who’s next’ 국내최초 파리초청 단독 패션쇼 △2011년 뉴욕패션위크 데뷔 △2012년 GS홈쇼핑 콜라보레이션 브랜드 ‘SJ. WANT’ 런칭 △2021년 남성복 라인 ‘와니니 (WANINI) 출시</p>
	<p><u>[○○○○=○○○ 기자]</u> 12일 아침 최저기온이 영하 7도까지 내려간다. 한낮에도 최고 기온이 12도에 불과할 정도로 당분간 ‘동장군’이 맹위를 떨치겠다.</p> <p>11일 기상청 예보에 따르면 내일 아침 최저기온은 -7~5도, 낮 최고기온은 4~12도다. 북쪽의 찬 공기가 빠르게 내려오면서 당분간 아침기온이 영하로 내려가는 곳이 많겠다. 강한 바람에 체감온도는 더 낮아지겠다.</p> <p>남부지방에는 비 소식이 예정됐다. 제주도에는 12일 오후부터 13일 오<u>[○○○○=○○○ ○ 기자][○○○○=○○○ 기자]</u>전 사이에, 전라 서해안은 12일 밤부터 13일 새벽 사이 5mm 안팎의 비가 내리겠다. 제주도 산지에는 1~3cm의 눈이 내릴 수도 있다.</p>

	<p>=====</p> <p>기자 정보가 기사 상단, 하단이 아닌 곳에 위치한 경우이다. 기자 정보는 이처럼 기사의 곳곳에 남아 있을 수 있으므로 주의가 필요하다.</p> <p>하 작가는 "쌀이 생산되지 않던 제주에서 메밀은 보리와 함께 중요한 곡식이었다"며 "제주 사람들은 '배지근한 맛'이라고 표현했는데 '속부터 차오르는 깊고 구수한 맛'이란 뜻"이라고 했다. 잔치음식으로도 즐겼던 두 가지 음식은 모두 8~10인분 기준이다.</p> <p>◆옥수수 기정떡</p> <p>재료 : 찹옥수수알(생것) 800g, 찹옥수수알(말린 것) 800g, 찹잎 20장, 완두콩 200g, 소금 2g, 설탕 100g, 참깨 200g</p> <p>① 찹옥수수 말린 것을 물에 불리고 생것과 함께 넣어 곱게 간다.</p> <p>② 볶은 참깨를 잘 빵아 설탕, 소금으로 간한다.</p> <p>③ 찹잎을 골라 씻어 물기를 뺀다.</p> <p>④ ①②에 완두콩을 섞은 후 한 수저씩 떠서 ③의 찹잎에 찐다.</p> <p>⑤ ④를 찜통에 넣고 20분간 쪄 식힌 후 먹는다.</p> <p>◆메밀 닭반대기</p> <p>재료 : 닭살 3kg, 마른 두부 3kg, 메밀가루 500g, 계란 15개, 당근 400g, 홍고추 300g, 청양고추 150g, 간 생강 50g, 간 마늘 400g, 양파 800g, 대파 200g, 양조간장 200mL, 고춧가루 50g, 깨소금·소금</p> <p>① 닭살만 갈아서 준비한다.</p> <p>② ①에 마른 두부·메밀가루·계란과 각종 채소를 넣고 간을 맞추고 두께 2cm, 지름 15~20cm 정도 넓적하게 만들어 쪄 후 먹음직스럽게 잘라서 낸다.</p> <p>③ 간장·마늘·파·고춧가루로 양념장을 만들어 ②를 찍어 먹는다.</p>
기사와 상관 없는 광고	<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>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.</p> <p>[출처: ○○○○에서 제공하는 기사입니다.]</p> <p><a href="https://○○○○○.co.kr/news/newsView.php?id=○○○○○○○#csidx4dab120876716f7a9506745d61f1391">https://○○○○○.co.kr/news/newsView.php?id=○○○○○○○#csidx4dab120876716f7a9506745d61f1391</a></p>
반복 오류	<p>굵직굵직한 브랜드와 협업한 것도 전속 회사 역할이 컸다. 올 하반기엔 글로벌 럭셔리 브랜드와 협업도 <a href="#">선보일</a> <a href="#">선보일</a> 예정이다.</p> <p>바드는 구글 언어 모델 '람다(LAMDA)'를 <a href="#">기반으로</a> <a href="#">기반으로</a> 한다. 사용자가 질문과 요청을 바드에 대화 형태로 입력하면 이에 맞는 답을 제시한다.</p>
[기타] 유의 사항	<p>'3년이 지난 지난해 10월', '비정규직 정규직화' 같은 문장들은 바르게 사용된 경우이므로, 단어 반복 오류로 보고 삭제하지 않도록 유의한다.</p>





### <부록 3>

말뭉치 종류별 구축 예시

원시 데이터	<p>[서울=○○○○]○○○ 기자 = 독일에서 코로나19 백신 부작용을 호소하는 환자들이 주요 백신 제조사들을 상대로 소송을 시작했다.</p> <p>12일(현지시간) 외신을 종합하면 독일 법원은 코로나19 4대 백신 제조업체를 상대로 한 피해배상 소송 185건을 확인했다.</p> <p>송무를 맡는 법률회사인 뒤셀도르프와 비스바덴은 각각 135건, 50건의 소송을 관할 지방 법원에 제기했다고 밝혔다. 이들은 먼저 4대 제조사를 상대로 소송을 시작해 향후 확대할 계획이다.</p> <p>첫 소송은 바이오엔테크에 여성들이 제기한 백신 피해 건으로 다음달 초 첫 재판이 열린다. 이 여성들은 화이자·바이오엔테크 코로나19 백신 접종으로 심근염과 폐질환, 생리 불순 등의 피해를 봤다고 주장하고 있다.</p> <p>이번 재판의 핵심 쟁점은 후유증과 백신 사이 인과관계 증명이다. 바이오엔테크 측 변호인단은 "지금까지 바이오엔테크가 조사한 사례 중 백신 접종과 건강 이상 사이 인과 관계를 입증한 사례는 없었다"며 "우린 백신 제조업체로서 책임에 진지하게 임하고 있다"고 밝혔다.</p> <p>법조계는 재판 결과에 따라 향후 백신 피해 관련 소송은 크게 늘어날 것이라고 내다봤다. 토비아스 울브리히 변호사는 "이번 소송은 결과에 따라 막대한 법적 결과가 일어날 가능성이 있어 '괴물 재판'이라고 불리고 있다"고 말했다.</p> <p>○○○○ ----@----.com</p>
신문 기사 말뭉치	<p>독일에서 코로나19 백신 부작용을 호소하는 환자들이 주요 백신 제조사들을 상대로 소송을 시작했다.</p> <p>12일(현지시간) 외신을 종합하면 독일 법원은 코로나19 4대 백신 제조업체를 상대로 한 피해배상 소송 185건을 확인했다.</p> <p>송무를 맡는 법률회사인 뒤셀도르프와 비스바덴은 각각 135건, 50건의 소송을 관할 지방 법원에 제기했다고 밝혔다. 이들은 먼저 4대 제조사를 상대로 소송을 시작해 향후 확대할 계획이다.</p> <p>첫 소송은 바이오엔테크에 여성들이 제기한 백신 피해 건으로 다음달 초 첫 재판이 열린다. 이 여성들은 화이자·바이오엔테크 코로나19 백신 접종으로 심근염과 폐질환, 생리 불순 등의 피해를 봤다고 주장하고 있다.</p> <p>이번 재판의 핵심 쟁점은 후유증과 백신 사이 인과관계 증명이다. 바이오엔테크 측 변호인단은 "지금까지 바이오엔테크가 조사한 사례 중 백신 접종과 건강 이상 사이 인과 관계를 입증한 사례는 없었다"며 "우린 백신 제조업체로서 책임에 진지하게 임하고 있다"고 밝혔다.</p> <p>법조계는 재판 결과에 따라 향후 백신 피해 관련 소송은 크게 늘어날 것이라고 내다봤다. 토비아스 울브리히 변호사는 "이번 소송은 결과에 따라 막대한 법적 결과가 일어날 가능성이 있어 '괴물 재판'이라고 불리고 있다"고 말했다.</p>

인용 부호 수정 말뭉치	<p>&lt;p&gt;독일에서 코로나19 백신 부작용을 호소하는 환자들이 주요 백신 제조사들을 상대로 소송을 시작했다.&lt;/p&gt;</p> <p>&lt;p&gt;12일(현지시간) 외신을 종합하면 독일 법원은 코로나19 4대 백신 제조업체를 상대로 한 피해배상 소송 185건을 확인했다.&lt;/p&gt;</p> <p>&lt;p&gt;송무를 맡는 법률회사인 뒤셀도르프와 비스바덴은 각각 135건, 50건의 소송을 관할 지방법원에 제기했다고 밝혔다. 이들은 먼저 4대 제조사를 상대로 소송을 시작해 향후 확대할 계획이다.&lt;/p&gt;</p> <p>&lt;p&gt;첫 소송은 바이오엔테크에 여성들이 제기한 백신 피해 건으로 다음달 초 첫 재판이 열린다. 이 여성들은 화이자·바이오엔테크 코로나19 백신 접종으로 심근염과 폐질환, 생리 불순 등의 피해를 봤다고 주장하고 있다.&lt;/p&gt;</p> <p>&lt;p&gt;이번 재판의 핵심 쟁점은 후유증과 백신 사이 인과관계 증명이다. 바이오엔테크 측 변호인단은 “지금까지 바이오엔테크가 조사한 사례 중 백신 접종과 건강 이상 사이 인과 관계를 입증한 사례는 없었다”며 “우린 백신 제조업체로서 책임에 진지하게 임하고 있다”고 밝혔다.&lt;/p&gt;</p> <p>&lt;p&gt;법조계는 재판 결과에 따라 향후 백신 피해 관련 소송은 크게 늘어날 것이라고 내다봤다. 토비아스 울브리히 변호사는 “이번 소송은 결과에 따라 막대한 법적 결과가 일어날 가능성이 있어 ‘괴물 재판’이라고 불리고 있다”고 말했다.&lt;/p&gt;</p>
문장 말뭉치	<p>&lt;p&gt;&lt;s&gt;독일에서 코로나19 백신 부작용을 호소하는 환자들이 주요 백신 제조사들을 상대로 소송을 시작했다.&lt;/s&gt;&lt;/p&gt;</p> <p>&lt;p&gt;&lt;s&gt;12일(현지시간) 외신을 종합하면 독일 법원은 코로나19 4대 백신 제조업체를 상대로 한 피해배상 소송 185건을 확인했다.&lt;/s&gt;&lt;/p&gt;</p> <p>&lt;p&gt;&lt;s&gt;송무를 맡는 법률회사인 뒤셀도르프와 비스바덴은 각각 135건, 50건의 소송을 관할 지방법원에 제기했다고 밝혔다.&lt;/s&gt; &lt;s&gt;이들은 먼저 4대 제조사를 상대로 소송을 시작해 향후 확대할 계획이다.&lt;/s&gt;&lt;/p&gt;</p> <p>&lt;p&gt;&lt;s&gt;첫 소송은 바이오엔테크에 여성들이 제기한 백신 피해 건으로 다음달 초 첫 재판이 열린다.&lt;/s&gt; &lt;s&gt;이 여성들은 화이자·바이오엔테크 코로나19 백신 접종으로 심근염과 폐질환, 생리 불순 등의 피해를 봤다고 주장하고 있다.&lt;/s&gt;&lt;/p&gt;</p> <p>&lt;p&gt;&lt;s&gt;이번 재판의 핵심 쟁점은 후유증과 백신 사이 인과관계 증명이다.&lt;/s&gt; &lt;s&gt;바이오엔테크 측 변호인단은 “지금까지 바이오엔테크가 조사한 사례 중 백신 접종과 건강 이상 사이 인과 관계를 입증한 사례는 없었다”며 “우린 백신 제조업체로서 책임에 진지하게 임하고 있다”고 밝혔다.&lt;/s&gt;&lt;/p&gt;</p> <p>&lt;p&gt;&lt;s&gt;법조계는 재판 결과에 따라 향후 백신 피해 관련 소송은 크게 늘어날 것이라고 내다봤다.&lt;/s&gt; &lt;s&gt;토비아스 울브리히 변호사는 “이번 소송은 결과에 따라 막대한 법적 결과가 일어날 가능성이 있어 ‘괴물 재판’이라고 불리고 있다”고 말했다.&lt;/s&gt;&lt;/p&gt;</p>
<p>- 신문 기사 말뭉치: 원시데이터에서 캡션 정보 등 불필요한 요소를 제거한 말뭉치</p> <p>- 인용 부호 수정 말뭉치: 신문 기사 말뭉치에서 인용 부호를 수정한 말뭉치</p> <p>- 문장 말뭉치: 인용 부호 수정 말뭉치에서 단락을 문장 단위로 분할한 후 문장 단위로 &lt;s&gt;태그를 부착한 말뭉치</p>	



#### <부록 4>

신문 기사 말뭉치 오류 검색 목록

## 신문 기사 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" style="width: 100%;"> <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	<p>떠나보는 것은 어떨까? <u>○○○ 기자</u>"</p> <p><u>_____</u> }</p> <p><u>_____</u> ]</p> <p><u>_____</u> },</p> <p><u>_____</u> {</p> <p><u>_____</u> "id"</p>						

정규식	\\.[2-9] [1-4]{1}[0-9]{1})",\n ("form").*?[가-힣][^\."'\n				
설명	<ul style="list-style-type: none"> <li>기사 내에서 마침표(.) 없이 줄바꿈된 행을 검색합니다.</li> <li>- 불필요하게 줄바꿈된 문장, 마침표를 찍지 않은 문장, 소제목이 검색됩니다.</li> <li>- \\.[2-9] 부터 [^\."'\n 까지 중간 공백을 포함한 전체를 하나의 정규식으로 검색하여 사용합니다.</li> </ul> <table border="1" style="width: 100%;"> <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	<p>"id": "NZRW2400000002.3.<u>18</u>",</p> <p><u>"form": "◇과감한 투자·사업 영역 확대 이후 시장 안착·안정화"</u></p>				

정규식	[a-z]"\n [a-z][^\.]"\n		
설명	<ul style="list-style-type: none"> <li>영어 이후 줄바꿈된 문장을 검색합니다.</li> <li>기자 정보(이메일), 줄바꿈된 문장 등이 검색됩니다.</li> <li>json 형식에서는 " 기호 이후 모든 문장이 구분되어 있어 위와 같이 검색이 가능하나, 문장이 구분되어 있지 않은 문서라면 " 기호를 삭제한 정규식으로 검색합니다.</li> </ul> <table border="1"> <tr> <td>[a-z] ]</td><td>‘a’ ~ ‘z’ 까지 알파벳 검색 찾기 기능에서 대/소문자 구분(C)를 체크하지 않았다면 대소문자 구분 없이 모두 검색</td></tr> </table>	[a-z] ]	‘a’ ~ ‘z’ 까지 알파벳 검색 찾기 기능에서 대/소문자 구분(C)를 체크하지 않았다면 대소문자 구분 없이 모두 검색
[a-z] ]	‘a’ ~ ‘z’ 까지 알파벳 검색 찾기 기능에서 대/소문자 구분(C)를 체크하지 않았다면 대소문자 구분 없이 모두 검색		
예시	판테온 · 진주문화원 · 파크골프협회 진주 파크골프 홍보 · 발전 MOU"(☞ 정상적인 경우) 추진이 필요하다"고 역설했다. ---@-----.com"(☞ 기자 정보가 남아있는 경우)		

정규식	<table><tr><td>[^맞]이하는</td><td>전문이다.</td><td>서문</td><td>[^에]서 전문</td></tr><tr><td>다음은</td><td>전문\.</td><td>원문</td><td>[^뒀]담화</td></tr><tr><td>아래는</td><td>전문은[^행]</td><td>글 전문</td><td>아래와 같다</td></tr><tr><td>[^시]장 전문</td><td>공식입장</td><td>문 전문</td><td>다음과 같다</td></tr></table>	[^맞]이하는	전문이다.	서문	[^에]서 전문	다음은	전문\.	원문	[^뒀]담화	아래는	전문은[^행]	글 전문	아래와 같다	[^시]장 전문	공식입장	문 전문	다음과 같다
[^맞]이하는	전문이다.	서문	[^에]서 전문														
다음은	전문\.	원문	[^뒀]담화														
아래는	전문은[^행]	글 전문	아래와 같다														
[^시]장 전문	공식입장	문 전문	다음과 같다														
설명	<div><div><div><div>· 외부 전문이 그대로 사용된 내용을 검색합니다.</div><div>- 입장문, 연설문, 담화문, 공식입장, SNS글, 영문기사 원문, 인사/승진, 날씨 등이 검색됩니다.</div><div>- ‘아래와 같다’ , ‘다음과 같다’ 이후에는 삭제요소가 등장할 확률이 높으므로 주의해야 합니다.</div></div></div><div><table><tr><td>[^맞 ]</td><td>^ 기호가 대괄호 [ ] 안에 있다면, 괄호 안에 있는 단어를 제외하고 검색</td></tr></table></div></div>	[^맞 ]	^ 기호가 대괄호 [ ] 안에 있다면, 괄호 안에 있는 단어를 제외하고 검색														
[^맞 ]	^ 기호가 대괄호 [ ] 안에 있다면, 괄호 안에 있는 단어를 제외하고 검색																
예시	<div><div>‘1시군 1품(一品) 축제’로 선정된 축제는 <b>다음과 같다.</b></div><div>▲논산딸기축제 ▲금산삼계탕축제 ▲서산해미읍성축제</div></div>																

정규식	<div> <div>명예기자</div> <div>칼럼니스트</div> <div>지역종합</div> <div>취재단</div> <div>에디터</div> </div> <div> <div>객원기자</div> <div>연합</div> <div>시민기자</div> <div>특별취재팀</div> <div>이코노미스트</div> </div> <div> <div>논설위원</div> <div>연합뉴스</div> <div>도민기자</div> <div>특별취재반</div> <div>리포터</div> </div> <div> <div>칼럼</div> <div>전국종합</div> <div>공동취재단</div> <div>전문기자</div> <div>아나운서</div> </div>
설명	<ul style="list-style-type: none"> <li>· 저작권과 관련하여 삭제가 필요한 기사, 외부 기고문을 검색합니다.</li> <li>- 작성자를 확인해야 하며 단순히 내용에 포함된 것만으로는 삭제하지 않습니다.</li> </ul>
예시	<p>호국보훈의 달을 가슴속에 되새기는 작은 기폭제가 됐으면 하는 바람이다. <u>[글/사진 ㅇㅇㅇ기자, 사진공동취재단]</u></p>

정규식	<p style="text-align: center;">ㅇ   ㅣ   ㅓ   ㅕ</p> <p style="text-align: center;">ㄱ ㄴ ㄷ ㄹ ㅁ ㅂ ㅅ ㅇ ㅈ ㅊ ㅋ ㅌ ㅍ ㅎ 가 다 배 쓰 쥬              ㅏ ㅑ ㅓ ㅕ ㅗ ㅛ ㅜ ㅠ ㅡ ㅣ ㅞ ㅟ ㅠ ㅡ 과 내 겨 게 괴 거 너</p>
설명	<ul style="list-style-type: none"> <li>• 텍스트가 깨진 문장, 문장에 포함된 오타 등을 검색합니다.</li> <li>- 기자 정보에도 오타가 있을 수 있으므로 체크해야 합니다.</li> <li>- ‘제보자 ㄱ씨’, ‘ㄹ자형 건물’ 처럼 문맥상 필요한 자음과 모음 사용은 수정하지 않습니다.</li> </ul>
예시	<div style="font-family: monospace; font-size: 1.2em;"> <u>ㅇ</u>ㅏ    ㅈ    ㅓ     ㅔㅓ                 </div>

정규식	이 기사는      본 기획물은      위 텍스트는      대답      정리=
설명	<ul style="list-style-type: none"> <li>인공 지능이 작성한 기사, 협찬을 받아 작성된 기사를 검색합니다.</li> <li>- 인공 지능 작성, 협찬/지원 여부를 명백히 밝히고 있으므로 기사를 삭제 처리합니다.</li> <li>대답을 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>[<b>이 기사는</b> ○○○○ AI 로봇 기자가 작성한 글입니다.]</p> <p>※ <b>이 글은</b> ○○○의 편집방향과 일치하지 않을 수 있습니다</p> <p>[○○○ ○○○○ 전국부장 <b>대답, 정리</b>=○○○ 기자]</p>



정규식	<div> "사진 이미지= 자료 : 포스터 제공"  "&gt;사진 출처 자료=, 조감도 캡처"  "자료: 그림 투시도 장면" </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.  <b>사진</b> 한국관광공사</p> <p>큰 역할을 해오고 있다.  <b>[사진제공=</b>모코이엔티, 티모넷]</p> <p>획득하겠다”라며 각오를 전했다.  ○○○○○ <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><b>"&lt;승진&gt;</b> ◇전무이사 ▷동원산업</p> <p>순례를 다녀온 사람들은 오래 머문다. &lt;계속<b>&gt;"</b></p> <p>▶아시아나항공 <b>&lt;전무 승진&gt;</b> ▷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>전속계약을 맺었으나, <u>이후</u> <u>작품</u> <u>활동</u>은 없었다. (☞ 유니코드 2008)</p> <p>전속계약을 맺었으나, 이후 작품 활동은 없었다. (☞ 유니코드 0020)</p>

정규식	유니코드 FF0E ‘ . ’      유니코드 FF0C ‘ , ’
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>대통령실 행정관을 지냈다.<u>.</u></p> <p>대통령실 행정관을 지냈다.</p> <p>공매도(空賣渡,<u>.</u> Short Selling)는 말 그대로 ‘없는 걸 판다’는 뜻이다.</p> <p>공매도(空賣渡, Short Selling)는 말 그대로 ‘없는 걸 판다’는 뜻이다.</p>

정규식	^\n                      ^[.]                      ".
설명	<ul style="list-style-type: none"> <li>· 마침표를 잘못 찍은 경우를 검색합니다.</li> </ul>
예시	<u>".</u> 1882년 창단된 베를린 필은

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	진심 어린 사과를 건네 안방극장을 뭉클하게 만들었다."

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> <li>{2,4}에서 앞의 2는 어절의 최소 음절수, 4는 최대 음절수를 뜻합니다.</li> </ul>	
	\1	앞에서 검색한 소괄호 ( )의 내용이 동일하게 반복됨을 의미
예시	"이 같은 기대감이 기대감이 청약 시장에 반영된 것"이라고 말했다.	

정규식	디\. 다\. 따\. 썬\. 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	연결기준 매출의 6.41% 해당하 <del>날</del> .	

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	산연에 따르면 최근 국내 <del>건설?부동산</del> 시장은 산연에 따르면 최근 국내 건설·부동산 시장은				

정규식	<div> <div>■ 진행</div> <div>영상취재</div> <div>[영상]</div> <div>PD</div> <div>[앵커]</div> </div> <div> <div>■ 진 행</div> <div>영상편집</div> <div>촬영:  촬영</div> <div>프로듀서</div> <div>앵커</div> </div>				
설명	<ul style="list-style-type: none"> <li>영상/라디오 뉴스를 그대로 옮긴 기사를 검색합니다. 해당 기사는 사용하지 않습니다.</li> </ul>				
예시	■ <del>진행</del> : 김현정 앵커 (노컷뉴스 CBS라디오<김현정의 뉴스쇼>)				

## □ 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	강릉시 관계자는 이번 산불로 강릉 여행을

정규식	$\text{'}\wedge\text{'?}\$$ $\text{“}\wedge\text{”?}\$$
설명	<ul style="list-style-type: none"> <li>여는 따옴표만으로 끝난 문장을 검색합니다. (∼), (∼)</li> </ul>
예시	지난해 12월부터 올해 9월을 목표로 무등산 정상 상시 개방을 추진해왔다.

정규식	$\text{“\_ \_” ‘\_ \_’}$
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	웨스트 교수는 민주당 소속인 조 바이든 대통령은 '변변치 못한 신자유주의자'로, 공화당 소속인 도널드 트럼프 전 대통령은 '신파시스트'로 규정하며 비판했다.



정규식	<div> “” ”“ ‘’ ‘  ‘)’  ““  ””  ’  ’ </div>
설명	<ul style="list-style-type: none"> <li>잘못 쓰인 따옴표를 검색하고 확인합니다.</li> </ul>
예시	스위치용 ‘스컬 <sup>red</sup> 아카’ 특별 할인 이벤트  초콜릿 상자는 ‘분홍’, <sup>red</sup> ‘빨간’, ‘파란’색으로  <sup>red</sup> “작품활동도 하고 싶고“라고 전했다.

<기획·연구>

국립국어원 언어정보과장 강미영

국립국어원 학예연구관 김문오

국립국어원 연구원 이선영

<사업 참여자>

사업 책임자 윤종웅(㈜윤즈정보개발 소장)

사업 참여자 남가윤(㈜윤즈정보개발 연구원)

서경찬(㈜윤즈정보개발 책임연구원)

안소연(㈜윤즈정보개발 연구원)

윤종성(㈜윤즈정보개발 팀장)

이승철(㈜윤즈정보개발 수석연구원)

임승락(㈜윤즈정보개발 연구원)

최원수(㈜윤즈정보개발 연구원)

---

발행인: 국립국어원장

발행처: 국립국어원

서울시 강서구 금남화로 154

전화 02-2669-9775, 전송 02-2669-9757

인쇄일: 2024년 10월 31일

발행일: 2024년 10월 31일

인 쇄: 다큐팩토리

---

※ 이 보고서는 국립국어원의 용역비로 수행한 ‘2024년 신문 기사 원문  
자료 수집 및 정제’ 사업의 결과물을 발간한 것입니다.