

국립국어원 2025-01-35

발간등록번호
11-1371028-100033-01

2025년 신문 기사 원문 및 온라인 매체 게시 자료 구축

사업책임자
서 경 찬



제 출 문

국립국어원장 귀하

국립국어원과 체결한 연구용역 계약에 따라 ‘2025년 신문 기사 원문 및 온라인 매체 게시 자료 구축’ 사업에 관한 연구 보고서를 작성하여 제출합니다.

■ 사업 기간: 2025년 4월 23일 ~ 2025년 11월 7일

2025년 11월 7일

사업책임자: 서경찬(㈜윤즈정보개발)

연구 기관: (주)윤즈정보개발

사업 책임자: 서경찬

사업 참여자: 남가윤, 안소연, 윤종성

윤종웅, 이동민, 이승철

임순영, 지의선, 최원수

〈국문 요약〉

2025년 신문 기사 원문 및 온라인 매체 게시 자료 구축

국립국어원은 인공지능(AI) 및 거대언어모델(LLM) 기술의 고도화에 따라, 현대 한국어의 실제 사용 양상을 폭넓고 깊이 있게 반영하는 고품질 말뭉치를 구축하여 산·학·연이 자유롭게 활용할 수 있도록 제공하고 있다. 본 사업은 대규모 한국어 학습 자료 구축의 일환으로, 기존에 개별적으로 추진되던 신문 기사와 온라인 게시글 수집 사업을 통합 수행하여 정제된 고품질 텍스트 데이터를 공공재로 제공함으로써 국가 인공지능 데이터 경쟁력을 강화하는 데 그 목적이 있다.

본 사업의 수행 범위는 2024년 생산된 신문 기사(월평균 1,800만 어절 이상)와 온라인 게시글(총 5천만 어절 이상) 원문 자료 수집, 저작 권리자와의 직접 계약을 통한 법적 분쟁 소지 해소, 매체 특성에 최적화된 정제 및 비식별화, 그리고 3종의 고품질 말뭉치(원시, 인용 부호 수정, 문장) 구축으로 구성된다. 주요 사업 수행 내용과 성과는 다음과 같다.

첫째, 신문 기사 원문 자료 수집 및 정제를 통해 뉴스 기사 데이터를 확보하였다. 전국 종합지, 통신사, 지역 일간지 등 총 20개 이상의 다양한 매체와 직접 계약을 체결하여 저작권을 확보하였다. 수집된 자료는 기사 내 불필요한 요소(광고, 캡션, 비문 등)를 제거하고, 인용 부호와 한자 코드를 국제 표준에 맞게 통일하였으며, 인공지능 학습 효율을 높이기 위해 단락을 문장 단위로 분할한 ‘문장 말뭉치’를 구축하여 데이터의 활용성을 극대화하였다.

둘째, 온라인 게시글 원문 자료 수집 및 정제를 통해 실생활 언어와 주제의 다양성을 확보하였다. 블로그와 누리소통망 게시글을 대상으로 자료를 수집하였다. 특히 단순히 구어체적 특성뿐만 아니라, 정치, 경제, 사회, 문화, 아이티(IT) 등 다양한 주제의 글을 폭넓게 수집하여 특정 분야에 치우치지 않는 균형 잡힌 데이터세트를 구축하였다. 또한, 온라인 매체의 특성을 고려하여 욕설, 혐오 표현 등 표현을 정제하여 데이터의 윤리적 안정성을 확보하였다.

셋째, 법률 자문 및 비식별화 체계를 대폭 강화하여 공공 데이터로서의 무결성을 입증하였다. 전문 법무법인의 자문을 통해 매체별 저작권 계약의 법적 위험성을 사전 차단하였으며, 비식별화 대상을 기존의 개인정보(이름, 연락처 등)에서 대폭 확대하였다. 특히 공공 데이터 개방 시 발생할 수 있는 사회적 논란을 미연에 방지하고 소상공인 등의 잠재적 피해를 막기 위해, 특정 상호명, 브랜드명 등 민감할 수 있는 모든 정보를

포괄적으로 비식별화 처리하여 데이터 유통 과정에서의 안전성과 윤리성을 최우선으로 확보하였다.

최종적으로 본 사업을 통해 총 2.7억 어절 규모의 신문 및 온라인 통합 말뭉치가 구축되었다. 이 자료는 현대 한국어의 표준적 용법과 다양한 주제를 아우르는 실생활 언어를 동시에 담고 있어, 한국어 인공지능 모델의 성능 고도화 및 관련 학술 연구에 핵심적인 기초 자료로 활용될 것으로 기대된다.

주요어: 신문 말뭉치, 온라인 말뭉치, 인공지능 학습 데이터, 저작권 확보, 개인정보 비식별화, 상호명 비식별화, 문장 말뭉치, 정제 지침, 데이터 품질 관리

<Abstract>

Collection and Refinement of Original Newspaper Articles and Online Posts in 2025

The National Institute of Korean Language (NIKL) is developing high-quality linguistic corpora that comprehensively reflect the actual usage patterns of contemporary Korean, responding to rapid advances in Artificial Intelligence (AI) and Large Language Models (LLM). These corpora are provided as an open public resource accessible to industry, academia, and research institutions. As part of a national initiative to construct large-scale Korean training datasets, this project aims to enhance Korea's AI data competitiveness by integrating previously independent newspaper-article and online-post collection projects and supplying refined text data as a public asset.

The scope of the project includes: (1) collecting original texts from newspaper articles (more than 18 million eojeols per month) and online posts (over 50 million eojeols in total) produced in 2024; (2) eliminating legal risks through direct copyright agreements with rights holders; (3) performing medium-specific refinement and de-identification; and (4) constructing three types of high-quality corpora—raw corpus, quotation-mark-corrected corpus, and sentence-segmented corpus. The major outcomes are summarized as follows.

First, news article data was secured through the collection and refinement of original newspaper text. Copyright ownership was obtained through direct agreements with more than 20 media outlets, including national dailies, news agencies, and regional newspapers. Non-textual components such as advertisements, captions, and malformed textual segments were removed, and quotation marks and Hanja character codes were standardized according to international conventions. Additionally, a sentence-segmented corpus was developed by splitting paragraphs into sentence units, significantly improving usability for AI model training.

Second, linguistic realism and thematic diversity were strengthened through the collection and refinement of original online posts drawn from blogs and social networking platforms. A balanced dataset was constructed by extensively collecting texts across a wide range of subjects—including politics, economics, society, culture, and IT—thus preventing topical bias. To ensure ethical soundness, expressions involving profanity, hate speech, and other problematic content were refined according to the characteristics of online media.

Third, the project reinforced legal and de-identification frameworks to demonstrate the integrity of the corpus as a public resource. Legal risks inherent in copyright contracting were preemptively mitigated through consultations with professional legal counsel. The scope of de-identification was significantly expanded beyond personal information (e.g., names, phone numbers) to include potentially sensitive business names and brand names to prevent social disputes upon public release and safeguard small business owners. This approach prioritized safety and ethical responsibility throughout the data distribution process.

As a result, an integrated corpus comprising approximately 277 million eojeols was constructed. This corpus encompasses both standard contemporary Korean usage and diverse real-life language phenomena, and is expected to serve as a foundational resource for advancing Korean AI model performance and supporting linguistic and computational research.

Keywords: Korean Newspaper Corpus, Online Corpus, AI Training Dataset, Copyright Securance, De-identification, Business Name Masking, Sentence-Segmented Corpus, Refinement Standards, Data Quality Management

차 례

제1장 서론

1. 사업 목적 및 필요성	3
2. 사업 수행 범위	4
3. 사업 수행 절차	6
4. 사업 추진 경과	9

제2장 사업 수행 내용

1. 신문 기사 말뭉치 구축	13
2. 온라인 매체 게시 자료 말뭉치 구축	39

제3장 사업 수행 결과

1. 신문 기사 구축 결과	59
2. 온라인 매체 게시 자료 구축 결과	63
3. 매체별 파일명 규칙	66

<부록 1> 국가 언어 자원(말뭉치) 구축 및 활용 저작권 이용 허락 계약서	68
<부록 2> 데이터 정제 작업 지침	77
<부록 3> 말뭉치 종류별 구축 예시	84
<부록 4> 신문 기사 말뭉치 오류 검색 목록	87
<부록 5> 온라인 매체 게시 자료 비식별화 지침	97
<부록 6> 온라인 매체 게시 자료 문장 분할 지침	103

표 차례

<표 1> 사업 공정표	9
<표 2> 선정된 매체 구분	14
<표 3> 데이터 특징 예시 (a 엘리먼트)	14
<표 4> 유사도 비교를 통해 사용하지 않는 기사의 예	16
<표 5> 저작권 이용 문제로 인해 사용하지 않는 기사의 특징	18
<표 6> ‘한·중·일 호환용 한자 영역’ 한자 치환 표	19
<표 7> 치환 코드 목록	21
<표 8> 불필요한 요소 제거 내용	25
<표 9> 원시 데이터와 정제된 데이터 비교 1	28
<표 10> 원시 데이터와 정제된 데이터 비교 2	29
<표 11> 인용 부호 치환 표	35
<표 12> 인용 부호 수정 데이터 정제 전후	36
<표 13> 문장 말뭉치 데이터 정제	38
<표 14> 온라인 매체 선정 기준	39
<표 15> 온라인 매체 배제 이유	40
<표 16> 참여자 후보군 선정 기준	41
<표 17> 각 카테고리별 최초 확보 계정 수(블로그 기준 분류)	46
<표 18> 데이터 구조 분석	47
<표 19> 주제 분류 체계 대응(국립국어원·네이버)	49
<표 20> 특수 문자의 처리 방안	52
<표 21> 비식별화 대상 태그·항목	54
<표 22> 비식별화 적용(예시)	55
<표 23> 선정된 매체 구분	59
<표 24> 신문 기사 수집 총괄표	60
<표 25> 신문 기사 정제 총괄표	61
<표 26> 매체 유형별 구축 어절수	62
<표 27> 월별 구축 어절 수	62
<표 28> 주제별 기사 및 구축 어절 수	63
<표 29> 온라인 매체 게시 자료 수집 총괄표	63
<표 30> 온라인 매체 게시 자료 정제 총괄표	63
<표 31> 온라인 매체 게시 자료 구축 주제별 구축량	64
<표 32> 온라인 매체 게시 자료 구축 상세 주제별 구축량	65
<표 33> 신문 기사 말뭉치 파일명	66
<표 34> 온라인 매체 게시 자료 말뭉치 파일명	66

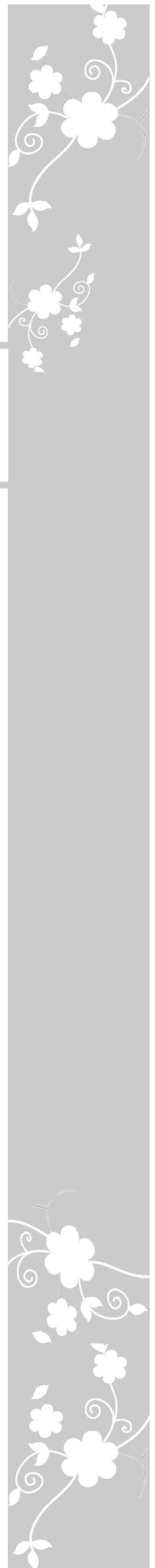
그림 차례

<그림 1> 신문 기사 말뭉치 구축 절차	6
<그림 2> 온라인 매체 게시 자료 구축 절차	7
<그림 3> 사용하지 않는 기사의 예 (저작권)	23
<그림 4> 사용하지 않는 기사의 예 (품질)	24
<그림 5> 기사 수정 예 ①	30
<그림 6> 기사 수정 예 ②	30
<그림 7> 기사 수정 예 ③	31
<그림 8> 작업 프로그램 화면	32
<그림 9> 정규식을 통한 작업 예	33
<그림 10> 데이터 작업 및 검수 공정	33
<그림 11> 문장 말뭉치 개념	37
<그림 12> 참여자 1:1 직접 섭외 과정	42
<그림 13> 국립국어원 공고를 통한 신뢰성 확보	43
<그림 14> 법률 검토를 통한 계약서 보완	44
<그림 15> 사용하지 않는 온라인 게시 자료의 예	48
<그림 16> 특수 문자 ‘ □□□□ ’ (U+2800)’를 사용하여 단락을 구분한 예	53



제 1 장

서 론



제1장 서론

1. 사업 목적 및 필요성

본 사업은 인공 지능(AI) 기술 혁신 및 국어 연구 활성화를 위한 고품질의 한국어 말뭉치를 국가 차원에서 안정적으로 구축하는 것을 목적으로 한다. 이를 위해 대규모의 신문 기사와 온라인 매체 게시 자료를 수집·정제하고, 활용에 필요한 저작권 이용 허락을 확보하여 국가 공공재로 제공하는 것을 목표로 하고 있다.

가. 국가 인공 지능 경쟁력 확보 기여

국내외 인공 지능 시장은 대규모 고품질 언어 데이터를 기반으로 발전하고 있다. 본 사업을 통해 구축되는 신문 기사 및 온라인 게시 자료 말뭉치는 한국어의 실제 사용 양상을 반영한 최신 자료로서, 자연어 처리(NLP), 기계 번역, 언어 생성 모델 등 차세대 인공 지능 기술 개발의 핵심 기반 자료로 활용될 것이다.

나. 국어 자원 공공 활용 가치 제고

수집 자료의 법적 이용 권한을 확보하고 표준화된 형태로 구축함으로써, 대학·연구소·중소기업 등 민간 부문에서도 저작권 문제없이 자유롭게 국어 자원을 활용할 수 있는 안정적 공공 인프라를 제공한다. 이러한 기반은 국어 정보화와 연구 생태계 확충에 기여할 것으로 기대된다.

2. 사업 수행 범위

가. 신문 기사 원문 자료 수집(1.5억 어절 이상)

- ▶ 신문 기사 말뭉치 구축에 필요한 2024년 작성 신문 기사 원문자료를 수집함.
- ▶ 신문 기사 원시 말뭉치는 매체별, 월별, 기사 주제별로 균형성을 갖춰 1.5억 어절 이상 구축함.
 - 기사 주제(사회, 경제, 생활, 과학 등)별 최소 구축 비율 4% 이상임.
 - 각 매체의 ‘사설(社說)’을 구축 대상에 필수 요소로 포함함.
- ▶ 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 원문자료 전체 활용에 필요한 저작권을 확보함.
- ▶ 수집한 기사 원문자료 중 국립국어원에서 말뭉치 구축 대상으로 선정하는 매체의 기사 원문에 대해서 저작권자와 저작물 이용 허락 계약을 체결함.
- ▶ 계약과 관련해 법률적인 검토를 받은 후 주관 기관이 제공한 계약서 양식에 따라 국립국어원과 협의하여 계약을 체결함.
- ▶ 저작권 이용 허락 대상 권리는 신문 기사 원문 자료 및 신문 기사 말뭉치의 저장, 복제, 전송, 2차적 저작물 작성권, 복제·변형물을 분석·처리·가공하여 사용할 수 있도록 허락할 권리를 포함함.
- ▶ 이용 허락 기간은 계약일로부터 최소 2036년 12월 31일까지로 함.

다. 중복 기사 제거 및 정제

- ▶ 수집된 기사 중에 동일 매체 내에서 기사 내용이 동일한 기사를 제거함.
- ▶ 신문 기사 내에 삽입되어 있는 사진, 표, 그래프, 그림 및 캡션, 불필요한 태그 등 기사 원문 텍스트 외의 요소들을 제거함.
- ▶ 기사 내용과 관련 없는 텍스트 및 저작권 침해 요소가 포함된 기사나 외부 작성자의 논설, 기고문 등을 제거함.
- ▶ 중복 기사, 길이가 너무 짧거나 긴 기사 등 말뭉치로 구축하기에 부적절한 기사 원문은 대상에서 제외하고, 정제된 신문 기사 원문을 대상으로 헤더 정보 부착 등의 표지 부착을 수행하여 원시 말뭉치 형태로 가공함.

라. 인용 부호 수정 말뭉치, 문장 분할 말뭉치 구축

- ▶ 최종 산출물은 ①신문 기사 말뭉치, ②인용 부호 수정 말뭉치, ③인용 부호 수정에 더해 문장 단위로 분할한 말뭉치로, 총 3종을 제출함.
- ▶ 인용 부호 수정 말뭉치는 신문 기사 원문에 포함된 다양한 인용 부호 오류를 수정한 말뭉치로 신문 기사 말뭉치와 함께 제출함.
- ▶ 문장 분할 말뭉치는 인용 부호 수정 말뭉치 자료에 단락으로 되어 있는 자료를 인공 지능 학습에 활용하기 위해 문장으로 분할한 자료로 인용 부호 수정 말뭉치와 함께 제출함.

마. 온라인 매체 게시 자료 원문 수집, 원시 말뭉치 구축(0.5억 어절 이상)

- ▶ 온라인 매체 게시글 말뭉치 구축에 필요한 온라인 매체 게시글 원문을 수집함.
- ▶ 온라인 매체 게시 자료 수집 대상 (2023~2024년 작성글 100,000건 이상): ❶ 블로그, ❷ 타유형
 - ❶ 블로그: 50,000건(게시물 건수) 이상, 건당 최소 200어절 이상
 - ❷ 타유형: 50,000건(게시물 건수) 내외, 건당 최소 10어절 이상
- 계정별 수집 한도: 작성자당 게시글 연 1,000건
- ▶ 주제별 비율 차이를 15%를 초과하지 않도록 하여 데이터 불균형을 방지함.
- ▶ 말뭉치로 구축하기에 부적절한 내용의 원문을 제외하고 정제된 온라인 매체 게시 자료의 원문을 원시 말뭉치로 0.5억 어절 이상을 구축함.
 - 중복 게시 자료, 펄글, 전문 외국어로 작성된 게시물이나 이미지나 동영상, 웹 주소, 핵심어 표시(해시태그) 등 비문서로만 구성된 웹 원문 자료, 너무 짧거나 긴 내용, 저작권 침해 문제가 있는 내용은 말뭉치 구축 대상에서 제외함
 - 이모티콘, 스티커, 사진, 동영상 및 파일 링크(URL 주소) 등 특수 메시지 및 시스템 메시지 정제, 개인정보를 비식별화 처리함
- ▶ 수집 날짜, 웹 주소 작성일 등 구축 대상 자료에 대한 메타 정보를 포함하여 구축.
- ▶ 저작권 이용 허락 대상 권리는 온라인 매체 게시 원문 자료 및 원시 말뭉치의 저장, 복제, 전송, 2차적 저작물 작성권, 복제·변형물을 분석·처리·가공하여 사용할 수 있도록 허락할 권리를 포함함.
- ▶ 이용 허락 기간은 계약일로부터 최소 2046년 12월 31일까지로 함.

3. 사업 수행 절차

2025년도 사업 수행 절차는 2024년의 신문 기사 단일 사업 수행 방식에서 확장되어, 신문 기사 말뭉치와 온라인 매체 게시 자료 말뭉치라는 성격이 상이한 두 유형의 사업을 통합적으로 수행하였다.

신문 기사 말뭉치 구축은 저작권 확보와 정형 데이터의 정밀 정제에 중점을 두는 반면, 온라인 매체 게시 자료 말뭉치 구축은 다수 참여자와의 계약 체결, 개인정보 비식별화, 비정형 데이터의 윤리성 검증 등 특화된 공정을 요구한다.

이에 따라 본 사업은 두 유형의 특수성을 극대화하고 품질과 법적 안정성을 완벽히 확보하기 위해, 서로 독립적인 두 개의 공정 흐름으로 구분하여 수행하였다. 각 절차의 세부 흐름도는 다음과 같다.

가. 신문 기사 말뭉치 구축

신문 기사 말뭉치 구축 절차는 총 5단계로 진행되며 각 단계는 다음과 같은 핵심 관리 요소를 포함한다.



그림 1. 신문 기사 말뭉치 구축 절차

- ① 저작권 확보: 저작권 사용 협의 및 계약서에 대한 법률 검토를 거쳐 매체별 계약을 체결한다. 이는 사업 결과물의 공공적 활용을 위한 법적 근거를 마련하는 핵심 초기 단계이다.
- ② 자료 수집·정제 및 정규화: 매체로부터 수령한 원시 데이터와 홈페이지에서 수집(크롤링)한 데이터를 데이터베이스(DB)에 저장한 뒤, 불필요한 태그를 제거한다. 이후 문자 코드 통일, 중복·유사 기사 제거, 저작권 이슈 기사 표시 등의 공정을 거쳐 원시 말뭉치를 정제한다.

다.

- ③ 타 정보 작성: 정제된 기사에 대해 국립국어원 지정 체계에 따른 주제 분류를 적용하고, 기사별 메타데이터(제목, 기자 정보, 어절 수 등)를 입력한다.
- ④ 인용 부호 수정 말뭉치·문장 말뭉치 생성: 정제된 원시 말뭉치를 기반으로 인용 부호 정제 작업을 수행하여 인용 부호 수정 말뭉치를 생성하고, 여기에 문장 분할 작업을 적용해 문장 말뭉치를 생성한다.
- ⑤ 3종 말뭉치 취합 및 납품: 최종적으로 3종 말뭉치(원시, 인용 부호 수정, 문장)를 취합하여 JSON 형식으로 생성한 뒤, 품질 검수 절차를 거쳐 최종 납품한다.

나. 온라인 매체 게시 자료 말뭉치 구축

온라인 매체 게시 자료 구축은 제한된 예산과 일정 속에서 이전 사업(2022년) 대비 약 3배 규모인 5천만 어절 이상의 데이터를 구축해야 하는 전략적 핵심 과제였다. 이에 본 사업은 장문형 텍스트가 풍부한 네이버 블로그를 주요 데이터원으로 선정하여 고효율 데이터 확보 체계를 수립하였으며, 동시에 개인정보·상호명·상표명 등 민감 개체를 비식별화하는 정밀 정제 공정을 핵심 관리 요소로 설정하였다.

온라인 매체 게시 자료 구축 절차는 총 4단계로 진행되며 각 단계별 주요 내용은 다음과 같다.

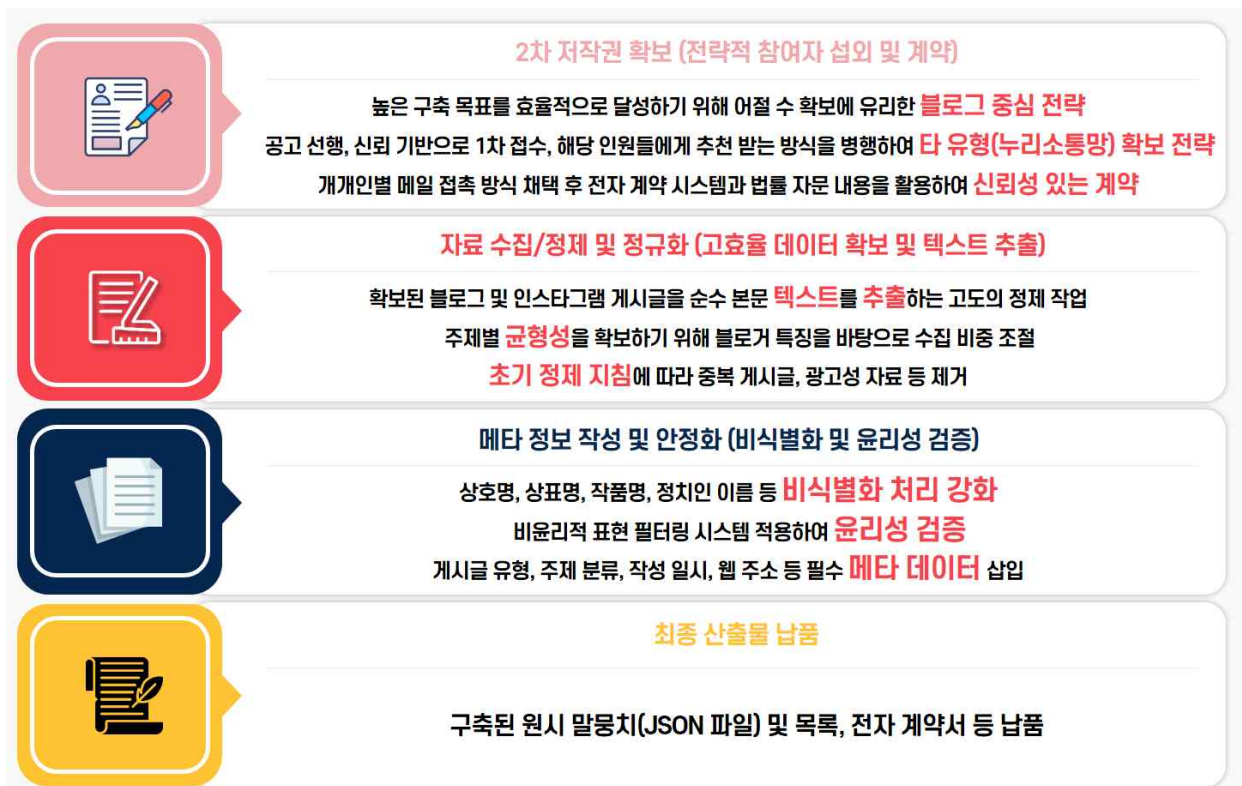


그림 2. 온라인 매체 게시 자료 구축 절차

① 2차 저작권 확보 (참여 대상 전략적 섭외 및 계약 체결)

- 블로그 중심 전략: 제한된 예산 속에서 높은 구축 목표를 효율적으로 달성하기 위해, 어절 수 확보에 유리한 네이버 블로그를 중심으로 구축을 진행하였다. 잠재 참여 계정을 선별하고, 어절 수가 많은 블로거를 1순위 대상으로 분류하여 접촉 우선순위를 설정하였다.
- 타 유형(누리소통망) 확보 전략: 국립국어원 홈페이지 공고를 선행하여 신뢰 기반의 1차 접수를 진행하고, 해당 참여자들의 추천을 병행하는 방식으로 게시글 수가 많은 계정을 선별하였다. 이를 통해 약 5만 건 내외의 인스타그램 게시글 확보를 목표로 계약을 체결하였다.
- 신뢰성 및 계약: 대량 메일 발송으로 인한 스팸 분류 위험을 방지하기 위해 개별 메일 접촉 방식을 채택하였다. 1차·2차·3차에 걸친 단계별 메일 발송을 통해 안정적으로 고품질 참여자를 확보하였으며, 전자 계약 시스템과 법률 자문 절차를 거쳐 저작권 이용 허락 계약을 체결하였다.

② 자료 수집·정제 및 정규화 (고효율 데이터 확보 및 텍스트 추출)

- 데이터 추출: 확보된 블로그 및 인스타그램 게시글을 수집하여 데이터베이스(DB)에 저장하였다. 네이버 블로그처럼 HTML 요소, 이모티콘, 광고·협찬 정보 등이 혼재된 비정형 데이터를 대상으로 순수 본문 텍스트만을 추출하는 고도의 정제 작업을 수행하였다.
- 균형성 및 초기 필터링: 주제별 균형성(제안 요청 기준 $\pm 15\%$)을 확보하기 위해 블로거의 특성과 게시 유형을 분석하여 수집 비중을 조정하였다. 초기 정제 지침에 따라 중복 게시글과 광고성 자료를 선별·제거하였다.

③ 메타 정보 작성 및 안정화 (비식별화 및 윤리성 검증)

- 비식별화 처리 강화: 기존 2019년 웹 말뭉치 구축 사업, 2022년 온라인 게시 자료 구축 사업, 2024년 일상대화 데이터 구축 사업의 기준을 참조하여 개인정보를 비식별 처리하였다. 특히 상호명, 상표명, 작품명, 정치인 이름 등 민감 정보를 모두 비식별화하여 데이터의 법적·윤리적 안정성을 확보하였다.
- 윤리성 검증: 비윤리적 표현 필터링 시스템을 적용하여 혐오, 차별 등 사회적으로 민감한 표현을 식별하고 제거하였다.
- 메타 정보 작성: 게시글 유형, 주제 분류, 작성 일시, 웹 주소 등 필수 메타데이터를 입력하고, 최종 검수를 통해 품질을 확정하였다.

④ 최종 산출물 납품

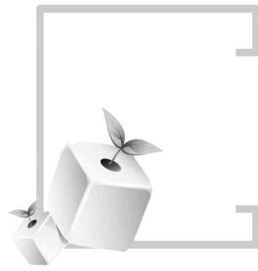
- 구축된 원시 말뭉치(JSON 파일)와 게시글 목록, 전자 계약서 등 관련 산출물을 일괄 납품.

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	11 월
준 비	신문 매체, 계정주 섭외								
수 집	저작권 사용 협의								
	법률 검토								
	계약								
정 제	데이터 수령(신문기사)								
	신문 기사 및 온라인 게시글 수집								
	HTML 태그 정리								
	메타데이터 정리								
메타데이터 생성	주제 분류								
	기사, 게시글 선택								
납품 및 종료	중복, 유사도 작업								
	불필요 요소 태그 부착								
	불필요 요소 태그 확인								
	인용부호 수정 말뭉치 생성(신문)								
	문장 말뭉치 생성(신문)								
	게시글 요소 본문 정리								
	개인정보 비식별화								
	비윤리적 표현								
	게시글 확인								
	메타 데이터 삽입								
	데이터 납품								
	구축 데이터 품질 검증								
	중간 보고 및 완료 보고								

<표 1> 사업 공정표



제 2 장

사업 수행 내용



제2장 사업 수행 내용

1. 신문 기사 말뭉치 구축

가. 매체 선정 및 계약

신문 기사 말뭉치 구축 사업은 대규모 신문 기사를 안정적으로 확보함과 동시에 저작권 침해 우려를 사전에 차단하는 것이 핵심 과제였다. 특히 올해는 사업이 통합 형태로 추진되면서 예산 효율성에 대한 요구가 높아졌으며, 신문 저작권 신탁 관리사인 한국언론진흥재단이 신문 말뭉치 사업에 지속적인 참여가 어렵게 되면서 저작권 계약 체계를 독자적으로 설계·운영해야 했다.

이에 수행사는 초기 단계부터 저작권 신탁을 거치지 않고 독립적으로 계약 체결이 가능한 비신탁 매체를 우선 식별하고 접촉하였다. 또한 기사 수와 공급 가능 기간을 기반으로 한 정량적 검토, 지역적·주제적 분포를 고려한 정성적 평가를 병행하여 매체 선정 기준을 고도화하였다.

주요 검토 항목은 다음과 같다.

- ▶ 저작권 독자 계약 가능 여부(비신탁 여부)
- ▶ 충분한 기사 확보 가능성과 공급 안정성
- ▶ 지역별·유형별 언론 다양성 확보
- ▶ 예산 대비 기사 단가 및 효율성

사전 검토를 통해 약 40여 개 매체를 후보군으로 선정하고, 단계별 협의 절차를 추진하였다. 이 과정에서 주요 매체의 참여 의사를 확인함으로써 말뭉치 구축을 위한 초기 기반을 마련하였다.

또한 지역 언론 생태계의 균형성과 언어 다양성 확보라는 사업 취지를 반영하여 지역 일간지를 중심으로 한 보완 전략을 수립·가동하였다. 단순한 물량 대체를 넘어, 지역 고유의 어휘·담론·문체를 반영하기 위해 다양한 권역의 매체를 우선 탐색하였다.

또한 예산 대비 기사 생산 효율이 높은 매체를 전략적으로 확보함으로써, 예상 손실분을 효과적으로 보전하였다. 이를 통해, 어절 수 확보와 더불어 지역 언어의 특성과 주제 다양성을 동시에 확장하는 이중의 성과를 달성하였다.

최종적으로 전국 종합지를 포함한 총 20개 매체와 저작권 이용 허락 계약을 체결하였으며, 1.5억 어절 이상의 기사 확보 기반을 마련하였다. 계약 체결 시 2036년 12월 31일까지의 이용 허락 및 연차 자동 갱신 조건을 설정하여 말뭉치의 장기적 활용 기반을 확보하였다.

구분	2025년 신문 기사 원문 자료 수집 및 정제 매체(총 20개)
인터넷(3)	뉴스투데이, 오마이뉴스, 오에스이엔
종합지(3)	경향신문, 디지털타임스, 일간투데이
지역지(11)	경기도민일보, 경남매일, 경인매일, 굿모닝충청, 대경일보, 대전투데이, 울산제일일보, 일간경기 전남매일, 충남일보, 충청매일
통신사(3)	뉴스1, 뉴스웍스, 뉴시스

<표 2> 선정된 매체 구분

나. 데이터 수집

신문 기사 말뭉치는 사업 목표인 1.5억 어절 이상 구축을 달성하기 위해 총 20개 매체로부터 원문을 수집하였으며, 최초 수집된 기사와 어절 수(본문과 제목 합계)는 각각 1,494,687건과 326,832,723어절이다.¹⁾

수집된 어절 수는 정제 전 데이터를 기준으로 산출하였으며, 문장 내 공백과 줄바꿈을 집계 기준으로 삼았다. 이후 정제 단계에서는 HTML 태그, 광고 스크립트, 표, 이미지 경로 등 비텍스트 요소를 일괄 삭제하여 본문만을 남겼다.

1) 원시 데이터 특징 분석

신문 매체는 기사 본문에 사용되는 요소(엘리먼트) 구조와 편집 체계가 매체별로 상이하다. 동일한 이름의 요소라도 어떤 매체에서는 본문 내용의 일부로 사용되지만, 다른 매체에서는 기자 정보나 광고 문구 등에 활용되는 경우가 있다. 이러한 구조적 차이를 고려하지 않을 경우, 본문 누락 또는 불필요한 정보 잔존과 같은 데이터 정제 오류가 발생할 수 있다.

이에 수행사는 매체별 기사 구조를 사전에 정밀 분석하여, 본문 텍스트를 구성하는 핵심 내용만 선별·추출하고, 비본문 영역(광고, 저작권 문구, 반복된 기사 헤더 등)은 선택적으로 제거하였다. 이 과정을 통해 매체별 특성을 유지하면서도 전체 기사 데이터의 무결성과 일관성을 확보한 본문 중심 데이터셋을 구축하였다.

a 엘리먼트의 삭제

(서울=뉴스1) 장아름 기자그룹 업텐션 출신 가수 겸 배우 김우석이 오랜 시간 동행해 온 소속사와 전속계약을 마무리하는 소감을 밝혔다. 그는 배우 전문 매니지먼트사인 블리즈웨이에서 새롭게 출발한다.(관련 기사 2024년 3월 5일 오전 뉴스1 단독 보도...[단독] 김우석, 10년 몸담은 티오피 떠난다...주지훈 소속사 行)

<표 3> 데이터 특징 예시 (a 엘리먼트)

1) 전체 수집 결과는 제3장 1. 신문 기사 구축 결과 <표 2> 신문 기사 수집 총괄표(60쪽) 참조.

이러한 오류는 기사를 세밀하게 검토하더라도 발견하기 어려우므로, 작업 초기 단계에서 각 매체의 구조적 특성을 사전에 분석하고 이를 반영하여 데이터를 처리하는 것이 중요하다. 이를 통해 보다 정확하고 신뢰성 높은 데이터를 생성할 수 있었다.

다. 데이터 1차 정제

1) 중복 기사, 유사한 데이터 제거

중복 및 유사 기사 제거는 원시 데이터 수령 이후 단계적으로 진행하였다. 먼저 전체 매체를 대상으로 본문 내용이 동일한 기사를 식별하고, 중복이 확인된 경우 최초 게재 기사를 기준으로 나머지를 제외하였다.

유사 기사 처리는 동일 매체 내에서 동일 이슈에 대한 반복 보도를 대상으로 수행하였다. 기사 게시 시점을 기준으로 전후 14일 범위에서 본문 유사도 검사를 실시하였으며, 본문 유사도가 85% 이상인 기사는 학습 편향과 표현 중복 가능성을 줄이기 위해 제외하였다.

이 과정을 통해 동일 매체의 반복 기사로 인한 데이터 밀집 현상을 방지하고 말뭉치의 다양성과 대표성을 확보하였다.

※ 좌우 비교를 위해 줄바꿈 임의 삽입

제목: 올드보이서 논란 인물까지 귀환...‘혁신 공천’ 빛바랜 민주당	제목: 올드보이·사법리스크 연루 의원 출마에... 커지는 민주당 공천 갈등
정동영 상임고문은 22일 YTN 라디오 인터뷰에서 당내 올드보이 불출마 요구에 대해 “오히려 한국 정치에는 OB(올드보이)가 너무 적다”며 “국회는 노·장·청 조화가 필요하다. 신병으로 전투력이 생기겠나”라고 반박했다. 전북 전주병에서 5선에 도전하는 정 상임고문은 지난 20일 선거사무소 개소식을 열고 “이재명 대표 지킴이의 선봉장이 되겠다”고 밝혔다. 전주병 현역 의원은 김성주 민주당 의원이다. 임종석 전 대통령비서실장은 이날 CBS 라디오에서 이번 총선에 출마하는 이유에 대해 “기조를 수정할 생각이 1도 없는 윤석열 정부의 폭주를 힘으로 멈춰 세우는 방법밖에 없다는 결론을 내렸다”고 말했다. 586 용퇴론에는 “특정 세대가 특정 세대를 배제하는 것도 뺄셈 정치로 가기 때문에 총선에서 윤석열 정부를 심판해야 한다는 대의와는 어긋난다”고 반박했다. 임 전 실장은 서울 중구성동구갑 출마를 준비하고 있다. 당 지도부는 홍익표 현 원내대표가 불출마를 선언한 중구성동구갑을 전략선거구로 지정했다. 민주당은 전략선거구에 청년과 여성을 우선 공천하기로 했다.	정동영 민주당 상임고문은 22일 YTN 라디오 인터뷰에서 당내 올드보이 불출마 요구에 대해 “오히려 한국 정치에는 OB(올드보이)가 너무 적다”며 “국회는 노·장·청 조화가 필요하다. 신병으로 전투력이 생기겠나”라고 반박했다. 전북 전주병에서 5선에 도전하는 정 상임고문은 출마 명분으로 ‘이재명 지킴이’를 자처하고 있다. 그는 지난 20일 전주에서 선거사무소 개소식을 열고 “이재명 대표 지킴이의 선봉장이 되겠다”고 밝혔다. 전주병 현역 의원은 김성주 민주당 의원이다. 임종석 전 대통령 비서실장은 이날 CBS 라디오에서 이번 총선에 출마하는 이유에 대해 “기조를 수정할 생각이 1도 없는 윤석열 정부의 폭주를 힘으로 멈춰 세우는 방법밖에 없다는 결론을 내렸다”고 말했다. 586 용퇴론에 대해서는 “특정 세대가 특정 세대를 배제하는 것도 뺄셈 정치로 가기 때문에 총선에서 윤석열 정부를 심판해야 한다는 대의와는 어긋난다”고 반박했다. 임 전 실장은 서울 중구성동구갑 출마를 준비하고 있다. 당 지도부는 홍익표 현 원내대표가 불출마를 선언한 중구성동구갑을 전략선거구로 지정했다. 민주당은 전략선거구에 청년과 여성을 우선 공천하

사범 리스크가 있거나 성비위·막말 논란에 휩싸인 인사들도 비명계 현역 의원 지역구에 출마를 선언했다. 이수진 의원(비례)은 이날 “이재명 대표의 심장을 뺏길 수 없다”며 비명계 운영찬 의원 지역구인 경기 성남 중원 출마를 발표했다. 전날 서울 서대문갑 출마를 철회한 지 하루 만이다. 이 의원은 라임 사태 핵심 인물로 꼽히는 김봉현 전 스타모빌리티 회장으로부터 금품을 받은 혐의로 재판받고 있다. 미투 논란에 휩싸였다가 명예훼손 혐의로 무죄 판결을 받은 친명계 정봉주 전 의원은 비명계 박용진 민주당 의원 지역구인 서울 강북을 에 출사 표를 냈다.	기로 했다. ‘사범 리스크’가 있거나 성비위·막말 논란에 휩싸인 인사들도 비명계 현역 의원 지역구에 출마를 선언했다. 이수진 의원(비례)은 이날 “이재명 대표의 심장을 뺏길 수 없다”며 비명계 운영찬 의원 지역구인 경기 성남 중원 출마를 선언했다. 전날 서울 서대문갑 출마를 철회한 지 하루 만이다. 이 의원은 라임 사태 핵심 인물로 꼽히는 김봉현 전 스타모빌리티 회장으로부터 금품을 받은 혐의로 재판받고 있다. 미투 논란에 휩싸였다가 명예훼손 혐의로 서 무죄 판결을 받은 친명계 정봉주 전 의원은 비명계 박용진 민주당 의원 지역구인 서울 강북을 출마를 선언했다.
---	--

<표 4> 유사도 비교를 통해 사용하지 않는 기사의 예

2) 기사 선택

① 기사 품질 및 길이 기준

- 어절 수가 100어절 미만인 기사
- 정제 전에도 최소 어절 수 충족이 어려운 것으로 판단되는 기사

② 기사 형식 요건 미충족

- 단순 광고, 이벤트 안내, 운세, 퀴즈 등 비기사성 콘텐츠
- 통계 수치 나열, 승진·부고 명단 등 정보 전달 위주 텍스트
- 스포츠 스코어 등 결과 수치만 포함된 기사
- 문학 작품(시, 소설 등)과 유사한 비보도형 텍스트

③ 언어 및 문체 기준

- 본문의 대부분이 영어·일본어 등 외국어로 구성된 기사
- 전체 문장이 구어체(“~했어요”, “~한대요” 등)로 작성된 기사

④ 저작권 및 출처 정책

- 저작권 이용에 문제가 될 소지가 있는 기사
- 번역 기사(기관 협의 기준 적용)

⑤ 기자 정보 및 출처 검증

- 기자 정보가 없는 기사(사설 제외)
- 외부 기고(대학생 기자, 시민기자, 타 기관 전문가 등)
- 기자 정보가 동일 계열사나 비매체 소속으로 확인된 경우
- 공동취재단 명의 기사

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 (전국체전 공동취재단, 외교부 공동취재단 등)
각 매체	특별취재팀	삭제	해당 매체와 계약 등을 일일이 확인 불가 (충선특별취재팀 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	라이프 기자	삭제	해당 매체에서 위촉한 다양한 직업군의 일반인
각 매체	해당 매체명	확인 및 삭제	url을 직접 확인하여 해당 매체 저작권 기사 여부 확인

<표 5> 저작권 이용 문제로 인해 사용하지 않는 기사의 특징

매체 소속 기자가 작성한 기사의 저작권은 해당 매체에 귀속되지만, 외부 기고가가 작성한 기사는 매체와의 계약 형태에 따라 권리 귀속이 달라질 수 있다. 따라서 저작권 귀속이 명확하지 않거나 권리 관계에 논란의 소지가 있는 기사는 모두 사용 대상에서 제외하였다.

일부 매체는 시민기자나 외부 전문가가 기사 작성에 참여하는 개방형 구조로 운영된다. 이러한

기사는 매체 측의 이용 허락을 받은 경우에도 공공 말뭉치로 활용 시 저작권 분쟁이 발생할 수 있어 전면 제외하였다.

또한 기자 이름이 표기되어 있으나 실제로는 외부 기고 형식으로 작성된 기사도 존재한다. 이러한 기사는 본문 내용 검토 과정에서 직접 확인 후 제외하였으며, 기자 정보가 불명확하거나 공동 취재단 명의로 기사 역시 동일한 기준에 따라 처리하였다.

이와 같이 본 사업은 저작권 분쟁의 가능성을 사전에 원천 차단하고, 권리 관계가 명확히 확인된 기사만을 선별하여 구축하였다.

3) 문자 및 기호 코드 통일

신문 기사 원문 데이터는 매체별 작성 환경, CMS 시스템 차이, 입력 방식의 다양성 등으로 인해 동일한 문장 내에서도 공백, 전각 문자, 특수 기호, 가운뎃점, 한자 코드 등이 혼재되어 있다. 이러한 문자 체계 불일치는 검색 및 색인 과정에서 오류를 유발하고, 형태소 분석 및 토큰화 과정의 일관성을 저해하며, 인공 지능 학습 데이터로 활용할 때 불필요한 잡음과 모델 성능 저하를 초래할 수 있다.

이에 본 사업에서는 원문의 의미를 변경하지 않는 범위 내에서 문자 및 기호 코드를 정규화하여 데이터의 일관성과 학습 안정성을 확보하였다. 정규화 대상은 전각·반각 문자, 스페이스 및 공백 문자, 한·중·일 호환용 한자 영역(F900-FAFF) 문자 등이며, 이를 모두 표준 유니코드 규격에 맞게 변환하였다. 특히 가운뎃점은 매체별로 다양한 코드가 혼용되어 있어(U+318D, U+22C5, U+30FB 등) 이를 단일 표준 문자(U+00B7)로 통일하였다.

① 한·중·일 호환용 한자 영역(F900-FAFF) 한자의 통일

❖ 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	果	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

코드	한자	치환	코드	한자	치환	코드	한자	치환	코드	한자	치환
F97E	量	91CF	F9FA	狀	72C0	F966	復	5FA9	F9E0	易	6613
F914	樂	6A02	F99A	連	9023	F905	串	4E32	F989	黎	9ECE
F91F	蘭	862D	F9A3	念	5FF5	F912	裸	88F8	F999	蓮	84EE
F94C	樓	6A13	F9CA	流	6D41	F915	洛	6D1B	F99B	鍊	934A
F902	車	8ECA	F988	麗	9E97	F916	烙	70D9	F99C	列	5217
F940	鹿	9E7F	F9C1	療	7642	F91A	駱	99F1	F99F	烈	70C8
F90F	羅	7F85	F997	聯	806F	F91D	欄	6B04	F9A2	廉	5EC9
F92E	冷	51B7	F9AE	瑩	7469	F949	雷	96F7	F9C9	柳	67F3
F972	沈	6C88	F9E7	裏	88CF	F955	凌	51CC	F9D1	六	516D
F92D	來	4F86	F9AB	嶺	5DBA	F976	略	7565	F9F1	隣	96A3
F97A	梁	6881	F9F6	臨	81E8	F90D	懶	61F6	F990	戀	6200
F918	落	843D	F99D	劣	52A3	F923	藍	85CD	F9A9	囹	56F9
F932	爐	7210	F9B2	零	96F6	F942	壘	58DF	F9C3	遼	907C
F984	瀘	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	荼	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	聾	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 6> ‘한·중·일 호환용 한자 영역’ 한자 치환 표

② 문장 부호 등 통일

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	₩	
FFE6	₩	20A9	₩	
FFEB	→	2192	→	
FF62	「	300C	「	
FF63	」	300D	」	
3000		0020		공백
0009		0020		공백
00a0		0020		공백

대상 코드	대상 문자	대상 코드	치환 문자	비고
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 7> 치환 코드 목록

이와 같은 문자 코드 통일 과정은 공공 말뭉치의 일관성, 신뢰성, 그리고 기계 처리 효율성을 향상시킬 뿐만 아니라, 향후 국가 언어자원 확충 및 인공 지능 학습 데이터로의 재활용 과정에서 재처리 비용과 품질 손실을 최소화한다는 점에서 중요한 의미를 지닌다.

라. 데이터 2차 정제

2차 정제는 인적 검수 중심의 핵심 공정으로, 가장 많은 인력이 투입되는 단계이다. 작업자가 기사를 검수하여 불필요한 요소를 최종 제거하며, 1차 정제에서 자동 및 반자동 처리된 결과를 기반으로, 기사 내용의 적합성·정확성·저작권 안정성을 재확인한다.

1) 저작권 검증 및 제외

다음 기준에 해당하는 기사는 저작권 위험 예방을 위해 불용 처리하였다.

- ▶ 외부 기고가(교수·의원·객원 기자 등)가 작성한 기사
- ▶ 방송 내용을 단순 전사한 기사
- ▶ 기자명이 있으나 실제 외부 기고 형식으로 작성된 기사
- ▶ 기자 정보가 본문에만 존재하고 출처가 명확하지 않은 기사
- ▶ 공동취재단 명의 기사

[군산·서울=] 기자 국방부 공동취재단 = 한국과 미국 양국은 4월 12일부터 26일까지 약 2주간 전북 군산기지에서 2024년 '연합편대군 종합훈련'(KFT)을 진행하고 있다. 한반도에서 실시되는 최대 규모의 연례 연합공중훈련인 KFT에는 25종 100여대의 항공전력과 장병 1400여명이 투입됐다.

19일 오후 국방부 공동취재단이 공군 군산기지 활주로에 들어서자 각종 계측장비와 연결된 채 출격전 점검을 하고 있는 미 공군의 무인공격기 MQ-9 리퍼가 취재진을 맞았다. 하늘의 암살자라 불리는 리퍼가 KFT에 참여한 것은 이번이 최초다.

정부가 조만간 내놓기로 한 기업 밸류 업 프로그램은 단기 처방에 그쳐서는 안 된다. 기업 체질을 개선하고 장기적으로 증시 투자자를 붙잡아둘 수 있는 근본 대책이라야 한다. 총선을 앞두고 인위적인 주가 부양만 꾀했다가는 시장을 왜곡해 심각한 후유증을 남기게 된다. 한국은 가계부채 비율이 GDP 대비 105.1%, 가처분소득 대비 203.7로 세계 상위권 가계부채국이다. 정부가 서둘러 내놓은 증시부양책이 빚 투자를 유인하지 않을까 우려스럽다. 저평가 원인 중 빼놓을 수 없는 것은 북한과 대치하고 있는 지정학적 리스크다. 북한을 자극하지 않고 상생 방안을 찾아야 한다. 최근 상황을 보면 정부가 오히려 리스크를 키우는 것만 같다.

사회경제연구원장

그림 3. 사용하지 않는 기사의 예 (저작권)

2) 품질 검증 및 제외

저작권 문제가 없더라도, 완결성을 갖추지 못한 기사는 활용 대상에서 제외하였다.

- ▶ 본문이 단절되어 문맥 이해가 불가능한 기사
- ▶ 문장이 과도하게 누락되거나 파손된 기사
- ▶ 동일 문단의 반복, 비문 등 오류가 다수 포함된 기사
- ▶ 캡션·링크·오류 문장만 남아 내용이 사실상 존재하지 않는 기사
- ▶ 정제 과정에서 작업자가 단순 오류 수정이 아닌 주관적으로 의미를 보완하거나 문장을 새로 작성해야 하는 기사

불법 스포츠 도박 단속의 실효성을 확보하기 위한 제도적인 장치가 필요하다는 의견이 나온다. 현재 불법 도박 사이트 운영자가 복제 사이트를 개설하는데는 약 1 한편 자원순환교육은 매주 수, 목 오전 10시부터 12시까지 진행된다. 매주 수요일은 관내 학생 대상, 매주 목요일은 관내 단체, 기업, 주민을 대상으로 교육을 실시한다. <u>교육 신청은 양평군청 청</u> 기사가 도중에 잘려 내용이 없거나 마무리 짓지 않은 기사는 사용할 수 없음
이번 접종에 활용되는 백신은 XBB 1.5 단가백신으로 해당 백신은 코로나19 유행변이에 대응해 새롭게 개발됐으며, 그간 1·2차 접종을 완료하지 않은 사람도 과거 접종력과 관계없이 1회 접종으로 완료된다. <u>및 마지막 코로나19 예방접종일로부터 3개월 이후 지정된 의료기관 및 보건소에서 접종이 가능하다.</u> 문장 수정이 불가능한 경우 사용하지 않음
<u>[작성중]</u> <u>인턴 기자</u> = 서울 성동구 용답동의 한 도로에서 경찰의 음주측정을 거부하고 도주하던 운전자가 추격 전 끝에 검거됐다. 데이터를 수정할 수 없는 경우 사용할 수 없음

그림 4. 사용하지 않는 기사의 예 (품질)

3) 불필요한 본문 내용 삭제

기사와 직접 관련 없는 정보는 모두 삭제하였다. 이러한 정보는 기사 본문의 흐름을 방해하거나 맥락을 왜곡할 수 있으며 인공 지능 학습 데이터의 품질을 저하시킬 우려가 있기 때문이다.

또한 기사 내에 연설문, 입장문, SNS 게시물 등 외부에서 작성된 전문(全文)이 포함된 경우, 이는 기사가 작성한 기사로 보기 어렵고 저작권 문제가 발생할 가능성이 있으므로 제거하였다.

외부 전문을 삭제할 때는 뒤이어 이어지는 문장이 문맥상 자연스럽도록 검토하였으며, 전문의 존재를 언급하는 문장 또한 기사 흐름 유지에 지장을 줄 경우 함께 삭제하였다.

이 과정을 통해 기사 본문만을 보존하고, 외부 텍스트로 인한 의미 왜곡을 방지하였다.

삭제 정보	예시
표, 그림, 그래프 등의 캡션 정보	[사진] '1박 2일' 시즌4 방송화면 캡처 ○○○ 단지 내 어린이집 개원식. 사진제공=○○구청 사진 설명: '제53회 전국소년체육대회' 참여한 세종대표선수들이 11일에 열린 시상식에 참여해 기념 촬영을 하고 있다.
기자의 이름, 계정(ID) 등의 정보	[○○○○=제주, ○○○ 기자] [서울=○○일보] 외교부 공동취재단·○○○ 기자 = 【서울=○○○】○○○ 기자 = /○○○ 기자·연합뉴스
저작권 관련 내용 ('Copyright©' 등)	마이크 존슨 미 하원의장. 2024.03.07 © 로이터 [사진]©Gettyimages(무단전재 및 재배포 금지)
전문	조교로 국방의 의무를 다했던 진은 특급전사를 따면서 두 번의 조기진급 소식을 전해 화제를 모았다. 모범적인 군 생활 중에서도 틈틈이 위버스에 등장해 팬들과 소통을 이어가기도 했다. 동생 멤버들이 차례로 입대할 땐 센스있는 반응으로 팬들을 웃음짓게 만들기도. 그랬던 진이 이제 곧 민간인이 된다. 소속사로서는 오랜만에 만나는 진과 팬들을 위해 특급 이벤트를 마련했지만 의도치 않게 원성이 커지자 재빨리 피드백을 수용했다. <u>다음은 방탄소년단 진 측 입장 전문이다.</u> <u>안녕하세요. 빅히트 뮤직입니다. '2024 FESTA' 진 오프라인 행사의 응모 기준이 변경 예정인 점 안내 드립니다.</u> <u>앞선 공지 내용이 불명확했던 점과 더불어 미리 세심하게 응모 기준을 설정하지 못한 점 사과 드립니다.</u> <u>금일(6/2) 오전 11시(KST) 오프라인 행사 응모 공지 이후 weverse shop 상세 페이지를 통해 오프라인 행사에 응모하신 앨범 구매자분들과 공지 이전 weverse shop에서 앨범을 구매한 이력이 있으신 분들 모두를 고려하는 방향으로 응모 기준을 보완할 예정입니다. 충분히 숙고한 뒤 빠른 시일 내 상세 내용을 공지하도록 하겠습니다.</u> <u>진의 마음이 팬 여러분께 잘 전달될 수 있게 최선을 다해 행사를 준비하겠습니다. 팬 여러분께 불편을 끼쳐 드린 점 다시 한번 사과의 말씀 드리며 오는 6월 13일 진행될</u>

삭제 정보	예시
	<u>진의 오프라인 행사에 많은 응원과 관심 부탁드립니다 감사합니다.</u> <u>/-----@----.co.kr</u> <u>[사진] ----- DB</u>
문장으로 볼 수 없는 정보	팔레스타인전은 2014년 7월 사퇴한 뒤 10년 만에 A대표팀 사령탑으로 돌아온 홍명보 감독의 복귀전이다. 다만 이를 대비해 2일 진행될 첫 훈련은 완전체로 진행되지 않는다. KFA는 "선수단은 오후 2시까지 고양종합운동장 인근 호텔로 소집된다"라며 "오늘 훈련은 K리거 12명을 포함해 총 19명이 훈련할 예정이다"라고 전했다. 2일 새벽 소속팀에서 경기를 치른 김민재, 이강인을 비롯해 손흥민, 황희찬, 황인범, 설영우, 이한범 등 해외파는 뒤늦게 참여한다. 이에 KFA는 "사정에 따라 인원은 변동될 수 있으며 그외 대부분의 선수는 2일 저녁 또는 3일 중 순차적으로 입국할 예정이다"라고 공지했다. 한편 KFA는 지난 달 30일 "권경원(32, 코르파칸 클럽) 좌측 종아리 부상으로 조유민(27, 샤르자FC)을 대체 발탁했다. 김문환(29, 대전하나) 좌측 내전근 부상으로 황재원(22, 대구FC, 이상 수비수)을 대체 선발 했다"라고 알렸다. 그러면서 "두 선수가 심각한 부상은 아니나 소집기간 최상의 컨디션으로 경기 출전이 어려울 것으로 판단돼 금번 소집에는 대체 됐다"라고 덧붙였다. <u>'홍명보호' 26명 소집 명단</u> <u>▲골키퍼</u> <u>조현우(울산HD), 송범근(쇼난벨마레), 김준홍(전북현대)</u> <u>▲수비수</u> <u>조유민(샤르자FC), 김민재(바이에른 뮌헨), 김영권(울산HD), 정승현(알 와슬FC), 이한범(미트윌란), 설영우(츠르베나즈베즈다), 이명재(울산HD), 최우진(인천 유나이티드), 황문기(강원FC), 황재원(대구FC)</u> <u>▲미드필더</u> <u>박용우(알 아인), 정우영(울산HD), 황인범(츠르베나즈베즈다), 이재성(마인츠), 손흥민(토트넘), 황희찬(울버햄튼), 이동경(김천상무), 정호연(광주FC), 양민혁(강원FC), 엄지성(스완지시티), 이강인(파리 생제르맹)</u> <u>▲공격수</u> <u>주민규(울산HD), 오세훈(마치다젤비아)</u> <u>/-----@----.co.kr</u>
	이날 첫 시상, 신인상의 주인공은 이성민이었다. 역시 참석하지 못한 이성민은 영상으로 "놀랍다. 신인상을 30년 전 연극할 때 받아본 이후 신인상을 받아본 적 없다. 내 인생에 신인상은 없다 생각했는데 느닷없이 이런 상을 주셔서 너무 감사하다"라며 기뻐했다. 특히 그는 가정집 벽지를 배경으로 하면서도 턱시도까지 갖춰 입고 진지하게 소감을 밝혔다. 그는 "이제 좀 인생이 나른해지는 그런 나이인데, 신인상을 받고 나니 새로운 목표가 생기고 활기가 생길 것 같다. 앞으로 기회가 된다면 이 분야에 좀 더 노력을 해서 인기상, 대상을 받을 수 있도록 한 번 해보겠다"라며 "2024년 그 어떤 상보

삭제 정보	예시
	다도 저에게 큰 에너지를 주는 상을 주셔서 너무 행복한 한 해를 마무리할 수 있을 것 같고 큰 상 주셔서 대단히 감사하다"라고 힘주어 말했다. 그런가 하면 유재석은 '특별대상'을 받았다. '핑크고' 모든 에피소드에서 활약한 그를 위해 제작진이 별도의 트로피를 마련한 것. 이에 "요새 제일 핫한 시상식", "방송사 안 부럽다"라는 호평을 받은 '핑크고 시상식'이 성대하게 마무리돼 의미를 더했다. <u>다음은 '제2회 핑계고 시상식' 수상자(작) 명단이다.</u> <u><대상> 황정민</u> <u><작품상> '돌아온 설 연휴는 핑계고'</u> <u><최우수상> 이동휘</u> <u><우수상> 이상인</u> <u><베테랑 인기스타상> 지석진, 홍진경</u> <u><라이징 인기스타상> 키, 김고은</u> <u><신인상> 이성민</u> <u><특별대상> 유재석</u> <u>[사진] 유튜브브 출처.</u>
기사와 관련 없는 내용	이경민 대신증권 연구원은 "대부분의 불확실성을 확인한 뒤 드디어 시장이 호재를 호재로 받아들이기 시작했다"며 "다음주 FOMC와 BOJ를 확인할 때 까지 불확실성이 잔존해 있으나 모든 불확실성과 변동성, 특히 엔화 강세의 영향을 확인한 뒤 저점을 확인할 시기가 가까워지고 있다"고 평가했다. <u>거래일</u> <u>wait for business day</u>
문장의 오류	도는 목표 달성을 위해 위해 신속 집행·소비투자 공통비목인 시설비(400) 중심으로 대규모 투자사업에 대한 선금 추가집행과 이월·계속비 사업 등 우기 전 상반기 최대 집행이 될 수 있도록 각별한 의지를 가지고 총력 대응하기로 했다.

<표 8> 불필요한 요소 제거 내용

작업 파일 데이터	정제 데이터
<u>[서울=○○일보]○○○ 기자 = <14시부터 보도 가능합니다></u> 대전시와 대전청년내일재단은 30일 대전시청 중회의실에서 각 분야 대표 장학생과 학부모가 참석한 가운데 ‘2024년 상반기 대전인재육성장학증서수여식’을 가졌다. (중략) 시는 그동안 대전인재육성장학재단을 통해 2009년부터 지금까지 총 1만 6명의 학생과 청년에게 82억 원의 장학금을 지원해 왔는데, 지난 6월 인재육성장학재단이 대전청년내일재단으로 새롭게 출범 함에 따라, 기존 장학재단의 업무는 물론 청년지원과 성장을 위한 다양한 정책을 발굴하여 추진할 계획이다. <u>-----, -----@-----.com</u>	대전시와 대전청년내일재단은 30일 대전시청 중회의실에서 각 분야 대표 장학생과 학부모가 참석한 가운데 ‘2024년 상반기 대전인재육성장학증서수여식’을 가졌다. (중략) 시는 그동안 대전인재육성장학재단을 통해 2009년부터 지금까지 총 1만 6명의 학생과 청년에게 82억 원의 장학금을 지원해 왔는데, 지난 6월 인재육성장학재단이 대전청년내일재단으로 새롭게 출범 함에 따라, 기존 장학재단의 업무는 물론 청년지원과 성장을 위한 다양한 정책을 발굴하여 추진할 계획이다.

<표 9> 원시 데이터와 정제된 데이터 비교 1

작업 파일 데이터	정제 데이터
김해시가 여름철 물놀이형 수경시설과 계곡 물놀이 지역 관리 계획을 수립하고 본격 관리에 돌입했다. 시에 따르면 물놀이장 수질과 시설 관리 전반을 점검해 안전한 물놀이 공간을 제공, 시민건강을 보호한다. 시는 물놀이형 수경시설 61곳과 주요 계곡 물놀이 지역 3곳을 대상으로 이달에는 운영상황을 점검하고 오는 6월부터 수질검사 등 현장점검을 추진한다. (중략) 김해 대표 계곡 물놀이 지역인 대청계곡, 신안계곡, 장척계곡의 수질 관리를 위해 6월은 2주에 1회, 7~8월은 주 1회, 9월은 1회 대장균 수질 검사를 해 권고기준을 초과하게 되면 물놀이 제한, 대시민 공지, 오염 원인 조사와 조치로 안전한 물놀이 환경을 조성한다. <u>○○○ 기자 김해시가 여름철 물놀이형 수경시설과 계곡 물놀이 지역 관리 계획을 수립하고 본격 관리에 돌입했다. 시에 따르면 물놀이장 수질과 시설 관리 전반을 점검해 안전한 물놀이 공간을 제공, 시민건강을 보호한다. 시는 물놀이형 수경시설 61곳과 주요 계곡 물놀이 지역 3곳을 대상으로 이달에는 운영상황을 점검하고 오는 6월부터 수질검사 등 현장점검을 추진한다.</u> (내용 반복 중략) <u>김해 대표 계곡 물놀이 지역인 대청계곡, 신안계곡, 장척계곡의 수질 관리를 위해 6월은 2주에 1회, 7~8월은 주 1회, 9월은 1회 대장균 수질 검사를 해 권고기준을 초과하게 되면 물놀이 제한, 대시민 공지, 오염 원인 조사와 조치로 안전한 물놀이 환경을 조성한다.</u>	김해시가 여름철 물놀이형 수경시설과 계곡 물놀이 지역 관리 계획을 수립하고 본격 관리에 돌입했다. 시에 따르면 물놀이장 수질과 시설 관리 전반을 점검해 안전한 물놀이 공간을 제공, 시민건강을 보호한다. 시는 물놀이형 수경시설 61곳과 주요 계곡 물놀이 지역 3곳을 대상으로 이달에는 운영상황을 점검하고 오는 6월부터 수질검사 등 현장점검을 추진한다. (중략) 김해 대표 계곡 물놀이 지역인 대청계곡, 신안계곡, 장척계곡의 수질 관리를 위해 6월은 2주에 1회, 7~8월은 주 1회, 9월은 1회 대장균 수질 검사를 해 권고기준을 초과하게 되면 물놀이 제한, 대시민 공지, 오염 원인 조사와 조치로 안전한 물놀이 환경을 조성한다.

<표 10> 원시 데이터와 정제된 데이터 비교 2

4) 오류 및 문장 수정

본문을 읽으며 불필요한 요소를 제거하고 오타나 기타 오류 등을 수정하여 사용할 수 있는 기사들은 수정을 진행하여 기사를 사용하였다.

의미없는 단어

설명한 부산역장은 "용기를 내어주셔서 직원 모두가 감동했다"며 "앞으로도 고객들의 신뢰와 사랑을 받는 부산역이 될 수 있도록 노력하겠다"고 말했다. 정석정 삭제

기자 이름/본문 구분 불가

삭제 [일간투데이 신형수 기자] 국민의힘은 17일 이재명 더불어민주당 대표가 재판지연 전략을 구사하고 있다면서 사법부의 신속한 재판을 촉구했다.

전체 기사 반복

김해시가 여름철 물놀이형 수경시설 관리계획을 수립하고 본격 관리에 돌입했다.

시에 따르면 물놀이형 수경시설 관리 전반을 점검해 안전한 물놀이 공간을 제공, 시민건강을 보호한다. 시는 물놀이형 수경시설 61곳과 주요 계곡 물놀이 지역 3곳을 대상으로 이달에는 운영상황을 점검하고 오는 6월부터 수질검사 등 현장점검을 추진한다.

수도물, 지하수 등을 인위적으로 저장, 순환해 이용하는 분수, 연못 등 인공 시설물 중 이용자의 신체와 직접 접촉해 물놀이를 하도록 설치한 물놀이형 수경시설은 △pH, 탁도, 대장균, 유리잔류염소 4개 항목에 대한 수질검사 △저류조 청소, 부유물침전물 제거 △소독시설, 살균소독제 투입 적정 여부 등을 점검한다.

시는 수질기준 초과와 시설 관리기준을 부적정하게 관리한 시설의 시설 개방을 중지하고 관리 점검 후 재개방하며 과태료 부과 등 행정조치를 시행한다.

김해 대표 계곡 물놀이 지역인 대창계곡, 신안계곡, 장척계곡의 수질 관리를 위해 6월은 2주에 1회, 7~8월은 주 1회, 9월은 1회 대장균 수질검사를 해 권고기준을 초과하게 되면 물놀이 제한, 대시민 공지, 오염 원인 조사와 조치로 안전한 물놀이 환경을 조성한다. 신청을 기사 김해시가 여름철 물놀이형 수경시설과 계곡 물놀이 지역 관리 계획을 수립하고 본격 관리에 돌입했다.

시에 따르면 물놀이형 수경시설과 시설 관리 전반을 점검해 안전한 물놀이 공간을 제공, 시민건강을 보호한다. 시는 물놀이형 수경시설 61곳과 주요 계곡 물놀이 지역 3곳을 대상으로 이달에는 운영상황을 점검하고 오는 6월부터 수질검사를 추진한다.

수도물, 지하수 등을 인위적으로 저장, 순환해 이용하는 분수, 연못 등 인공 시설물 중 이용자의 신체와 직접 접촉해 물놀이를 하도록 설치한 물놀이형 수경시설은 △pH, 탁도, 대장균, 유리잔류염소 4개 항목에 대한 수질검사 △저류조 청소, 부유물침전물 제거 △소독시설, 살균소독제 투입 적정 여부 등을 점검한다.

삭제

그림 5. 기사 수정 예 ①

동어 반복 수정

그는 "그는" 민주당이 190석의 의석을 앞세워 무제한 토론을 24시간마다 강제 중단시키는 것은 속의민주주의와 합의민주주의를 파괴하는 폭거"라며 "제 생각은 변함이 없다"고 했다.

이번 공연은 오는 오는 5월 4일 오전 10시30분과 오후 2시 총 2회차를 보령문화예술회관 대공연장에서 만나볼 수 있으며, 충남 보령시가 주최·주관, 기획은 나무예술기획, 충남일보의 후원으로 마련됐다.

한 선거운전자운동원 양경태(58)씨는 "바닥에 쓰러진 남성은 심장이 뛰지 않았고 혀가 말리면서 숨을 쉬지 못하는 위험한 상황이었다. 옆에 있던 시민과 CPR을 함께 하고 혈액순환이 잘 되도록 온몸을 주물렀다"면서 "그 자리에서는 제가 아니라 누구라도 뛰어들어 쓰러진 사람을 구하고 싶었을 것"이라고 말했다.

반복 되는 단어 하나로 수정

한글로 수정
(RPWPEHOTEK ▶ 게재됐다)

온라인 커뮤니티 '보배드림' 인스타그램에는 지난 30일 한 아파트 주민이 제보한 '실내 흡연 양해 요청' 메모 사진이 RPWPEHOTEK.

그림 6. 기사 수정 예 ②

오타 수정	세븐일레븐 비빔밥(22일), 롯데시네마 영화 관람권(23일), 롯데월드 어드벤처 입장권(24일) 등 롯데 계열사 상품을 최대 80% 할인 판매하는 '롯데 패밀리' 혜택도 선보인다. 이에 따라 만약 연 7% 고금리 신용대출을 5000만원 받은 소상공인이라면 연 4.5% 대환대출을 받을 경우 이자가 종전 연 350만원에서 연 225만원으로 125만원 경감된다. 마오 대변인은 "하나의 중국 입장은 거래 불가한 기본 준칙이자 국제사회의 보편적 합의"라며 "경제 대국인 중국과의 외교관계 회복으로 나우루에 전례없는 발전 기회를 가져올 것"이라고 덧붙였다. 지난해 우리나라의 GDP 대비 가계대출 비율은 100.8% 수준이다. 국제금융협회(IIF) 보고서에 따르면 지난해 3분기 기준 GDP 대비 가계부채 비율이 100%를 넘는 국가는 한국이 조사 대상 34개국 가운데 유일했다. 당국은 가계부채 증가율은 GDP 성장률 이내로 관리해 완만하게 비율을 줄여가겠다는 방침이다.	관람권 ▶ 관람권 대출 ▶ 대출 전례없는 ▶ 전례없는 가계대출 ▶ 가계대출
오타 수정	아버지는 아들을 바로잡아 보기 위해 미국 로스앤젤레스로 유학까지 보냈으나 사람 되기는 커녕 도박에 빠져 3700만 원 (현재 가치 3억 원대)의 빚을 졌다. 이번 주 투자자들의 시선은 미국 물가지수와 금리 결정을 향할 전망이다. 한국 경제가 두 분기 연속 역성장을 피했다. 코로나19 사회적 거리두기 종료 이후 민간 소비가 증거현 대 떠른 것이다, 지난 3월 11일 85명의 근로자와 33명의 농가주가 참여한 가운데 치러진 외국인 계절근로자 농가대면식에서 주한베트남대사관 따띠탄티 노무관사무소장은 연천군의 우수지자체 선정을 축하했으며, 대사관 측에서도 지속적인 관심으로 베트남 근로자 의 한국생활 적응을 돕겠다고 저 냉 쌀.	현재 가치 ▶ 현재 가치 향할 ▶ 향할 증거현 대 떠른 ▶ 증가한 데 따른 저냉쌀 ▶ 전했다

그림 7. 기사 수정 예 ③

```

<newsitemid>NISX20240229_0002644874</newsitemid>
<HeadLine>"7원에도 2000부 매진"...조선어사전, 삼일절 맞아 86년 만에 복간</HeadLine>
<url>https://www.newsis.com/view/NISX20240229_0002644874</url>
<used>Y</used>
<byline>이○○ 기자</byline>
<content>
<d>[서울=뉴시스] 이○○ 기자 = </d>최초 국어사전 조선어사전이 105주년 삼일절을 맞아 복간된다.
초판 발간 후 86년 만이다.
조선어사전은 우리말로 된 최초 국어사전으로 청람 문세영(1895~1952)이 편찬해 1938년 발간됐다.
최현배의 '우리말본'(1937), 김윤경의 '조선문자국어학사'(1938)와 함께 일제강점기 우리말 관련 3대
저술로 꼽힌다.
조선어학회가 1933년 제정한 '한글 맞춤법 통일안'에 의해 표기한 최초 사전이기도 해 당시의 표준어
보급에도 기여했다. 현재 온전한 실물은 거의 남아 있지 않다.
지식공작소는 "최초의 국어사전이 박물관에서도 구경하기 힘든 상황을 안타깝게 여겨 영인본으로 복
간하게 됐다"며 "국립한글박물관, 국립중앙도서관, 고려대학교 소장본과 비교·대조해 원형과 최대한
동일하게 재현했다"고 밝혔다.
이번에 복간된 '조선어사전' 영인본은 한글학회 '우리말 큰사전' 수석 편찬원을 지낸 국어학자 조재
수가 소장한 초판본을 저본으로 삼았다.
아니라 사고방식과 문화를 살필 수 있는 귀중한 자료다.
'모던걸', '모던보이' 등 근래 사전에는 수록되지 않은 신어가 실린 사례, '러버(Lover)'의 뜻풀이로
'마음속에 있는 사람. 戀人(연인)'을 제시하고 있다. '연인'은 올림말로 등재되지 않은 사례 등은 서구
문물이 유입되던 시대상을 짐작할 수 있게 한다.
1원이 넘는 책이 드물던 당시 7원에 달하는 값비싼 책이었음에도 초판 1000부, 재판 2000부가 매진
됐다. 당시 사람들이 품은 우리말 사전에 대한 열망이 얼마나 컸는지를 보여 준다.
<d>©공감언론 뉴시스 suejeeq@newsis.com</d>

```

그림 8. 작업 프로그램 화면

불필요한 요소 제거는 정규식 목록 적용과 수동 검수를 병행하여 수행하였다. 모든 기사는 XML 파일 단위로 처리하였으며, 정규식 규칙을 적용할 때 직접 삭제하지 않고 마크업 태그를 부여하는 방식으로 작업하였다. 이를 통해 제거 항목과 유지 항목을 명확히 구분하고, 후속 검수 및 오류 수정을 용이하게 하였다.

정제 과정에서는 'used' 속성을 활용하여 기사 사용 여부를 표시하였으며, 불용 처리된 기사는 시스템 상에서 명확히 분리하였다. 각 요소는 색상과 태그 형태로 시각화하여 작업자가 수정 사항을 직관적으로 확인하고 누락 없이 검수할 수 있도록 하였다.

작업 중 새롭게 발견되는 패턴이나 오류 유형은 실시간 공유하고 공용 정규식 목록에 반영하였다. 예를 들어 기사 말미의 기자명 탐지 정규식에서 철자 오류(예: '기자'→'기지', '가지')가 발견된 경우, 해당 패턴을 추가하여 유사 사례를 자동 탐지할 수 있도록 하였다. 이러한 방식으로 정제 규칙을 지속적으로 보완·축적하였다.

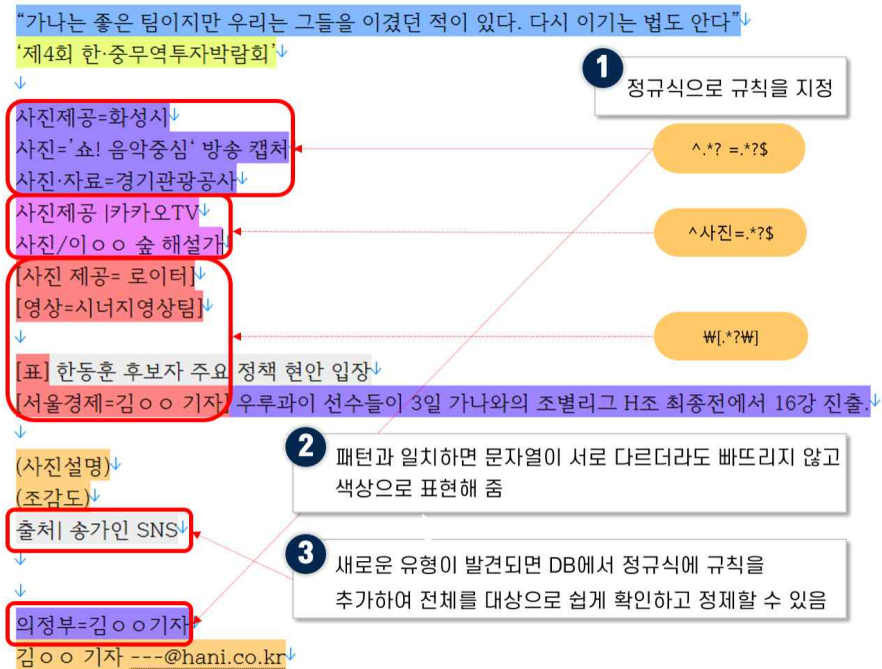


그림 9. 정규식을 통한 작업 예

2025년에는 전년도 사업에 참여했던 인력이 동일하게 참여하여, 별도의 추가 교육 없이 숙련된 방식으로 작업을 수행하였다. 정제된 데이터는 문자 수 비교와 내용 검수를 통해 본문이 소실되거나 누락되지 않았음을 확인하였으며, 소실이 발생한 경우 원본 기사와 대조하여 복구 작업을 진행하였다.

1차 정제를 마친 파일은 작업자가 자체 검수를 수행한 뒤 검수 전용 폴더에 업로드하였다. 이후 검수자가 2차 검수를 실시하였으며, 오류 유형 목록을 참고하여 항목별 확인 작업을 반복하였다. 검수 과정에서 오류가 발견될 경우 파일을 반려하고 즉시 피드백을 제공하였다. 동일한 유형의 오류가 반복될 때는 정규식 목록을 보완하거나 사내 공유 문서를 통해 작업 기준을 명확히 하였다. 이 과정을 통해 정제 기준의 일관성을 유지하고 데이터 품질의 안정성을 확보하였다.

작업 공정

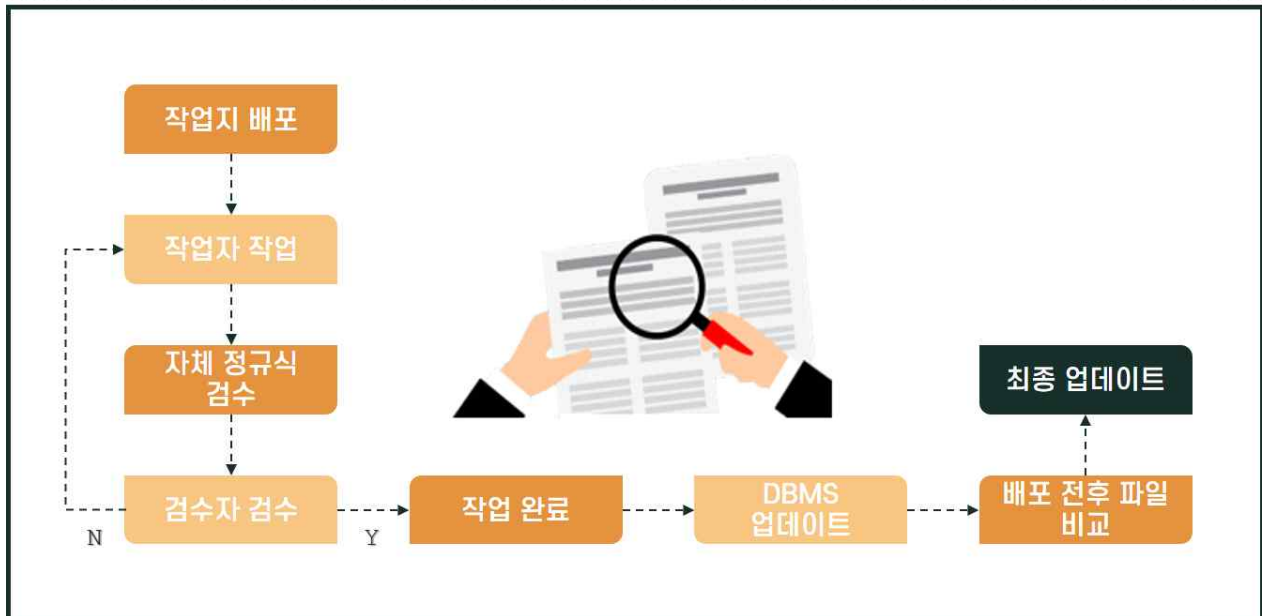


그림 10. 데이터 작업 및 검수 공정

마. 인용 부호 수정 말뭉치

2차 정제가 완료된 데이터는 ‘신문 말뭉치’ 1종으로 구축된다. 사업 제안 요청서에 포함된 ‘인용 부호 수정 말뭉치’는 신문 말뭉치에서 인용 부호를 정제·수정한 데이터를 의미하며, 이 말뭉치가 ‘모두의 말뭉치’ 공개 대상이다.

1) 인용 부호의 통일

인용 부호는 한국어 문장 구조와 발화 인식에서 핵심적인 역할을 수행한다. 그러나 실제 신문 기사 데이터에서는 매체별 입력 체계 차이로 인해 작은따옴표와 큰따옴표가 혼용되거나, 열림과 닫힘 짝이 맞지 않는 등 다양한 오류가 다수 발견되었다. 예를 들어 큰따옴표로 시작하여 작은따옴표로 닫히거나, 열림 기호만 존재하거나 닫힘 기호만 남아 있는 등 비표준적 사용 사례가 빈번하게 나타났다.

본 사업에서는 이러한 인용 부호 오류를 정규식 기반 규칙과 수동 검수 절차를 병행하여 수정하였다. 표준 인용 부호(‘ ’, “ ”) 외에 일반 키보드 따옴표(' U+0027, " U+0022)를 사용한 사례는 표준 기호 체계에 맞게 정정하였다. 이 과정에서 영어 아포스트로피(Apostrophe)처럼 소유격과 축약 기능으로 사용되는 부호는 의미 보존 원칙에 따라 원형을 유지하였다.

인용 부호는 한국어 텍스트에서 문장 경계 결정, 발화 주체 인식, 종속절 및 인용구 판별 등 의미 구조에 직접적인 영향을 미치므로, 인용 부호의 정합화는 후속 문장 분리, 형태소 분석, 토큰화 과정의 안정성을 확보하기 위한 필수 공정이다. 인용 부호를 정규화함으로써 불필요한 토큰 분할과 비정상 구문 해석을 예방하고, 대규모 언어모델 학습 단계에서 오류 전파를 방지하며 통계·규칙 기반 분석 모델의 정확도를 향상시킨다.

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 11> 인용 부호 치환 표

데이터 정제 전	데이터 정제 후
그러면서 “3년 전인가 2년 전에 타로 쌤과 사주 선생님께 갔는데, 그때 믿었어야 했다”라고 토로하며 “나의 날은 오로지 이 일, 엔터테인먼트다. 지금 빈털터리는 아닌데 다시 20년 전으로 돌아갔다”라고 전했다.	그러면서 “3년 전인가 2년 전에 타로 쌤과 사주 선생님께 갔는데, 그때 믿었어야 했다”라고 토로하며 “나의 날은 오로지 이 일, 엔터테인먼트다. 지금 빈털터리는 아닌데 다시 20년 전으로 돌아갔다”라고 전했다.
박혜민은 “2024 갑진년 새해는 더욱 하나가 되어 좋은 모습 많이 보여드리도록 최선을 다하겠다”며 의지를 전했다.	박혜민은 “2024 갑진년 새해는 더욱 하나가 되어 좋은 모습 많이 보여드리도록 최선을 다하겠다”며 의지를 전했다.
프리미어리그 토트넘은 소셜미디어(SNS) 계정을 통해 “토트넘에 오신 것을 환영합니다!”라며 양민혁의 영입 사실을 공개했다.	프리미어리그 토트넘은 소셜미디어(SNS) 계정을 통해 “토트넘에 오신 것을 환영합니다!”라며 양민혁의 영입 사실을 공개했다.
이날 한예슬은 “나는 연애에 대해 공개적으로 이야기 해왔다”라며 “지금 만나고 있는 사람과 정말 잘 맞고, 앞으로도 오래 함께 할 것 같다는 생각이 든다”고 쿨하게 전해 호감을 높였다.	이날 한예슬은 “나는 연애에 대해 공개적으로 이야기 해왔다”라며 “지금 만나고 있는 사람과 정말 잘 맞고, 앞으로도 오래 함께 할 것 같다는 생각이 든다”고 쿨하게 전해 호감을 높였다.

<표 12> 인용 부호 수정 데이터 정제 전후

바. 문장 말뭉치 구축

문장 말뭉치는 인용 부호 수정 말뭉치와 함께 활용될 수 있도록, 단락 구분을 표시하는 <p> 태그와 문장 구분을 표시하는 <s> 태그를 삽입하여 문장을 구분하였다. 다만, 복수의 인용문을 한꺼번에 인용하는 경우에는 각 인용문을 문장 종결 부호 단위로 분할하지 않았다. 이렇게 구성된 데이터는 학습 속도와 정확도, 문장 구조 파악 등 자연어 처리 모델의 성능 향상에 기여할 것으로 기대된다.

1) 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 활용

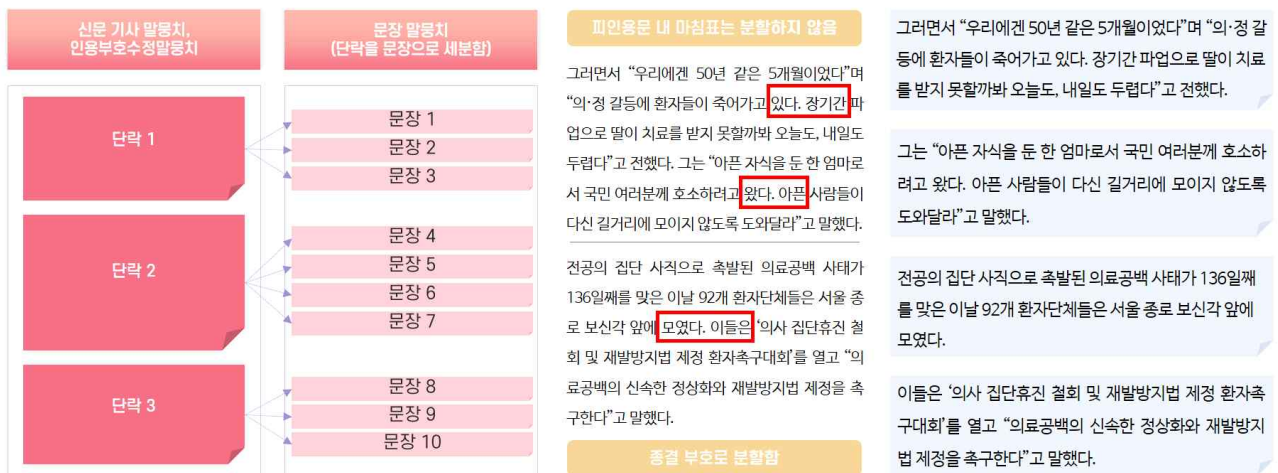


그림 11. 문장 말뭉치 개념

인용 부호 수정 말뭉치	문장 말뭉치
<p>그러면서 “우리에겐 50년 같은 5개월이었다”며 “의·정 갈등에 환자들이 죽어가고 있다. 장기간 파업으로 딸이 치료를 받지 못할까봐 오늘도, 내일도 두렵다”고 전했다. 그는 “아픈 자식을 둔 한 엄마로서 국민 여러분께 호소하려고 왔다. 아픈 사람들이 다신 길거리에 모이지 않도록 도와달라”고 말했다.</p>	<p><s>그러면서 “우리에겐 50년 같은 5개월이었다”며 “의·정 갈등에 환자들이 죽어가고 있다. 장기간 파업으로 딸이 치료를 받지 못할까봐 오늘도, 내일도 두렵다”고 전했다.</s> <s>그는 “아픈 자식을 둔 한 엄마로서 국민 여러분께 호소하려고 왔다. 아픈 사람들이 다신 길거리에 모이지 않도록 도와달라”고 말했다.</s></p>
<p>전공의 집단 사직으로 촉발된 의료공백 사태가 136일째를 맞은 이날 92개 환자단체들은 서울 종로 보신각 앞에 모였다. 이들은 ‘의사 집단휴진 철회 및 재발방지법제정 환자촉구대회’를 열고 “의료공백의신속한 정상화와 재발방지법제정을 촉구한다”고 말했다.</p>	<p><s>전공의 집단 사직으로 촉발된 의료공백 사태가 136일째를 맞은 이날 92개 환자단체들은 서울 종로 보신각 앞에 모였다.</s> <s>이들은 ‘의사 집단휴진 철회 및 재발방지법제정 환자촉구대회’를 열고 “의료공백의신속한 정상화와 재발방지법제정을 촉구한다”고 말했다.</s></p>

<표 13> 문장 말뭉치 데이터 정제

2. 온라인 매체 게시 자료 말뭉치 구축

이번 사업은 기존에 별도 과제로 진행되던 신문 기사 말뭉치 구축과 온라인 게시글 말뭉치 구축이 통합 발주된 첫 사례이다. 따라서 신문 기사 정제와 병행하여 온라인 게시 자료를 수집·정제해야 했으며, 한정된 예산과 기간 내에서 두 영역을 동시에 성공적으로 수행하는 것이 핵심 목표였다.

온라인 텍스트는 형식, 품질, 저작권 구조가 신문 기사와 전혀 다르므로, 수집 플랫폼 분석부터 저작권 확보 방식, 비식별화 체계까지 각 단계별로 별도의 전략을 수립하여 추진하였다.

가. 플랫폼 선정

본 사업은 신문 기사와 온라인 게시 자료가 통합된 첫 사업으로, 온라인 영역에서만 5천만 어절 이상을 확보해야 했다. 단순히 ‘플랫폼 다양성 확보’에 그치지 않고, ‘어절 확보 가능성’을 절대 기준으로 삼았다.

“5천만 어절을 기한 내에, 저작권을 안전하게 확보하면서 품질을 유지할 수 있는가?”라는 목표 아래 수행하는 후보 플랫폼 풀을 폭넓게 검토하였다.

평가 기준	세부 내용
어절 확보력	장문 텍스트 비중, 게시글 규모, 계정당 생산량
정제 가능성	HTML 구조, 수집 후 텍스트 추출 난이도
저작권 구조	권리 확인 가능 여부, 계약 방식
비식별화 부담	개인정보·상표·정치적 발화 노출 정도
행정/기술 위험	접근성, 폐쇄성, API 제한, 법적 위험

<표 14> 온라인 매체 선정 기준

1) 블로그 선정

검토 결과, 네이버 블로그 단독 집중 전략이 사업 목표 달성(5천만 어절 확보)과 데이터 품질 및 법적 안정성 측면에서 가장 타당한 대안으로 판단되었다.

네이버 블로그는 국내에서 가장 안정적으로 장문 자연어 데이터가 대량 생산되는 플랫폼으로, 하나의 계정에서 다량의 텍스트를 확보할 수 있으며, 게시글 간 맥락, 문단 구조, 주제의 폭이 넓어 인공 지능 학습에 최적화된 언어 자원을 제공한다.

또한 검색 접근성과 필터링 구조가 체계적이며, 수집 후 정제 과정까지 안정적으로 관리할 수 있어, 수집 → 정제 → 비식별화 → 메타 태깅의 전 과정을 효율적으로 수행할 수 있었다.

플랫폼	배제 이유
티스토리	계정당 게시량·어절량 부족
카페/밴드	폐쇄형 → 법적·기술 접근 제한, 비식별 난이도 높음
페이스북	정치/광고 비중 높음, 어절 밀도 낮음
X(트위터)	단문 위주, 비속어·비식별 요소의 출현
스레드	서비스 초기, 데이터량 부족

<표 15> 온라인 매체 배제 이유

2) 누리소통망 선정 배경

블로그 기반 말뭉치는 대부분 장문 서술형·정보 중심의 텍스트로 구성되어 있어, 구어체, 신조어, 누리소통망 스타일 발화 등 최신 온라인 언어 특성을 보완할 필요가 있었다.

이에 수행사는 타 SNS와 비교할 때 비식별화 난이도, 콘텐츠 품질, 정책 위험성이 상대적으로 낮고, 실제 텍스트 생성 규모가 안정적으로 유지되는 인스타그램을 수집 대상으로 선정하였다.

인스타그램은 어절 확보 목적이 아닌, 언어 다양성, 시의성, 구어체 표현 확보를 위한 플랫폼으로 활용되었다.

그 결과 인스타그램 데이터는 누리소통망 특유의 발화 패턴 확보에 기여하였으며, 실제 구축 어절은 약 420만 어절(45,413건 수) 내외로 제안 요청서 상 요구된 온라인 타 유형 수집 요건(약 5만 건 내외)을 충족하였다.

나. 참여자 섭외 및 저작권 계약 체계 구축

온라인 게시 자료는 작성자 개인이 저작권을 보유하고 있으므로, 플랫폼 선정 이후에는 반드시 계정 소유자와 직접 접촉하여 저작권 이용 허락 및 계약을 체결해야 했다.

수행사는 이에 따라 사전 후보군 식별 → 신뢰 기반 섭외 → 계약 체결 → 데이터 수집의 4단계 체계를 수립하고 운영하였다.

1) 참여자 후보군 식별 및 우선순위 체계 수립

플랫폼 특성 분석 결과, 어절 생산량과 주제 다양성 확보가 사업 목표 달성의 핵심 요인으로 확인되었다. 이에 총 1만여 개 이상의 계정 후보군을 수집하고 다음 기준에 따라 우선순위를 설정하였다.

구분	선정 기준	목적
장문·어절 기여도	게시글 평균 어절수, 누적 작성량, 최근(2023~24년) 활동량	5천만 어절 목표 대응
주제 다양성	여행·일상·육아·건강·요리·경제·문화 등 균형	다양한 언어 도메인 확보
창작성	직접 작성 비중, 인용·스크랩 비중	저작권 안정성
품질·윤리	과도한 광고, 비방·정치 편향 회피	공공데이터 신뢰 확보

<표 16> 참여자 후보군 선정 기준

블로그의 경우 주제별로 1~4순위로 세분화하여 순차적으로 접촉하였으며, 장문형·지속적 작성 계정을 우선 섭외하였다. 인스타그램은 국립국어원 홈페이지 공고 후, 사회 관계망 서비스의 영향력자(인플루언서) 섭외와 직접 신청 경로를 병행 운영하였다.

2) 1:1 직접 섭외 원칙

대량 메시지 발송은 스팸 신고 및 신뢰도 저하 위험이 있으므로, 모든 계약 대상자는 개별 접촉 방식으로 섭외하였다.

주요 연락 수단은 네이버 쪽지, 이메일, 인스타그램 DM을 활용하였다.

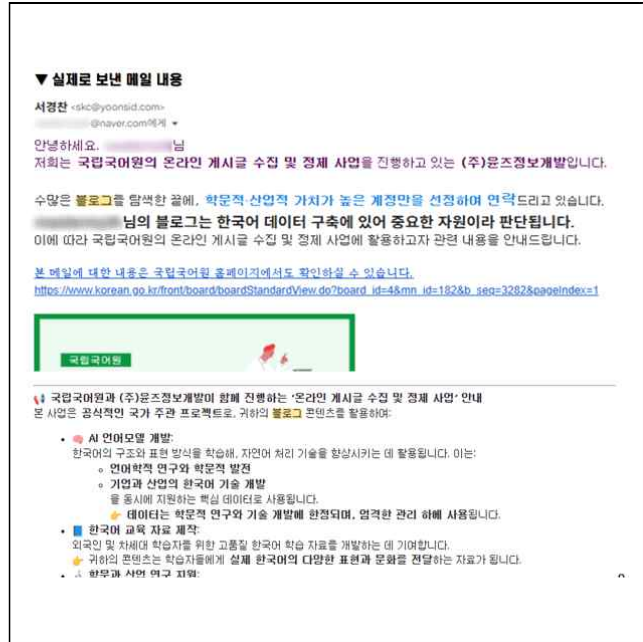
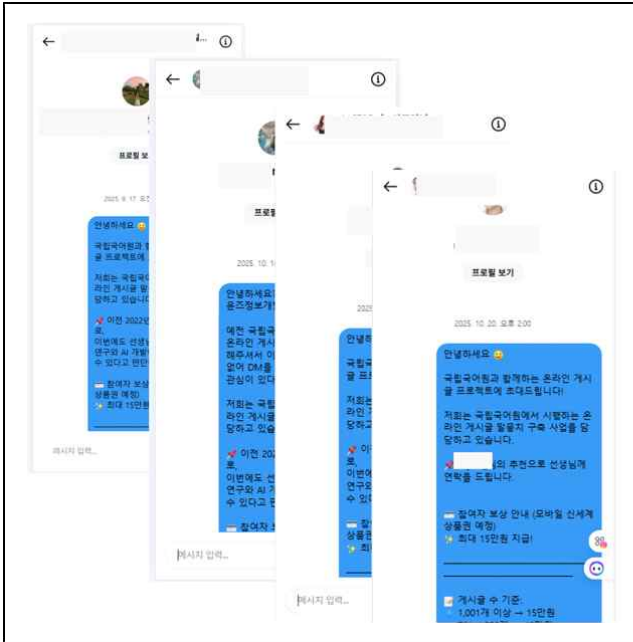


그림 12. 참여자 1:1 직접 협의 과정

3) 국립국어원 공고 연계 및 신뢰성 확보

온라인 창작자는 지적재산권에 민감하여 단순 안내만으로는 참여를 유도하는 데 한계가 있었다. 이에 이 사업에서 가장 중요한 요소는 공공기관인 국립국어원에 대한 신뢰를 기반으로 한 참여자 신뢰 확보였으며, 이를 위해 다음과 같은 절차로 운영하였다.

- ▶ 최종 모집 공고를 국립국어원 홈페이지에 공식 공지하고 포스터 게시를 요청함.
- ▶ 이후 개별 섭외를 진행하며, 공고 링크를 함께 공유함.
- ▶ 참여 희망자 중 일부는 국립국어원에 직접 문의하여 사업의 신뢰성을 확인함.

그 결과 다수의 창작자가 “공고 확인 후 참여를 결정했다”는 의견을 표명하였으며, 공공기관 기반 신뢰 확보 전략의 효과성이 입증되었다.

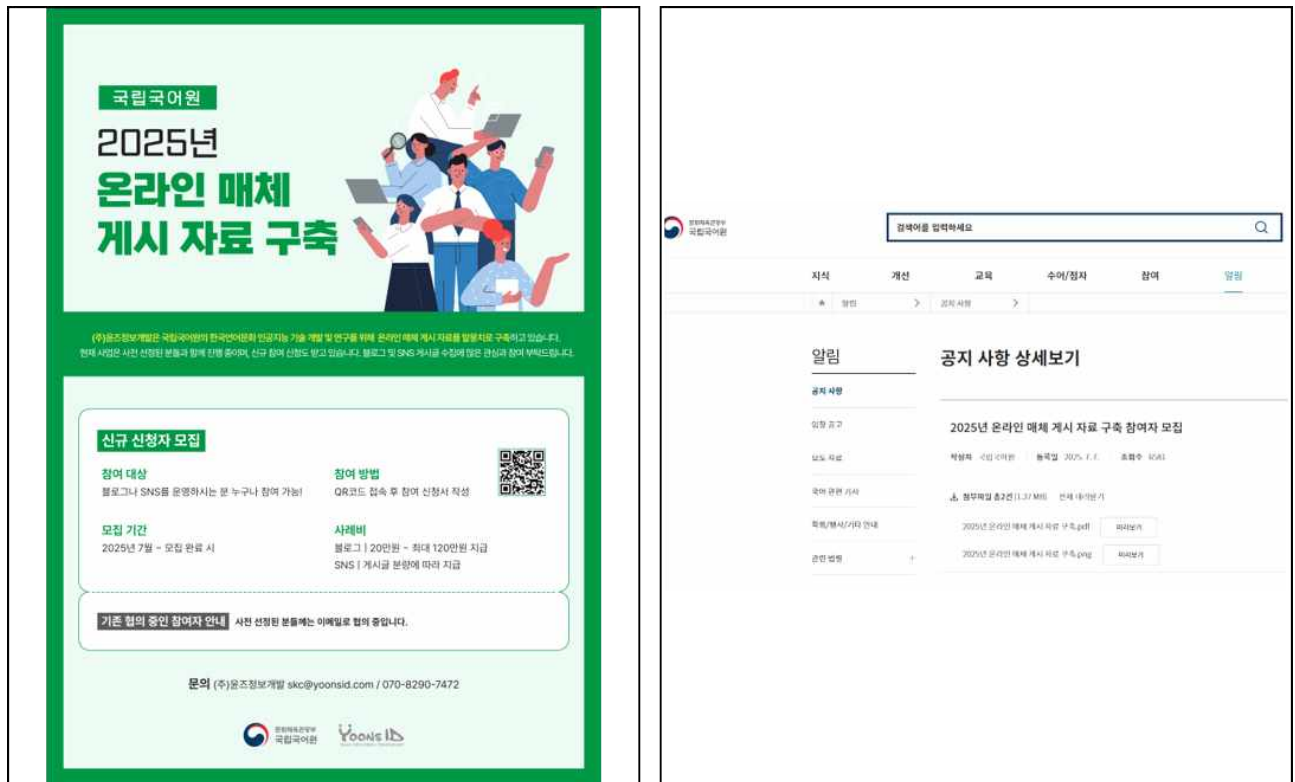


그림 13. 국립국어원 공고를 통한 신뢰성 확보

4) 전자계약 기반 저작권 확보 체계

대규모 개인 창작자와의 계약을 신속하고 안전하게 진행하기 위해, 수행사는 검증된 상용 전자계약 플랫폼을 활용하였다.

별도의 시스템을 구축하지 않고, 공공기관에서도 널리 사용되는 상용 전자계약 서비스를 이용함으로써 법적 효력, 보안성, 행정 효율성을 동시에 확보하였다.

참여자는 이메일 또는 링크를 통해 계약 내용을 확인한 뒤, 전자 서명 방식으로 즉시 체결할 수 있었으며, 이를 통해 서면 계약 대비 절차가 대폭 단축되고 참여자의 거부감과 행정적 부담이

최소화되었다.

또한 계약서는 기존 온라인 게시 자료 구축 사업에서 사용하던 문안을 기반으로 법률 검토를 진행하고 필요한 부분을 보완하였다.

 법무법인 민후 서울 강남구 테헤란로 134, 포스코타워 역삼 11층 Tel. 02-532-3483 Fax. 02-532-3486 수 신 : 주식회사 홈즈정보개발 발 신 : 법무법인 민후(民後) 담당번호사 앞진영, 현수진 일 자 : 2025. 6. 19. 제 목 : 국립국어원 국가 언어 자원 (말뭉치) 사업 관련 법률검토 귀사의 일익번창을 기원합니다. 본 법무법인은 위 제목의 것과 관련한 귀사의 질의에 대하여 다음과 같이 의견을 드리며, 아울러 아래의 의견에 대한 질문이 있으면 주저하지 마시고 언제든지 문의해 주시기 바랍니다. — 다 음 — I. 배경사실 및 질의의 요지 귀사는 국립국어원의 국가 언어 자원 (말뭉치) 구축 및 활용·용역 사업의 수행자로서, 해당 용역 사업을 수행하기 위하여 블로그 또는 SNS 게시글(2023년 및 2024년 사이 업데이트 된 게시글)을 수집하여, 적합한 게시글을 선정한 뒤 복제·번역하여 원시 및 분석 말뭉치로 구축하고자 합니다. 귀사는 블로그 또는 SNS 게시글에 대한 적법한 권리 확보 및 용역 사업 수행에 필요한 개인정보의 적법한 수집을 위하여 블로그 또는 SNS 운영자(정보주체)와 「저작권 이용 허락 계약서」를 체결하고 「개인정보 수집·이용 동의서」 및 「개인정보 제3자 제공(공개) 동의서」를 받을 예정인바, 본 사업과 관 MINWHO Law Group 언하여 저작권법 및 개인정보 보호법에서 발생할 수 있는 주요 쟁점을 검토하고 제공하여 주신 시식을 통해 본건 용역 사업에 필요한 적법 요인이 충족되는지 여부에 대한 검토를 진행하겠습니다. II. 본 법무법인의 의견 I. 저작권법 쟁점 및 「저작권 이용 허락 계약서」 분석 가. 블로그 및 SNS 게시글의 저작물성 여부 저작권법상 "저작물"로 인정받기 위해서는 저작물성, 즉 창작성이 인정되어야 합니다. 저작권법은 모든 글에 대해 저작권을 부여하지 않으며, "저작물"로 보호 받으려면 해당 글이 단순한 사실의 전달이나 누구나 동일하게 쓸 수밖에 없는 표현이 아니라, 창작자의 개성과 창의성이 드러난 창작물이어야 합니다. 귀사는 2023년 및 2024년 사이 대상 블로그 또는 SNS의 게시글(2023년 및 2024년 사이 업데이트 된 게시글) 중 연구 및 말뭉치 구축에 적합한 게시글을 선정하여 이용하고자 합니다. 이렇게 선정된 게시글은 블로그나 SNS 사용자가 자신의 생각, 경험, 정보 등을 독창적으로 표현한 게시글이므로, 저작물로 인정되어 저작권이 발생합니다. 나. 블로그 및 SNS 게시글의 적법한 이용 방안 저작재산권은 저작자가 자신의 저작물에 대해 가지는 재산적 권리로서, 저작
--

그림 14. 법률 검토를 통한 계약서 보완

다. 온라인 데이터 수집 및 정제 공정

본 사업의 온라인 매체 게시 자료 구축은 데이터 수집, 정제 및 정규화, 비식별화 및 검증의 3단계 공정으로 수행되었다.

온라인 데이터는 비정형성이 높고 개인의 자유로운 표현이 포함되어 있으므로, 법적 안정성과 품질 완성도를 동시에 확보하기 위한 다단계 기술·윤리적 필터링 절차를 적용하였다.

1) 자료 수집 단계: 플랫폼 특성 기반의 체계적 수집

가) 수집 대상 및 기준

선정된 핵심 플랫폼(네이버 블로그, 인스타그램)의 게시글을 중심으로 2023년~2024년 사이 작성된 게시글만을 수집하였다.

블로그의 경우 장문형 서술 콘텐츠, 인스타그램의 경우 구어체 콘텐츠를 중심으로 하여 어절 수가 많고, 게시글을 많이 보유한 계정주를 우선 접촉하였다.

이 과정을 통해 두 플랫폼의 언어적 특성을 상호 보완하며 데이터의 다양성과 품질을 확보하였다.

나) 계정 기반 구조적 수집

온라인 게시자료는 창작자 개인의 저작물로 구성되어 있으므로, 수행사는 무단 수집을 지양하고 사전 저작권 확보 후 수집 원칙을 철저히 준수하였다.

블로그의 경우 플랫폼 내 공개 게시글을 무작위로 수집하는 방식이 아니라, 사전 협의 및 계약 체결을 전제로 한 계정 기반 접근 방식을 적용하였다. 총 1만여 개 계정을 확보하고, 카테고리별 주제 균형을 고려하여 각 분야별로 약 200명 내외의 후보군을 선별하였다. 이후 게시글 수, 어절 수, 주제 다양성 등을 종합적으로 평가하여 1순위부터 4순위까지의 우선순위 체계를 수립하고, 순위에 따라 단계적으로 접촉 및 계약을 진행하였다.

계약이 체결된 계정에 한해서만 본문 데이터를 수집하였으며, 계약 이전의 게시글은 일절 수집하지 않았다. 이러한 절차는 법적·윤리적 안정성을 전제로 한 데이터 구축 모델로서, 공공 말뭉치 사업의 투명성과 신뢰성을 동시에 확보하는 기반이 되었다.

인스타그램의 경우에도 동일한 원칙 하에 국립국어원 공식 홈페이지 공고를 통한 공개 모집과 인플루언서 중심의 직접 섭외 방식을 병행하였다.

구분	세부 구분	확보 계정 수	구분	세부 구분	확보 계정 수
엔터테인먼트/ 예술	공연_전시	575	취미/여가/여행	게임	263
	드라마	323		국내여행	400
	만화_애니	364		맛집	376
	문화_책	508		사진	468
	미술_디자인	462		세계여행	377
	방송	223		스포츠	316
	스타_연예인	300		자동차	330
	영화	242		취미	490
	음악	294		IT_컴퓨터	193
생활/노하우/쇼 핑	반려동물	425	지식/동향	건강_의학	361
	상품리뷰	542		교육_학문	336
	요리_레시피	205		비즈니스_경제	389
	원예_재배	428		사회_정치	349
	육아_결혼	438		어학_외국어	276
	인테리어_div	373			
	일상_생각	588			
	좋은글_이미지	246			
	패션_미용	242			

<표 17> 각 카테고리별 최초 확보 계정 수(블로그 기준 분류)

다) 데이터 수집 관리 체계

데이터 수집 과정에서는 각 게시글의 수집 시점과 주요 메타정보를 함께 기록하여 데이터의 관리 및 추적이 가능하도록 하였다. 또한 동일 게시글의 중복 수집을 방지하기 위해 URL을 기준으로 중복 여부를 검사하고, 중복된 데이터는 자동으로 제거하였다. 이를 통해 데이터의 효율성과 정확성을 동시에 확보하였다.

2) 정제 및 정규화 단계 : 비정형 데이터의 표준화

가) HTML 기반 본문 정제

온라인 데이터는 매체별 구조와 형식이 상이하며, 특히 블로그의 경우 HTML 태그와 기능성 코드가 혼재되어 있어 정제 과정에서 고도의 기술적 접근이 요구된다. 수행사는 네이버 블로그의 구조적 특성을 분석하여 본문 이외의 불필요한 요소를 정밀하게 제거하였다. 네이버 블로그는 콘텐츠 본문 외에 광고 배너, 지도 및 위치 정보, 사진 캡션, 태그, 스크립트, 링크, 댓글 영역 등 다양한 HTML 요소를 포함하고 있으며, 이러한 비텍스트 영역은 인공 지능 학습에 불필요한 노이즈로 작용한다. 이에 HTML DOM 구조를 세분화하여 분석하고, 본문이 위치하는 주요 콘텐츠 블록만을 식별하여 순수 텍스트를 추출하였다. 이를 통해 각 게시글의 문장 흐름은 온전하게 유지하면서도 비정형 요소가 완전히 제거된 고품질 텍스트 데이터를 확보할 수 있었다.

요소 구분	HTML클래스/태그 예시	내용
핵심 본문 컨테이너	<div class='se-main-container'>	모든 콘텐츠 요소(텍스트, 이미지, 광고, 링크 등)를 포함하는 최상위 논리적 블록. 이 블록 내에서 필요한 요소만을 선별
순수 텍스트 블록	<div class='se-text'>, <p class='se-text-paragraph'>, <div class='se-quotation'>	블로거가 직접 작성한 핵심 문단 및 인용문. 이 영역의 텍스트가 인공지능 학습을 위한 최종 데이터
구조화된 미디어 /메타데이터	<div class='se-image'>, <div class='se-video'>, 	이미지, 동영상, 외부 링크(OG Tag)와 같이 본문 흐름을 끊는 비텍스트/반정형 요소
기능/광고성 노이즈	<div class='ssp-adcontent'>, <div class='se-horizontalLine'>	블로그 서비스 운영을 위한 기능성 코드(se-horizontalLine, 구분선) 또는 외부 광고(ssp-adcontent) 영역. 인공지능 학습에 치명적인 노이즈를 유발

<표 18> 데이터 구조 분석

3) 데이터 선별 및 사용 제한 기준

온라인 게시자료는 계정 단위로 수집된 이후, 학습 데이터로 활용하기 적합한 게시글만을 선별하였다. 이 단계에서는 플랫폼별 특성과 어절 분포를 고려하여 사용 제외 기준을 명확히 설정하였다.

블로그의 경우 수집된 게시글 중 어절 수가 현저히 적은 글은 학습 데이터로서 활용도가 낮다고 판단하여 제외하였으며, 인스타그램 또한 유사한 기준을 적용하여 어절 수가 지나치게 적은 게시글은 사용하지 않았다. 또한 제목이나 본문 내용에 비윤리적 표현이 강하거나, 특정 인물에 향한 저격성 발언 등이 심한 경우, 사회적 갈등을 유발할 우려가 있는 게시글은 비식별화 조치 대신 선제적으로 사용 대상에서 제외하였다.

블로그의 특성상 공유 게시글, ‘1년 전 오늘’과 같이 과거 게시글을 자동 재게시하는 콘텐츠, 전문(全文) 인용 형태의 게시글, 가사 또는 시나리오(대본)가 주를 이루는 게시글 등 저작권 침해 가능성이 있는 게시글 또한 모두 제외하였다.

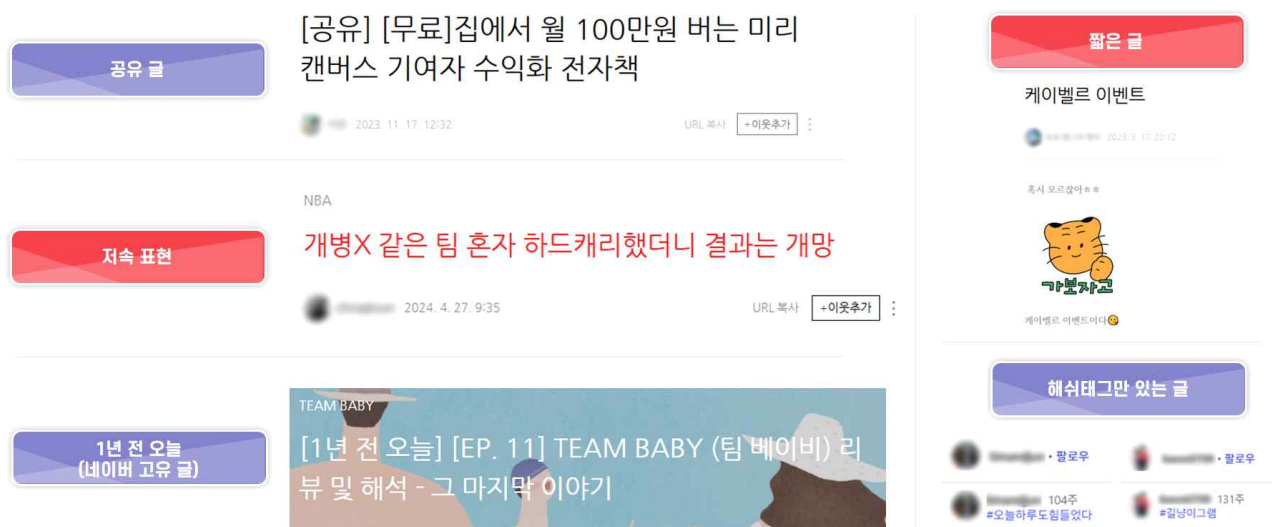


그림 15. 사용하지 않는 온라인 게시 자료의 예

4) 메타데이터 삽입

메타데이터 항목은 ID(추후 비식별화), URL, 게시일자, 수집 시간, 어절 수, 플랫폼 구분, 주제 분류 등으로 구성되었으며, 각 항목은 자동 추출과 수동 검증 절차를 병행하였다.

가) 국립국어원 말뭉치 기반 1차 학습

우선, 2022년도 「모두의 말뭉치」에 공개된 2022년 온라인 게시 자료 구축 데이터를 기반으로 1차 학습을 수행하였다.

국립국어원의 분류체계는 ‘문화/예술’, ‘생활/쇼핑’, ‘지식/교육’, ‘취미/여행’ 등 4개 대분류와 20여 개 세부 분류로 구성되어 있으며, 공공데이터 간 상호운용성을 확보하기 위한 표준 체계로 활용된다.

나) 블로그 주제 분류 학습 및 분류 항목 연계

이후, 네이버 블로그의 고유한 대·소분류 데이터를 통합하여 분류 정확도를 보완하였다.

네이버는 실제 이용자가 작성한 게시글에 주제 분류 체계를 자동으로 부여하고 노출에 활용하므로, 실제 콘텐츠의 맥락을 반영하는 데 유리하다.

이에 수행사는 네이버의 대·소분류 체계를 국립국어원의 분류 항목과 일대일(1:1)로 대응시켜 자동 분류 모델의 정합성을 확보하였다.

네이버 대분류	네이버 소분류	국립국어원 대분류	국립국어원 소분류
엔터테인먼트/예술	드라마	문화/예술	영화/드라마/방송
생활/노하우/쇼핑	요리_레시피	생활/쇼핑	요리/음식
취미/여가/여행	국내여행	취미/여행	여행/관광
지식/동향	건강_의학	지식/교육	의료/건강
지식/동향	사회_정치	지식/교육	정치/경제/사회

(이하 생략)

<표 19> 주제 분류 체계 대응(국립국어원·네이버)

다) 복합 학습 및 정확도 향상

국립국어원(1차)과 네이버 블로그(2차)의 분류체계를 결합한 복합 학습 모델을 구축하였다.

1차 학습 단계에서는 국립국어원 데이터셋으로 주제별 언어적 특징을 학습시키고, 2차 보정 단계에서는 네이버 데이터셋을 병합하여 문맥 기반 확률 분류의 정확도를 향상시켰다.

이를 통해 자동 분류 정확도는 약 89% 수준으로 향상되었으며, 분류 신뢰도가 70% 이하인 데이터는 작업자가 검토하여 재분류하였다.

이러한 수동 검증 절차를 통해 자동화 과정에서의 오분류 가능성을 최소화하고, 데이터 전체의

주제 일관성과 품질을 확보하였다.

5) 데이터 본문의 정제

가) 해시태그의 처리

해시태그는 본문의 일부로 사용하는 경우도 있고 검색용 키워드로 추가하는 경우도 있다. 본문 내용의 일부로 판단이 되는 경우에는 그대로 유지하며, 검색용 키워드는 삭제한다.

본문 중에 내용의 일부로 사용된 해시태그는 그대로 인정한다.

#월미도레미카페 #바이킹 뷰 가 인상적인 곳! (인정)
#루프탑 잔디 아이들 뛰어놀기 너무 좋아요! (인정)
#월미도 #제인언니 #인싸인서 #생각대로사는여자 #월미도카페 (삭제)

① 검색용 키워드로 사용되는 해시 태그는 제거한다.

#월미도레미카페 #바이킹 뷰 가 인상적인 곳!
#루프탑 잔디 아이들 뛰어놀기 너무 좋아요!
#월미도 #제인언니 #인싸인서 #생각대로사는여자 #월미도카페 (삭제)

나) URL 주소의 처리

온라인 게시글 내 많은 URL은 모두 비식별화 처리를 한다.

① http, https로 시작하는 일반적인 URL주소는 비식별화 처리를 한다.

"id": "ESRW2500000045.69.23",
"form": "https://smartstore.naver.com/micojinishop/products/8056302891",
"original_form": "➡https://smartstore.naver.com/micojinishop/products/8056302891"

② 인스타그램의 계정 등도 비식별화 처리한다.

@kissme.kr @glowpick는 인스타그램이나 트위터 같은 SNS에서 특정 계정을 부를 때 쓰는 '언급 (Mention)' 형식이다. 언급은 인스타그램, 트위터, 카톡 같은 곳에서 아이디 앞에 @를 붙이는 행위를 말한다. 이렇게 하면 알림보내거나 링크 연결이 된다.

다) 문장의 분할

본문 내용이 긴 경우에 문장 단위로 분할한다. 문장의 분할은 아래의 원칙으로 진행한다.

- ① '문장'은 주어와 서술어의 문장 성분을 갖추고 있어야 하고, 서술어에 종결어미가 실현되어 있어야 한다.
- ② 하나의 독립된 생각이나 정보가 완결되는 지점을 문장의 경계로 본다.

- ③ 문장의 기본 단위는 종지 부호 . ? !를 기준으로 한다.
- ④ 문장은 문법 구조를 중심으로 해서 분할한다.
- ⑤ 종지 부호가 사용되었어도 단순 수식 형태는 하나의 문장으로 인정하지 않는다.

■ 우선 ♥초저가! 홈플러스딜 상품♥

❖ '♥초저가!'는 주어와 서술어가 갖춰진 완벽한 문장이 아니라, 뒤에 나올 상품명을 강조하기 위한 감탄사형 수식어로 보고 분할하지 않는다.

■ 내 손 안의 편리함! B.R.월렛~

❖ !는 문장의 경계로 인정하지 않는다.

■ 맥세이프 케이스 많이들 쓰길래,, 편해? 물어보면 10명이면 다다다 동일한 답을 하더라구요!!

❖ '편해?'를 문장의 경계로 인정하지 않는다.

■ 이런맛은 알려야돼. 하며 (뽀뽀) 회사에도 싸갔는데 인기만점! 검증완료!! □ ♀ □ ♀ □ ♀ 먹어보신분들도 어떻게 만드냐고 많이 물어봐서 릴스 냉큼 만들어왔쥬! ♪(ˊ ˋ)>"

☞ 이런맛은 알려야돼. 하며 (뽀뽀) 회사에도 싸갔는데 인기만점!

검증완료!! 🍷♀🍷♀🍷♀

먹어보신분들도 어떻게 만드냐고 많이 물어봐서 릴스 냉큼 만들어왔쥬! ♪(ˊ ˋ)>"

- ⑥ 문장 종지 부호 뒤에 오는 이모지는 독립적인 문장으로 인정하지 않는다.

■ 근데 선생님! 그러는데 튀르키예에서도 이거 집에서 잘안만드는데요. 힘들어서.ㅋㅋㅋ 다들사먹는다고ㅋㅋㅋ 여러분도 사드세요□□

☞ 근데 선생님! 그러는데 튀르키예에서도 이거 집에서 잘안만드는데요.

힘들어서.ㅋㅋㅋ

다들사먹는다고ㅋㅋㅋ

여러분도 사드세요😊□□

- ⑦ 'ㅋㅋㅋㅋ'를 문장의 끝을 맺는 종결어미나 문장 부호처럼 취급한다. 마찬가지로 이모지, 이모티콘은 앞 문장에 붙여 분할한다.

■ 편안하고 즐겁게 시청합니다 ㅋㅋㅋㅋㅋ

■ "어? 내 얘기! ㅋㅋㅋㅋㅋㅋ" 공감 버튼 백개를 누르고 싶었어요□

☞ "어? 내 얘기! ㅋㅋㅋㅋㅋㅋ"

공감 버튼 백개를 누르고 싶었어요👉

■ 실패할 걱정 없어요!! ㅋㅋㅋ

■ “캐주얼 러블리룩” 으로 보여요 ㅋㅋㅋ□□

- ⑧ 인스타그램의 특성상 텍스트 내에 !를 습관적으로 많이 사용한다. 아래와 같은 '지금!'은 감탄사적 성격이 지니고 있어 하나의 독립된 문장으로 볼 수 있으나 문장 전체를 수식하거나 주의를 환기하는 독립어 역할로 본다.

■ 바삭바삭! 고소하고 맛있는 간식을 찾는다면, 파머스칩□

■ 지금! 정관장 네이버 공식스토어에서

- 동아제약 디판버그젤! 벌레몰림, 가려움, 땀띠에

⑨ 단순 나열은 문장으로 인정하지 않는다.

- ✓임대분쟁 2. 층간소음 3. 고장시설 4. 난방 5. 곰팡이, 결로 6. 누수,침수 7. 수압,배수 8. 냄새 9. 벌레 10. 채광, 통풍 11. 전기, 누전 12. 기타소음 13. 인테리어 14. 사생활침해 15. 주차 16. 쓰레기배출 17. 보안 18. 혐오시설 19. 상대평가 20. 추천여부 등

⑩ 아래와 같이 종결어미 뒤에 심표를 사용했지만 마침표(.)의 오류로 보고 문장으로 분할할 수 있다.

- &online-account& 네에서 ‘&brand&’ 선물 받았어요♡,, &brand& 입고 다리 각선미도 뽐내 보았어요 ‘ㄷㄷ’. 피부에 닿는 모든 부분이 부드러워 기분이 날아갈 것 같나봐요~✿.< 마지막 영상에 &art& 노래도 들려주네요♥

☞ &online-account& 네에서 ‘&brand&’ 선물 받았어요♡,, &brand& 입고 다리 각선미도 뽐내 보았어요 ‘ㄷㄷ’. 피부에 닿는 모든 부분이 부드러워 기분이 날아갈 것 같나봐요~✿.< 마지막 영상에 &art& 노래도 들려주네요♥

라) 특수 문자의 처리

인스타그램 텍스트 내부에는 ‘␣(U+2800, BRAILLE PATTERN BLANK)’가 사용된 경우가 많다. 인스타그램은 연속 줄바꿈과 공백만 있는 줄을 자동으로 제거한다. 이 때문에 ‘␣’를 사용하여 개행을 표현한다. ‘␣’는 눈에 보이지는 않으나 인스타그램에서는 “내용이 있는 문자”로 인식한다. 그래서 ‘␣’를 사용하여 빈 줄이 유지가 된다. 인스타그램에서 사용된 특수 문자는 아래와 같은 원칙에 의해 처리한다.

코드포인트	설명	처리
U+2800	점자 빈칸(BRAILLE PATTERN BLANK) - 인스타그램에서 빈 줄 만들기에 가장 많이 악용되는 문자. 공백처럼 보이지만 일반 공백이 아님	삭제
U+3164	한글 채움 문자 - 완전히 투명, 폭만 차지	공백 치환
U+1160	한글 중성 채움 문자 - 완전히 투명	삭제
U+FE0F	텍스트 변형 선택자 - 단독으로는 보이지 않음	삭제
U+FFFD	깨진 문자 - 원본 데이터가 손상/인코딩 오류	삭제

<표 20> 특수 문자의 처리 방안

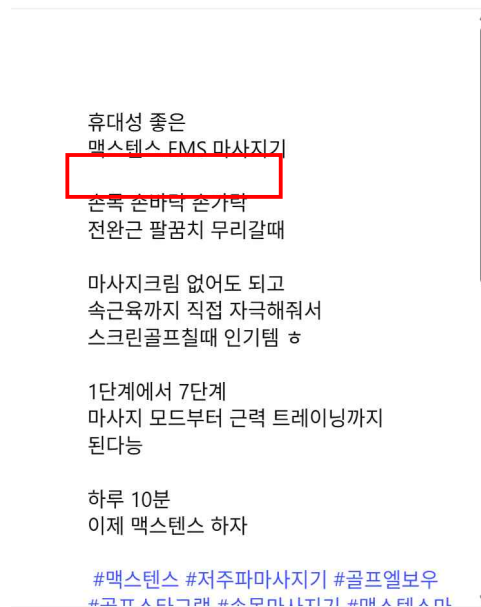


그림 16. 특수 문자 '꺆'(U+2800)'를 사용하여 단락을 구분한 예

6) 개인정보 및 개체명 비식별화 체계

온라인 게시 자료, 특히 블로그와 인스타그램에는 개인의 일상 리뷰, 제품 후기, 체험기 등 상업적·문화적 맥락이 결합된 데이터가 대부분이다. 이에 따라 기업명, 상호명, 브랜드명, 제품명, 서비스명 등이 텍스트 전반에 빈번히 등장하며, 이러한 요소를 그대로 남길 경우 공공데이터로서의 중립성이 훼손될 우려가 있었다. 수행사는 국립국어원과의 협의를 통해 기존 개인정보 중심의 비식별화 기준을 확장하여, 공공데이터의 신뢰성과 윤리성을 확보하기 위한 ‘확장형 비식별화 원칙’을 적용하였다.

가) 확대 비식별화의 필요성 및 배경

① 광고·협찬성 콘텐츠의 간접 노출 방지

온라인 게시글에는 상업적 목적의 광고·협찬(PPL) 콘텐츠가 포함되어 있다. 이들은 대부분 긍정적 표현만 포함하기 때문에, 명칭을 그대로 공개하면 특정 브랜드가 공공 말뭉치를 통해 간접적으로 홍보되는 부작용이 발생할 수 있다. 이는 공공데이터의 비영리·비상업적 성격과 상충하므로, 긍정적 표현이라도 명칭은 비식별화하는 것이 원칙으로 정립되었다.

② 데이터의 중립성 확보 및 편향 방지

‘부정적 언급만 비식별화’ 처리할 경우, 결과적으로 긍정적 표현만 남게 되어 공공 말뭉치 전체가 상업 주체의 “화이트리스트”처럼 작동할 위험이 있다. 이러한 불균형은 학문적·산업적 인공지능 학습 과정에서 상업적 편향을 유발하고, 특정 브랜드나 서비스에 대한 왜곡된 긍정 감성을 학습시킬 수 있다.

③ 공공데이터의 가치와 비판 언어의 다양성 보호

모든 상호명과 브랜드명을 중립적으로 가명 처리함으로써, 공공 데이터가 특정 기업이나 개인의 명예에 영향을 미치지 않도록 하면서도 비판적 언어 표현의 자유와 다양성을 보존하였다. 이는 비식별화는 단순 보호가 아니라 공공적 신뢰 확보 과정임을 보여준다.

온라인 매체 게시 자료 구축 비식별화 대상		
대분류	태그	항목
이름	&name&	일반인의 실명, 정치인명 (역사적 인물, 공인, 가명, 이칭 별칭 제외)
출신	&affiliation&	비식별화 대상의 출신 (학교, 지역, 소속 등)
번호	&social-security-num&	주민등록번호
	&tel-num&	전화번호, 팩스번호
	&card-num&	카드 번호
	&bank-account&	계좌 번호
	&num&	기타 번호 (사업자 번호 등)
온라인계정	&account&	온라인 계정, 이메일
주소/장소	&address&	동 이하의 상세 주소
상호명/회사명	&company&	전체
상표명/제품명	&brand&	전체 (기술/서비스, 행사/이벤트명 등)
작품명	&art&	전체 (공연/전시 등)
기타	&other&	비윤리적 표현, 욕설 등

<표 21> 비식별화 대상 태그·항목

나) 비식별화 적용 사례

구분	원문(예시)	비식별화 처리 결과
이름	몸무게 12kg정도 나가는 0석 ²⁾ 이 5단계 사용하니	몸무게 12kg정도 나가는 &name&이 5단계 사용하니
	11월에는 우크라이나 재건주로 묶이면서 급등. 12월은 계엄이후 이재명 테마주로 묶이면서 상승.	11월에는 우크라이나 재건주로 묶이면서 급등. 12월은 계엄이후 &name& 테마주로 묶이 면서 상승.
출신	제가 대전이 고향이라 어릴 때 시내 나가 면 성심당 한 번씩은	제가 &affiliation&이 고향이라 어릴 때 시 내 나가면 성심당 한 번씩은
	전 그래서 지인 추천으로 업체를 주렸어요!! 제가 직장이 대전이고,	전 그래서 지인 추천으로 업체를 주렸어요!! 제가 직장이 &affiliation&이고,
번호	정기휴무 매주 월요일 **-*****-****	정기휴무 매주 월요일 &tel-num&
상호명	여기가 베스트 웨스턴 프리미어 사파이어 하룻이예요. 딱 보면 아파트같이 생지지 않았나 싶어요.	여기가 &company&이예요. 딱 보면 아파트같이 생지지 않았나 싶어요.
브랜드명	먼저 파리바게뜨 4종 신제품 후기 시작은	먼저 &company& 4종 신제품 후기 시작은
	명성에 먹칠을 할 정도로 이상한 카메라들이 좀 보입니다. 그 대표적인 카메라가 캐논 V10 브이로그 카메라입니다.	명성에 먹칠을 할 정도로 이상한 카메라들이 좀 보입니다. 그 대표적인 카메라가 &brand&입니다.
작품명	오징어 게임을 다시 봤어요.	&ar&을 다시 봤어요.
	얼마 전에 소개해 드린 '단어의 집'은 제가 요즘 필사하기 바쁜 책이기도 합니다.	얼마 전에 소개해 드린 '&ar&'은 제가 요즘 필사하기 바쁜 책이기도 합니다.
기타(욕설)	전각 대부분이 헐렸고 (스비니들 무슨 코엑스 박람회도 아니고....	전각 대부분이 헐렸고 (&other&들 무슨 코엑스 박람회도 아니고....
	내가 만났던 놈들은 다 개새끼고. 날 추앙 해	내가 만났던 놈들은 다 &other&고. 날 추 앙해

<표 22> 비식별화 적용(예시)

2) 보고서에 설명과 실제 전화번호를 기입할 수 없어 일부 비식별화 하였다.

다) 수행 방식

비식별화는 1차 자동 탐지 프로그램 + 2차 전수 검수(100%)의 2단계 체계로 수행하였다.

1차: 정규식 기반 자동 탐지 프로그램을 이용하여 기본 개인정보를 식별하고 임시 마스킹

2차: 모든 결과를 사람이 전수 확인하여 문맥에 따라 적절히 태깅 수정

→ 문맥판단이 필수적인 상표명·기관명·정치인명 등은 전수 수작업 검수로 처리

검수 체계: '작업자'와 '검수자'를 분리하여 2중 확인

이 방식은 자동화의 효율성과 수작업의 정확성을 동시에 확보하였으며, 특히 기업명/브랜드명/서비스명 혼용 사례에 대해 문맥 기반으로 상호와 브랜드를 구분하였다.



제 3 장

사업 수행 결과



제3장 사업 수행 결과

1. 신문 기사 구축 결과

2025년 신문 기사 원문 자료 수집과 정제 대상으로 선정된 매체의 종류와 매체 수는 아래와 같다.

구분	수집 매체명	매체 수
인터넷	뉴스투데이, 오마이뉴스, 오에스이엔	3
종합지	경향신문, 디지털타임스, 일간투데이	3
지역지	경기도민일보, 경남매일, 경인매일, 굿모닝충청, 대경일보, 대전투데이, 울산제일일보, 일간경기, 전남매일, 충남일보, 충청매일	11
통신사	뉴스1, 뉴스웍스, 뉴시스	3
합계	20개 매체	

<표 23> 선정된 매체 구분

- 신문 기사 수집 현황

매체명	최초 수집 기사 수	최초 수집 어절 수
경기도민일보	18,792	3,388,480
경남매일	22,425	3,640,096
경인매일	39,473	6,981,979
경향신문	71,687	23,326,400
굿모닝충청	10,870	3,089,144
뉴스1	352,504	67,082,913
뉴스웍스	29,719	7,759,342
뉴스투데이	23,093	8,820,069
뉴시스	378,317	96,944,181
대경일보	45,485	8,581,331
대전투데이	14,692	2,845,204
디지털타임스	79,600	18,744,800
오마이뉴스	15,998	5,084,277
오에스이엔	213,074	36,357,292
울산제일일보	18,736	3,683,411
일간경기	17,844	3,900,101
일간투데이	49,935	10,107,660
전남매일	24,593	4,355,568
충남일보	50,926	9,117,951
충청매일	16,924	3,022,524
합계	1,494,687	326,832,723

<표 24> 신문 기사 수집 총괄표

- 신문 기사 정제 현황

매체명	최종 기사 수	최종 구축 어절 수
경기도민일보	12,275	2,056,403
경남매일	10,998	1,917,016
경인매일	29,322	4,849,436
경향신문	48,594	12,695,523
굿모닝충청	6,189	1,483,748
뉴스1	239,590	49,378,033
뉴스웍스	21,001	4,859,703
뉴스투데이	13,415	3,372,770
뉴시스	298,473	61,762,925
대경일보	27,681	4,922,640
대전투데이	9,045	1,716,373
디지털타임스	61,743	14,859,088
오마이뉴스	8,584	3,018,822
오에스이엔	76,039	20,184,194
울산제일일보	8,273	1,598,154
일간경기	12,127	2,275,484
일간투데이	36,812	7,074,455
전남매일	15,313	3,002,006
충남일보	28,527	4,940,647
충청매일	10,033	1,762,314
합계	974,034	207,729,734

<표 25> 신문 기사 정제 총괄표

- 매체 구분별 정제 현황

구분	수집 매체명	매체 수	총기사수	총어절수
인터넷	뉴스투데이, 오마이뉴스, 오에스이엔	3	98,038	26,575,786
종합지	경향신문, 디지털타임스, 일간투데이	3	147,149	34,629,066
지역지	경기도민일보, 경남매일, 경인매일, 굿모닝충청, 대경일보, 대전투데이, 울산제일일보, 일간경기, 전남매일, 충남일보, 충청매일	11	169,783	30,524,221
통신사	뉴스1, 뉴스웍스, 뉴시스	3	559,064	116,000,661
합계		20	974,034	207,729,734

<표 26> 매체 유형별 구축 어절수

- 월별 구축 현황

월	목표 어절 수	구축 어절 수
1월	12,500,000어절 이상	17,602,915
2월	12,500,000어절 이상	15,851,781
3월	12,500,000어절 이상	16,913,055
4월	12,500,000어절 이상	17,959,582
5월	12,500,000어절 이상	18,082,216
6월	12,500,000어절 이상	17,458,866
7월	12,500,000어절 이상	19,063,031
8월	12,500,000어절 이상	15,620,795
9월	12,500,000어절 이상	16,632,838
10월	12,500,000어절 이상	18,192,150
11월	12,500,000어절 이상	17,818,231
12월	12,500,000어절 이상	16,534,274
합계	1.5억 어절 이상	207,729,734

<표 27> 월별 구축 어절 수

- 주제별 구축 현황

주제별	기사 수	비율	어절 수	비율
경제	127,466	13%	27,958,541	13%
문화	39,649	4%	7,828,788	4%
미용/건강	44,413	5%	9,741,335	5%
사회	297,103	30%	59,417,057	29%
생활	138,821	14%	26,640,109	13%
스포츠	70,570	7%	17,281,326	8%
연예	76,895	8%	16,813,418	8%
정치	113,659	12%	27,162,710	13%
정보통신/과학	65,458	7%	14,886,450	7%
합계	974,034	100	207,729,734	100

<표 28> 주제별 기사 및 구축 어절 수

2. 온라인 매체 게시 자료 구축 결과

구분	수집 게시글 수	계정 수
블로그	73,176	109
누리소통망	42,937	93
합계	116,113	202

<표 29> 온라인 매체 게시 자료 수집 총괄표

- 정제 구축 결과

구분	목표량(건수)	구축 건수	목표 어절	구축 어절
블로그	50,000 이상	73,176	50,000,000	48,185,315
누리소통망	50,000 내외	42,937		3,558,753
합계		116,113		51,744,068

<표 30> 온라인 매체 게시 자료 정제 총괄표

- 주제별 구축 건 / 어절 분포

주제별	게시글 수	비율	어절 수	비율
문화/예술	20,165	17.4%	12,579,125	24.6%
생활/쇼평	51,738	44.6%	15,751,129	30.3%
지식/교육	17,677	15.2%	9,100,601	17.5%
취미/여행	26,531	22.8%	14,312,157	27.6%
특정 주제 없음/주제 판단 불가	2	0.0%	1,056	0.0%
합계	116,113	100%	51,744,068	100%

<표 31> 온라인 매체 게시 자료 구축 주제별 구축량

- 상세 주제별 구축 건 / 어절 분포

대분류	소분류	건수	어절 수	건수 비율	어절 수 비율
문화/ 예술	게임/캐릭터	3,272	2,160,486	2.8%	4.2%
	공연/전시/박람회	3,957	1,715,268	3.4%	3.3%
	도서/문학	6,185	3,786,218	5.3%	7.3%
	영화/드라마/방송	4,605	3,494,068	4.0%	6.8%
	음악/음반/콘서트	1,099	858,354	0.9%	1.7%
	인물/연예인	1,047	564,731	0.9%	1.1%
생활/ 쇼핑	가족/연애/모임	7,904	1,688,612	6.8%	3.3%
	반려동물/키우기	3,766	2,002,290	3.2%	3.9%
	요리/음식	5,693	1,825,932	4.9%	3.5%
	인테리어/집꾸미기	2,720	885,206	2.3%	1.7%
	일상생활	10,574	3,241,905	9.1%	6.3%
	제품구매/소개	12,995	3,899,171	11.2%	7.5%
	패션/미용/뷰티	8,086	2,208,013	7.0%	4.3%
지식/ 교육	교육/취업/직업	3,195	1,555,509	2.8%	3.0%
	기타지식고유/전문 지식	4,831	3,064,475	4.2%	5.9%
	기타취미활동	3,195	1,379,361	2.8%	2.7%
	의료/건강	2,992	811,858	2.6%	1.6%
	정치/경제/사회	3,464	2,289,398	3.0%	4.4%
취미/ 여행	기타취미활동	2,873	1,319,563	2.5%	2.6%
	여행/관광	9,200	5,296,314	7.9%	10.2%
	요리/음식	12,425	6,171,875	10.7%	11.9%
	운동/스포츠	2,033	1,524,405	1.8%	2.9%
기타	특정주제없음/주제 판단불가	2	1,056	0.0%	0.0%
합계		116,113	51,744,068	100%	100%

<표 32> 온라인 매체 게시 자료 구축 상세 주제별 구축량

3. 매체별 파일명 규칙

신문 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	25	00000001	뉴스투데이
N	I	RW	25	00000002	오마이뉴스
N	I	RW	25	00000003	오에스이엔
N	L	RW	25	00000001	경기도민일보
N	L	RW	25	00000002	경남매일
N	L	RW	25	00000003	경인매일
N	L	RW	25	00000004	대경일보
N	L	RW	25	00000005	대전투데이
N	L	RW	25	00000006	울산제일일보
N	L	RW	25	00000007	일간경기
N	L	RW	25	00000008	전남매일
N	L	RW	25	00000009	충남일보
N	L	RW	25	00000010	충청매일
N	P	RW	25	00000002	디지털타임스
N	W	RW	25	00000001	경향신문
N	W	RW	25	00000003	일간투데이
N	Z	RW	25	00000001	굿모닝충청
N	Z	RW	25	00000002	뉴스1
N	Z	RW	25	00000003	뉴스웍스

<표 33> 신문 기사 말뭉치 파일명

온라인 매체 게시 자료 말뭉치의 구분 정보는 다음과 같다.

말뭉치 유형 구분	매체 분류	분석 층위 구분	구축 연도	8자리 일련번호
E: 온라인 매체 게시 자료 말뭉치	B: 블로그 S: 누리소통망	RW: 원시 말뭉치	25	00000001~ 99999999

<표 34> 온라인 매체 게시 자료 말뭉치 파일명

<부록 1>

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

■ 저작권 이용 허락 계약서 양식(안): 신문 기사 원문 자료

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

저작자 및 저작재산권 이용허락자 ○○ (이하 “권리자”라 함)와/과 저작재산권 이용자 국립국어원(이하 “이용자”라 함)은 아래 저작물에 관한 저작재산권 이용허락과 관련하여 다음과 같이 계약을 체결한다.

제1조 (계약의 목적)

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

제2조 (정의)

이 계약에서 사용하는 용어의 뜻은 다음과 같다.

- ① ‘전체 기사’란 권리자가 제공하는 ____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 저작재산권 중 이 조에 명시한 이용허락 범위에 필요한 권리로 한다.

저작물: ____년 ____월 ____일 ~ ____년 ____월 ____일까지의 기사 중 권리자가 저작재산권을 이용허락할 권리를 보유한 기사

저작재산권 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 저작권을 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 등이 존재하는 경우, 이용자에게 그 사실을 사전에 알려야 한다.

제6조 (이용자의 권리 및 의무)

- ① 이용자는 ‘대상저작물’을 제4조의 이용허락 기간 동안 제3조의 이용 허락을 받은 범위 내에서 비독점적으로 자유롭게 이용할 수 있다.
- ② 이용자는 과업수행자를 통해 별지 이용료를 지급하되 지급방법은 부속합의서로 정한다. 이용허락 기간 자동 갱신에 따른 추가적인 이용료는 발생하지 않는다.
- ③ 이용자는 관례적으로 저작자 및 저작권자의 성명 등 표시를 허용하는 ‘대상저작물’을 이용하는 경우, 그 저작자 및 저작권자의 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 저작권 이용허락 계약을 체결하는 데 필요한 권리 및 권한을 적법하게 보유하고 있다는 것
 2. ‘대상저작물’에 대하여 이용자에게 사전에 알린 제3자의 권리 외에는 이용자의 이용을 제한할 수 있는 부담이 더 이상 존재하지 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 적절한 기간을 정하여 상대방에게 그 시정을 최고하고, 상대방이 그 기간이 지나도록 이행하지 않는 경우에는 계약을 해지할 수 있다. 다만, 상대방이 명백한 시정 거부 의사를 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 권리자와 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 권리자와 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)

이 계약의 효력은 계약 체결일로부터 발생한다.

2025년 월 일

권리자

주소

성명

(인)

국립국어원

서울특별시 강서구 금남화로 154

국립국어원장

(인)

제안요청서에 삽입된 계약서 형태

■ 저작권 이용 허락 계약서 양식(안): 온라인 매체 게시 자료

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

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

제1조 (계약의 목적)

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

제2조 (계약의 대상)

이 계약의 이용허락 대상이 되는 권리는 아래의 저작물(이하 “대상저작물”)에 대한 저작재산권 중 당사자가 합의한 권리로 한다.

저작물: 저작자 및 저작재산권자가 국립국어원의 온라인 매체 게시 자료 수집 및 정제 사업 기간(0000년 0월 0일부터 0000년 0월 0일까지) 동안 위 사업에 제공하는 모든 온라인 매체 게시 자료

저작자 및 저작재산권자: _____

종별: ☐ 어문저작물

권리: ☐ 복제권 ☐ 공중송신권 ☐ 배포권 ☐ 2차적저작물작성권 ☐ 편집저작물작성권 ☐ 부차권

저작재산권 이용허락 범위

1. 국립국어원 및 과업수행자가 ‘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 ‘대상저작물’ 및 ‘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 ‘대상저작물’ 및 ‘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
5. 국립국어원 및 과업수행자가 ‘대상저작물’에 대한 요약본을 만드는 등 ‘대상저작물’의 부차적 이용을 허락하는 일

제3조 (이용허락 기간)

‘대상저작물’ 및 ‘복제·변형물’의 이용허락 최소 기간은 계약체결일부터 ____년 __월 __일까지로 한다. 최소 기간 만료되기 3개월 전까지 권리자가 이용허락 중지 의사를 밝히지 않으면 이용허락이 5년 단위로 자동 갱신되며, 권리자가 이용허락 중지 의사를 밝히면 그 의사 내용에 따라 이용허락이 중지된다.

제4조 (권리자의 의무)

- ① 권리자는 이용자에게 이 계약서 제2조에 따른 저작재산권을 이용할 권리 및 제3자에게 재이용을 허락할 권리를 제3조의 기간 동안 비독점적으로 허락한다.
- ② 권리자는 이용자에게 계약 체결일부터 10일 이내에 ‘대상저작물’의 이용을 위해 필요한 적절한 자료를 인도하여야 한다. 이때 자료를 인도하는 형식과 방법은 부속합의서에 따른다. 다만, 대상저작물이 한국저작권위원회에 등록되어 있지 않은 경우 이용자가 요청하면 이용 허락자는 대상저작물의 저작재산권을 등록한 후 위 의무를 이행한다.
- ③ 권리자는 ‘대상저작물’에 이 계약 이행에 지장을 주는 제3자의 이용허락권, 질권 등이 존재하는 경우, 이용자에게 그 사실을 사전에 알려야 한다.
- ④ 권리자는 ‘대상저작물’의 저작재산권 전부 또는 일부를 제3자에게 양도하거나 이에 대하여 질권을 설정하고자 하는 경우, 사전에 이용자에게 이 사실을 통보하여야 한다.

제5조 (이용자의 권리 및 의무)

- ① 이용자는 ‘대상저작물’을 제3조의 이용허락 기간 동안 제2조의 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 저작자의 특별한 의사가 있는 경우에는 이에 따라 표시하며, 별도의 특별한 의사가 부존재하는 경우에는 저작자가 실명 또는 이명을 표시한 바에 따라 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.

제6조 (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’의 내용이 제3자의 저작권, 상표권, 인격권을 비롯한 일체의 권리를 침해하지 아니한다는 것
 3. ‘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 제2조에 규정된 저작권재산권 이용허락 범위와 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 제2조에 규정된 저작권재산권 이용허락 범위와 조건 및 재양도 금지, 목적 외 사용 금지 등 주의사항을 고지할 것

제7조 (계약 내용의 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.

제8조 (계약의 해지)

- ① 당사자는 천재지변 또는 기타 불가항력으로 계약을 유지할 수 없는 경우에 이 계약을 해지할 수 있다.
- ② 당사자는 상대방이 정당한 이유 없이 이 계약을 위반하는 경우에 적절한 기간을 정하여 상대방에게 그 시정을 최고하고, 상대방이 그 기간이 지나도록 이행하지 않는 경우에는 계약을 해지할 수 있다. 다만, 상대방이 명백한 시정 거부 의사를 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.

제9조 (손해 배상)

당사자가 정당한 이유 없이 이 계약을 위반하는 경우, 그로 인하여 상대방에게 발생한 모든 손해를 배상할 책임이 있다. 다만, 제8조 1항의 사유로 이 계약을 이행하지 못한 경우에는 손해배상책임을 면한다.

제10조 (비용의 부담)

계약 체결에 따른 비용은 이용자가 전부 부담한다.

제11조 (분쟁 해결)

- ① 이 계약에서 발생하는 모든 분쟁은 권리자와 이용자가 상호 원만한 합의에 이르도록 노력하여야 하며, 분쟁이 원만히 해결되지 않는 경우에는 소 제기에 앞서 한국저작권위원회에 조정을 신청할 수 있다.
- ② 제1항에 따라 해결되지 않을 때에는 대한민국의 민사소송법 등에 따른 관할법원의 소송에 의해 해결토록 한다.

제12조 (비밀 유지)

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

제13조 (기타 부속 합의)

① 권리자와 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)

이 계약의 효력은 계약 체결일부터 발생한다.

2025년 월 일

주소

국립국어원

주민등록번호

서울특별시 강서구 금남화로 154

성명

(인)

국립국어원장

(인)

제안요청서에 삽입된 계약서 형태

<부록 2>

신문 기사 데이터 정제 작업 지침

데이터 정제 작업 지침

□ 사용하지 않는 기사의 표시

삭제 기사 구분	내용
저작권 관련 검토 필요 기사	- 외부 기고가가 작성한 기사 - 교수, 박사, 의원, 소장, 대표, 변호사 등 - 해당 언론 소속 기자 이외의 모든 직업 및 직책 - 동일 미디어 그룹 내의 다른 매체 소속 기자가 작성한 기사 - 명예 기자, 객원 기자, 시민 기자, 도민 기자, 학생 기자가 작성한 기사 - 명예 기자나 객원 기자라고 표기되지 않고, 이름만 나오는 기사는 기고문 형식을 확인한 후 사용 - 방송, 라디오 방송, 유튜브 라이브를 그대로 옮겨 적은 기사 - 외부 기고문임을 의미하는 단어를 포함하는 기사 [기고], [특별기고], [발언대], [확대경], [아침의 창] 등 - 매체별로 다양하게 쓰이고 있으므로 반드시 확인 - 기자명이 있으나 실제 외부 기고 형식으로 작성된 기사 - 기자 정보가 본문에만 존재하고 출처가 명확하지 않은 기사 - 기자명이 비어있는 기사 - 외국 기사를 번역한 기사 - 다른 매체의 머리기사나 단신, 일정을 모아 놓은 기사 00월 00일 언론 동정, 00일 국회 일정 등 - 공동 취재단이 작성한 기사 - 공동 취재단, 공동 취재팀, 공동 취재반, 특별 취재단 등 ‘공동 000, 공동000’과 같이 띄어쓰기를 달리한 경우가 있으므로 확인 필수 - 연합뉴스 등 통신사발 기사
구어체 기사	- 기사의 대부분이 구어체로 이루어진 기사는 사용하지 않음 - 친근하게 전달하기 위해 구어체로 쓴 기사, 인터뷰를 구어체 그대로 옮긴 기사 등
불필요한 정보를 삭제한 후 기사 내용이 짧은 기사	불필요한 요소를 삭제하고 남은 기사가 최소 어절 수에 못 미치는 경우, 기사를 사용하지 않음

불완전하게 종료되는 기사	• ‘관계자는...라고 말’, ‘정부는...예정’ 처럼 ‘했다.’, ‘다.’가 빠진 것으로 유추할 수 있는 경우에는 사용 • 누락된 글자를 유추하여 기사의 마지막 문장을 완성할 수 있는 경우에는 사용하지만, 문장의 대부분 또는 주요 성분이 누락되어 유추가 어려운 수준으로 불완전한 기사는 사용하지 않음
한글이 모아쓰기 되지 않은 기사	• 한글이 모아쓰기 되지 않은 경우, 원래의 단어를 확인할 수 있으면 사용하고, 수정 후보가 많으면 사용하지 않음
명확한 광고 기사	• 기사에서 명확히 광고라고 표기하는 경우에는 사용하지 않음 – [AD], <광고> 등의 표시가 있는 기사, 협찬, 제공, 지원, 협조를 받아 작성하였음을 밝힌 기사도 사용하지 않음
인공 지능이 작성한 기사	• 인공 지능이 작성하였음을 밝힌 기사는 사용하지 않음 – “본 기사의 제목은 챗GPT가 작성한 것임을 밝힙니다.” 등의 문구로 인공 지능 작성 여부를 밝힌 기사
단순 기사	• 날씨, 승진, 부고, 운세, 전보, 임용, 스포츠 득점 정보, 여론 조사 결과, 출구 조사 결과, 어록 모음

□ 기사 본문 내 불필요한 정보의 삭제

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

삭제 정보	예시
표, 그림, 그래프 등의 캡션 정보	[사진 제공= 로이터] 사진 제공=ENA, SBS Plus 로이터=○○○ [사진설명] 사진/계양구 [그래픽] <그래픽> 일러스트 KBS 연예대상 캡처 (사진설명) 표> 공정위 망 이용대가 불공정 조사 쟁점 (자료 제공=학도의용군전승기념관) (사진은 기사 내용과 무관함) ○○○ 단지 내 어린이집 개원식. 사진제공=○○구청 ◎ 사진 설명 : '제53회 전국소년체육대회' 참여한 세 종대표선수들이 11일에 열린 시상식에 참여해 기념 촬영을 하고 있다. [표] 2024년 상반기 요양기관(서울) 관외진료현황
기자의 이름, ID 등	○○○○=제주, ○○○ 기자] [서울=○○일보] 외교부 공동취재단·○○○ 기자 = 【서울=○○○】○○○ 기자 = /○○○ 기자·연합뉴스
‘Copyright©’ 등 저작권 관련 내용	마이크 존슨 미 하원의장. 2024.03.07 © 로이터 [사진]©Gettyimages(무단전재 및 재배포 금지)
전문	조교로 국방의 의무를 다했던 진은 특급전사를 따면서 두 번의 조기진급 소식을 전해 화제를 모았다. 모범적인 군 생활 중에서도 틈틈이 위버스에 등장해 팬들과 소통을 이어가기도 했다. 동생 멤버들이 차례로 입대할 땐 센스있는 반응으로 팬들을 웃음짓게 만들기도. 그랬던 진이 이제 곧 민간인이 된다. 소속사로서는 오랜만에 만나는 진과 팬들을 위해 특급 이벤트를 마련했지만 의도치 않게 원성이 커지자 재빨리 피드백을 수용했다. 다음은 방탄소년단 진 측 입장 전문이다. 안녕하세요, 빅히트 뮤직입니다. ‘2024 FESTA’ 진 오프라인 행사의 응모 기준이 변경 예정인 점 안내 드립니다. 앞선 공지 내용이 불명확했던 점과 더불어 미리 세심하게 응모 기준을 설정하지 못한 점 사과드립니다. 금일(6/2) 오전 11시(KST) 오프라인 행사 응모 공지 이후 weverse shop 상세 페이지를 통해 오프라인 행사에 응모하신 앨범 구매자분들과 공지 이전 weverse shop에서 앨범을 구매한 이력이 있으신 분들 모두를 고려하는 방향으로 응모 기준을 보완할 예정입니다. 충분히 숙고한 뒤 빠른 시일 내 상세 내용을 공지하도록 하겠습니다. 진의 마음이 팬 여러분께 잘 전달될 수 있게 최선을 다해 행사를 준비하겠습니다. 팬 여러분께 불편을 끼쳐 드린 점 다시 한번 사과의 말씀 드리며 오는 6월 13일 진행될 진의 오프라인 행사에 많은 응원과 관심 부탁드립니다 감사합니다. /-----@----.co.kr [사진] ----- DB

기사 본문으로 볼 수 없는 부가 정보 의 나열 등	김 당선인은 미래 청년들에 대한 애착도 드러냈다. 그는 "청년들이 우리 사회에 더욱 희망을 가질 수 있도록 미래 세대에 보탬이 되는 방향으로 저부터 의정 활동을 펼쳐나가겠다"고 강조한 뒤 "특히 저출산, 지방소멸 등 미래 세대가 보다 나은 삶을 살아갈 수 있는 미래 담론들을 다루는데 주저하지 않을 것"이라고 약속했다. 김해시민들에게 전하고 싶은 말에 대해 "3선 중진위원이 돼서 그동안 추진해왔던 김해와 경남 현안 해결에 박차를 가하겠다. 조선의원 때의 치열한 마음가짐과 겸허한 자세로 3선 중진의 지혜와 경륜을 더해 지역발전과 함께 민주주의와 민생 회복을 위해 열심히 뛰겠다"고 강조했다. ◇김정호 당선인 약력 △1960년 제주 △부산남고 △부산대 경제학과 △부산민족민주연합 사무차장 △청와대 총무비서관실 선임행정관, 대통령기록관리 비서관 △(주)봉하마을 대표이사 △제19대 대통령선거 문재인 후보 농업정책 특보 △제20, 21대 국회의원 한편 구하라는 지난 2008년 그룹 카라에 합류해 가요계에 데뷔했다. 이후 '프리티 걸' '허니' '미스터' '맘마미아' '루팡' '판도라' 등 다수의 히트곡을 냈고, 일본에도 진출해 큰 사랑을 받았다. 또한 2015년에는 솔로 앨범을 발매하고 다수의 드라마 OST 참여하는 등 솔로 가수로도 두각을 나타내며 한국과 일본에서 활발히 활동해 왔다. ※ 우울감 등 말하기 어려운 고민이 있거나 주변에 이런 어려움을 겪는 가족·지인이 있을 경우 자살예방 상담전화 ☎1393, 정신건강 상담전화 ☎1577-0199, 희망의 전화 ☎129, 생명의 전화 ☎1588-9191, 청소년 전화 ☎1388, 청소년 모바일 상담 '다 들어줄게' 앱, 카카오톡 등에서 24시간 전문가의 상담을 받을 수 있습니다. 권 수석대변인은 "해당 차량은 김씨를 숙소내내려주고 바로 현장을 떠난 것으로 확인됐다"며 "부산경찰청은 지난 3일 브리핑에서 '피의자가 공범은 없고 개인적으로 범행했다고 진술했다'고 밝힌 바 있다"고 했다. 이어 "하지만 김씨를 모텔까지 데려다준 차량이 존재했다는 사실이 드러남에 따라 해당 차량 운전자의 신원과 김씨와의 관계를 반드시 확인해야 한다"며 "경찰은 이번 사건의 범행 동기와 공범여부에 대해 한 점의 의혹도 남지 않도록 엄정하고 신속하게 수사할 것을 촉구한다"고 말했다. 앞서 헤럴드경제는 이 대표 피습사건 피의자 김모씨가 지난 1일 경남 창원시 진해구 한 모텔 앞에 서 있는 모습이 폐쇄회로(CC)TV에 포착됐다고 밝혔다. 당시 김씨는 회색 벤츠로 추정되는 차량에서 내려는데 이 차량은 김 씨를 내려준 후 곧바로 떠났다. <!--/data/user/0/com.samsung.android.app.notes/files/clipdata/clipdata_bodytext_240106_124742_640.sdocx--> ◎-----@---.com 특히 올해 축제의 메인 프로그램이었던 'GXG 2024 게임음악 경연대회 : The 1st GXG SOUND TRACK'에서는 총상금 2,300만원을 걸고 총 10팀이 본선 경연을 펼쳤다. 모든 팀의 공연이 끝나고 발표된 최후의 승자는 라이엇게임즈의 리그오브레전드를 모티브 게임으로 참가한 '리프(Lif)'팀으로, 상금 1,000만원의 주인공이 되었다. 리프(Lif)팀은 특색있는 락 장르의 창작곡 '슈리마'로 열정과 개성 넘치는 공연을 선보이며 심사위원 평가 및 현장 관객 투표에서 높은 점수를 획득하였고, 결국 영예의 대상을 차지했다. [경연대회 최종 순위] ▲대상 : 리프 ▲금상 : Run into Galaxy ▲은상 : 키프로&수자인, e-dust ▲동상 : 목화, 데마시아 ▲본상(가나다 순) : dizzy, DPS, 백안, Unknown tree / 이상 총 10팀
--	--

	이외에도 일러스타 페스 스테이지, WCG 레전드 슈퍼 매치, 갈라 콘서트 등 다채로운 무대 프로그램을 선보인 GXG 2024는 게임음악 경연대회 참가팀들의 갈라 콘서트를 마지막으로, 축제의 막을 내렸다. 특히 올해 축제는 공간 뿐만 아니라, 참가 기관·기업 수에서도 작년 대비 확장된 규모로 진행된 만큼, 더욱 많은 즐거움을 선사했다. <u>[○○○○=○○○ 기자]</u> 12일 아침 최저기온이 영하 7도까지 내려간다. 한낮에도 최고기온이 12도에 불과할 정도로 당분간 ‘동장군’이 맹위를 떨치겠다. 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. 남부지방에는 비 소식이 예정됐다. 제주도에는 12일 오후부터 13일 오<u>[○○○○=○○○ ○ 기자]</u><u>[○○○○=○○○ 기자]</u>전 사이에, 전라 서해안은 12일 밤부터 13일 새벽 사이 5mm 안팎의 비가 내리겠다. 제주도 산지에는 1~3cm의 눈이 내릴 수도 있다. ===== 기자 정보가 기사 상단, 하단이 아닌 곳에 위치한 경우이다. 기자 정보는 이처럼 기사의 곳곳에 남아 있을 수 있으므로 주의가 필요하다. 현대자동차가 고객이 원하는 일을 할 수 있도록 가능성을 열어주고, 그 일을 스마트하게 할 수 있도록 비즈니스에 최적화된 모빌리티를 선보였다. 개발 단계에서부터 고객사의 니즈를 적극 수용해 최적화한 솔루션을 만들었으며, 확장성까지 갖춰 지속적으로 서비스를 업데이트할 수 있게 했다. 현대차는 이를 통해 소프트웨어 중심 자동차(SDV)로의 전환을 가속화할 계획이다. <u>&lt;iframe width="100%" height="465"</u> <u>src="https://www.youtube.com/embed/KEG-guY8Tr8" title="[임주희 기자의 통통통 차차차] &quot;나만의 일터&quot; 무한확장 전기차</u> <u>&#39;ST1&#39;" frameborder="0" allow="accelerometer; autoplay;</u> <u>clipboard-write; encrypted-media; gyroscope; picture-in-picture; web-share"</u> <u>referrerpolicy="strict-origin-when-cross-origin" allowfullscreen&gt;&lt;/iframe&gt;</u> <u>&lt;iframe src='https://tv.naver.com/embed/50407343?autoPlay=true'</u> <u>frameborder='no' scrolling='no' marginwidth='0' marginheight='0' WIDTH='100%'</u> <u>HEIGHT='466' allow='autoplay' allowfullscreen&gt;&lt;/iframe&gt;</u> 현대차는 새로운 전동화 비즈니스 플랫폼인 ST1의 물류 특화 모델 '카고'와 '카고 냉동'을 출시한다고 24일 밝혔다. ===== 기사 본문 가운데 영상이나 표 등 비텍스트적 요소가 포함된 경우이다. 다양한 유형이 있으므로 작업에 주의하여 삭제한다.
기사와 관련 없는 문장	봉사단원 10명은 1개월간의 사전교육을 끝마치고 22일부터 내달 1일까지 2주간 키르기스스탄 비슈케크지역 초·중등학생을 대상으로 한국어 교육 및 한국문화 체험(K-Pop, 태권도 등) 봉사를 한다. ○○○ 사범대학장은“예비교사들이 해외교육봉사를 통해 타인에 대한 배려와 책임을 실천하는 정신을 배우고, 타 문화에 대한 새로운 안목을 갖출 수 있는 소중한 시간이 되길 바라며 모범적인 교육봉사가 이루어지도록 적극 지원하겠다”고 말했다. <u>국립○○대학교 홍보센터 ○ ○ ○</u> <u>사무실: ***-***-**** / 폰 ***-***-****</u> <u>◎-----@---.com</u>

	김태규 국민권익위 부위원장은 “최근 공동주택에서 안전사고가 자주 발생, 국민들이 안전에 대한 우려가 매우 큰 상황”이라며 “앞으로도 주민 의견을 반영한 해결책을 찾는 데 최선의 노력을 다하겠다”고 말했다. <u><사진은 오후 늦은 시간에 보낼 예정입니다></u> ===== 기사 작성시 삭제되었어야 할 정보가 그대로 포함된 경우이다. 기사 본문에 해당하지 않는 부분으로 보고 삭제한다.
반복 오류	당시 눈이 <u>허리까지</u> <u>허리까지</u> 차올라 차량 진입도 어려운 데다, 통행로가 보이지 않아 자칫하면 대원들까지 절벽 아래로 추락할 수 있는 아찔한 상황이었다. 강원도는 대체로 흐린 가운데 강릉과 영서 <u>일부지역에</u> <u>일부지역에</u> 빗방울이 떨어지는 곳이 있겠다.
[기타] 유의 사항	‘민주·개혁·진보연<u>합의 합의</u> 정신’, ‘계<u>다가 다가</u>올 보궐선거를’ 같은 문장들은 바르게 사용된 경우이므로, 단어 반복 오류로 보고 삭제하지 않도록 유의한다.

<부록 3>

신문 기사 말뭉치 종류별 구축 예시

원시 데이터	<u>[서울=○○○○]○○○ 기자</u> = 과학기술정보통신부가 기관별로 산재된 국가 나노랩 서비스를 온라인에서 한 곳에 모아 제공하는 서비스를 본격 시작한다. 과기정통부는 '랩 서비스 통합정보시스템(MoaFab·모아랩)' 서비스를 개시했다고 25일 밝혔다. 모아랩은 지난 1월8일부터 두 달여 간 나노종합기술원, 한국나노기술원, 나노융합기술원 3개 기관을 대상으로 시범 운영하는 기간을 가졌다. 이를 통해 ▲사용자 편의성 향상 ▲장비 관리 지원기능 추가 ▲모아-나노랩 기관 시스템 간 인터페이스 오류 최소화 등을 추진하여 시스템 안정화를 진행했다. 시범서비스 기간 중 모아랩 서비스를 신청한 건수는 총 3377건이다. 1월 169건, 2월 767건, 3월 2441건으로 서비스 이용이 점차 확대됐다. 하반기엔 한국전자통신연구원(ETRI), 서울대학교 반도체공동연구소(ISRC), 대구경북과학기술원(DGIST)으로 지원기관을 확대할 예정이다. 이종호 과기정통부 장관은 "앞으로 전국의 나노랩 서비스는 MoaFab으로 통합 제공하여 이용자들이 필요로 하는 서비스를 편리하게 제공할 것"이라며 "MoaFab 서비스 제공을 기점으로 국가 나노·반도체 경쟁력이 한 단계 도약할 것으로 기대한다"고 말했다. <u>©-----@-----.com</u>
신문 기사 발문치	과학기술정보통신부가 기관별로 산재된 국가 나노랩 서비스를 온라인에서 한 곳에 모아 제공하는 서비스를 본격 시작한다. 과기정통부는 '랩 서비스 통합정보시스템(MoaFab·모아랩)' 서비스를 개시했다고 25일 밝혔다. 모아랩은 지난 1월8일부터 두 달여 간 나노종합기술원, 한국나노기술원, 나노융합기술원 3개 기관을 대상으로 시범 운영하는 기간을 가졌다. 이를 통해 ▲사용자 편의성 향상 ▲장비 관리 지원기능 추가 ▲모아-나노랩 기관 시스템 간 인터페이스 오류 최소화 등을 추진하여 시스템 안정화를 진행했다. 시범서비스 기간 중 모아랩 서비스를 신청한 건수는 총 3377건이다. 1월 169건, 2월 767건, 3월 2441건으로 서비스 이용이 점차 확대됐다. 하반기엔 한국전자통신연구원(ETRI), 서울대학교 반도체공동연구소(ISRC), 대구경북과학기술원(DGIST)으로 지원기관을 확대할 예정이다. 이종호 과기정통부 장관은 "앞으로 전국의 나노랩 서비스는 MoaFab으로 통합 제공하여 이용자들이 필요로 하는 서비스를 편리하게 제공할 것"이라며 "MoaFab 서비스 제공을 기점으로 국가 나노·반도체 경쟁력이 한 단계 도약할 것으로 기대한다"고 말했다.

인용 부호 수정 말뭉치	<p>과학기술정보통신부가 기관별로 산재된 국가 나노랩 서비스를 온라인에서 한 곳에 모아 제공하는 서비스를 본격 시작한다.</p> <p>과기정통부는 ‘앱 서비스 통합정보시스템(MoaFab·모아앱)’ 서비스를 개시했다고 25일 밝혔다.</p> <p>모아앱은 지난 1월8일부터 두 달여 간 나노종합기술원, 한국나노기술원, 나노융합기술원 3개 기관을 대상으로 시범 운영하는 기간을 가졌다. 이를 통해 ▲사용자 편의성 향상 ▲장비 관리 지원기능 추가 ▲모아-나노랩 기관 시스템 간 인터페이스 오류 최소화 등을 추진하여 시스템 안정화를 진행했다.</p> <p>시범서비스 기간 중 모아앱 서비스를 신청한 건수는 총 3377건이다. 1월 169건, 2월 767건, 3월 2441건으로 서비스 이용이 점차 확대됐다.</p> <p>하반기엔 한국전자통신연구원(ETRI), 서울대학교 반도체공동연구소(ISRC), 대구경북과학기술원(DGIST)으로 지원기관을 확대할 예정이다.</p> <p>이종호 과기정통부 장관은 “앞으로 전국의 나노랩 서비스는 MoaFab으로 통합 제공하여 이용자들이 필요로 하는 서비스를 편리하게 제공할 것”이라며 “MoaFab 서비스 제공을 기점으로 국가 나노·반도체 경쟁력이 한 단계 도약할 것으로 기대한다”고 말했다.</p>
문장 말뭉치	<p><s>과학기술정보통신부가 기관별로 산재된 국가 나노랩 서비스를 온라인에서 한 곳에 모아 제공하는 서비스를 본격 시작한다.</s></p> <p><s>과기정통부는 ‘앱 서비스 통합정보시스템(MoaFab·모아앱)’ 서비스를 개시했다고 25일 밝혔다.</s></p> <p><s>모아앱은 지난 1월8일부터 두 달여 간 나노종합기술원, 한국나노기술원, 나노융합기술원 3개 기관을 대상으로 시범 운영하는 기간을 가졌다.</s> <s>이를 통해 ▲사용자 편의성 향상 ▲장비 관리 지원기능 추가 ▲모아-나노랩 기관 시스템 간 인터페이스 오류 최소화 등을 추진하여 시스템 안정화를 진행했다.</s></p> <p><s>시범서비스 기간 중 모아앱 서비스를 신청한 건수는 총 3377건이다.</s> <s>1월 169건, 2월 767건, 3월 2441건으로 서비스 이용이 점차 확대됐다.</s></p> <p><s>하반기엔 한국전자통신연구원(ETRI), 서울대학교 반도체공동연구소(ISRC), 대구경북과학기술원(DGIST)으로 지원기관을 확대할 예정이다.</s></p> <p><s>이종호 과기정통부 장관은 “앞으로 전국의 나노랩 서비스는 MoaFab으로 통합 제공하여 이용자들이 필요로 하는 서비스를 편리하게 제공할 것”이라며 “MoaFab 서비스 제공을 기점으로 국가 나노·반도체 경쟁력이 한 단계 도약할 것으로 기대한다”고 말했다.</s></p>
- 신문 기사 말뭉치: 원시데이터에서 기자 정보, 캡션 정보 등 불필요한 요소를 제거한 말뭉치 - 인용 부호 수정 말뭉치: 신문 기사 말뭉치에서 인용 부호를 수정한 말뭉치 - 문장 말뭉치: 인용 부호 수정 말뭉치에서 단락을 문장 단위로 분할한 후 문장 단위로 <s> 태그를 부착한 말뭉치	

<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	- 기사의 마지막 문장이 마침표(.)로 끝나지 않은 기사를 검색합니다. - 처리되지 않은 기자 정보, 사진 캡션 또는 마지막에 끊긴 기사, 마침표 이외의 기호로 끝난 기사들이 검색됩니다. <table border="1"> <tr> <td>[가-힣]</td><td>‘가’ ~ ‘힣’ 까지 한글 검색</td></tr> <tr> <td>\\n</td><td>엔터키에 의한 줄바꿈</td></tr> <tr> <td>[^\.]</td><td>마침표(.)가 아님(대괄호 안에서 ^는 not을 의미)</td></tr> </table>	[가-힣]	‘가’ ~ ‘힣’ 까지 한글 검색	\\n	엔터키에 의한 줄바꿈	[^\.]	마침표(.)가 아님(대괄호 안에서 ^는 not을 의미)
[가-힣]	‘가’ ~ ‘힣’ 까지 한글 검색						
\\n	엔터키에 의한 줄바꿈						
[^\.]	마침표(.)가 아님(대괄호 안에서 ^는 not을 의미)						
예시	착공을 시작했다. ○○○ 기자 _____} _____] _____, ____{ _____ "id"						

정규식	\\([2-9] [1-4]{1}[0-9]{1})",\n ("form").*?[가-힣][^\.]"\n				
설명	- 기사 내에서 마침표(.) 없이 줄바꿈된 행을 검색합니다. - 불필요하게 줄바꿈된 문장, 마침표를 찍지 않은 문장, 소제목이 검색됩니다. - \\([2-9] 부터 [^\.]"\n 까지 중간 공백을 포함한 전체를 하나의 정규식으로 검색하여 사용합니다. <table border="1"> <tr> <td>([2-9] [1-4]{1}[0-9]{1})</td><td>숫자 2부터 49까지 검색(1번 문장은 기사 제목이므로 제외)</td></tr> <tr> <td>{1}</td><td>바로 앞의 대괄호 [1-4], [0-9]가 지정하는 범위에서 하나를 선택하여 검색 [1-4]{1}[0-9]{1}는 1과 4 사이의 숫자 하나를, 0과 9 사이의 숫자 하나를 검색하게 되며, 각각 십의 자리, 일의 자리가 되어 10~49까지를 검색</td></tr> </table>	([2-9] [1-4]{1}[0-9]{1})	숫자 2부터 49까지 검색(1번 문장은 기사 제목이므로 제외)	{1}	바로 앞의 대괄호 [1-4], [0-9]가 지정하는 범위에서 하나를 선택하여 검색 [1-4]{1}[0-9]{1}는 1과 4 사이의 숫자 하나를, 0과 9 사이의 숫자 하나를 검색하게 되며, 각각 십의 자리, 일의 자리가 되어 10~49까지를 검색
([2-9] [1-4]{1}[0-9]{1})	숫자 2부터 49까지 검색(1번 문장은 기사 제목이므로 제외)				
{1}	바로 앞의 대괄호 [1-4], [0-9]가 지정하는 범위에서 하나를 선택하여 검색 [1-4]{1}[0-9]{1}는 1과 4 사이의 숫자 하나를, 0과 9 사이의 숫자 하나를 검색하게 되며, 각각 십의 자리, 일의 자리가 되어 10~49까지를 검색				
예시	"id": "NZRW2400000002.3. 18 ", "form": " ◇과감한 투자·사업 영역 확대 이후 시장 안착·안정화 "				

정규식	[a-z]"\n [a-z][^\.]"\n	
설명	<	

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>[^뒷]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	[^뒷]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	[^뒷]담화																	
아래는	전문은[^행]	글 전문	아래와 같다																	
[^시]장 전문	공식입장	문 전문	다음과 같다																	
설명	· 외부 전문이 그대로 사용된 내용을 검색합니다. - 입장문, 연설문, 담화문, 공식입장, SNS글, 영문기사 원문, 인사/승진, 날씨 등이 검색됩니다. - ‘아래와 같다’, ‘다음과 같다’ 이후에는 삭제요소가 등장할 확률이 높으므로 주의해야 합니다. <table><tr><td>[^맞]</td><td>^ 기호가 대괄호 [] 안에 있다면, 괄호 안에 있는 단어를 제외하고 검색</td></tr></table>				[^맞]	^ 기호가 대괄호 [] 안에 있다면, 괄호 안에 있는 단어를 제외하고 검색														
[^맞]	^ 기호가 대괄호 [] 안에 있다면, 괄호 안에 있는 단어를 제외하고 검색																			
예시	이번 우수 건축물 전시 일정은 다음과 같다. ▲경남도 본청 8월5~9일 ▲도의회 8월12~16일 ▲사천시 8월27~31일																			

정규식	명예기자 칼럼니스트 지역종합 취재단 에디터 객원기자 연합 시민기자 특별취재팀 이코노미스트 논설위원 연합뉴스 도민기자 특별취재반 리포터 칼럼 전국종합 공동취재단 전문기자 아나운서
설명	- 저작권과 관련하여 삭제가 필요한 기사, 외부 기고문을 검색합니다. - 작성자를 확인해야 하며 단순히 내용에 포함된 것만으로는 삭제하지 않습니다.
예시	<u>[파리=공동취재단]○○○ 기자</u> = 전세계 장애인 스포츠 선수들의 축제 파리 패럴림픽이 개막한다.

정규식	ㅇ ㅏ ㅣ ㄱ ㄴ ㄷ ㄹ ㅁ ㅂ ㅅ ㅇ ㅈ ㅊ ㅋ ㅌ ㅍ ㅎ ㄱ ㄷ ㅂ ㅅ ㅈ ㅏ ㅑ ㅓ ㅕ ㅗ ㅛ ㅜ ㅠ ㅡ ㅣ ㅞ ㅟ ㅠ ㅡ 과 내 저 제 니 거 너
설명	• 텍스트가 깨진 문장, 문장에 포함된 오타 등을 검색합니다. - 기자 정보에도 오타가 있을 수 있으므로 체크해야 합니다. - ‘제보자 ㅏ씨’, ‘ㅁ자형 건물’ 처럼 문맥상 필요한 자음과 모음 사용은 수정하지 않습니다.
예시	ㅇ ㅏ ㅈ ㅏ ㄷ ㅏ ㅏ ㅏ ㅏ ㅏ ㅏ 답글 것으로 기대해볼 수 있다"ㅏ녀서 21일까지(현지시간) 영ㅏ리는 이번 행사는

정규식	이 기사는 본 기획물은 위 텍스트는 대담 정리=
설명	- 인공 지능이 작성한 기사, 협찬을 받아 작성된 기사를 검색합니다. - 인공 지능 작성, 협찬/지원 여부를 명백히 밝히고 있으므로 기사를 삭제 처리합니다. - 대답을 나눈 것을 옮긴 기사를 검색하여 삭제 처리합니다.
예시	# <u>이 기사</u>는 제주특별자치도의 지원을 받아 작성했습니다. [<u>이 기사</u>는 해당 언론사의 자동 생성 알고리즘을 통해 작성되었습니다.] ※ <u>이 글</u>은 ○○○의 편집방향과 일치하지 않을 수 있습니다 [○○○ ○○○○ 전국부장 대담, 정리=○○○ 기자]

정규식	"사진 이미지= 자료 : 포스터 제공" >사진 출처 자료=, 조감도 캡처" 자료: 그림 투시도 장면"
설명	· 사진 캡션을 검색합니다. - 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.
예시	잔류 의지를 다졌다. [사진] 한국프로축구연맹 제공. 매주 월요일 오후 9시 10분에 방송된다. [사진제공]=크레아 스튜디오] 시청자들의 반응도 폭발했다. ○○○○○ 제공

정규식	" < > " _ <
설명	· 사진 캡션, 인사/승진 정보 등을 검색합니다. 정규식에서 or를 의미 (키보드에서 엔터 위 원화기호(W)+Shift 키를 눌러 입력)
예시	"<미래에셋그룹 승진 인사> ◇ 부사장 ▲ IB2부문 세 번째 전시회를 가진 바 있다. <계속>"

정규식	유니코드 2008 ‘ ’ 유니코드 2028 ‘ ’ 유니코드 3164 ‘ ’ 유니코드 202F ‘ ’ 유니코드 3000 ‘ ’
설명	· 여러가지 공백을 일반적인 공백(space bar, 유니코드 0020)으로 수정합니다. - 따옴표 사이의 공백을 복사+붙여넣기 하여 검색합니다.
예시	개선_현황을_시장에_적극_알리는데_집중할_것으로_전망된다. (☞ 유니코드 2008) 개선 현황을 시장에 적극 알리는데 집중할 것으로 전망된다. (☞ 유니코드 0020)

정규식	유니코드 FF0E ‘. ’ 유니코드 FF0C ‘, ’
설명	· 코드가 다른 마침표와 쉼표를 검색하여 일반적인 마침표와 쉼표로 수정합니다. - 따옴표 사이의 기호를 복사+붙여넣기 하여 검색합니다.
예시	재평가가 진행된다.<u>.</u> 재평가가 진행된다. 산업,<u>.</u> 문화예술,<u>.</u> 돌봄,<u>.</u> 관광,<u>.</u> 체육,<u>.</u> 산림 등 산업, 문화예술, 돌봄, 관광, 체육, 산림 등

정규식	<u>.</u> \n ^[.] ".
설명	· 마침표를 잘못 찍은 경우를 검색합니다.
예시	<u>.</u> 이와 관련해 롯데이노베이트는

정규식	다," \w,"
설명	· 마침표가 와야 할 자리에 쉼표를 잘못 사용한 것을 검색합니다. - 잘못 나뉜 문장이라면 붙여주고, 마침표가 와야 할 곳에는 부호를 마침표로 바꿔 줍니다.
예시	일반고의 교육개선 방안을 모색하겠 <u>.</u> "

정규식	([가-힣]{2,4})_1	
설명	- 반복되는 어절을 검색합니다. - 문맥에 유의하며 반복되는 어절을 삭제합니다. ‘민주·개혁·진보연합의 합의 정신’, ‘게다가 다가올 보궐선거를’같은 경우를 삭제하지 않도록 유의합니다. {2,4}에서 앞의 2는 어절의 최소 음절수, 4는 최대 음절수를 뜻합니다.	
	\1	앞에서 검색한 소괄호 ()의 내용이 동일하게 반복됨을 의미
예시	"다각적인 노력을 노력을 기울이겠다"이라고 말했다.	

정규식	다\ 다\ 따\ 쎄\ 낱\.	
설명	이후 공정에 지장을 줄 수 있으므로 문장 끝에서 자주 발생하는 오타를 검색하여 수정합니다.	
예시	긍정적 영향이 간 것으로 풀이되낱. 확대에 실질적으로 기여했디.	

정규식	[?][가-힣] [?][a-z] [?][0-9] ?? \\				
설명	특수기호, 한자 등 글자가 깨진 부분을 검색하여 수정합니다.				
예시	비상 대책에 따른 휴학 승인 원칙?절차 이행 비상 대책에 따른 휴학 승인 원칙·절차 이행				

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

□ 문장부호 수정 말뭉치

정규식	$\wedge[\wedge\text{“}]\text{“}?\text{”}$ $\wedge[\wedge\text{'}]'\text{'?}'$
설명	- 여는 따옴표(“, ‘)없이 닫는 따옴표만 있는 문장을 검색합니다. (∼”), (∼’) - 빠진 따옴표를 넣어주거나 잘못 쓰인 따옴표를 바꿔 줍니다.
예시	연구원은 15일 보고서에서 ”지난 주말 사이 미국 증시는

정규식	$\text{'}\wedge\text{'}\text{“}?\text{”}\$$ $\text{“}\wedge\text{”}\text{'}\text{'?}\text{'}\$$
설명	여는 따옴표만으로 끝난 문장을 검색합니다. (‘∼), (“∼)
예시	자체진행 사회공헌프로그램 ‘따뜻한청년들도 진행중이다.

정규식	$\text{“}_ _ \text{”} \text{'}_ _ \text{'}$
설명	- 따옴표 앞뒤로 띄어쓰기가 잘못된 문장을 검색합니다. - 띄어쓰기가 잘못된 경우 또는 따옴표의 짝이 맞더라도 뒤집혀 있는 경우가 검색됩니다.
예시	삼성전자가 약 2년여 만에 ’8만 전자’를 기록하면서, 증권가에서는 ’10만 전자’ 기대감이 한층 커지는 모양새다.

정규식	“” ”“ ‘’ ’ ‘)’ “ ” “ ” ’ ’
설명	• 잘못 쓰인 따옴표를 검색하고 확인합니다.
예시	“같이’의 가치를 느낄 수 있었다‘ “전세사기 피해자에 대한 법률복지 서비스를 강화하고 법적 권리구제를 위해 최선을 다하겠습니다"고 말했다. “”1990년대 보스니아 전쟁 당시 겪었던 위협과 협박을 떠올리게 했다“고 말했다.

<부록 5>

온라인 매체 게시 자료 비식별화 지침

온라인 매체 게시 자료 비식별화 지침

원칙적으로 비식별화의 대상은 ‘개인정보’이다. 「개인정보 보호법」 제2조 제1호에 따르면, ‘개인정보’는 살아 있는 개인에 관한 정보로서 특정 개인을 알아볼 수 있는 정보(성명, 주민등록번호 등)를 의미한다.

그러나 공공기관에서 구축하는 말뭉치로서 활용 과정에서 발생할 수 있는 모든 잠재적 논란과 분쟁을 방지하기 위해, 본 사업에서는 국립국어원의 요구 사항을 반영하여 비식별화 대상을 확대하였다. 이는 특정 개인, 단체, 상호, 상품명, 작품명 등 모든 명칭을 그대로 노출하여 발생할 수 있는 법적 위험성 및 윤리적 논란의 여지를 최소화하기 위함이다.

부정적 표현 또한 개인의 감상과 비평 같은 단순 의견을 넘어 인격 모독, 모멸감, 혐오감을 유발하는 표현은 모욕죄에 해당할 수 있으며, 주관적 평가라도 상대방의 인격권을 침해하는 수준이라면 문제가 될 수 있다. 이러한 표현들은 비식별화 처리를 통한 관리가 필요하다.

본 사업에서는 2022년 「온라인 게시자료 수집 및 정제」 사업의 비식별화 유형을 기본 방향으로 하고, 2024년 「일상대화 말뭉치 구축」 사업의 일상대화 비식별화 지침을 따라 아래와 같은 지침을 작성하고, 개인의 담론과 표현의 자유를 존중하는 방식으로 정제를 진행하며 데이터의 법적·윤리적 무결성을 확보하고자 하였다.

2022 온라인 게시자료 수집 및 정제			2024 일상대화 말뭉치 구축			2025 신문기사 원문 및 온라인 매체 게시 자료 구축		
대분류	태그	항목	대분류	태그	항목	대분류	태그	항목
이름	&name&	개인의 실명 (공인 제외)	이름	&name&	실명, 애칭, 별명, 대학교 명, 필명, 그룹명 등	이름	&name&	정치인, 일반 인의 실명
			이름	&politician&	정치인의 실 명, 애칭, 별명, 변형된 형태 등			
출신 및 소속	&affiliation&	개인의 출신 및 소속	출신	&affiliation&	학교, 지역, 온라인 커뮤니티명, 팬클럽명, 동아리 명	출신	&affiliation&	비식별화 필요 대상의 출신 (학교, 지역, 소속 등)
			출신	&party&	정치 조직/정 당 정보			
고유 식별 번호 (주민등록번호)	&social-security-num&	개인의 주민등록번호	번호	&social-security-num&	주민등록번호	번호	&social-security-num&	주민등록번호
전화번호	&tel-num&	휴대폰 번호, 사업장 번호 등	번호	&tel-num&	전화번호, 팩스번호	번호	&tel-num&	전화번호, 팩스번호
카드 번호	&card-num&	신용카드 번호 등	번호	&card-num&	카드 번호	번호	&card-num&	카드 번호
			번호	&bank-account&	계좌 번호	번호	&bank-account&	계좌 번호
기타 번호	&num&	비밀번호 등 기타 비식별화 대상 번호	번호	&num&	기타 번호 (사업자 등록 번호, 비밀번호 등)	번호	&num&	기타 번호 (사업자 번호 등)
온라인 계정(아이디)	&account&	특정 사이트 의 온라인 계정	온라인 계정	&online-account&	아이디, 이메일 주소	온라인계정	&account&	온라인 계정, 이메일
주소		동 이하의 상세 주소	주소/장소	&address&	주소	주소/장소	&address&	주소
			상호명/회사 명	&company&	상호명/회사 명	상호명/회사 명	&company&	상호명/회사 명
			주소/장소	&location&	장소명			
			상표명/제품 명	&brand&	상표명/제품 명	상표명/제품 명	&brand&	상표명/제품 명
			작품명	&art&	작품명 (소설, 영화, 드라마, 만화 등)	작품명	&art&	작품명
기타 비식별화 필요 항목	&address&	위 항목 외 비식별화 대 상	기타	&other&	위 항목에 포함되지 않 은 비윤리적 표 현, 욕설 등	기타	&other&	비윤리적 표 현, 욕설 등

□ 이름

지침	• 이름의 비식별화 처리는 저명성을 가지지 않는 일반인을 대상으로 한다. 일반인의 실명인 경우에만 비식별화 처리를 진행한다. • 모든 정치인은 비식별화 처리한다. • 정치인을 제외하고 연예인 등 일반적으로 공인(公人)으로 분류되는 저명한 인물이나 역사적 인물, 가상 인물은 비식별화 대상에서 제외한다. • 이칭, 별칭 등은 비식별화하지 않는다. 유명인이라면 별명만으로도 해당 인물을 특정할 수 있지만, 일반인은 별명만으로 제3자가 해당 인물을 특정할 수 없기 때문이다. • 비식별화 처리가 필요한 이름은 긍정/부정 맥락에 관계없이 모두 비식별화한다. • 나이, 생년 등 기타 개인 정보는 비식별화 하지 않는다. • 해당 이름이 가명임이 분명한 경우는 비식별화 처리하지 않는다. • 비식별화 표지 : &name&
----	--

지금은 그의 아드님인 □□□ ³⁾ (58) 대표가 후계자로 가업을 이어 받아 영업을 이어가고 있는 장인정신의 가게입니다.	⇨	지금은 그의 아드님인 &name&(58) 대표가 후계자로 가업을 이어 받아 영업을 이어가고 있는 장인정신의 가게입니다.
□□이가 적은 소원은 “□□□ 사랑해~ “라며 자기에 충만한 소원메모를 적었다. 나는 로또 당첨을 소원으로 적어 걸었는데, 포커싱이 □□이 메모에만 집중되어 버렸다.	⇨	&name&이가 적은 소원은 “&name& 사랑해~ “라며 자기에 충만한 소원메모를 적었다. 나는 로또 당첨을 소원으로 적어 걸었는데, 포커싱이 &name&이 메모에만 집중되어 버렸다.
비식별화 사례 (이름)		

□ 출신

지침	• 출신 학교나 지역, 소속 등 개인 정보와 관련 사항은 노출되지 않도록 비식별화한다. (모든 인물의 모든 출신, 소속을 비식별화 처리하는 것이 아니라, 비식별화가 필요한 대상을 특정할 수 있는 출신 정보만 비식별화 처리한다. 해당 정보는 한 텍스트 내에서 모두 비식별화한다.) • 비식별화 표지 : &affiliation&
----	--

제가 □□□□로 학생들에게 국어수업을 합니다. 어려운 중세국어 문법을 가르치고, 개념적인 것을 설명할 때는 아이들 눈높이에 맞춰	⇨	제가 &affiliation&로 학생들에게 국어수업을 합니다. 어려운 중세국어 문법을 가르치고, 개념적인 것을 설명할 때는 아이들 눈높이에 맞춰
제가 □□□ 태생이라 감자요리에 진심인데 이곳은 직접 감자를 갈아 정성스럽게 만든 웅심이와 감자전이 진짜 제 입맛에 딱!	⇨	제가 &affiliation& 태생이라 감자요리에 진심인데 이곳은 직접 감자를 갈아 정성스럽게 만든 웅심이와 감자전이 진짜 제 입맛에 딱!
비식별화 사례 (출신)		

3) 지침에 개인 정보나 비속어를 기입할 수 없어 비식별화 하였다.

□ 번호

지침	- 개인 혹은 단체를 식별할 수 있는 정보로 활용될 가능성이 있는 번호들은 비식별화 처리한다. - 비식별화 표지			
	&social-security-num&	주민등록번호	&tel-num&	전화번호
	&bank-account&	계좌 번호	&num&	기타 번호 (사업자 번호 등)
	&card-num&	카드 번호		

기초노령연금지급신청서 성명 □□ 주민등록번호 □□□□-□□□□ □□ □□□ □□□ □□-□ 근로소득 무 사업소득 무	⇒	기초노령연금지급신청서 성명 &name& 주민등록번호 &social-security-num& &address& 근로소득 무 사업소득 무
참가 방법: OT 참가비 입금 후 비밀 댓글로 아래 정보를 작성해 주세요. [성함 · 나이 · 성별 · 전화번호 · 블로그 주소 · 목표] 계좌번호: □□□□□□□□□□ (카카오뱅크 □□□) OT 참가비: 3만 원 (장소 대여 및 다과 제공)	⇒	참가 방법: OT 참가비 입금 후 비밀 댓글로 아래 정보를 작성해 주세요. [성함 · 나이 · 성별 · 전화번호 · 블로그 주소 · 목표] 계좌번호: &bank-account& (카카오뱅크 &name&) OT 참가비: 3만 원 (장소 대여 및 다과 제공)
비식별화 사례 (주민등록번호, 전화번호, 계좌 번호)		

□ 온라인 계정

지침	- 개인이나 특정 단체 등을 식별할 수 있는 온라인 계정, 이메일은 비식별화 처리한다. - 비식별화 표지 : &account&
----	---

인스타 □□□□□□□□ 매일 11:00-22:00 라스트오더 21:00 □□-□□-□□□ 국내최초!!! 편백점과 샤브샤브를 동시에!!	⇒	인스타 &account& 매일 11:00-22:00 라스트오더 21:00 &tel-num& 국내최초!!! 편백점과 샤브샤브를 동시에!!
○ 제출된 작품은 반환하지 않으며 홍보물, 책자발간 등에 이용될 수 있습니다. 11. 문 의 처 : 농촌진흥청 국립원예특작과학원 기획조정과 담당자 □□□ □□-□□-□□□ □□□□□	⇒	○ 제출된 작품은 반환하지 않으며 홍보물, 책자발간 등에 이용될 수 있습니다. 11. 문 의 처 : 농촌진흥청 국립원예특작과학원 기획조정과 담당자 &name& &tel-num&, &account&
비식별화 사례 (온라인 계정)		

□ 주소, 상호명/회사명/상표명/제품명/작품명

지침	• 주소의 경우, 동 이하의 상세주소는 모두 비식별화 한다. • 상호명/회사명/상표명/제품명/작품명은 모두 비식별화 한다. • 기업의 기술/서비스, 행사/이벤트는 ‘상표/제품명’ 으로 분류한다. 회사명과 상표명/제품명 등이 동일한 경우, 해당 단어가 쓰인 맥락에 따라 분류한다. • 공연/전시는 ‘작품명’ 으로 분류한다. • 제품명 등에서 일반명사는 비식별화 하지 않는다. • 비식별화 표지			
	&address&	주소	&brand&	상표명/제품명
	&company&	상호명/회사명	&art&	작품명

그러다 보니 저가 항공 위주로 탑승하게 되었고, □□□□□□ 항공이 선택 아닌 필수였습니다. 요즘은 패키지 여행도 많이 하지만 좋아하는 편이 아니라 이번에는 다낭 자유 여행 4박 5일 일정으로 준비하고 계획을 세웠습니다.	⇨	그러다 보니 저가 항공 위주로 탑승하게 되었고, &company& 항공이 선택 아닌 필수였습니다. 요즘은 패키지 여행도 많이 하지만 좋아하는 편이 아니라 이번에는 다낭 자유 여행 4박 5일 일정으로 준비하고 계획을 세웠습니다.
그냥 제품 설명 해주는 게 더 편한데요. 그냥 좀 보다 나왔습니다. 작년보다 더 초라한 □□ □□ □□□이네요.	⇨	그냥 제품 설명 해주는 게 더 편한데요. 그냥 좀 보다 나왔습니다. 작년보다 더 초라한 &art&이네요.
비식별화 사례 (주소, 회사명, 작품명)		

□ 비윤리적표현

지침	• 공공 말뭉치에 적합하지 않은 비속어와 혐오 표현들은 비식별화 처리한다. • 비식별화 표지 : &other&	
글고 누가 목욕탕가서 남의 □□□□ 보냐고... 솔직히 관계할때도 별 신경 안쓰였음 그냥 사람들 생김새가 다 다른것처럼 □□□□ 모양도 다 다른거라고 여겼음	⇨	글고 누가 목욕탕가서 남의 &other& 보냐고... 솔직히 관계할때도 별 신경 안쓰였음 그냥 사람들 생김새가 다 다른것처럼 &other& 모양도 다 다른거라고 여겼음
와중에 근처 혈벗은 □□들 이 인파속에서 전담피길래 훈장님 병의해서 뭐라캬 — 개념없나 진짜 아니 흡연구역 따로있는데 거 가서 피지	⇨	와중에 근처 혈벗은 &other&들 이 인파속에서 전담피길래 훈장님 병의해서 뭐라캬 — 개념없나 진짜 아니 흡연구역 따로있는데 거 가서 피지
비식별화 사례 (비윤리적표현)		

<부록 6>

온라인 매체 게시 자료 문장 분할 지침

온라인 매체 게시 자료 문장 분할 지침

온라인 게시글 중 본문 내용이 긴 경우에 문장 단위로 분할한다. 온라인 게시글은 작성자에 의해 문장 중간에 개행을 넣은 경우가 많다는 특징이 있다. 이러한 개행을 기본 단위로 하여 문장을 분할하되 아래의 원칙으로 진행한다.

- ① 문장의 분할은 개행을 기본 단위로 하여 진행한다.
- ② ‘문장’은 주어와 서술어의 문장 성분을 갖추고 있어야 하고, 서술어에 종결어미가 실현되어 있어야 한다.
- ③ 하나의 독립된 생각이나 정보가 완결되는 지점을 문장의 경계로 본다.
- ④ 문장의 기본 단위는 종지 부호 . ? !를 기준으로 한다.
- ⑤ 문장은 문법 구조를 중심으로 해서 분할한다.
- ⑥ 종지 부호가 사용되었어도 단순 수식 형태는 하나의 문장으로 인정하지 않는다.

■ 우선 ♥초저가! 홈플러스딜 상품♥

❖ ‘♥초저가!’는 주어와 서술어가 갖춰진 완벽한 문장이 아니라, 뒤에 나올 상품명을 강조하기 위한 감탄사형 수식어로 보고 분할하지 않는다.

■ 내 손 안의 편리함! B.R.월렛~

❖ !는 문장의 경계로 인정하지 않는다.

■ 맥세이프 케이스 많이들 쓰길래,, 편해? 물어보면 10명이면 다다다 동일한 답을 하더라구요!!

❖ ‘편해?’를 문장의 경계로 인정하지 않는다.

■ 이런맛은 알려야돼. 하며 (뽀뽀ㄱ) 회사에도 싸갔는데 인기만점! 검증완료!!□ ♀□ ♀□ ♀ 먹어보신분들도 어떻게 만드냐고 많이 물어봐서 릴스 냉큼 만들어왔쥬! ♪(ˆ)>”

☞ 이런맛은 알려야돼. 하며 (뽀뽀ㄱ) 회사에도 싸갔는데 인기만점!

검증완료!!👩♀👩♀👩♀

먹어보신분들도 어떻게 만드냐고 많이 물어봐서 릴스 냉큼 만들어왔쥬! ♪(ˆ)>”

- ⑦ 문장 종지 부호 뒤에 오는 이모지는 독립적인 문장으로 인정하지 않는다.

■ 근데 손님님이 그러는데 튀르키예에서도 이거 집에서 잘안만드는데요. 힘들어서.ㅋㅋㅋ 다들사먹는다고ㅋㅋㅋ 여러분도 사드세요□□

☞ 근데 손님님이 그러는데 튀르키예에서도 이거 집에서 잘안만드는데요.

힘들어서.ㅋㅋㅋ

다들사먹는다고ㅋㅋㅋ

여러분도 사드세요☺□

- ⑧ 'ㅋㅋㅋㅋ'를 문장의 끝을 맺는 종결어미나 문장 부호처럼 취급한다. 마찬가지로 이모지, 이모티콘은 앞 문장에 붙여 분할한다.

■ 편안하고 즐겁게 시청합니다 ㅋㅋㅋㅋ

■ "어? 내 얘기! ㅋㅋㅋㅋㅋㅋ" 공감 버튼 백개를 누르고 싶었어요□

☞ "어? 내 얘기! ㅋㅋㅋㅋㅋㅋ"

공감 버튼 백개를 누르고 싶었어요☹

■ 실패할 걱정 없어요!! ㅋㅋ

■ “캐주얼 러블리룩” 으로 보여요 ㅋㅋ□□

⑨ 인스타그램의 특성상 텍스트 내에 !를 습관적으로 많이 사용한다. 아래와 같은 ‘지금!’은 감탄사적 성격이 지니고 있어 하나의 독립된 문장으로 볼 수 있으나 문장 전체를 수식하거나 주의를 환기하는 독립어 역할로 본다.

■ 바삭바삭! 고소하고 맛있는 간식을 찾는다면, 파머스칩□

■ 지금! 정관장 네이버 공식스토어에서

■ 동아제약 디판버그젤! 벌레물림, 가려움, 땀띠에

⑩ 단순 나열은 문장으로 인정하지 않는다.

■ ✓임대분쟁 2. 층간소음 3. 고장시설 4. 난방 5. 곰팡이, 결로 6. 누수,침수 7. 수압,배수 8. 냄새 9. 벌레 10. 채광, 통풍 11. 전기, 누전 12. 기타소음 13. 인테리어 14. 사생활침해 15. 주차 16. 쓰레기배출 17. 보안 18. 혐오시설 19. 상대평가 20. 추천여부 등

⑪ 아래와 같이 종결어미 뒤에 쉼표를 사용했지만 .의 오류로 보고 문장으로 분할할 수 있다.

■ &online-account& 네에서 ‘&brand&’ 선물 받았어요♡,, &brand& 입고 다리 각선미도 뽐내 보았어요 ˆ ˆ . 피부에 닿는 모든 부분이 부드러워 기분이 날아갈 것 같나봐요~➡.< 마지막 영상에 &art& 노래도 들려주네요♥

☞ &online-account& 네에서 ‘&brand&’ 선물 받았어요♡,,

&brand& 입고 다리 각선미도 뽐내 보았어요 ˆ ˆ .

피부에 닿는 모든 부분이 부드러워 기분이 날아갈 것 같나봐요~➡.<

마지막 영상에 &art& 노래도 들려주네요♥

<기획·연구>

국립국어원 언어정보과장 이현주

국립국어원 학예연구관 박미영

국립국어원 연구원 장연지

<사업 참여자>

사업 책임자 서경찬(㈜윌즈정보개발 책임)

사업 참여자 남가운(㈜윌즈정보개발 연구원)

안소연(㈜윌즈정보개발 연구원)

윤종성(㈜윌즈정보개발 팀장)

윤종웅(㈜윌즈정보개발 소장)

이동민(㈜윌즈정보개발 연구원)

이승철(㈜윌즈정보개발 수석)

임순영(㈜윌즈정보개발 연구원)

지의선(㈜윌즈정보개발 선임)

최원수(㈜윌즈정보개발 선임)

발행인: 국립국어원장

발행처: 국립국어원

서울시 강서구 금남화로 154

전화 02-2669-9775, 전송 02-2669-9757

인쇄일: 2025년 11월 7일

발행일: 2025년 11월 7일

인 쇄: 다큐팩토리

※ 이 보고서는 국립국어원의 용역비로 수행한 ‘2025년도 신문 기사 원문 및 온라인 매체 게시 자료 구축’ 사업의 결과물을 발간한 것입니다.