국립국어원 2021-01-11

발 간 등 록 번 호
11-1371028-000861-01

2021년 신문 기사 원문 자료 수집 및 정제

사업 책임자

윤 종 웅

국립국어원

제 출 문

국립국어원장 귀하

국립국어원과 체결한 연구용역 계약에 따라 '2021년 신문 기사 원문 자료 수집 및 정제' 에 관한 연구 보고서를 작성하여 제출합니다.

■ 사업기간: 2021년 4월 7일 ~ 2021년 10월 7일

2021년 10월 7일

사업 책임자: 윤 종 웅((주)윤즈정보개발)

사업 수행 기관 ㈜윤즈정보개발
사업 책임자 윤 종 웅
사업 참여자 강성준, 김보희,
김하은, 남가윤,
박지영, 서경찬,
안소연, 윤여민,
이승철, 이재용,
지의선, 최원수

사업 수행자 ㈜윤즈정보개발

사업 책임자	윤종웅((주)윤즈정보개발 소장)
사업 참여자	강성준((주)윤즈정보개발 연구원)
	김보희((주)윤즈정보개발 연구원)
	김하은((주)윤즈정보개발 연구원)
	남가윤((주)윤즈정보개발 연구원)
	박지영((주)윤즈정보개발 연구원)
	서경찬((주)윤즈정보개발 책임연구원)
	안소연((주)윤즈정보개발 연구원)
	윤여민((주)윤즈정보개발 연구원)
	이승철((주)윤즈정보개발 수석연구원)
	이재용((주)윤즈정보개발 연구원)
	지의선((주)윤즈정보개발 연구원)
	최원수((주)윤즈정보개발 연구원)

〈국문 초록〉

2021년 신문 기사 원문 자료 수집 및 정제

본 사업은 올해 3년 차로, 다양한 분야의 신문 기사 원문을 수집하여 저작권을 해결하고 해당 데이터를 4차 산업혁명 대비 인공지능 기술 개발 및 학계 연구에 활용하기 위한 말뭉치를 구축하는 사업이다.

인공지능 학습에는 데이터가 중요하지만 개인이나 기업, 학계에서는 학습에 필요한 대량의 말뭉치를 확보하는 데 어려움이 있다. 국립국어원에서는 이러한 문제를 해결하기 위해서 신문 기사 말뭉치를 구축하여 누구나 자유롭게 활용할 수 있도록 제공한다.

국립국어원에서는 현재의 실제 언어 사용을 반영하는 신문 기사 말뭉치를 구축하기 위해 2020년 신문 기사 원문을 수집하고, 구축한 말뭉치의 산업계 및 학계 기술 개발, 연구 활용 이용권을 확보한다.

본 사업의 수행 범위는 신문 기사 원문 자료 수집(월별 1,000만 어절 이상), 저작권리자와의 이용 허락 계약을 통한 저작권 해결, 중복 기사 제거 및 정제, 신문 기사 원시 말뭉치 구축, 기사별 메타 정보 작성 및 목록 작성으로 구분되어 있다.

매체 선정은 한국언론진흥재단과 조선일보 두 기관과 협의하였으며 총 35개 매체를 기관과 협의하여 선택하였다. 두 기관과 계약서 및 부속합의서를 작성하고, 해당 내용을 공증하여 저작권 해결을 진행하였다.

원시 말뭉치 구축의 경우 말뭉치의 활용성을 높이기 위해 기존 방식의 신문 기사 말뭉치와 인공지능 학습 등을 위하여 단락을 세분한 문장 말뭉치, 문장에서 맞춤법을 수정한 문장 교정 말뭉치 등 3벌의 말뭉치를 구축하기로 제안하였으며 총 3종류의 말뭉치를 구축하였다.

현재까지 공개된 신문 기사 말뭉치는 기사의 단락을 최소 단위로 하고 있다. 그러나 대부분의 인공지능 학습은 문장을 기본 단위로 하고 있다. 특히 형태소 분석과 기계 번역은 대부분 문장을 기본 단위로 하고 있다. 따라서 단락을 문장으로 세분하여 문장 말뭉치를 구축한다. 또한, 인공지능으로 학습할 때 본문 중의 오자와 띄어쓰기 오류는 학습에 지장을 준다. 이런 점을 고려하여 인공지능 학습에 영향을 미치는 심각한 오류를 교정한 문장 교정 말뭉치를 추가로 구축하였다.

35개 매체로부터 3,013,829건의 기사와 561,739,209개의 어절 데이터를 확보하였다.

각 매체별로 특성을 파악하여 누락된 데이터를 점검하였고, 특정 매체에서는 한국언론진흥재단의 원시 데이터에서 글자 누락이 있는 것을 발견하였다.

데이터 1차 정제 공정에서 중복기사, 유사기사 등을 제거하였고, 특정 어절 수 초과이

거나 미만인 데이터, 오류가 많은 데이터 등을 확인하여 제거하였다.

데이터 2차 정제 공정은 작업자가 직접 기사들을 읽어가면서 문장으로 볼 수 없는 기사 내용과 본문에 불필요한 내용을 삭제하고, 저작권에 문제가 되는 기사, 사용할 수 없는 기사를 직접 삭제하면서 작업을 진행하였다. 그 후 기사에 사용되는 큰따옴표, 작은따옴표 인용부호의 통일 작업과 한중일 호환용 한자 영역(F900-FAFF) 한자를 통일하는 작업을 진행하였다.

문장 말뭉치 구축 공정은 신문 기사 말뭉치 데이터에서 단락으로 되어 있는 정보를 종결부호를 기준으로 하여 문장으로 나누었다. 피인용문 내 종결부호는 나누지 않았으며 평균 1개의 단락이 약 1.7개의 문장으로 분할되었다. 또한, 통일되지 않는 가운뎃점, 공백 문자 등 문장 부호를 통일하였다.

문장 교정 말뭉치는 문장 말뭉치 데이터에서 명백한 띄어쓰기 오류와 맞춤법 등을 정제한 데이터이다. 전체 데이터에서 많이 등장하는 오류를 중점적으로 수정하였으며, 문장 말뭉치와 함께 이용하게 된다면 인공지능 학습에 유용한 데이터가 될 것으로 보인다.

최종적으로 선정된 기사는 730,017건이고 203,585,743개의 어절이다.

구축된 데이터는 국립국어원의 9가지 분야의 주제로 분류하였으며, 어절 수, 날짜, 기사 제목 등 국립국어원이 요구하는 메타데이터를 작성하였다.

최종 말뭉치 파일은 제이슨(JSON) 파일 형태로 납품하였다.

저작권이 해결된 대규모의 말뭉치를 누구나 이용하여 인공지능 기술 개발 및 학계 연구 등 다양한 분야에서 활용할 수 있을 것으로 기대된다.

주요어: 신문 말뭉치, 인공지능, 신문 기사, 학습용 데이터, 현대 한국어

차 례

표 차례

그림 차례

제 1 장

서 론

제 1장

1. 사업목적

올해로 3년 차인 신문 기사 원문 자료 수집 및 정제 사업은 2019년에 시작된 10년 치 신문 말뭉치 기사 수집 및 정제 사업의 연장선에 있다. 2020년 발행된 신문의 1년 치 기사 저작권을 확보하여 월별로 약 1,000만 어절씩 총 1억 어절 이상의 데이터를 수집하여 말뭉치로 구축하는 사업이다.

인공지능 학습에는 데이터의 양이 중요하다. 그러나 개인이나 기업, 학계에서는 대량의 유의미한 말뭉치를 확보하는 데 어려움이 많다. 때문에 국립국어원에서는 이러한 문제를 해결하고자 신문 기사 말뭉치를 구축, 누구나 활용할 수 있도록 제공하고 있다.

신문 기사 원문을 수집하여 실제 언어 사용을 반영한 신문 기사 말뭉치를 구축하는 본 사업을 통해 생성된 결과물은 4차 산업혁명 대비 인공지능 기술 개발 및 학계 연구 등 여러 분야에서 활용될 수 있을 것으로 기대된다.

2. 사업 수행 범위

본 사업의 범위는 네 부분으로 나눌 수 있다. 첫 번째, **신문 기사의 원문 자료 수집**이다. 기사의 원문 자료 수집은 2020년에 작성된 기사로 월별 1,000만 어절 이상의 수집을 목표로 한다. 매체는 25개 이상 선정해야 하며, 이 중 인터넷 기반 매체는 전체 매체 수의 10% 이내로 해야 한다.

두 번째는 **해당 매체 기사의 저작권을 확보**해야 한다. 선정된 매체 기사의 저작권을 확보하여 사업 수행 결과물을 누구나 자유롭게 이용할 수 있어야 한다.

세 번째는 **기사 데이터의 정제화**로 기사의 불필요한 요소(이미지, 도표, 문장으로 볼 수 없는 정보 등)를 정제하는 것이다. 데이터 정제를 통해 인공지능 학습 및 학계에서 활용할 수 있는 데이터를 생성해야 한다. 본 수행사는 기존 사업과 동일한 형태의 데이터를 생산하고, 해당 데이터를 효율적으로 활용할 수 있는 문장 말뭉치와 문장 교정 말뭉치를 함께 생산하였다.

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

또한 인공지능으로 학습할 때 본문 중의 오자와 명백한 띄어쓰기 오류는 학습에 지장

을 준다. 이런 점을 고려하여 인공지능 학습에 영향을 미치는 심각한 오류를 교정한 문장 교정 말뭉치를 추가로 제안하였다.

이로써 현재 한국어 사용자의 일반적인 사용 양상이 반영된 1억 어절의 신문 기사 원시 말뭉치, 같은 데이터를 정제한 문장 말뭉치, 문장 교정 말뭉치 등 3종류의 말뭉치를 구축하였다. 이러한 3종류의 말뭉치는 인공지능 학습을 통해서 문장을 자동으로 분할하거나 맞춤법 오류를 교정하고 기계 번역에 사용할 병렬말뭉치 구축에 활용하는 등 기존 말뭉치에 비해 훨씬 다양하게 활용할 수 있다.

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

가. 신문 기사 원문 자료 수집(2020년 작성 기사, 1억 어절 이상)

- 신문 기사 말뭉치 구축에 필요한 신문 기사 원문자료를 수집.
- 대상은 2020년 기사로 월별 1,000만 어절 이상.
- 전국 종합지는 3개 이상의 매체를 포함하고, 인터넷 기반 매체는 수집하는 전체 매체 수의 10% 이내로 한정(매체 25개 이상: 기술협상서).
- 현재 한국어 사용자의 일반적인 사용 양상이 반영된 신문 기사 원시 말뭉치는 매체별, 월별, 기사 주제별로 균형성을 갖춰 1억 어절 이상 구축.
- 파일명과 표지의 종류 및 부착 형식 등은 국립국어원의 지침을 따름.

나. 신문 기사 저작 권리자와의 저작권 이용 허락 계약 체결

- 국립국어원 및 사업 수행자가 수집한 기사 원문자료 전체 활용에 필요한 저작권을 확보.
- 수집한 기사 원문자료 중 국립국어원에서 말뭉치 구축 대상으로 선정하는 매체의 기사 원문에 대해서 저작권자와 저작물 이용 허락 계약을 체결.
- 계약과 관련해 법률적인 검토를 받은 후 주관 기관이 제공한 계약서 양식에 따라 국립국어원과 협의하여 계약을 체결.
- 저작권 이용 허락 대상 권리는 신문 기사 원문자료 및 신문 기사 말뭉치의 저장, 복제, 전송, 배포, 2차적 저작물 작성권을 포함.
- 이용 허락 기간은 계약일로부터 최소 2032년 12월 31일까지로 함.

다. 기사 데이터의 정제화

- 수집된 기사 중에 동일 매체 내에서 기사 내용이 동일한 중복 기사는 제거해야 함.

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

라. 추가 제안 – 문장 말뭉치와 문장 교정 말뭉치 구축

- 본 사업의 목적은 “4차 산업혁명 대비 인공지능 기술 개발 및 학계 연구 활용을 위한 대규모 신문 기사 말뭉치 구축”에 있음.
- 양질의 말뭉치를 구축하여 학계와 산업계에 데이터를 유용하게 제공하는 것이 목적.
- 코드의 통일, 명백한 오자 등을 교정한 말뭉치를 함께 구축하여 활용성을 크게 높임.
- 최종 선정된 기사를 총 3종류로 구축.
- 데이터 활용성을 높이기 위해 기존 방식과 동일한 신문 기사 말뭉치, 신문 기사 말뭉치의 단락을 문장으로 쪼갠 문장 말뭉치, 문장 말뭉치에서 인공지능 학습에 심각한 지장을 주는 오자, 맞춤법, 띄어쓰기 오류를 바로잡은 문장 교정 말뭉치 등 3종류의 말뭉치를 구축.
- 이러한 3종류의 말뭉치는 인공지능 학습을 비롯하여 문장의 자동 분할과 자동 교정 등의 분야에서 활용할 수 있음.

<그림 1> 구축 말뭉치 종류별 개념

신문 기사 말뭉치	문장 말뭉치	문장 교정 말뭉치
▪ 불필요한 요소 제거 ▪ 작은따옴표, 큰따옴표 변환 ▪ 한중일 호환용 한자 영역 (F900-FAFF)[1]	▪ 신문 기사 말뭉치의 원 단락을 문장으로 분할 ▪ 종결부호로 문장을 분할 ▪ 피인용문 내 문장은 분할하지 않음 ▪ 코드 통일	▪ 문장 말뭉치에서 띄어쓰기, 오자, 맞춤법 등이 교정된 데이터

<표 1> 각 말뭉치 종류별 개념 설명

마. 메타 데이터 작성

❍ 국어원이 지정하는 9가지 분류 체계로 신문 기사 주제 재분류

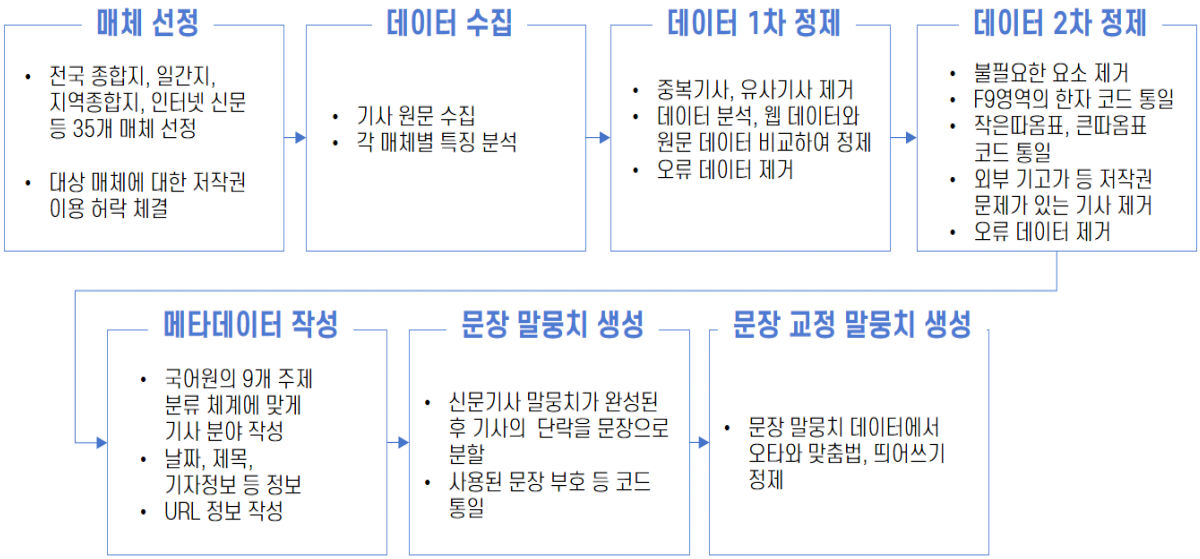
❍ 신문사명, 기사 작성일, 주제 분류, 기사 제목, 어절 수 등 국립국어원이 지정하는 항목과 형식으로 기사별 메타 정보 입력 및 수집 기사 목록 작성

1) 최초 제안은 문장 말뭉치에서 코드 통일 공정에 들어가는 내용이었으나, 국립국어원의 요청으로 신문 기사 말뭉치에 해당 공정을 도입하기로 함.

3. 사업 수행 절차

공정은 크게 7단계로 구분된다. 먼저 한국언론진흥재단, 조선일보와 계약을 체결하여 기사를 확보하고 저작권을 해결하였다. 원시 데이터를 분석하고 가공하여 쓸 수 있는 데이터를 1차적으로 구분하여 정제하였고, 해당 데이터를 가공하여 신문 기사 말뭉치 데이터를 생성하였다. 최종 선정된 기사를 바탕으로 메타데이터를 작성하였다.

추가 제안인 문장 말뭉치와 문장 교정 말뭉치 작업을 위해 단락 단위 데이터를 문단 단위로 분할하고 문장부호를 통일하였으며, 명백한 오타와 의미 전달에 문제를 주는 띄어쓰기 부분을 정제하였다.



<그림 2> 공정별 구축 내용

4. 사업 추진 경과

본 사업의 추진 경과는 다음과 같다.

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

<표 2> 사업공정표

제 2 장

사업 수행 내용

제 2장 사업 수행 내용

1. 매체 선정

가. 매체 선정 및 계약

매체를 선정하기 전 한국언론진흥재단과 접촉하여 선정된 매체의 기사 수를 먼저 확보하고 기사 수가 적은 매체는 제외하였으며, 최소 25개 이상의 매체를 선정하여야 한다는 조건에 따라 최종적으로 35개의 매체를 선정하였다. 최소 200만 건 이상의 기사와 5억 어절 이상의 데이터를 확보하기 위해 사전에 한국언론진흥재단에 매체별로 기사 수를 문의하여 기사 수가 많은 매체 위주로 선정하였다. 전국 종합지 5개 매체를 선정하였고, 인터넷 매체는 전체 매체 수 대비 10% 이하로 선정하였다.

저작물은 2020년 1월 1일부터 2020년 12월 31일까지의 기사였으며, 계약대상자는 한국언론진흥재단과 조선일보로 저작물 계약을 통해 최대한 많은 기사 수와 어절 수를 확보하였다.

본 사업은 작년 사업과 동일하게 국립국어원과 매체 간의 2자 간 저작권 이용 허락 계약과 국립국어원, 매체, 사업 수행사 3자 간의 부속합의서 계약으로 진행하였다. 계약서와 부속합의서에 대해 공증 절차를 진행하였다.

금액과 기간을 제외하고는 큰 쟁점이나 이슈 없이 계약을 체결하였다. 이용 허락 최소 기간은 2032년 12월 31일까지로 하였고, 저작자인 언론사가 이용 허락 중지 의사를 밝히지 아니하면 이용 허락이 1년 단위로 자동 갱신되도록 하였다.

구분	매체명
▪ 경제일간	▪ e대한경제, 머니투데이, 서울경제, 아시아경제, 아주경제, 파이낸셜뉴스, 한국경제, 헤럴드경제
▪ 스포츠일간	▪ 스포츠서울
▪ 인터넷신문	▪ EBN산업뉴스, 노컷뉴스, 뉴스핌
▪ 전국종합일간	▪ 국민일보, 서울신문, 조선일보, 한겨레, 한국일보
▪ 전문일간	▪ 전자신문, 환경일보
▪ 지역종합일간	▪ 강원도민일보, 경기일보, 경남도민일보, 경북일보, 남도일보, 대구신문, 대전일보, 매일신문, 부산일보, 인천일보, 전남일보, 전북도민일보, 중도일보, 중부일보, 충청일보, 충청투데이

<표 3> 선정된 매체 구분

나. 원시 데이터 엑스엠엘(XML) 특징 분석

한국언론진흥재단에서 구매하여 제공받는 데이터는 기사 하나가 하나의 엑스엠엘(XML) 파일로 되어 있다.

```
    <Metadata>
     <Property FormalName="PublishDate" Value="20200303" />
     <Property FormalName="PublisherCode" Value="" />
     <Property FormalName="PaperEdition" Value="" />
     <Property FormalName="PageCategory" Value="" />
     <Property FormalName="PageCategoryId" Value="" />
     <Property FormalName="PrintingPage" Value="0" />
     <Property FormalName="PrintingPageNo" Value="0" />
     <Property FormalName="GenreInfo" Value="" />
     <Property FormalName="KsOrgan" Value="" />
<Property FormalName="KsCompany" Value="" />
     <Property FormalName="KsPeople" Value="" />
<Property FormalName="UciCode" Value="G703:RA101-01200101.20200303215836001:1"
/>
     <Property FormalName="ModifyInfo" Value="" />
     <Property FormalName="CharacterCounter" Value="751" />
     <Property FormalName="NewsItemOrgId" Value="2249110" />
     <Property FormalName="LinkPage"
Value="www.kyeonggi.com/news/articleView.html?idxno=2249110" />
     <Property FormalName="LinkmPage" Value="" />
     <Property FormalName="SubjectInfo" Value="의왕시" />
     <Property FormalName="SubjectInfo1" Value="" />
     <Property FormalName="SubjectInfo2" Value="" />
     <Property FormalName="SubjectInfo3" Value="" />
     <Property FormalName="SubjectInfo4" Value="" />
     <Property FormalName="MapSubjectInfo" Value="" />
     <Property FormalName="AutoSubjectInfo" Value="지역,경남|지역,대전|" />
     <Property FormalName="AutoSubjectCode"
Value="006000000,006003000|006000000,006007000|" />
     <Property FormalName="AutoKeywordInfo" Value="장학생,고등부,대학부,의왕시,의왕시
인재육성재단,학업성적,고등부 특기장학생,고등부 복지,또는,월평균" />
     <Property FormalName="Latitude" Value="" />
     <Property FormalName="Longitude" Value="" />
     <Property FormalName="PageCoordinate" Value="" />
     <Property FormalName="GroupCoordinate" Value="" />
     <Property FormalName="ScrapPage" Value=""/>
```

```
    <Property FormalName="ScrapCategoryId" Value=""/>
    <Property FormalName="ScrapCoordinate" Value="" />
    <Property FormalName="ScrapPdfFileName" Value=""/>
   </Metadata>
   <NewsComponent>
    <Role FormalName="Main" />
    <ContentItem>
     <MediaType FormalName="Text"/>
     <MimeType FormalName="text/plain" />
     <DataContent><![CDATA[의왕시 인재육성재단은 상반기 대학부·고등부 장학생
54명을 선발해 7천만 원의 장학금을 지급할 예정이라고 2일 밝혔다.

분야별로는 대학부 희망드림 장학생 12명에게 1인당 150만 원씩, 고등부 성적우수
장학생 15명에게 1인당 100만 원씩, 고등부 복지 장학생 20명에게 1인당 150만
원씩, 고등부 특기장학생 5명과 효행(선행) 장학생 2명에게 1인당 100만 원씩의
장학금을 지급할 예정이다.

대학부 모집대상은 의왕시에 2년 이상 거주하는 대학생으로 학업성적이 백분율
환산 80점 이상, 부모소득 전체 월평균 400만 원 이하이어야 한다.

고등부는 학업성적 3과목 이상이 2등급 이상인 관내 고등학교 재학생 등
성적우수자와 학업성적 3과목 이상이 4등급 이상이며 부모소득 전체 월평균
400만 원 이하에 해당하는 관내 고등학교 재학생(복지)이다. 또 특기장학생은
기능·예능·과학 등 단위 규모 대회에서 1위에 입상하거나 전국 규모 대회에서
3위 이내 입상한 관내·외 고등학생이어야 한다. 효행·선행 장학생은 각종
기관이나 단체의 표창 또는 언론에서 칭송을 받은 학생으로 학교장 또는 관할
동장의 추천을 받은 관내 고등학교 재학생이다.

신청은 오는 23일부터 27일까지 5일간이며, 분야별 제출 서류는 의왕시
인재육성재단 홈페이지(www.uwinjae.or.kr) 공고사항을 확인 후 신청서와 기타
증빙서류 등을 장학재단으로 제출하면 된다.

자세한 사항은 의왕시인재육성재단(031-345-2590)으로 문의하면 된다.
의왕=임진흥기자]]></DataContent>
</ContentItem>
```

<표 4> 원시데이터 예시

아래의 본문을 살펴보면 본문 내용(DataContent)이 두 가지 형태로 되어 있다.

<DataContent><![CDATA[신종 코로나바이러스감염증(코로나19) 사태가 장기화하면서 어려움을 겪는 농어촌교회가 많아지고 있다. 도시에선 많은 교회가 온라인 예배에 동참했지만, 농어촌교회는 대응에 애를 먹고 있다.

“요즘 맘이 참 거시기허죠. 교회 주변엔 코로나19 감염자가 없어서 예방 수칙을 지키며 예배를 드렸는데 도시 사는 자녀들이 하루에도 몇 번씩 전화를 준다네요. 교회 가지 말라고요. 지난주일 강단에선 ‘정 불안하고 자녀들 걱정시킬 거 괴로우시면 한두 번 쉬셔요’ 하고 안내했는데 가슴팍이 꽉 맥히는 거 같더라고요(한숨).”

전북 김제에서 사역하는 A목사의 목소리엔 막막함이 느껴졌다. 수십 년째 동네 주민들의 사랑방이었던 예배당은 최근 몇 주 사이 추수 끝난 논처럼 을씨년스러워졌다. 주일예배 후 음식을 먹으며 도란도란 정을 나누던 모습도 사라졌다. A목사는 “사회적 거리 두기가 확산되는 동안 농촌교회에선 ‘공동체적 거리감’이 생겼다”고 했다.

-- 중간 생략 --

최기영 기자 ky710@kmib.co.kr]]></DataContent>
</ContentItem>
</NewsComponent>
<NewsComponent>
<Role FormalName="OriginMain" />
<ContentItem>
<MediaType FormalName="Text"/>
<MimeType FormalName="text/plain" />
<DataContent><![CDATA[신종 코로나바이러스감염증(코로나19) 사태가 장기화하면서 어려움을 겪는 농어촌교회가 많아지고 있다. 도시에선 많은 교회가 온라인예배에 동참했지만, 농어촌교회는 대응에 애를 먹고 있다.

“요즘 맘이 참 거시기허죠. 교회 주변엔 코로나19 감염자가 없어서 예방 수칙을 지키며 예배를 드렸는데 도시 사는 자녀들이 하루에도 몇 번씩 전화를 준다네요. 교회 가지 말라고요. 지난주일 강단에선 ‘정 불안하고 자녀들 걱정시킬 거 괴로우시면 한두 번 쉬셔요’ 하고 안내했는데 가슴팍이 꽉 맥히는 거 같더라고요(한숨).”

-- 중간 생략--
지금은 성도들을 보듬으며 얼른 이 사태가 지나가길 기도할 뿐입니다.”

최기영 기자 ky710@kmib.co.kr

GoodNews paper ⓒ

국민일보(www.kmib.co.kr), 무단전재 및 재배포금지]]></DataContent> </ContentItem>

<표 5> 원시 데이터 본문과 메타데이터 형태

다. 데이터 특징과 오류 유형 분석

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

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

<표 6> 데이터 특징

한국언론진흥재단에 데이터 중 일부 자료에서는 소제목과 다음 단락이 붙어 버리는 문제가 있었다. 엑스엠엘(XML) 변환 과정에서 발생된 것으로 보인다. 하나의 문장이 앞의 소제목과 붙어버림으로써 불완전한 문장이 된다. 이 부분은 작업할 때 해당 기사를 정독하지 않으면 놓치게 된다. 서울신문 외 몇 개의 매체에서 이러한 문제가 있었다.

또한, 캡션 정보가 본문과 구분되지 않는 경우가 존재한다. 캡션 정보가 마치 본문인 것처럼 등장하는데, 캡션 정보는 불필요한 요소로 삭제 대상이다. 환경일보, 인천일보 외 다수의 매체에서 발견되었다. 오류의 유형은 다음과 같다.

<그림 3> 매체별 데이터 특징

위의 오류 유형은 웹 페이지의 데이터를 확인하면서 소제목 뒤에 엔터를 삽입하는 방식과 캡션 정보를 찾아 태그로 감싸는 방식으로 데이터를 정제하였다.

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

경기일보 실제 웹 화면

용인 뮤지엄그라운드, 작가의 개성과 익숙한 캐릭터들 결합한 그래피티展 <My Space> 오는 12일까지 연다

권재민 기자 ohtaku@kyeonggi.com | 입력 2019.12.31 오후 2:37 | 댓글 0

용인 뮤지엄그라운드는 그래피티를 주 매체로 한 전시 <My Space>를 오는 12일까지 연다.

그래피티의 사전적 정의는 길거리 여기저기 벽면에 낙서처럼 그리거나 페인트를 분무기로 내뿜어서 그리는 그림으로 현대 미술에서는 작가적 개성을 드러내는 매체이자 형식과 내용에 얽매이지 않는 매체로 인식되고 있다.

이번 전시는 지난 1990년대 후반부터 2000년대 초중반 사이 그래피티 작업을 시작한 아티스트 알타임 죠, 제바, 세미, 켄지 차이의 작품을 선보인다. 이들은 작가의 독자적 해석과 표현방식에 아무런 제한도 두지 않았다.

그 예로 알타임 죠의 작업에는 익숙한 애니메이션, 게임, 영화의 주인공들이다. <스누피>의 '찰리 브라운', <드래곤볼>의 '손오공' 등이 알타임 죠 특유의 그림체와 그만의 공간 속에 펼쳐져 등장한다. 과거와 현재가 조화롭게 뒤섞여 있어 관람객에겐 어린 시절을 떠오르게 함과 동시에 지금을 공유하는 매개체로 자리한다. 이어 제바는 일상을 둘러싼 세계인 감각적 공간의 한계를 벗어나려는 작가의 발상을 선보인다. 추상과 반추상의 독창적 이미지는 경험해 볼 수 없는 상상의 공간에서 자유롭게 변화하고 비상한다.

전반적으로 4명의 작가는 미술관이라는 공간의 안과 밖의 경계를 허물었다. 이들은 전시 공간을 벽면 전체를 채우는 뮤럴, 다양한 오브제, 캐릭터와 레터스타일과 같은 다양한 형태로 확장했다.

경기TV 더보기

경기일보 보도, 그 후

#2021년 8월 #45호선 용인 ...
용인시, 마평교차로 도로에 '화물차 전용 주차장' 조성

#2021년 8월 #경기도의회 웹...
성차별 없애고 의정 공감대 높인 '명품' 대변신 기대감

#2021년 9월 #성범죄자 알림...
정부 부처 '성범죄자 신상정보 관리 체계' 개편작업 착수

#2021년 9월 #경기남부 통합...
수원군공항 이전 추진 7년 만에 돌려온 낭보

<그림 4> 경기일보 서명 안의 텍스트가 삭제된 예

한국언론진흥재단으로부터 받은 원시 데이터

<NewsComponent>
<Role FormalName="Main" />
<ContentItem>
<MediaType FormalName="Text"/>
<MimeType FormalName="text/plain" />
<DataContent><![CDATA[용인 뮤지엄그라운드는 그래피티를 주 매체로 한 전시 를 오는 12일까지 연다.
그래피티의 사전적 정의는 길거리 여기저기 벽면에 낙서처럼 그리거나 페인트를 분무기로 내뿜어서 그리는 그림으로 현대 미술에서는 작가적 개성을 드러내는 매체이자 형식과 내용에 얽매이지 않는 매체로 인식되고 있다.
이번 전시는 지난 1990년대 후반부터 2000년대 초중반 사이 그래피티 작업을 시작한 아티스트 알타임 죠, 제바, 세미, 켄지 차이의 작품을 선보인다. 이들은 작가의 독자적 해석과 표현방식에 아무런 제한도 두지 않았다.
그 예로 알타임 죠의 작업에는 익숙한 애니메이션, 게임, 영화의 주인공들이다. 의 '찰리 브라운', 의 '손오공' 등이 알타임 죠 특유의 그림체와 그만의 공간 속에 펼쳐져 등장한다. 과거와 현재가 조화롭게 뒤섞여 있어 관람객에겐 어린 시절을 떠오르게 함과 동시에 지금을 공유하는 매개체로 자리한다. 이어 제바는 일상을 둘러싼 세계인 감각적 공간의 한계를 벗어나려는 작가의 발상을 선보인다.

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

2. 데이터 수집

35개의 매체에서 수집된 기사 수와 어절 수는 각각 3,013,829건, 561,739,209개이고 한 기사의 평균 어절 수는 186어절로 집계되었다. 어절 수의 집계는 문장의 공백 수와 줄바꿈 수로 카운트 하였다.

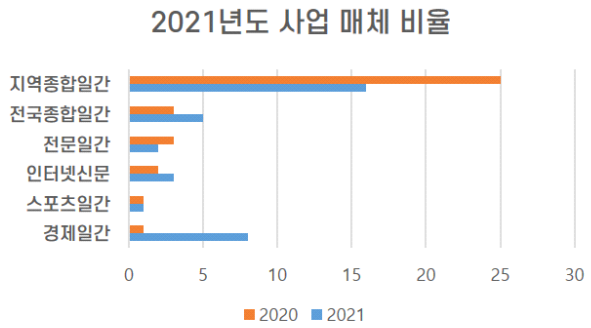
최초 목표인 2백만 기사 이상 5억 어절 이상의 조건을 충족하였다. 매체 수는 2020년도 사업과 동일하였지만, 최초 수집된 기사는 약 100만 건을 더 수집하였다. 매체별 기사 수와 어절 수에 대한 내용을 정리하면 다음과 같다.

상위 5 기사수

매체	기사 수
뉴스핌	284,167
아시아경제	215,433
한국경제	192,832
머니투데이	186,723
서울경제	171,187

하위 5 기사수

매체	기사 수
e대한경제	15,478
경남도민일보	29,492
대구신문	29,716
전남일보	30,379
강원도민일보	32,429



<그림 5> 매체별 비율, 상위, 하위 기사 수

매체명	기사 수	어절 수	매체명	기사 수	어절 수
한국경제	192,832	39,554,000	부산일보	85,321	15,652,601
서울경제	171,187	32,508,732	충청투데이	54,810	8,176,411
아시아경제	215,433	37,333,297	대전일보	48,955	7,892,885
파이낸셜뉴스	49,871	9,111,932	경북일보	34,184	6,386,017
아주경제	118,435	27,052,649	뉴스핌	284,167	39,218,664
머니투데이	186,723	39,098,450	중도일보	88,023	13,239,031
스포츠서울	125,433	18,848,466	헤럴드경제	142,459	30,128,952
노컷뉴스	142,690	25,376,435	중부일보	52,951	9,163,532
EBN산업뉴스	48,947	9,767,159	인천일보	48,534	8,641,210
전자신문	61,528	12,098,643	e대한경제	15,478	3,239,142
한겨레	39,935	11,856,729	전북도민일보	38,769	6,364,152
국민일보	122,074	25,924,246	전남일보	30,379	6,220,459
서울신문	120,787	25,641,537	대구신문	29,716	5,433,994
한국일보	112,864	26,369,560	경남도민일보	29,492	4,635,045
경기일보	45,824	8,083,866	남도일보	37,181	6,940,911
강원도민일보	32,429	4,010,182	환경일보	43,409	8,360,492
충청일보	60,837	8,522,189	조선일보	49,864	11,310,071
매일신문	52,308	9,577,568	**총 합**	**3,013,829**	**561,739,209**

<표 7> 최초 수집 기사 수, 어절 수

3. 데이터 1차 정제

가. 중복 기사, 유사한 데이터 제거

원시 데이터를 수령한 뒤 중복 기사와 유사 기사를 제거한다. 이후 불필요한 요소를 제거(데이터 2차 정제)하고 중복, 유사 데이터 제거 공정을 한 차례 더 진행하게 된다. 데이터 2차 정제 공정 후 기사의 불필요한 부분이 삭제되어 내용이 중복되거나 유사한 경우가 존재하기 때문이다. 작업 기준은 다음과 같다.

❍ 중복체크를 통해 모든 매체의 기사 안의 내용이 일치하는 데이터는 제외함.

❍ 같은 매체 기사 전후 14일을 비교하여 유사도가 85% 이상인 기사는 제외함.

ㅇㅇ일보 매체 기사 중 제목은 다르나 내용이 중복인 기사의 예	
제목: 공동종합법률 로펌 보담, 착한가게 가입	제목: 올해 첫 착한가게 공동종합법률 로펌 보담
공동종합법률 로펌 보담이 올해 들어 첫 번째 착한가게에 가입했다. 15일 오전 10시 공동종합법률 로펌 보담 사무실에서 보담 백홍기 대표변호사와 대전사회복지공동모금회 박용훈 사무처장이 참석한 가운데 대전사회복지공동모금회 2020년도 첫 번째 착한가게 현판 전달식이 진행됐다. 공동종합법률 로펌 보담 백홍기 대표변호사는 "누군가 도움을 주는 직업을 가진 사람으로서 주변의 어려운 이웃을 돕는 일은 우리의 사명"이라며, "도움이 필요한 어려운 이웃들에게 소중히 전달되길 바란다"고 말했다. 이 날 착한가게에 가입한 '공동종합법률 로펌 보담'의 백홍기 대표변호사는 지역의 나눔문화를 선도하는 개인 기부자인 '나눔리더'에도 가입했다. 안기호 회장은 “사회복지공동모금회가 연중 추진하고 있는 착한가게는 업종에 상관없이 모든 자영업자가 하루 1000원(매달 3만 원 이상) 정기기부를 할 경우 자격이 주어진다”며, “기부자 예우 차원으로 착한가게 현판을 전	공동종합법률 로펌 보담이 올해 들어 첫 번째 착한가게에 가입했다. 15일 오전 10시 공동종합법률 로펌 보담 사무실에서 보담 백홍기 대표변호사와 대전사회복지공동모금회 박용훈 사무처장이 참석한 가운데 대전사회복지공동모금회 2020년도 첫 번째 착한가게 현판 전달식이 진행됐다. 공동종합법률 로펌 보담 백홍기 대표변호사는 "누군가 도움을 주는 직업을 가진 사람으로서 주변의 어려운 이웃을 돕는 일은 우리의 사명"이라며, "도움이 필요한 어려운 이웃들에게 소중히 전달되길 바란다"고 말했다. 이 날 착한가게에 가입한 '공동종합법률 로펌 보담'의 백홍기 대표변호사는 지역의 나눔문화를 선도하는 개인 기부자인 '나눔리더'에도 가입했다. 안기호 회장은 “사회복지공동모금회가 연중 추진하고 있는 착한가게는 업종에 상관없이 모든 자영업자가 하루 1000원(매달 3만 원 이상) 정기기부를 할 경우 자격이 주어진다”며, “기부자 예우 차원으로 착한가게 현판을 전

달하고 있다"고 밝혔다. 한편 착한가게 문의는 대전사회복지공동모금회 모금사업팀 042-347-5176으로 하면 된다. 한성일 기자 hansung007@	달하고 있다"고 밝혔다. 한편 착한가게 문의는 대전사회복지공동모금회 모금사업팀 042-347-5176으로 하면 된다. 한성일 기자 hansung007@

ㅇㅇ일보 매체 기사 중 유사도 비교를 통해 사용하지 않는 기사의 예(유사도 85%)	
제목: 순천시, 2천만 보장 시민안전보험 가입	제목: 순천시, 전 시민 안전보험 가입…최대 2천만원 보장
순천시, 2천만 보장 시민안전보험 가입 내년 4월 9일까지 11개 항목 보장 순천시는 일상 속 뜻밖의 각종 재난과 사고로부터 피해를 당한 시민들의 생활안정을 돕기 위한 '시민안전보험'을 올해 첫 시행한다. '시민안전보험'은 관내에 주민등록을 두고 거주하는 등록외국인 포함 모든 시민이 대상이 된다. 별도의 가입 절차와 보험료 없이 순천시민이면 누구나 혜택을 받을 수 있다. 또 개개인의 다른 보험 가입여부와 상관없이 중복 보상이 가능하다. 보장기간은 2020년 4월 10일부터 2021년 4월 9일까지 1년간이다. 보장항목은 △ 일사병, 열사병 등을 포함한 자연재해사망 △ 폭발·화재·붕괴·산사태 상해사망 및 후유장애 △ 대중교통이용 상해사망 및 후유장애 △ 강도상해·사망 및 후유장애 △ 스쿨존 교통상해 부상치료비 △ 온열 질환 진단비 △ 침몰사고 사망 △ 익사, 농기계, 추락, 화재 등 상해의료비 지원 등 11개 항목이다. 보장금액은 사망 시 2천만 원, 후유장애 시 후유장애 비율(3~100%)에 따라 최대 1천만 원까지 보장되며, 특히 온열질환 진단비(10만원), 상해의료비 지원(200만원 한도)도 포함해 보장 혜 넓다. 보험료 청구는 청구 사유 발생 시 피해자 또는 법정 상속인이 청구서 등 관련 서류를 첨부해 현대해상화재보험(주)으로 청구하면 되고, 청구 소멸시효는 사고일로부터 3년이다. 순천시관계자는"시민들이 뜻밖의 재난과 사고로부터 안심하고 생활할 수 있도록	순천시, 전 시민 안전보험 가입…최대 2천만원 보장 순천시는 일상 속 각종 재난과 사고로부터 피해를 당한 시민들의 생활안정을 돕기 위한 '시민안전보험'을 올해 첫 시행한다고 27일 밝혔다 시민안전보험은 순천시에 주민등록을 두고 거주하는 등록 외국인 포함 모든 시민이 대상이다. 별도의 가입 절차와 보험료 없이 순천시민이면 누구나 혜택을 받을 수 있다. 개개인의 다른 보험 가입여부와 상관없이 중복 보상이 가능하다. 보장기간은 내년 4월 9일까지 1년이다. 보장항목은 자연재해사망 폭발·화재·붕괴·산사태 상해사망 및 후유장애 대중교통이용 상해사망 및 후유장애 강도상해사망 및 후유장애 스쿨존교통상해부상치료비 온열질환 진단비 침몰사고 사망 상해의료비 지원 등 11개 항목이다. 보장금액은 사망 시 2천만원, 후유장애 시 후유장애 비율(3~100%)에 따라 최대 1천만원까지 보장된다. 특히 온열질환 진단비(10만원), 상해의료비 지원(200만원 한도)도 포함해 보장 혜택 넓다. 보험료 청구는 청구 사유 발생 시 피해자 또는 법정상속인이 청구서 등 관련서류를 첨부해 현대해상화재보험으로 청구하면 된다. 청구 소멸시효는 사고일로부터 3년이다. 시 관계자는"뜻밖의 재난과 사고로부터 시민들이 안심하고 생활할 수 있도록 최소한의 경제적 생활안정을 지원하는 데 중점을 뒀다"며"매년 보험에 가입할 예정

최소한의 경제적 생활안정을 지원하는 데 중점을 뒀다"며"시민안전보험을 적극 운영하고 매년 보험에 가입할 예정이다"고 말했다. 동부취재본부/유홍철 기자 yhc@namdonews.com	이다"고 말했다. 동부취재본부/장봉현 기자 coolman@namdonews.com

나. 기사 선택

원시 데이터의 메타정보를 활용하여 사용할 수 없는 기사를 선별하고 구축 대상 기사에서 제외하였다. 기준은 아래와 같다.

- 기자 정보가 없는 데이터는 제외함.(기자 정보가 없는 데이터는 한국언론진흥재단 측에 문의한 결과 해당 정보를 얻을 수 없다고 답변받음.)
- 기사 길이가 1,000어절 이상이거나 180어절 이하는 제외함.
- 단순 광고, 띠별 오늘의 운세, 퀴즈 등 기사로 보기 어려운 것은 제외함.
- 승진자나 부고 명단, 스포츠 경기의 결과 수치만으로 구성된 기사는 제외함.
- 저작권 문제의 가능성이 있는 타 매체의 기사는 제외함.
- 번역된 기사는 사용하지 않음.(기관 협의)
- 뉴스 기사의 특성이 전혀 없는 시(詩)나 소설 등 문학 작품은 제외함.
- 기사의 대부분이 영어나 일어 등 다른 언어로 된 것은 제외함.
- '~했어요.', '~란다.', '~할까요?' 등 기사 전체가 구어체로 이루어진 기사는 제외함.
- 인공지능 로봇이 작성한 기사는 제외함.

4. 데이터 2차 정제[2)]

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

가. 웹 페이지 데이터 확인

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

---

2) 데이터 2차 정제 과정은 신문 기사 말뭉치 구축 단계이며, 해당 단계에서는 불필요한 요소 제거 외에 인용부호(작은따옴표, 큰따옴표) 통일, F9영역의 한자 치환 공정이 포함되어 있음. 일반 부호의 통일 과정은 문장 말뭉치 과정에서 적용됨.

소제목과 소제목 다음에 오는 단락이 붙어 버리는 문제의 경우에는 작업자가 기사를 정독하면서 해당 기사의 유알엘(URL)을 확인하지 않으면 발견하기 어려운 경우가 많다. 아래 예시와 같이 캡션 정보에 아무런 표기가 되어 있지 않다면 하나의 본문처럼 인식하기 쉽다. 이런 오류 등은 웹에서 문서를 직접 확인하고 해결해야 한다. 웹페이지 데이터에는 한국언론진흥재단으로부터 받은 데이터에 없는 정보가 표시되어 있어 이를 활용할 수 있다. 아래 내용은 중간 제목이 본문 사이에 들어간 경우이다.

웹에서 확인한 실제 기사 내용

이보다 자세한 '검찰개혁안'은 이전 대선인 2012년12월에 발표한 바 있다. 문 대통령은 당시 "이명박 정권 5년 동안 대통령 및 청와대가 검찰 수사와 인사에 관여했던 악습을 완전히 뜯어 고치겠습니다"고 공언했다. 이를 위해 대통령에게 주어졌던 검찰총장 임명권을 국민에게 돌려주겠다고 했다. 구체적으로는 총장추천위에 검찰 내부의 의견이 수렴될 수 있도록 제도적 장치를 마련하고 시민단체 등 외부인사가 과반수 이상 참여하게 한다고 했다. 검찰인사위는 외부 인사가 과반수 이상 참여하는 형태로 확대 개편하고 검사장급 인사에 대해 인사청문회를 시행하겠다고 했다.

공약한 검찰인사 제도 개선, 감감무소식

또 문 대통령은 취임 직후 취임사에서 "대통령의 제왕적 권력을 최대한 나누겠습니다"며 "권력기관은 정치로부터 완전히 독립시키겠습니다"고 밝혀 이같은 검찰인사 개혁 조치를 해나갈 것을 시사했다. 2017년8월 내놓은 '100대 과제'에서는 공약 사항인 총장추천위와 검찰인사위의 중립성·독립성 확보를 위한 제도 정비를 당해부터 들어가겠다고 명시했다.

<그림 6> 소제목과 본문 내용 확인 실제 URL 기사

한국언론진흥재단으로부터 받은 데이터 내용

이보다 자세한 '검찰개혁안'은 이전 대선인 2012년12월에 발표한 바 있다. 문 대통령은 당시 "이명박 정권 5년 동안 대통령 및 청와대가 검찰 수사와 인사에 관여했던 악습을 완전히 뜯어 고치겠습니다"고 공언했다. 이를 위해 대통령에게 주어졌던 검찰총장 임명권을 국민에게 돌려주겠다고 했다. 구체적으로는 총장추천위에 검찰 내부의 의견이 수렴될 수 있도록 제도적 장치를 마련하고 시민단체 등 외부인사가 과반수 이상 참여하게 한다고 했다. 검찰인사위는 외부 인사가 과반수 이상 참여하는 형태로 확대 개편하고 검사장급 인사에 대해 인사청문회를 시행하겠다고 했다.

— 공약한 검찰인사 제도 개선, 감감무소식 또 문 대통령은 취임 직후 취임사에서 "대통령의 제왕적 권력을 최대한 나누겠습니다"며 "권력기관은 정치로부터 완전히 독립시키겠습니다"고 밝혀 이같은 검찰인사 개혁 조치를 해나갈 것을 시사했다. 2017년8월 내놓은 '100대 과제'에서는 공약 사항인 총장추천위와 검찰인사위의 중립성·독립성 확보를 위한 제도 정비를 당해부터 들어가겠다고 명시했다.

❍ 캡션 정보가 본문과 구분되지 않아 본문처럼 보이는 내용

웹에서 확인한 실제 기사 내용

또한 해양수산부는 이달의 축제로 '2020 양평 빙어축제'를 선정했다. 이 축제는 2020년 1월3일(금)부터 2월16일(일)까지 경기도 양평군 단월면 백동저수지 일대에서 열리며, 빙어낚시와 얼음 미끄럼틀, 아이스 범퍼카 등 다양한 겨울놀이기구를 즐길 수 있다.

1992년을 시작으로 양평의 깊은 산중에 자리 잡은 백동저수지에서 매년 빙어자원을 꾸준히 조성한 양평빙어축제2020이 2019년 12월20일(금)부터 2020년 2월16일(일)까지 59일간 개최된다. <자료제공=해양수산부>

해양수산부 관계자는 "새해 첫 이달의 수산물로 선정된 송어와 김은 겨울철에 특히 맛이 좋고 영양도 풍부하니, 많이 드시고 희망찬 새해를 든든하게 시작하시길 바란다"라고 말했다.

<그림 7> 캡션 정보와 본문이 구분되지 않는 예

한국언론진흥재단으로부터 받은 데이터 내용

또한 해양수산부는 이달의 축제로 '2020 양평 빙어축제'를 선정했다. 이 축제는 2020년 1월3일(금)부터 2월16일(일)까지 경기도 양평군 단월면 백동저수지 일대에서 열리며, 빙어낚시와 얼음 미끄럼틀, 아이스 범퍼카 등 다양한 겨울놀이기구를 즐길 수 있다.

1992년을 시작으로 양평의 깊은 산중에 자리 잡은 백동저수지에서 매년 빙어자원을 꾸준히 조성한 양평빙어축제2020이 2019년 12월20일(금)부터 2020년 2월16일(일)까지 59일간 개최된다. <자료제공=해양수산부>해양수산부 관계자는 "새해 첫 이달의 수산물로 선정된 송어와 김은 겨울철에 특히 맛이 좋고 영양도 풍부하니, 많이 드시고 희망찬 새해를 든든하게 시작하시길 바란다"라고 말했다.

나. 기사 선택 및 불필요한 요소 제거

데이터 정제 2차 작업 중에서 작업자들이 선택된 기사를 읽어가며 불필요한 요소를 삭제하고 사용하지 말아야 하는 기사를 표기하는 공정이다. 데이터 1차 정제를 통해 사용하지 않는 기사를 걸러냈지만, 이는 내용을 전부 파악하고 선별한 것이 아니기에 작업자는 내용을 읽으며 사용하지 않는 기사를 표시하게 된다. 사용하지 않는 기사 선택 기준은 다음과 같다.

1) 제외 대상 기사 표시

- 저작권 문제가 있는 기사
- 문장이 도중에 잘렸거나, 오류가 많은 기사
- 본문 기사 안에 광고라고 표기된 기사
- 외부기고가가 작성한 기사
- 기사 내부에 명확히 광고라고 표기한 기사
- 구어체로 된 기사
- 일반적인 신문 기사로 볼 수 없는 기사(승진, 부고, 스포츠 스코어, 출구조사, 여론조사 등등)

사업의 목적에 맞는 문어체 말뭉치 구축이 목표이기 때문에 해당 목표에 맞지 않는 기사는 직접 읽으며 제외하였다. 또한, 연합뉴스 등 타 매체에 저작권이 있는 기사는 본문에 매체명이 표시된 경우도 존재하여 데이터 2차 정제 작업 과정에서 직접 확인하여 제외하면서 저작권에 위배되는 기사에 유의하였다. 명예기자, 객원기자, 시민기자의 기사도 제외하였다.

2) 불필요한 요소 제거

작업자가 직접 기사를 읽어가며 불필요한 요소를 삭제하는 공정으로 기사 내에는 표, 그림, 그래프와 기사와 무관한 정보들이 그대로 남아 있다. 이 정보들은 전체 맥락을 해치므로 제거해 주어야 한다. 또한 연설문, 입장발표문, 사과문 등의 외부 전문이 실린 경우, 기자가 작성한 기사와는 다른 문체로 쓰이고 있으므로 신문기사 말뭉치를 구축하는 본 사업의 목표에 맞지 않는다. 따라서 제거해 주어야 하며, 이때 다음에 전문이 존재함을 알리는 기사 문장 또한 기사의 완결성에 유의하며 제거한다.

아래 예시에서 굵은 붉은색 글꼴은 삭제되어야 할 대상이다.

삭제 정보	예시	
표, 그림, 그래프 등의 캡션 정보	(사진제공=건국대학교) (사진제공=SBA) 사진제공=tvN 사진=CJ엔터테인먼트 제공 [그래픽] <그래픽> 일러스트| 화면 캡처 〈	표〉 공정위 망 이용대가 불공정 조사 쟁점 〈표〉 한상혁 후보자 주요 ICT 정책 현안 입장 ▲영상제공= 사진=빅펀치이엔티 제공 사진=FNC엔터테인먼트 제공 출처: 라디오타임스 / 굿모닝브리튼,사진=스타쉽 제공
기자의 이름, ID 등의 정보	[서울경제TV=배ㅇㅇ기자] 동양네트웍스(030790)가 강세다. 글·사진=양ㅇㅇ 기자 ----@kmib.co.kr 김ㅇㅇ ------@kmib.co.kr.사진=인터파크 제공	
'Copyright©' 등 저작권 관련 내용	<저작권자(c) 연합뉴스, 무단 전재-재배포 금지> ondol@yna.co.kr/2019-08-29 10:14:05/<저작권자 ⓒ 1980-2019 ㈜연합뉴스. 무단 전재 재배포 금지.>	
전문	대검찰청 정책관 등 중간간부들이 26일 윤석열 검찰총장 직무배제가 부당하다며 추미애 법무부 장관에게 재고를 요청하는 성명을 냈다. 손준성 수사정보정책관, 이창수 대검 대변인 등 대검 중간간부 27명은 이날 검찰게시판에 '대검찰청 중간 간부들의 입장'이라는 제목의 글을 올렸다. 이들은 "검찰총장에 대한 직무집행정지는 적법절차를 따르지 않고, 충분한 직상확인 과정도 없이 이뤄진 것으로 위법부당하다"며 "이는 검찰의 중립성은 물론이고 검찰개혁, 나아가 소중하게 지켜온 대한민국의 법치주의 원칙을 크게 훼손하는 것"이라고 강하게 비판했다. 이어 "검찰이 헌법과 법률에 따라 책임과 직무를 다 할 수 있도록 (윤 총장에 대한) 징계청구와 직무집행 정지를 재고해주실 것을 간곡히 요청드린다"고 적었다. **아래는 대검 중간간부들의 성명서 전문이다.** **<대검찰청 중간 간부들의 입장>** **○ 코로나19 등으로 인한 국가적 위기 상황 속에서 검찰과 관련된 각종 논란으로 국민들께 심려를 끼치고 있어 송구스럽게 생각합니다.** **○ 검찰이 변화해야 한다는 국민의 뜻에 부응하기 위해 노력하고 있으나, 여전히 많이 부족하다는 것을 잘 알고 있습니다.** **○ 다만, 최근 검찰을 둘러싸고 진행되고 있는 상황들에 대해 침묵하는 것은 공직자로서 올바른 자세가 아니라는 데에 뜻을 함께 한 대검찰청 중간 간부들은** **2020. 11. 26. 아래와 같이 의견을 모았습니다.** **○ 검찰공무원은 범죄로부터 우리 국민들을 보호하고, 온전한 법치주의 실현을 통해 자유롭고 안정된 민주사회를 구현해야 할 사명이 있습니다.**	

삭제 정보	예시
	○ 검찰총장에 대한 11. 24. 징계청구와 직무집행정지는 적법절차를 따르지 않고, 충분한 진상확인 과정도 없이 이루어진 것으로 위법, 부당합니다. ○ 이는 검찰의 정치적 중립성은 물론이고, 검찰개혁, 나아가, 소중하게 지켜온 대한민국의 법치주의 원칙을 크게 훼손하는 것이기도 합니다. ○ 검찰이 헌법과 법률에 따라 책임과 직무를 다 할 수 있도록 징계청구와 직무집행 정지를 재고해 주실 것을 법무부장관께 간곡하게 요청드립니다. ○ 저희들도 국민과 함께 하는 검찰공무원으로서 본연의 임무를 충실히 수행해 나가겠습니다. 2020. 11. 26. 손준성 이정봉 최성국 이창수 박기동 강범구 전무곤 고필형 구승모 임승철 이만흠 반종욱 최창민 진현일 박혁수 김용자 김우 백수진 한기식 김승언 김종현 신준호 추혜윤 장준호 손진욱 김연아 정태원 배ㅇㅇ 기자 ----@hani.co.kr
문장으로 볼 수 없는 정보	■ 인천·경기지역 시급 현안 '인천·경기에서 가장 우선적으로 해결해야 할 사안이 무엇이라고 생각하느냐'는 질문에 '일자리 창출'이 28.0%로 가장 높았다. 이어 '지역간 균형발전'이 19.1%, '부동산 가격 안정화'가 15.0%, '광역교통망 구축' 13.6%, '미세먼지 대책마련'이 10.7%, '수도권 규제완화'가 3.6% 순이다. '기타'가 7.5%, '잘 모름'이 2.4%다. 지역별로는 계양·부평권과 남동·연수·미추홀권은 일자리 창출이 각각 32.0%와 30.0%로 가장 높은 반면, 동·서·중구·강화·옹진권은 지역간 균형발전이 22.9%로 가장 높았다. 연령대별로 대부분은 일자리 창출을 시급한 현안으로 꼽았지만, 유일하게 40~49세에서만 지역간 균형발전이 가장 높았다. ㅇㅇㅇ기자 어떻게 조사했나 이번 조사는 경기일보의 의뢰로 조원씨앤아이가 2019년 12월28일(土)부터 30일(月)까지 사흘간, 인천광역시 거주 만19세 이상 남녀를 대상으로 ARS 여론조사(유선전화 RDD 12%+통신사 제공 휴대전화 가상번호 88% 방식, 성,연령,지역별 비례할당무작위추출)를 실시한 결과이며,표본수는 805명(총 통화시도 17,366명, 응답률 4.6%), 표본오차는 95% 신뢰수준에 ±3.5%p임. 그 밖의 사항은 중앙선거여론조사심의위원회 홈페이지 참조 ※오차보정방법 : [림가중] 성별, 연령별, 지역별 가중값 부여(2019년 11월말 행정안전부 발표 주민등록인구기준)
	지난 25일 서울 성동구에서 23년째 PC방을 운영하고 있는 이모씨(47)는 올해 추석 계획을 묻는 기자의 지문에 "가게를 지키는 일"이라고 답했다.

삭제 정보	예시
	코로나19로 적자가 너무 심해져 하루라도 가게를 비울 수 없다는 것이다.(관련 기사 ☞ "나라도 돈 벌겠다" 중2 아들말에…PC방 사장님 3일째 집에 못갔다)
	특히 생체리듬으로 알려진 '써카디안(circadian) 리듬'은 간헐적 단식에서 아주 중요한 요소다. 써카디안 리듬과 중추시계, 말초시계가 일치돼야 건강한 일상이 가능하기 때문. 햇볕이 비출 때 일어나고 일정한 시간에 건강한 음식을 취하며 해가 지면 잠자리에 드는 '원시 인류'의 생활을 따라야 한다고 저자는 강조한다. ◇호르메시스와 간헐적 단식=박용우 지음. 블루페가수스 펴냄. 276쪽/1만5000원.
	이후 2020대한민국지속가능혁신리더대상 조직위 운영 사무국으로 이메일 또는 우편(서울시 중구 청계천로 11(서린동, 청계한국빌딩 16층))을 통해 6월 30(화)(오후 6시까지 도착분에 한함)까지 제출하면 된다. 응모 자격은 정부 상훈 관련법에 부합하는 지자체·기관·법인 및 단체·개인으로 접수된 신청서류는 반환되지 않는다. 평가는 1차 서류심사, 2차 실사를 포함한 심층심사, 3차 최종평가를 거쳐 최종 수상자를 선정한다. 자세한 내용은 아래 개요를 참고하시기 바랍니다. 대한민국을 이끄는 혁신리더들의 많은 참여 바랍니다. [2020 대한민국지속가능혁신리더대상 개요] 주 최 : 2020 대한민국지속가능혁신리더대상 조직위원회 주 관 : 머니투데이, 더리더 접수마감 : 2020년 6월 30(화) 접수문의 : 02-724-0952(머니투데이 더리더) 접 수 처 : 이메일(awards@mt.co.kr) 시상일시 : 2020년 7월 중 시상장소 : 여의도 켄싱턴호텔 신청대상 : 정치·사회·경제·교육·체육·문화·예술·환경 등 각 분야의 지속적인 혁신 공로가 인정되는 지자체 및 우수 기관, 단체, 개인리더 등 신청양식 : 더리더 홈페이지 우측 상단 배너 클릭, 기사하단 신청서 다운로드에서 클릭 후 내려받기 가능

<표 8> 불필요한 요소 제거 내용

<그림 8> 원시 데이터와 정제된 데이터

캡션 정보가 본문과 붙어 있는 경우도 있으며, 본문과 구분이 되지 않는 것들도 존재하였다. 해당 기사의 웹페이지를 확인하고 비교하면서 작업을 진행하였다. 사용하지 않는 기사의 경우 표기를 따로 하여 데이터를 정제하였다.

데이터 정제 전	정제 데이터
[인천일보] 과천시 '갈등전환 퍼실리테이터 양성과정' 교육생 23명이 수료식을 마치고 기념촬영을 하고 있다. 교육생은 10월 22일부터 11월 19일까지 5주간 교육과정을 거쳤다./사진제공=과천시 과천시는 '갈등전환 퍼실리테이터 양성과정' 교육생 23명에 대한 수료식을 시청 대강당에서 진행했다고 22일 밝혔다. 이번 교육에서는 갈등전환 촉진토론 등 문제를 발굴하고 해결해 나가는 숙의 과정을 익히고 지역 현안을 주제로 한 공론장 운영을 실습했다. 이날 수료식에 참석한 한 교육생은 "다양한 갈등 상황에 대한 접근, 조정 등 효과적인 소통 기술을 익힐 수 있어 너무 유익했다"고 말했다. 시는 수료생들을 지역 문제 해결을 위한 토론회 개최 시 보조 퍼실리테이터로 참여할 기회를 줘 민관협치의 저변을 확대해 나갈 방침이다. 김종천 과천시장은 "정책을 추진하는 과정에서 발생하는 다양한 갈등과 문제를 지역주민들이 모여 고민하고 조율해 나가는 시작점이 될 것이며 수료생들께서 지역 내에서 갈등조정의 선도적 역할을 해달라"고 당부했다. /과천=ㅇㅇㅇ 기자 -------@incheonilbo.com	과천시는 '갈등전환 퍼실리테이터 양성과정' 교육생 23명에 대한 수료식을 시청 대강당에서 진행했다고 22일 밝혔다. 이번 교육에서는 갈등전환 촉진토론 등 문제를 발굴하고 해결해 나가는 숙의 과정을 익히고 지역 현안을 주제로 한 공론장 운영을 실습했다. 이날 수료식에 참석한 한 교육생은 "다양한 갈등 상황에 대한 접근, 조정 등 효과적인 소통 기술을 익힐 수 있어 너무 유익했다"고 말했다. 시는 수료생들을 지역 문제 해결을 위한 토론회 개최 시 보조 퍼실리테이터로 참여할 기회를 줘 민관협치의 저변을 확대해 나갈 방침이다. 김종천 과천시장은 "정책을 추진하는 과정에서 발생하는 다양한 갈등과 문제를 지역주민들이 모여 고민하고 조율해 나가는 시작점이 될 것이며 수료생들께서 지역 내에서 갈등조정의 선도적 역할을 해달라"고 당부했다.

<표 9> 원시 데이터와 정제된 데이터 비교 1

데이터 정제 전	정제 데이터
[서울=뉴스핌] OOO 기자 = 한국공예디자인문화진흥원(원장 김태훈)은 2021년 제101회 전국체전이 개최되는 경북 구미시 체육시설에 유니버설 안내 체계를 적용한다고 11일 밝혔다. 유니버설 디자인은 성별, 언어, 연령에 관계 없이 누구나 균등한 혜택을 제공받을 수 있다. 진흥원은 지난해 9월 구미시와 업무협약을 체결하고 올해 시민경기장과 보조경기장, 복합스포츠센터 등에 유니버설디자인 안내체계를 시범 조성한다. [서울=뉴스핌] ○○○ 기자 = (개선안) 출입구 표기 체계화, 파생되는 정보 노출하여 현 위치 정보 제공 강화 [사진=한국공예디자인문화진흥원] 2020.08.11. ------@newspim.com 주요 대상지인 시민경기장의 반복되는 구조물 특성상 현 위치와 게이트의 인지가 어려웠던 점을 개선하고 주요 지점의 안내판 추가 및 가독성을 더욱 높일 예정이다. 또한 다국어 표기, 장애인 좌석 표기나 시각장애인 배려 시설 등 현재 부재한 시설을 전반적으로 개선한다. (중략) 코로나19의 여파로 전국체전의 개최는 2021년으로 늦췄으나 구미시 체육 시설의 유니버설 안내체계 적용은 올해 11월까지 완료된다. 향후 안내체계 구현 사례를 수록한 가이드라인도 발행될 예정이다. ------@newspim.com	한국공예디자인문화진흥원(원장 김태훈)은 2021년 제101회 전국체전이 개최되는 경북 구미시 체육시설에 유니버설 안내 체계를 적용한다고 11일 밝혔다. 유니버설 디자인은 성별, 언어, 연령에 관계 없이 누구나 균등한 혜택을 제공받을 수 있다. 진흥원은 지난해 9월 구미시와 업무협약을 체결하고 올해 시민경기장과 보조경기장, 복합스포츠센터 등에 유니버설디자인 안내체계를 시범 조성한다. 주요 대상지인 시민경기장의 반복되는 구조물 특성상 현 위치와 게이트의 인지가 어려웠던 점을 개선하고 주요 지점의 안내판 추가 및 가독성을 더욱 높일 예정이다. 또한 다국어 표기, 장애인 좌석 표기나 시각장애인 배려 시설 등 현재 부재한 시설을 전반적으로 개선한다. (중략) 코로나19의 여파로 전국체전의 개최는 2021년으로 늦췄으나 구미시 체육 시설의 유니버설 안내체계 적용은 올해 11월까지 완료된다. 향후 안내체계 구현 사례를 수록한 가이드라인도 발행될 예정이다.

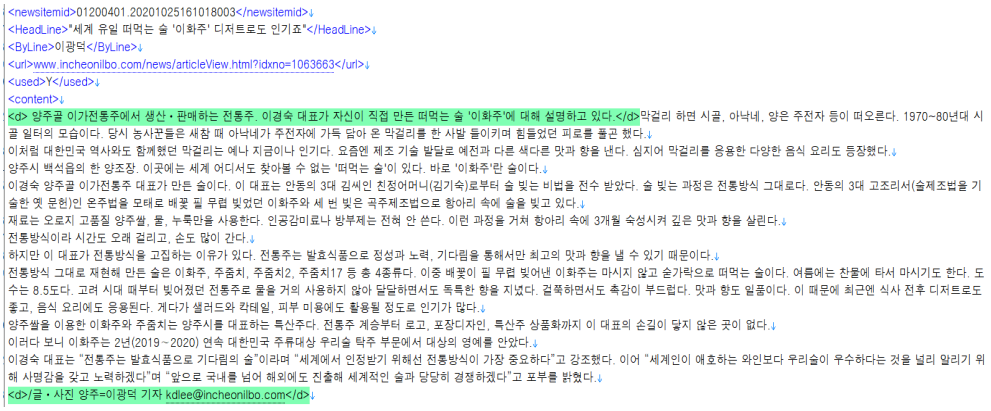
<표 10> 원시 데이터와 정제된 데이터 비교 2

데이터 정제 전	정제 데이터
(생략) A씨는 최근 인사발령에 따라 올해 3월부터 한 기관에서 지출(경리) 업무를 담당해왔다. 경찰은 유서 내용 등을 토대로 A 소방사가 극단적 선택을 한 것으로 보고 유족 등을 상대로 정확한 사망원인을 조사하고 있다. **※ 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.**	(생략) A씨는 최근 인사발령에 따라 올해 3월부터 한 기관에서 지출(경리) 업무를 담당해왔다. 경찰은 유서 내용 등을 토대로 A 소방사가 극단적 선택을 한 것으로 보고 유족 등을 상대로 정확한 사망원인을 조사하고 있다.

<표 11> 원시 데이터와 정제된 데이터 비교 3(본문 기사와 상관 없는 내용 삭제)

사용하지 않는 기사 예
술잔이 정신없이 오간다. '64년생' 자리만 조용하다. 오는 술잔도, 가는 술잔도 없다. 맘 속으로 내가 말한다. '차라리 집에 가라.' 하지만 '64년생'은 계속 앉아 있다. 30, 40대의 광적인 노래가 이어진다. 30여분 지났을까, 40대가 배려한다. "자, 64년생 어르신 모십니다." 맘 속으로 내가 또 말한다. '제발 옛날 노래는 하지 마라.' '64년생'은 또 기대를 저버린다. "바람에 날려버린...앗싸." '안동역에서'다. 앵콜이 없다. 억지로 의미를 부여하는 게 아니다. 2019년 시작된 '386'의 현실이다. 세대 중심이라던 자부심이 무너졌다. '2030세대'의 공격이 시작됐다. 무능해서 세상을 이렇게 만들었다고 추궁한다. 자리를 비켜 달라는 '88만원세대'의 삿대질이다. '6070세대'의 비난도 시작됐다. 겨우 이러려고 그 난리를 쳤냐며 비웃는다. 20, 30년 전에 밀려났던 '유신 세대'의 역공이다. 이날 모습이 그랬다. 중심에서 밀려나는 386의 현실이었다. 무너짐의 조짐은 정치에서 나왔다. 386 불출마 선언이 잇따랐다. 표창원(66년생)·이철희(64년생)·임종석(66년생)이 떠났다. 저마다 멋들어진 이유를 댄다. 현실 정치 실망·책임 정치 실현·통일사업 전념. 진실일 수도, 거짓일 수도 있다. 하지만, 언론은 간단히 정리했다. '386 퇴진 시작'이라고 제목을 뽑았다. 그러면서 나머지 386에도 마이크를 댄다. 사퇴할 생각 없냐고 따져 묻는다. 아마 몇은 더 떠날 듯싶다.

문어체 문장 집합으로 볼 수 없는 기사는 사용하지 않는 기사로 표기하였다.



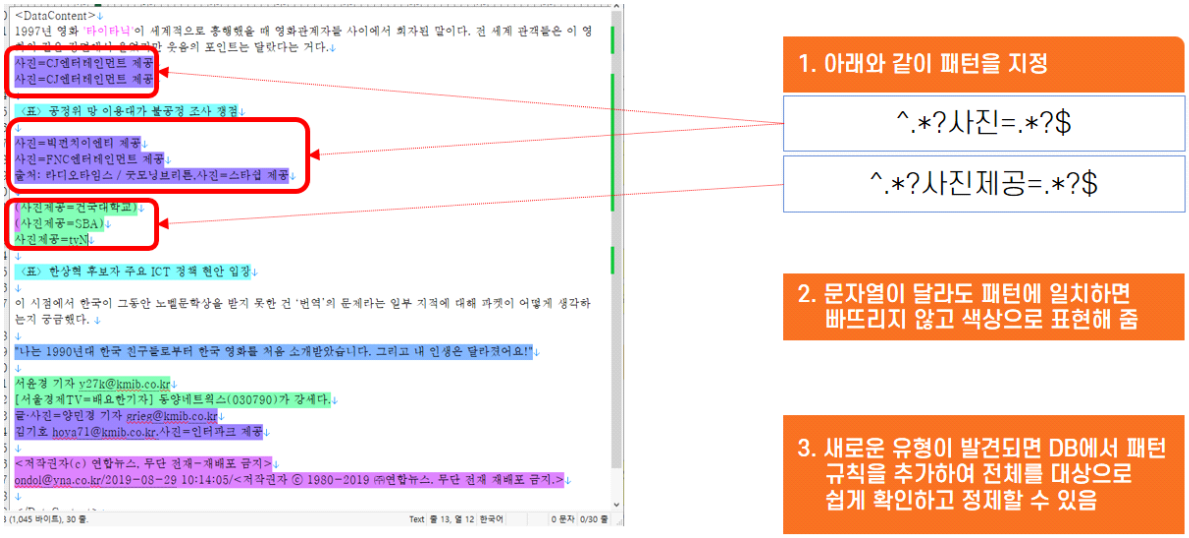
<그림 9> 작업 편집 화면

불필요한 요소는 리스트를 활용하여 처리함으로써 확인요소임을 분명하게 하였다. 수행사가 가지고 있는 프로그램을 사용하였으며, 모든 데이터는 기사 단위로 하여 데이터베이스관리시스템(DBMS)으로 관리하였다. 작업자들은 해당 기사를 엑스엠엘(XML) 데이터로 수령 받아 확인목록을 이용하여 직접 삭제가 아닌 마크업을 부여하여 표기하였고, 사용하지 않는 기사는 'used'라는 엘리먼트에 표기하여 작업하였다.

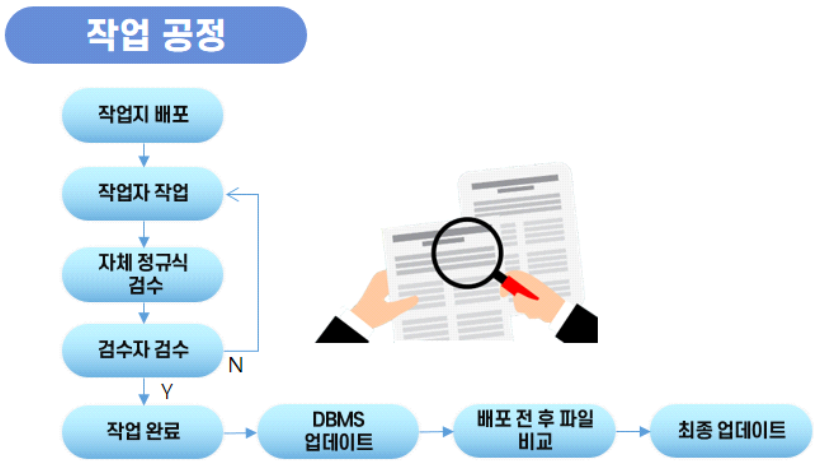
새로운 유형이 나오는 경우 작업자들이 구글 시트를 활용하여 해당 유형을 공유해서 나타날 수 있는 유형 요소들을 축적하였다.

인천일보의 경우 웹 문서로도 캡션 정보를 얻을 수 없어서 본문과 웹페이지를 비교하여 직접 확인하며 정제 작업을 거쳤다. 해당 매체가 가장 난도가 높은 매체라고 할 수 있다.

불필요한 요소 작업을 마친 데이터는 소실 비교를 통해 문자 데이터의 누락 등을 검수하였다.



<그림 10> 작업 프로그램 화면



<그림 11> 데이터 정제 2차 검수 공정

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

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

다. 한중일 호환용 한자 영역(F900-FAFF) 한자의 통일

인공지능 학습과 데이터 유통에 있어 통일되지 않은 코드는 크고 작은 문제를 일으킬 수 있다. 따라서 불필요한 요소를 제거한 후 해당 한중일 호환용 한자 영역에 대한 코드 통일을 진행하였다. 해당 데이터에서는 아래와 같은 한중일 호환용 한자 영역의 한자가 등장하였다.

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

최종 선정된 기사에서 사용된 한중일 호환용 한자 영역의 한자의 수는 아래와 같다. 약 3400건의 한자가 치환되어 통일되었다.

코드	한자	사용횟수	코드	한자	사용횟수	코드	한자	사용횟수
F9E1	李	393	F9F6	臨	31	F9F4	林	14
F978	兩	392	F941	論	27	F90F	羅	13
F90A	金	334	F997	聯	27	F92E	冷	13
F967	不	256	F93D	綠	25	F972	沈	13
F981	女	143	F9B6	禮	25	F9D3	陸	13
F9E4	理	140	F9CA	流	25	F9EA	離	13
F9FE	茶	122	F9B3	靈	22	F92D	來	12
F95C	樂	115	F9AA	寧	20	F97A	梁	12
F92F	勞	114	F97E	量	18	F98C	歷	12
F934	老	109	F999	蓮	18	F9A8	令	11
F98E	年	100	F9DD	利	18	F9C3	遼	11
F933	盧	71	F914	樂	17	F918	落	10
F9D1	六	65	FA04	宅	17	F932	爐	10
F9C7	劉	59	F91F	蘭	16	F984	濾	10
F9C4	龍	55	F990	戀	16	F9DA	栗	10
F9AE	瑩	50	F94C	樓	15	F98A	力	9
F9F7	立	50	F902	車	14	F9B4	領	9
F99A	連	47	F940	鹿	14	F9D8	律	9
F91B	亂	31	F9C9	柳	14	F93F	錄	8

<표 12> 최종 선정 기사 한중일 호환용 한자 영역의 한자 수(이하 생략)

❍ 기존 한중일 호환용 한자 영역의 한자는 아래 표의 정보로 치환함.

코드	한자	치환 코드	코드	한자	치환 코드	코드	한자	치환 코드	코드	한자	치환 코드
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
F93F	錄	9304	F9BF	樂	6A02	F946	牢	7262	F9C8	杻	677B
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
F959	陵	9675	F9D3	陸	9678	F971	辰	8FB0	F9E4	理	7406
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	凉	51C9	F9F4	林	6797

코드	한자	치환 코드	코드	한자	치환 코드	코드	한자	치환 코드	코드	한자	치환 코드
F96B	參	53C3	F9E2	梨	68A8	F985	礪	792A	FA08	行	884C
F986	閭	95AD									

<표 13> 한중일 호환용 한자 영역 한자 치환 표

라. 인용부호의 통일

인용부호의 통일은 최초 문장 교정 말뭉치에서 통일하는 것으로 제안하였으나 국어원의 요청으로 신문 기사 말뭉치에 적용하기로 하였다. 현재 대부분의 매체에서는 표준이 아닌 인용부호로 '(0027)와 "(0022)를 사용하는 경우가 많고 실제 데이터에서도 대부분 위와 같은 부호가 사용되었다. 각 매체별 편집기에서 문서를 작성할 때 열고 닫는 인용부호를 사용하기 어려워서 '와 "를 사용한 것으로 보인다.

인용부호의 통일 작업은 생각보다 쉽지 않았다. 부호가 열리고 닫히지 않는 기사, 다르게 열고 닫힌 부호, 부호가 통일되지 않는 기사들이 다수 존재하였다.

아래 예시는 부호 짝이 맞지 않는 경우이다. 다음과 같이 짝이 맞지 않는 경우가 상당히 많이 존재하였으며 큰따옴표로 열리고 작은따옴표로 닫히는 경우, 큰따옴표로 열리고 닫히지 않는 경우, 작은따옴표로 열리고 닫히지 않는 경우, 닫히는 부호만 있는 경우 등 다수의 사례가 존재하였다. 인용부호에서만 해당 내용을 수정하였으며, 영어에 등장하는 ‘Apostrophe’의 경우에는 그대로 살려 주었다.

수행사는 데이터베이스관리시스템(DBMS)을 이용하여 부호의 수가 맞지 않는 단락을 찾아내어 패턴 등을 통해 해당 문제를 해결하였다.

엔터 옆 기호를 사용한 부호를 표준 기호에 맞게 수정하였다.

코드	문자	치환 코드	치환 문자	비고
0027	'	2018	‘	여는 내용
0022	"	2019	’	닫는 내용
0027	'	201C	“	여는 내용
0022	"	201D	”	닫는 내용

<표 14> 인용부호 치환 표

데이터 정제 전	데이터 정제 후
그러나 결국 강남은 땅을 받지 않았다. 태진아는 지난해 12월9일 방송된 KBS 쿨FM '박명수의 라디오쇼'에 출연해 "시골에 땅이 하나 있다. 300평에서 600평 정도 될 거다. 내 이름으로 돼 있다" 며 "내가 강남한테 '결혼하면 300평 줄게'했는데 강남이 3000평으로 알아들은 거다"라고 설명했다. 이어 "결혼할 때 이 땅 주기로 했으니까 가져가라고 했더니, 300평이면 안된다고 한다더라. 3000평이나 돼야 스케이트장 만든다고, 너무 멀어서 싫다더라"고 덧붙였다.	그러나 결국 강남은 땅을 받지 않았다. 태진아는 지난해 12월9일 방송된 KBS 쿨FM ‘박명수의 라디오쇼’ 에 출연해 “시골에 땅이 하나 있다. 300평에서 600평 정도 될 거다. 내 이름으로 돼 있다”며 “내가 강남한테 ‘결혼하면 300평 줄게’ 했는데 강남이 3000평으로 알아들은 거다”라고 설명했다. 이어 “결혼할 때 이 땅 주기로 했으니까 가져가라고 했더니, 300평이면 안된다고 한다더라. 3000평이나 돼야 스케이트장 만든다고, 너무 멀어서 싫다더라”고 덧붙였다.

<표 15> 인용 부호 수정 데이터 정제 전 후

데이터 정제 전	데이터 정제 후
사업책임자인 머니투데이 권현수 부장은 “리테일테크(Retail Technology)가 국내 유통업계의 트렌드로 부상, 이에 발맞춰 기존의 ‘유통·물류 전문가 양성과정’을 ‘ICT 기반 SCM 리테일테크 전문가 양성과정’으로 진화시켰다’며 이번 교육생 모두가 교육과정을 잘 수료해 유통·물류와 정보통신기술을 융합할 수 있는 리테일테크 전문가로 성장하길 소망한다”고 말했다.	사업책임자인 머니투데이 권현수 부장은 “리테일테크(Retail Technology)가 국내 유통업계의 트렌드로 부상, 이에 발맞춰 기존의 ‘유통·물류 전문가 양성과정’을 ‘ICT 기반 SCM 리테일테크 전문가 양성과정’으로 진화시켰다”며 “이번 교육생 모두가 교육과정을 잘 수료해 유통·물류와 정보통신기술을 융합할 수 있는 리테일테크 전문가로 성장하길 소망한다” 고 말했다.

<표 16> 인용 부호 수정 데이터 정제 전 후 2

매체명	기사 수	어절 수	매체명	기사 수	어절 수
한국경제	56,918	16,446,248	부산일보	23,631	6,451,426
서울경제	52,907	15,025,365	충청투데이	9,011	2,396,700
아시아경제	60,090	16,660,172	대전일보	10,802	2,748,143
파이낸셜뉴스	13,504	3,863,152	경북일보	8,005	2,145,656
아주경제	33,292	9,460,900	뉴스핌	35,101	9,787,773
머니투데이	56,155	16,232,239	중도일보	12,163	3,001,989
스포츠서울	21,701	5,894,401	헤럴드경제	45,065	12,273,735
노컷뉴스	25,267	7,669,377	중부일보	1,129	278,568
EBN산업뉴스	18,082	5,127,369	인천일보	9,126	2,284,191
전자신문	19,069	5,039,492	e대한경제	5,156	1,413,814
한겨레	17,337	5,300,538	전북도민일보	6,335	1,607,993
국민일보	39,143	11,045,271	전남일보	5,337	1,519,470
서울신문	35,742	10,118,015	대구신문	8,227	2,088,152
한국일보	22,226	6,201,601	경남도민일보	6,572	1,857,029
경기일보	10,907	2,681,205	남도일보	10,304	2,720,209
강원도민일보	3,795	976,544	환경일보	12,232	3,325,166
충청일보	8,143	1,968,647	조선일보	14,164	4,547,807
매일신문	13,379	3,427,386	**총 합**	**730,017**	**203,585,743**

<표 17> 최종 선정 기사 수

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

5. 메타데이터 작성

메타데이터 작성 공정은 기사의 제목, 저자, 발행자, 기사 작성일, 원 주제분류, 국어원에서 제시한 9가지 주제로 기사를 재분류하는 주제분류, 어절 수 등을 메타데이터로 작성하는 공정이다.

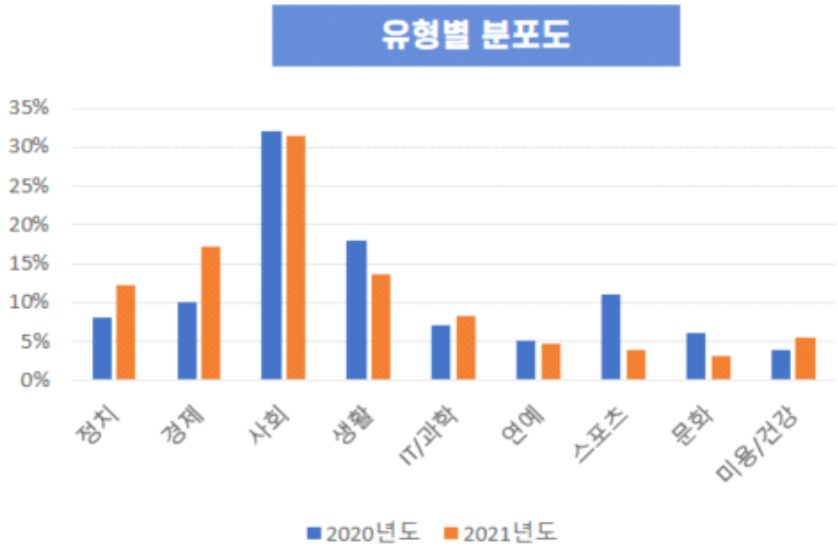
신문사별로 카테고리를 분류하게 되는데 이는 각 매체마다 구분 짓는 방법도 다르고 카테고리명 또한 다르다. 메타 정보에는 신문사에서 분류한 원 주제가 분류되어 들어가고, 이를 통합하여 관리하기 위해 국립국어원이 제시한 9가지 분류체계로 모든 기사를 재분류하여야 한다. 정치, 경제, 사회, 생활, IT/과학, 연예, 스포츠, 문화, 미용/건강의 통합 분류 체계로 최종 선정된 기사를 재분류하였다.

수행사가 가지고 있는 인공지능 모델을 신문 기사 학습에 최적화시켜 기사 분류를 진행하였다. 정확도 93.7%의 모델을 생성하여 사용하였으며, 기존 공개된 약 350만 개의 기사와 주제 분류를 학습시켜 정확도 80% 이상인 데이터만을 선별하였다.

<그림 12> 인공지능을 활용한 주제 분류

정치	경제	사회	생활	IT/과학	연예	스포츠	문화	미용/건강
12.2%	17.2%	31.5%	13.7%	8.2%	4.7%	4.0%	3.1%	5.4%

<표 18> 2021년 신문기사 사업 주제 분류 비율



<그림 13> 2020년, 2021년 유형별 분포 비교

6. 문장 말뭉치

신문 기사 내에는 수많은 오류와 일관성 없이 사용된 문자 등이 있어 인공지능 학습에 나쁜 영향을 준다.

ＡＢＣａ 등 전각 알파벳, ［？＠；（＇＆ 등 전각 부호, ０１２ 등 전각 숫자 등은 데이터의 일관성 및 정보처리 효율성을 위해 모두 반각 문자로 치환하였다.

·(MIDDLE DOT)는 ・(318D), ・(22C5), ･(30FB), •(2219), •(2022), ·(0387), ·(1427), ·(2024), ·(2027), ・(2981), ･(FF65) 등의 가운뎃점은 ·(00B7)로 치환하였다.

또한 대부분의 인공지능 학습은 문장을 기본 단위로 하고 있다. 특히 형태소 분석과 기계 번역은 대부분 문장을 기본 단위로 하고 있다. 따라서 단락을 문장으로 세분한 문장 말뭉치를 구축한다.

말뭉치의 활용성을 높이기 위해 단락을 문장으로 한 단계 더 분할하였다. 하나의 문장은 보통 마침표(.), 느낌표(!), 물음표(?) 등의 문장 부호를 기본 단위로 한다. 그러나 문장이 끝나는 부분이 아닌 곳에 사용되는 예가 많기 때문에 반드시 예외 처리를 해주어야 한다. 주의해야 할 점은 문장으로 분할 할 때 피인용문 내부에 사용된 마침표(.), 느낌표(!), 물음표(?) 등의 문장 부호는 분할하지 않는 것을 원칙으로 한다.

기존 신문 기사 말뭉치와 함께 활용할 수 있도록 기존 ‘<p>’태그는 기존 단락의 태그이고 해당 태그 안에 ‘<s>’태그를 삽입하여 문단을 구분하였다.

가. 코드 통일

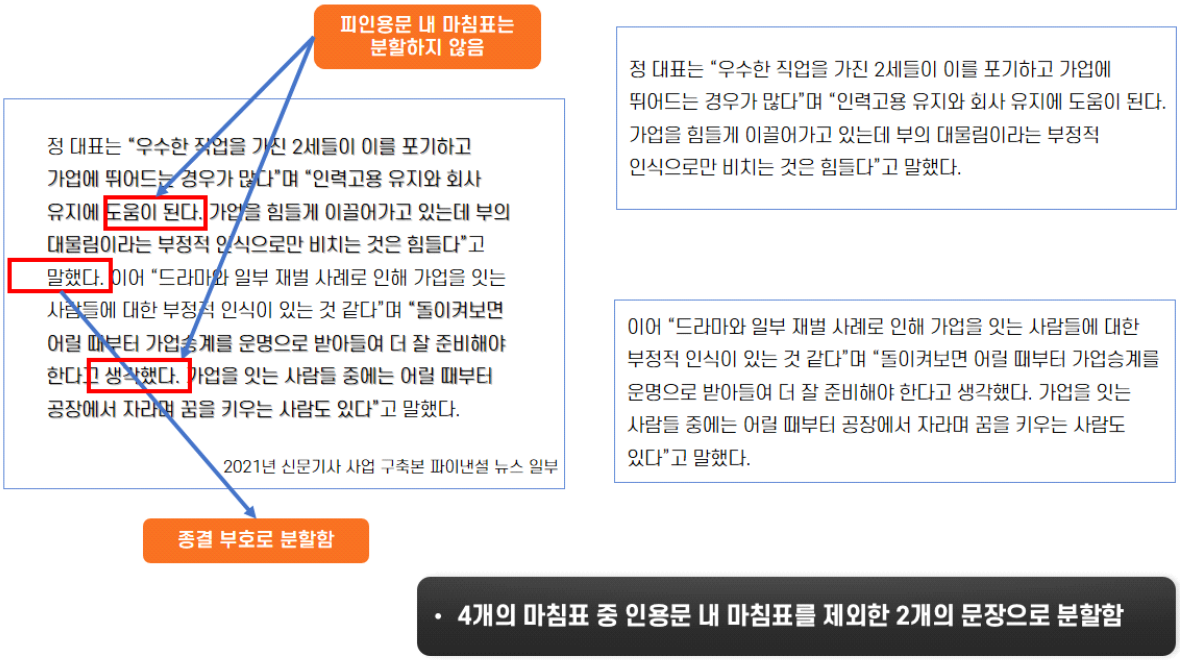
❍ 아래 목록은 치환 코드 리스트

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	/	

대상 코드	대상 문자	대상 코드	치환 문자	비고
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
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	·	

나. 문단 분할

- 문장의 분할은 수행사가 가지고 있는 문단 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 진행.
- 피인용문 내 마침표(.), 느낌표(!), 물음표(?) 등의 문장부호는 분할하지 않음.



<그림 14> 문장 말뭉치 개념

신문기사 말뭉치	문장 말뭉치
공효진은 "드라마를 함께한 배우들이 상을 받을 때마다 내가 받은 것보다 더 울컥하고"라고 한 뒤 말을 잇지 못했다. 그는 "향미(손담비 분)하고 눈이 마주쳤는데"라며 눈물을 흘렸고 "덤덤할 거라고 주변에 이야기했는데, 이 자리가 그냥 막 마음을 이렇게 만든다. 같이 했던 배우들이 눈앞에 있어서 더 그런 기분이 드는 거 같다"고 말했다.	공효진은 "드라마를 함께한 배우들이 상을 받을 때마다 내가 받은 것보다 더 울컥하고"라고 한 뒤 말을 잇지 못했다. 그는 "향미(손담비 분)하고 눈이 마주쳤는데"라며 눈물을 흘렸고 "덤덤할 거라고 주변에 이야기했는데, 이 자리가 그냥 막 마음을 이렇게 만든다. 같이 했던 배우들이 눈앞에 있어서 더 그런 기분이 드는 거 같다"고 말했다.

<표 19> 문장 말뭉치 데이터 정제 예

매체	단락	문장	매체	단락	문장
강원도민일보	25,403	57,272	인천일보	17,099	23,620
경기일보	76,511	115,727	전남일보	10,846	15,570
경남도민일보	43,282	77,628	전북도민일보	11,169	13,789
경북일보	52,063	61,323	전자신문	28,584	51,401
국민일보	170,061	338,823	조선일보	15,676	45,426
남도일보	57,497	76,520	중도일보	17,242	21,692
노컷뉴스	130,067	167,961	중부일보	1,805	2,079
뉴스핌	136,193	200,852	충청일보	13,170	14,286
대구신문	25,433	37,684	충청투데이	13,171	17,673
대전일보	31,211	45,267	파이낸셜뉴스	19,820	34,215
매일신문	37,050	55,387	한겨레	14,467	37,449
머니투데이	160,780	275,829	한국경제	59,227	129,480
부산일보	59,055	95,210	한국일보	21,767	42,960
서울경제	85,663	197,897	헤럴드경제	47,898	81,040
서울신문	76,935	134,484	환경일보	14,530	17,738
스포츠서울	34,530	74,673	e대한경제	6,374	9,352
아시아경제	108,817	189,513	EBN 산업뉴스	20,782	32,033
아주경제	66,334	108,651	**총 합**	**1,710,512**	**2,900,504**

<표 20> 단락 단위를 문장 단위로 분할한 수치

상위 5 문장분할수

매체	문장분할 수
조선일보	2.9
한겨레	2.6
서울경제	2.3
강원도민일보	2.3
한국경제	2.2

하위 5 문장분할수

매체	문장 분할 수
충청일보	1.1
중부일보	1.2
경북일보	1.2
환경일보	1.2
전북도민일보	1.2

<그림 15> 문장분할 상위 5개, 하위 5개 매체

조선일보의 경우 문장 분할 평균이 한 단락당 2.9개로 가장 높았으며 충청일보의 경우 1.1개로 가장 낮은 수치를 보였다. 평균 문장 분할 수는 1.7개이다.

7. 문장 교정 말뭉치

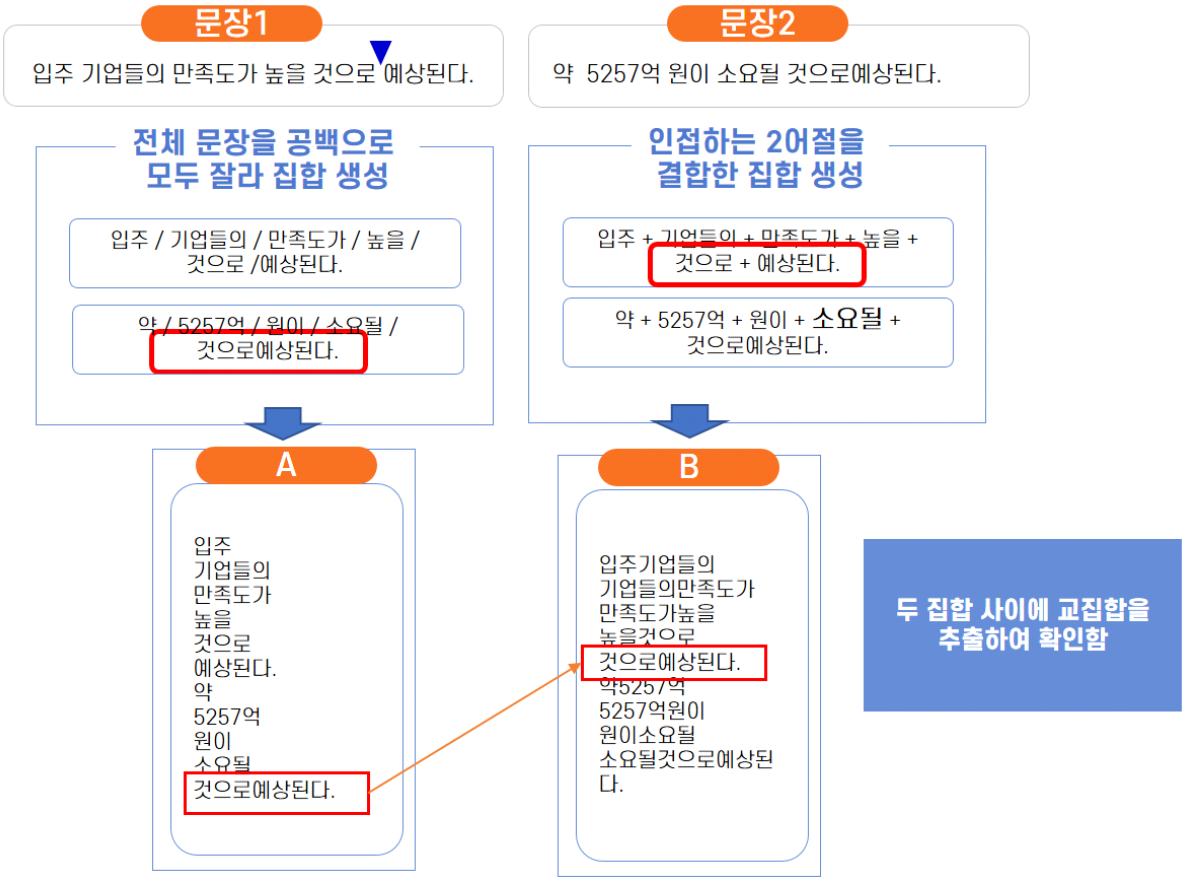
인공 지능 학습에 있어 맞춤법 오류, 띄어쓰기 오류는 장애를 일으킬 수 있다. 이를 바로잡아 데이터의 일관성 및 품질을 높여야 한다. 맞춤법 교정은 인공지능 학습에 심각한 영향을 주는 오류를 중심으로 수정 진행하였다. 오자나 의미 전달에 지장을 주는 띄어쓰기 오류를 중점적으로 수정하였으며. 신문 기사 데이터를 어절로 분할하여 A그룹을 만든 뒤 A그룹의 앞, 뒤 어절을 합한 어절들로 B그룹을 생성하였다. A그룹과 B그룹을 비교하여 교집합을 생성하고 내용을 비교 확인 후 교정을 진행하였다.

교정은 아래와 같은 기준으로 진행하였다.

오자 같은 경우 전체 데이터에서 사용된 글자를 전부 조회하여 오류 후보 글자를 추출하여 해당 글자를 전체 데이터에서 조회하면서 기사를 확인한 후 수정하였다.

적용 원칙
❍ 본 용언, 보조 용언의 띄어쓰기는 원 형태를 그대로 인정함
– 예) 늙어 간다, 되어 간다
❍ 복합명사 등 띄어쓰기에 의미 전달에 큰 문제가 되지 않는 것은 그대로 인정함
– 예) 칸영화제, 학생운동
❍ 한 단어로 등재되지 않았으나 여러 단어가 관용적으로 한 단어처럼 사용된다면 그 형태를 가능한 한 유지함
– 예) 핵심지표, 시민친화적, 최종후보
❍ 고유명사, 동의어, 유의어, 외래어는 원 형태를 그대로 인정함
❍ 연도나 숫자의 띄어쓰기는 그대로 인정하되, 유형을 찾아 수정 할 수 있는 표현은 수정함.
– 예) 2018년12월 -> 2018년 12월, 3월말->3월 말 , 8월이후->8월 이후
❍ 문장 부호의 앞 뒤 띄어쓰기는 수정함
– 예) ‘“”]) ?!, . 등의 부호 앞 뒤 공백
❍ 학습에 장애가 되는 심각한 오류는 수정함

<표 21> 띄어쓰기 적용 원칙



<그림 16> 검증을 통한 확인 띄어쓰기 확인 방법

후보 대상	변경 대상	후보 대상	변경 대상
보고없이	보고 없이	제일먼저	제일 먼저
뿌리깊은	뿌리 깊은	상황인만큼	상황인 만큼
필요해보인다.	필요해 보인다.	부드러운탄닌감과	부드러운 탄닌감과
휴양및	휴양 및	봄이오는	봄이 오는
성공할수	성공할 수	부담때문에	부담 때문에
등강화된	등 강화된	받고있다.	받고 있다
지역및	지역 및	놓고간	놓고 간
는요청을	는 요청을	타개하기위해	타개하기 위해
가지고있던	가지고 있던	역할을해온	역할을 해온
힘을내야	힘을 내야	부여하여총	부여하여 총
병원내	병원 내	지키고있다.	지키고 있다.
큰그림을	큰 그림을	붙을것으로	붙을 것으로
서로다른	서로 다른	있다"고전했다.	있다"고 전했다.
이또한	이 또한	뒤를이었다.	뒤를 이었다.
검토해야한다"고	검토해야 한다"고	필요없는	필요 없는
다바쳐	다 바쳐	낮은것으로	낮은 것으로
네이버는글로벌	네이버는 글로벌	것아니냐는	것 아니냐는
그때오라	그때 오라	아침일찍	아침 일찍
설명하기도했다.	설명하기도 했다.	크지않다.	크지 않다.
나쁜개는	나쁜 개는	따르면지난해	따르면 지난해
있는것을	있는 것을	내린비로	내린 비로
인기가높다.	인기가 높다.	보호하겠다고한	보호하겠다고 한
사이에둔	사이에 둔	방역수칙을꼭	방역수칙을 꼭
마련에온	마련에 온	어린이집등	어린이집 등
부처가운데	부처 가운데	부담아닌	부담 아닌
봇물터지듯	봇물 터지듯	계약시	계약 시
복구하기위해선	복구하기 위해선	지급받을수	지급받을 수
발표뒤	발표 뒤	봤을법한	봤을 법한
플랫폼을구축해	플랫폼을 구축해	보인만큼	보인 만큼
소설속	소설 속	한것도	한 것도

<표 22> 띄어쓰기 오류 후보 목록 추출 표

❍ 위의 후보 목록은 기사에서 사용된 어절들을 추출한 일부임.

맞춤법 오류와 오타 등의 수정 예시는 다음과 같다.

유형	오류	수정
맞춤법 오류	괜시리	괜스레
	깡총	깡충
	꺽다	꺾다
	뇌졸증	뇌졸중
	닥달	닦달
	단발마	단말마
	단촐	단출
	무릎쓰	무릅쓰
	설겆이	설거지
	쉽상	십상
	씼다	씻다
	어짜피	어차피
	얼키고 섥힌	얽히고 설킨
	요세	요새
	움추	움츠
	일찍이	일찍이
	있슴	있음
	장농	장롱
	죘	좼
	찌게	찌개
	통털어	통틀어
	희안	희한
오타	돼기고기	돼지고기
	들끊	들끓
	때	때
	스마프폰	스마트폰
	빠저	빠져
	찾아	찾아
	햇다	했다
	했따	했다

유형	세부유형	오류	수정
띄어쓰기	어미+명사	할수	할 수
		될만큼	될 만큼
		넘볼수	넘볼 수
		한관계자는	한 관계자는
	명사+명사	기업내	기업 내
		고민중인	고민 중인
	조사+용언	것으로보인다.	것으로 보인다.
		등을통해	등을 통해
		이에따라	이에 따라
		이와함께	이와 함께
	명사+용언	강도높은	강도 높은
		바있다.	바 있다.
		숨쉴	숨 쉴
		전례없는	전례 없는
	부사+용언	공고히할	공고히 할
		못받은	못 받은
		더나은	더 나은
	어미+용언	하고있다.	하고 있다.
		라고말했다.	라고 말했다.
	관형사+명사	다른사람들에게	다른 사람들에게
붙여쓰기	명사+지정사	것 이다.	것이다.
	접두사+명사	저 품질	저품질
	명사+접미사	지도 상	지도상
	한 단어 용언	사로 잡았다.	사로잡았다.
		조리 돌리고	조리돌리고
		드러 났다.	드러났다.
		발가 벗고	발가벗고
		큰코 다치다.	큰코다치다.
		합 치다.	합치다.
		남 부끄러운	남부끄럽지
		두드러 졌다.	두드러졌다.
		보잘것 없다.	보잘것없다.

말뭉치 데이터에서 사용된 글자를 전부 조회하여 사용 오류 확률이 높은 글자를 추출하고 해당 글자를 데이터에서 조회하여 확인하고 수정하는 절차를 진행하였다.

오류 후보 목록 글자를 추출하여 데이터를 조회하여 직접 보며 수정하였다.

❍ 명백한 오자 등을 검증하여 문장 교정 말뭉치 생성을 진행함.

오류 후보 글자														
좠	꽐	붑	켬	쥠	뙨	쬬	덨	쇘	켭	궘	닶	륻	붚	앝
쳥	낏	뽂	쿳	쫘	뚬	쳰	뎨	슁	퀌	궨	닼	릏	붰	앍
퀭	닯	슼	퀑	쫜	뚯	츕	돎	싼	큄	귑	댬	맊	븅	얐
횰	뒈	쒔	틉	쬠	뜽	켱	돘	쌥	킯	귱	뎄	먓	븕	얖
껜	땟	엶	폅	쳘	떡	쾡	됙	쎽	킾	긾	덬	먿	븝	옪
딹	떪	옿	퓔	쵤	뢸	텥	듧	앍	톻	꺳	뎟	멂	뺃	옫
룔	뜹	웟	햤	췻	룅	패	뒬	앻	퉅	껶	돟	몃	뺘	옺
쑴	먕	윱	횅	츱	륩	폽	딫	엏	튭	꼉	됀	뫈	뺸	왌
쥣	믓	졔	흼	탹	맆	핬	떢	옸	팔	꼍	됭	묑	뻇	욍
촙	밞	좔	겻	텉	맜	햡	떱	왤	펈	꾶	둉	묠	뻉	욺
퉤	뵘	찿	곬	툥	먈	횄	뗬	왰	폠	꾜	뒬	묭	뻴	웚
튐	븟	콕	팼	퍅	묫	횬	뙜	욀	폣	꾿	뒺	뭎	뽜	웗
풂	븨	큔	궝	퍈	벱	홇	뛉	웽	폮	끛	듲	뭏	섵	웛
꾐	빍	텬	깆	푄	볌	훵	뜷	읏	푠	끰	듴	뭬	셑	웤
뗑	뺴	틜	꿏	픙	뷕	휸	랟	읔	핚	낧	듷	믁	셗	윘
멉	쒼	폔	꿕	핱	뷔	흣	럅	읨	핟	낻	땁	믜	셱	윚
붜	옾	햑	놯	혤	빋	횝	롓	읶	핧	냪	땡	믱	솀	윺
솃	웝	걘	뇔	ㄸ	셕	갂	뤌	잍	핶	넑	뗸	밣	쇈	읃
슉	읖	댱	덛	ㅆ	숖	걲	륫	잛	홥	넔	뗼	밨	쉠	읇
좇	쟉	덞	됫	ㅛ	쌨	겝	륽	쟘	횝	넖	똸	뱐	쉡	읉
쯧	챨	떽	딮	ㄶ	썪	곈	맗	쟨	횳	넢	뙈	뱝	쉣	읗
쳅	큇	뜔	룁	ㅂ	썻	곌	먁	졉	횟	넫	뙬	벇	싀	읫
췐	퀑	떱	릐	ㅎ	쒸	굔	먻	죕	횔	넷	뚸	벍	싁	읱
쾃	큭	릍	먄	겯	앚	귁	밪	줕	걈	녋	랒	벎	싷	있
꼈	퐈	벛	몀	굄	얉	깉	뱍	줫	걺	녑	랼	벝	쌜	잆
궃	햔	삣	몆	귝	엤	꽥	붸	쥡	겍	녓	럄	벴	쌴	쟐
됬	휄	섑	뷘	깄	옜	꾄	븥	쥼	곙	녺	렿	볓	쌥	쟜
밎	겆	솬	븍	꺠	옝	뀨	빝	짌	곩	놏	롁	볘	쌍	젅
쇗	괼	슙	빽	꽜	옯	끋	쀠	짛	곷	놭	롟	볻	쎔	젇
슙	긔	쑬	삠	놘	욌	낢	삷	챂	괂	뇠	롷	봈	쎘	졀

오류 후보 글자														
잌	깠	옘	삡	늫	읎	넀	샇	챰	괍	눚	롹	봍	쎙	졲
짦	끅	왬	섈	닸	읗	넗	샮	챱	괕	눞	뢀	봑	쏫	좃
쨔	끎	젔	솓	돨	읻	넞	색	챵	굇	눤	룃	봔	쏱	죡
쬔	넘	좍	솨	뒷	젹	뇬	섄	쳑	굏	눳	룐	봰	쐤	죨
횔	닢	줸	앟	뒨	죔	늗	섥	쳣	굘	눼	룦	봴	쓩	짔
괏	묜	쫴	졈	땐	쥼	닞	섧	춍	굻	늡	뤈	뵀	쓷	짼
굥	믕	첬	졘	뗴	짯	닺	쇘	췰	궘	늣	뤕	뷺	쓋	쨩
귿	봏	췰	죵	똣	짹	댛	쉉	컾	궨	뉜	륑	붗	씅	쩎

<표 23> 선정 기사에서 등장하는 오류 후보 글자 목록

위의 오류 후보 글자를 조회하여 해당 기사를 조회하고 수정한 내용 중 일부는 다음과 같다.

오류 후보 글자	해당 내용	교정 내용
옸	5·18 왜곡 처벌법은 더불어민주당이 당론으로 추진해옸으며, 이 개정안은 5·18 민주화 운동에 대한 허위사실을 유포해 명예훼손을 할 경우 처벌하는 규정을 신설하는 것이다.	5·18 왜곡 처벌법은 더불어민주당이 당론으로 추진해왔으며, 이 개정안은 5·18 민주화 운동에 대한 허위 사실을 유포해 명예훼손을 할 경우 처벌하는 규정을 신설하는 것이다.
횝	부산시는 코로나19로 인한 고용위기를 극복하고 침체한 지역경제를 횝고하기 위해 취약계층에 공공일자리를 제공하는 '코로나19 극복 부산희망일자리사업'을 추진한다고 13일 밝혔다.	부산시는 코로나19로 인한 고용위기를 극복하고 침체한 지역경제를 회복하기 위해 취약계층에 공공일자리를 제공하는 '코로나19 극복 부산희망일자리사업'을 추진한다고 13일 밝혔다.
궙	삼성전자는 이번 주총부터 주주권리 강화 일환으로 전자투표제를 도입해 주주들이 보다 손궙게 의결권을 행사할 수 있도록 했다.	삼성전자는 이번 주총부터 주주권리 강화 일환으로 전자투표제를 도입해 주주들이 보다 손쉽게 의결권을 행사할 수 있도록 했다.
뛸	지난 1일부터 한.중간에 실시하고 있는 경제인입국간소화절차가 본보기가 뛸 것으로 내다봤다.	지난 1일부터 한.중간에 실시하고 있는 경제인입국간소화절차가 본보기가 될 것으로 내다봤다.

오류 후보 글자	해당 내용	교정 내용
뭐	또 전체생존기간(OS)도 43.2개월로 뭐어나다고 회사측은 설명했다.	또 전체생존기간(OS)도 43.2개월로 뛰어나다고 회사측은 설명했다.
랒	브랒리 보건부는 '더 많은 의사들'(Mais Medicos) 프로그램을 통해 5천800여 명의 의사를 선발해 다음 달 초부터 보건소 등 공공보건 시설에 투입한다는 계획이다.	브라질 보건부는 '더 많은 의사들'(Mais Medicos) 프로그램을 통해 5천800여 명의 의사를 선발해 다음 달 초부터 보건소 등 공공보건 시설에 투입한다는 계획이다.
롟	마인크래프트 던전스, 헤일로 5: 가디언스, 오리와 도깨비불 등을 비롟해 100여 종의 게임을 모두 즐길 수 있다.	마인크래프트 던전스, 헤일로 5: 가디언스, 오리와 도깨비불 등을 비롯해 100여 종의 게임을 모두 즐길 수 있다.
롼	이단이나 사이비종교의 위장포교가 정상적인 포교롼동이 아니라는 첫 판결이어서 다른 이단 피해자들의 소송이 이어질 것으로 보인다.	이단이나 사이비종교의 위장포교가 정상적인 포교활동이 아니라는 첫 판결이어서 다른 이단 피해자들의 소송이 이어질 것으로 보인다.
곩	한편, 이번 학술대회에 국회의장, 국무총리, 과학기술부장관, 문화체육관광부장관, 중소벤처기업부장관 등이 영상을 곩해 축사를 할 예정이다.	한편, 이번 학술대회에 국회의장, 국무총리, 과학기술부장관, 문화체육관광부장관, 중소벤처기업부장관 등이 영상을 통해 축사를 할 예정이다.
셕	듀얼 퓨얼 인젝션 시스템을 적용했다. 듀얼 퓨얼 인젠셕 시스템을 통해 MPI(다중분사)와 GDi(가솔린 직분사) 방식을 상황에 따라 유동적으로 사용할 수 있다.	듀얼 퓨얼 인젝션 시스템을 적용했다. 듀얼 퓨얼 인젝션 시스템을 통해 MPI(다중분사)와 GDi(가솔린 직분사) 방식을 상황에 따라 유동적으로 사용할 수 있다.
륩	아모레퍼시픽그륩은 3~5년 목표로 중장기 전략을 수립해 현재 37% 수준인 해외사업 비중을 2023년까지 50%로 늘린다	아모레퍼시픽그룹은 3~5년 목표로 중장기 전략을 수립해 현재 37% 수준인 해외사업 비중을 2023년까지 50%로 늘린다

오류 후보 글자	해당 내용	교정 내용
	는 계획이었다	는 계획이었다
섧	‘현장 조류인플루엔자 서브타입 감별용 신속 키트 및 유전자칩 개발 및 산업화’을 지원한 결과라고 **섧명했다.**	‘현장 조류인플루엔자 서브타입 감별용 신속 키트 및 유전자칩 개발 및 산업화’을 지원한 결과라고 **설명했다.**

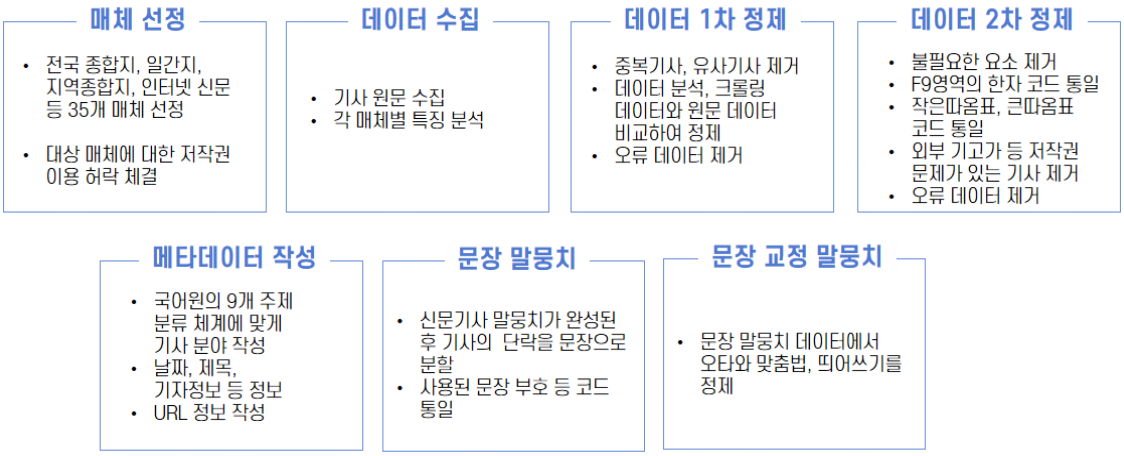
<표 24> 오류 후보 목록 글자 수정 전 후

제 3 장

사업 수행 결과

제 3장 사업 수행 결과

1. 신문 기사 정제 결과



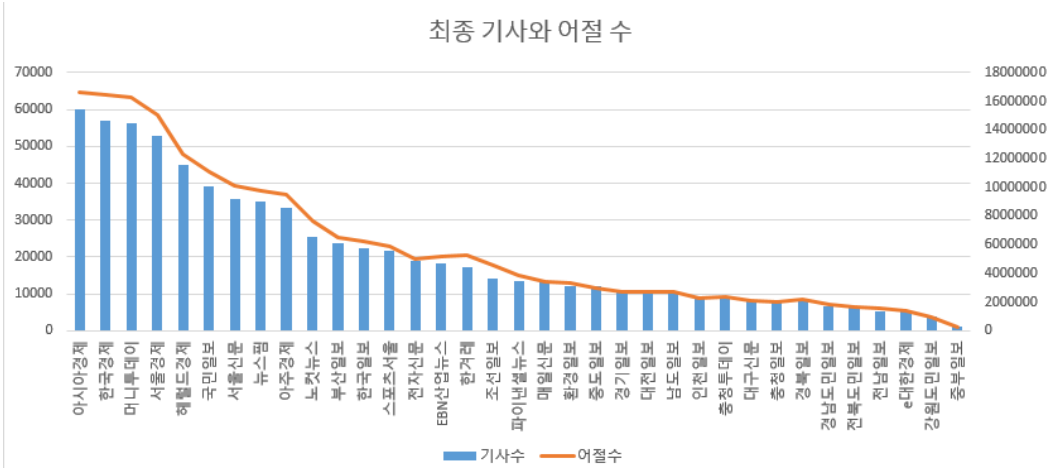
본 사업은 매체 선정부터 문장 교정 말뭉치 작업까지 총 7단계의 프로세스를 거쳐 수행되었다. 최종 정제 완료된 데이터는 730,017건의 기사와 203,585,743개의 어절로 구축되었다.

가장 많은 기사와 어절 수를 구축한 매체는 아시아경제였고, 중부일보는 가장 적은 기사와 어절 수를 구축하였다. 중부일보의 경우 기자 이름 데이터가 없는 기사가 대부분이었기에, 해당 데이터는 사용하지 않았다.

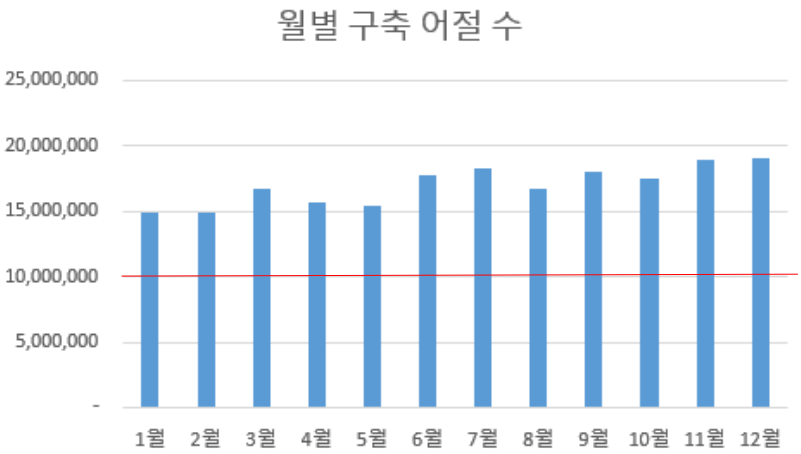
매체명	최초 수집 기사 수	최초 수집 어절 수	정제 수집 기사 수	정제 수집 어절 수
한국경제	192,832	39,554,000	56,918	16,446,248
서울경제	171,187	32,508,732	52,907	15,025,365
아시아경제	215,433	37,333,297	60,090	16,660,172
파이낸셜뉴스	49,871	9,111,932	13,504	3,863,152
아주경제	118,435	27,052,649	33,292	9,460,900
머니투데이	186,723	39,098,450	56,155	16,232,239
스포츠서울	125,433	18,848,466	21,701	5,894,401
노컷뉴스	142,690	25,376,435	25,267	7,669,377
EBN산업뉴스	48,947	9,767,159	18,082	5,127,369
전자신문	61,528	12,098,643	19,069	5,039,492
한겨레	39,935	11,856,729	17,337	5,300,538
국민일보	122,074	25,924,246	39,143	11,045,271
서울신문	120,787	25,641,537	35,742	10,118,015
한국일보	112,864	26,369,560	22,226	6,201,601
경기일보	45,824	8,083,866	10,907	2,681,205
강원도민일보	32,429	4,010,182	3,795	976,544
충청일보	60,837	8,522,189	8,143	1,968,647
매일신문	52,308	9,577,568	13,379	3,427,386
부산일보	85,321	15,652,601	23,631	6,451,426
충청투데이	54,810	8,176,411	9,011	2,396,700
대전일보	48,955	7,892,885	10,802	2,748,143
경북일보	34,184	6,386,017	8,005	2,145,656
뉴스핌	284,167	39,218,664	35,101	9,787,773
중도일보	88,023	13,239,031	12,163	3,001,989
헤럴드경제	142,459	30,128,952	45,065	12,273,735
중부일보	52,951	9,163,532	1,129	278,568
인천일보	48,534	8,641,210	9,126	2,284,191
e대한경제	15,478	3,239,142	5,156	1,413,814
전북도민일보	38,769	6,364,152	6,335	1,607,993
전남일보	30,379	6,220,459	5,337	1,519,470
대구신문	29,716	5,433,994	8,227	2,088,152
경남도민일보	29,492	4,635,045	6,572	1,857,029
남도일보	37,181	6,940,911	10,304	2,720,209
환경일보	43,409	8,360,492	12,232	3,325,166
조선일보	49,864	11,310,071	14,164	4,547,807
총 합	**3,013,829**	**561,739,209**	**730,017**	**203,585,743**

<표 25> 신문 기사 정제 총괄표

월별 1,000만 어절 이상의 데이터를 구축해야 하는 목표를 초과 달성하였다. 월별 평균 약 1,700만 어절의 데이터를 구축하였으며, 총 2억 어절 이상의 말뭉치를 구축하였다. 한 기사당 평균 어절 수는 279어절이다.



<그림 17> 매체별 최종 기사 수와 어절 수



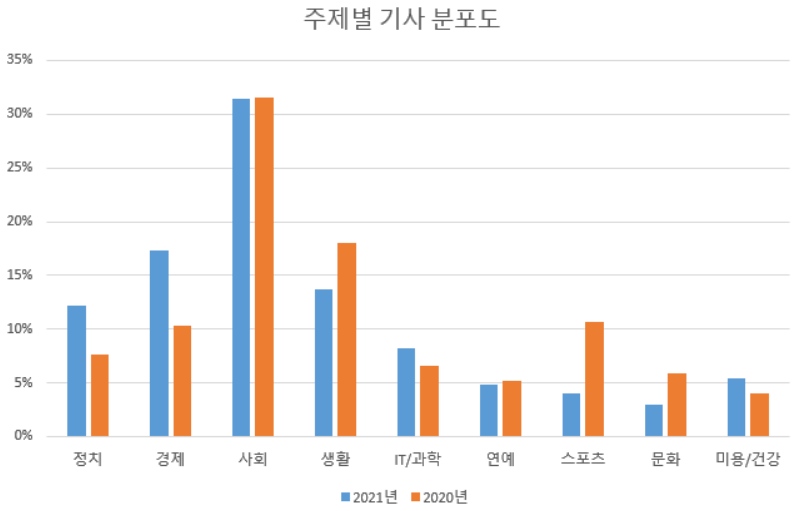
<그림 18> 월별 구축 어절 수 그래프

월	어절 수
1월	14,876,387
2월	14,882,822
3월	16,650,530
4월	15,632,412
5월	15,467,653
6월	17,696,850
7월	18,266,024

월	어절 수
8월	16,727,236
9월	17,971,301
10월	17,451,261
11월	18,875,816
12월	19,087,451
합계	203,585,743

<표 26> 월별 구축 어절 수

1년치의 기사가 다양한 주제로 선정되어야 한다는 과업 내용에 맞추어 구축된 주제별 분포는 아래와 같다. 2020년 주제별 분포 분석을 통해 정치 분야와 경제 분야의 기사 비중이 이전 연도에 비해 높은 것을 알 수 있다.



<그림 19> 주제별 기사 분포도 그래프

주제별	기사 수	어절 수	평균 어절 수
경제	126,282	35,994,232	285
문화	21,645	6,230,591	288
미용/건강	39,566	10,970,792	277
사회	230,132	63,552,358	276
생활	99,771	27,421,095	275
스포츠	29,531	8,075,218	273
연예	34,436	9,733,975	283
정치	89,124	24,980,394	280
IT/과학	59,530	16,627,088	279

<표 27> 주제별 기사 수 및 구축 어절 수

매체 구분	기사 수	어절 수	평균 어절 수
경제일간	323,087	91,375,625	282
스포츠일간	21,701	5,894,401	271
인터넷신문	78,450	22,584,519	287
전국종합일간	128,612	37,213,232	289
전문일간	31,301	8,364,658	267
지역종합일간	146,866	38,153,308	259

<표 28> 매체별 기사 수 및 구축 어절 수

2. 매체별 납품 파일명

말뭉치 유형 구분	매체 및 장르 분류	분석 층위 구분	구축 연도	매체일련 번호	매체명
N	I	RW	21	1	EBN산업뉴스
N	I	RW	21	2	노컷뉴스
N	I	RW	21	3	뉴스핌
N	W	RW	21	1	국민일보
N	W	RW	21	2	서울신문
N	W	RW	21	3	조선일보
N	W	RW	21	4	한겨레
N	W	RW	21	5	한국일보
N	P	RW	21	1	e대한경제
N	P	RW	21	2	머니투데이
N	P	RW	21	3	서울경제
N	P	RW	21	4	스포츠서울
N	P	RW	21	5	아시아경제
N	P	RW	21	6	아주경제
N	P	RW	21	7	전자신문
N	P	RW	21	8	파이낸셜뉴스
N	P	RW	21	9	한국경제
N	P	RW	21	10	헤럴드경제
N	P	RW	21	11	환경일보
N	L	RW	21	1	강원도민일보
N	L	RW	21	2	경기일보
N	L	RW	21	3	경남도민일보
N	L	RW	21	4	경북일보
N	L	RW	21	5	남도일보
N	L	RW	21	6	대구신문
N	L	RW	21	7	대전일보
N	L	RW	21	8	매일신문
N	L	RW	21	9	부산일보
N	L	RW	21	10	인천일보

말뭉치 유형 구분	매체 및 장르 분류	분석 층위 구분	구축 연도	매체일련 번호	매체명
N	L	RW	21	11	전남일보
N	L	RW	21	12	전북도민일보
N	L	RW	21	13	중도일보
N	L	RW	21	14	중부일보
N	L	RW	21	15	충청일보
N	L	RW	21	16	충청투데이

<표 29> 납품 데이터 파일명

<부록1>

국가 언어 자원(말뭉치) 구축 및 활용 저작권 이용 허락 계약서

국가 언어 자원(말뭉치) 구축 및 활용 저작권 이용허락 계약서

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

다 음

제1조 (계약의 목적)
본 계약은 국가 언어 자원(말뭉치) 구축 및 활용을 위한 저작재산권 이용허락과 관련하여 권리자와 이용자 사이의 권리관계를 명확히 하는 것을 목적으로 한다.

제2조 (정의)
본 계약에서 사용하는 용어의 뜻은 다음과 같다.
(1) ‘전체 기사’라 함은 권리자가 제공하는 2020년 1년 동안 생산된 신문 기사 원문 자료를 말한다.
(2) ‘수집 기사’ 라 함은 국립국어원 및 국립국어원이 발주한 용역 사업의 수행자(이하 “과업수행자”라 함)가 ‘전체 기사’에서 수집한 신문 기사 월별 1000만 어절 분량(총 1.2억 어절)에 포함된 기사를 말한다.
(3) ‘대상저작물’ 이라 함은 ‘수집 기사’ 중 국립국어원 및 과업수행자가 말뭉치 구축 대상으로 선정한 1억 어절 분량의 기사 원문을 말한다.
(4) ‘복제·변형물’이라 함은 국립국어원 및 과업수행자가 ‘대상저작물’에 편집 및 삭제, 언어 단위별 분리, 언어적·비언어적 정보 부착 등 처리를 더한 결과물인 원시 및 분석 말뭉치를 말한다.

제3조 (계약의 대상)
본 계약의 이용허락 대상이 되는 권리는 아래의 저작물에 대한 저작재산권 중 본조에 명시한 이용허락 범위로 한다.

저작물: 2020년 1월 1일 ~ 2020년 12월 31일까지(1년 간)의 기사 중 권리자가 저작권 또는 저작권 재이용을 허락할 권리를 보유한 기사
매체명:

저작권 이용 허락 범위

1. 국립국어원 및 과업수행자가 ‘수집기사’, ‘대상저작물’ 및 ‘복제·변형물’을 일정한 형식으로 전자적 기록 매체에 담아 보존하는 일
2. 국립국어원 및 과업수행자가 자모, 음절, 어휘, 어절, 구절, 문장 및 텍스트 단위의 국어 연구와 언어 정보 처리 분야에 응용하기 위해 ‘대상저작물’을 복제·변형(목차·머리말·도표·그림·각주 등의 편집 및 삭제, 언어 단위별 분리, 언어적·비언어적 정보 부착 등)착하여 원시 및 분석 말뭉치로 구축하는 일
3. 국립국어원이 ‘복제·변형물’을 국어 연구와 언어 정보 처리 분야 응용을 위하여 학계·연구기관·산업체 등이 이용할 수 있도록 홈페이지 등을 통해 제공하고, ‘복제·변형물’을 배포하는 일
4. ‘대상저작물 및 그 복제·변형물’을 제공·배포 받은 학계·연구기관·산업체 등이 국어 연구와 언어 정보 처리 분야 응용을 위하여 ’복제·변형물’을 분석 및 처리하여 사용하는 것을 허락하는 일

제4조 (이용허락 기간)

(1) ‘전체 기사’ 및 ‘수집 기사’의 이용허락 기간은 계약체결일부터 2021년 12월 31일까지로 한다.

(2) ‘대상저작물’ 및 ‘복제·변형물’의 이용허락 최소 기간은 계약체결일부터 2032년 12월 31일까지로 한다. 최소 기간 만료 후 권리자 또는 저작자인 언론사가 이용허락 중지 의사를 밝히지 아니하면 이용허락이 1년 단위로 자동 갱신되며, 권리자 또는 저작자인 언론사가 이용허락 중지 의사를 밝히면 그 의사 내용에 따라 이용허락이 중지된다.

제5조 (권리자의 의무)

(1) 권리자는 이용자에게 본 계약서 제3조에 따른 저작재산권을 이용할 권리를 제4조의 기간 동안 비독점적으로 허락한다.

(2) 권리자는 이용자에게 계약 체결일로부터 10일 이내에 ‘대상저작물’의 이용을 위해 필요한 상당한 자료를 인도하여야 한다. 이때 자료를 인도하는 형식과 방법은 부속합의서에 따른다.

(3) 권리자는 ‘대상저작물’에 본 계약 이행에 지장을 주는 제3자의 이용허락권,

질권 등이 존재하는 경우, 이용자에게 그 사실을 사전에 알려야 한다.

제6조 (이용자의 권리 및 의무)

(1) 이용자는 '대상저작물'을 제4조의 이용허락 기간 동안 제3조의 이용 허락을 받은 범위 내에서 비독점적으로 자유롭게 이용할 수 있다.

(2) 이용자는 과업수행자를 통해 별지 이용료를 지급하되 지급방법은 부속합의서로 정한다. 이용허락 기간 자동 갱신에 따른 추가적인 이용료는 발생하지 않는다.

(3) 이용자는 관례적으로 저작자 및 저작재산권자의 성명 등 표시를 허용하는 '대상저작물'을 이용하는 경우, 그 저작자 및 저작재산권자의 성명 등을 표시하여야 한다.

(4) 이용자는 '대상저작물'을 이용함에 있어서 저작인격권을 침해하지 아니한다. 다만, 본 계약의 목적에 따라 '대상저작물'의 본질적인 내용을 변경하지 않는 범위 내에서 변형할 수 있다.

제7조 (확인 및 보증)

(1) 권리자는 이용자에게 다음 각 호의 사항을 확인하고 보증한다.

1. 본 저작권 이용허락 계약을 체결하는 데 필요한 권리 및 권한을 적법하게 보유하고 있다는 것
2. '대상저작물'에 대하여 이용자에게 사전에 알린 제3자의 권리 외에는 이용자의 이용을 제한할 수 있는 부담이 더 이상 존재하지 아니한다는 것

(2) 이용자는 권리자에게 다음 각 호의 사항을 확인하고 보증한다.

1. '대상저작물' 및 '복제·변형물'에 적용된 이용허락 조건에 의해서만 재이용을 허락할 것
2. '대상저작물' 및 '복제·변형물'을 제3자의 명예권을 비롯한 인격적 권리를 침해하는 방식으로 이용하지 아니할 것
3. '대상저작물' 및 '복제·변형물'의 제공·배포 시 이용허락 조건 및 재배포 금지, 목적 외 사용금지 등 주의사항을 고지할 것

제8조 (계약내용의 변경)

본 계약 내용 중 일부를 변경할 필요가 있는 경우에는 권리자와 이용자의 서면합의에 의하여 변경할 수 있으며, 그 서면합의에서 달리 정함이 없는 한, 변경된 사

항은 그 다음날부터 효력을 가진다.

제9조 (계약의 해지)
(1) 당사자는 천재지변 또는 기타 불가항력으로 계약을 유지할 수 없는 경우에 본 계약을 해지할 수 있다.

(2) 당사자는 상대방이 정당한 이유 없이 본 계약을 위반하는 경우에 상당한 기간을 정하여 상대방에게 그 시정을 최고하고, 상대방이 그 기간이 지나도록 이행하지 아니하는 경우에는 계약을 해지할 수 있다. 다만, 상대방이 명백한 시정 거부의사를 표시하였거나 위반 사항의 성격상 시정이 불가능하다는 것이 명백히 인정되는 경우에는 위와 같은 최고 없이 계약을 해지할 수 있다.

(3) 본 계약에 대한 해지권의 행사는 상대방에 대한 손해배상청구권 행사에 영향을 미치지 아니한다.

제10조 (손해배상)
당사자가 정당한 이유 없이 본 계약을 위반하는 경우, 그로 인하여 상대방에게 발생한 모든 손해를 배상할 책임이 있다. 다만, 제9조 1항의 사유로 본 계약을 이행하지 못한 경우에는 손해배상책임을 면한다.

제11조 (분쟁해결)
(1) 본 계약에서 발생하는 모든 분쟁은 권리자와 이용자가 상호 원만한 합의에 이르도록 노력하여야 하며, 분쟁이 원만히 해결되지 않는 경우에는 소제기에 앞서 한국저작권위원회에 조정을 신청할 수 있다.

(2) 제1항에 따라 해결되지 아니할 때에는 대한민국의 민사소송법 등에 따른 관할법원에서의 소송에 의해 해결토록 한다.

제12조 (비밀유지)
양 당사자는 본 계약의 체결 및 이행과정에서 알게 된 상대방에 관한 정보, 본 계약의 내용을 상대방의 서면에 의한 승낙 없이 제3자에게 공개하여서는 아니 된다. 다만, 계약의 내용을 저작자에게 알리는 경우는 예외로 한다.

제13조 (기타부속합의)
(1) 권리자와 이용자는 본 계약의 내용을 보충하거나, 이 계약에서 정하지 아니

한 사항을 규정하기 위하여 부속합의서를 작성할 수 있다.

(2) 제1항에 따른 부속 합의는 본 계약의 내용과 배치되거나 위반하지 않는 범위 내에서 유효하다.

제14조 (계약의 해석 및 보완)
본 계약서에서 명시되어 있지 아니하거나 해석상 이견이 있을 경우에는 저작권법, 민법 등을 준용하고 사회 통념과 조리에 맞게 해결한다.

제15조 (계약 효력 발생일)
본 계약의 효력은 계약 체결일로부터 발생한다.

년 월 일

권리자 :		이용자 :
성명	(인)	성명 국립국어원장 (인)
주소		주소 서울특별시 강서구 금낭화로 154

<부록2>
데이터 정제 작업 지침

데이터 정제 작업 지침

□ 사용하지 않는 기사의 표시

삭제 기사 구분	내용
저작권 관련 검토 필요 기사	– 연합뉴스발 기사 – 외부 기고가가 작성한 기사(~위원, ~교수, 영화평론가 등) – 명예기자, 객원기자, 시민기자가 작성한 기사 (기자 이름에 명예기자나, 객원기자라고 표기되지 않고, 이름만 나오는 경우에는 그대로 사용함.) – 외국 기사를 번역한 기사 – 다른 매체의 헤드라인을 모아 놓은 기사
구어체 기사	– 대부분이 구어체로 이루어진 기사는 사용하지 않는다, – 구어체의 경우 ~입니다. 예정입니다. 등과 같은 기사는 사용해도 무방하다.
불필요한 정보를 삭제한 후 기사 내용이 짧은 기사	– 불필요한 요소를 삭제하고 남은 기사가 짧은 경우, 기사를 사용하지 않는다.
불완전하게 종료되는 기사	– 기사가 불완전하게 종료된 경우, 내용을 유추하여 명확하게 완성을 시킬 수 있는 경우에는 사용하지만, 유추가 어려운 수준으로 불완전한 기사는 사용하지 않는다.
한글이 조합형으로 깨진 기사	– 조합형으로 한글이 이루어져 있으나, 수정이 가능한 범위면 사용하고, 수정 대상이 많으면 사용하지 않는다.
명확한 광고 기사	– 광고의 경우 기사 내용 안에 명확히 광고라고 표기하는 경우에는 사용하지 않는다.
단순 기사	– 날씨, 승진, 부고, 운세, 전보, 임용, 스포츠 스코어, 여론조사결과, 출구조사결과, 어록 모음 – 매체의 헤드라인을 모아 놓은 기사

□ 기사 본문 내 불필요한 정보의 삭제

예시 안 굵은 글꼴이 삭제 대상

삭제 정보	예시
표, 그림, 그래프 등의 캡션 정보는 삭제함	(사진제공=건국대학교) (사진제공=SBA) 표〉 공정위 망 이용대가 불공정 조사 쟁점 사진제공=tvN 〈표〉 한상혁 후보자 주요 ICT 정책 현안 입장 사진=CJ엔터테인먼트 제공 ▲영상제공= [그래픽] 사진=FNC엔터테인먼트 제공 <그래픽> 출처: 라디오타임스 / 굿모닝브리튼,사진=스타쉽 일러스트| 제공 화면 캡처
기자의 이름, ID 등의 정보는 제거함	[서울경제TV=배요한기자] 동양네트웍스(030790)가 강세다. 글·사진=양ㅇㅇ 기자 ----@kmib.co.kr 김OO ----@kmib.co.kr.사진=인터파크 제공
'Copyright©' 등 저작권 관련 내용은 제거함	<저작권자(c) 연합뉴스, 무단 전재-재배포 금지> ondol@yna.co.kr/2019-08-29 10:14:05/ <저작권자 © 1980-2019 ㈜연합뉴스. 무단 전재 재배포 금지.>
전문	대검찰청 정책관 등 중간간부들이 26일 윤석열 검찰총장 직무배제가 부당하다며 추미애 법무부 장관에게 재고를 요청하는 성명을 냈다. 손준성 수사정보정책관, 이창수 대검 대변인 등 대검 중간간부 27명은 이날 검찰게시판에 '대검찰청 중간간부들의 입장'이라는 제목의 글을 올렸다. 이들은 "검찰총장에 대한 직무집행정지는 적법절차를 따르지 않고, 충분한 직상확인 과정도 없이 이뤄진 것으로 위법부당하다"며 "이는 검찰의 중립성은 물론이고 검찰개혁, 나아가 소중하게 지켜온 대한민국의 법치주의 원칙을 크게 훼손하는 것"이라고 강하게 비판했다. 이어 "검찰이 헌법과 법률에 따라 책임과 직무를 다 할 수 있도록 (윤 총장에 대한) 징계청구와 직무집행 정지를 재고해주실 것을 간곡히 요청드린다"고 적었다. 아래는 대검 중간간부들의 성명서 전문이다. <대검찰청 중간 간부들의 입장> ○ 코로나19 등으로 인한 국가적 위기 상황 속에서 검찰과 관련된 각종 논란으로 국민들께 심려를 끼치고 있어 송구스럽게 생각합니다. ○ 검찰이 변화해야 한다는 국민의 뜻에 부응하기 위해 노력하고 있으나, 여전히 많이 부족하다는 것을 잘 알고 있습니다. ○ 다만, 최근 검찰을 둘러싸고 진행되고 있는 상황들에 대해 침묵하는 것은 공직자로서 올바른 자세가 아니라는 데에 뜻을 함께 한 대검찰청 중간 간부들은 2020. 11. 26. 아래와 같이 의견을 모았습니다. ○ 검찰공무원은 범죄로부터 우리 국민들을 보호하고, 온전한 법치주의 실현을 통해 자유롭고 안정된 민주사회를 구현해야 할 사명이 있습니다. ○ 검찰총장에 대한 11. 24. 징계청구와 직무집행정지는 적법절차를 따르지 않고, 충분한 진상확인 과정도 없이 이루어진 것으로 위법, 부당합니다. ○ 이는 검찰의 정치적 중립성은 물론이고, 검찰개혁, 나아가, 소중하게 지켜온 대한민국의 법치주의 원칙을 크게 훼손하는 것이기도 합니다.

	○ 검찰이 헌법과 법률에 따라 책임과 직무를 다 할 수 있도록 징계청구와 직무집행 정지를 재고해 주실 것을 법무부장관께 간곡하게 요청드립니다. ○ 저희들도 국민과 함께 하는 검찰공무원으로서 본연의 임무를 충실히 수행해 나가겠습니다. 2020. 11. 26. 손준성 이정봉 최성국 이창수 박기동 강범구 전무곤 고필형 구승모 임승철 이만흠 반종욱 최창민 진현일 박혁수 김용자 김 우 백수진 한기식 김승언 김종현 신준호 추혜윤 장준호 손진욱 김연아 정태원 배ㅇㅇ 기자 ----@hani.co.kr
문장으로 볼 수 없는 정보의 나열 등	■ 인천·경기지역 시급 현안 '인천·경기에서 가장 우선적으로 해결해야 할 사안이 무엇이라고 생각하느냐'는 질문에 '일자리 창출'이 28.0%로 가장 높았다. 이어 '지역간 균형발전'이 19.1%, '부동산 가격 안정화'가 15.0%, '광역교통망 구축' 13.6%, '미세먼지 대책마련'이 10.7%, '수도권 규제완화'가 3.6% 순이다. '기타'가 7.5%, '잘 모름'이 2.4%다. 지역별로는 계양·부평권과 남동·연수·미추홀권은 일자리 창출이 각각 32.0%와 30.0%로 가장 높은 반면, 동·서·중구·강화·옹진권은 지역간 균형발전이 22.9%로 가장 높았다. 연령대별로 대부분은 일자리 창출을 시급한 현안으로 꼽았지만, 유일하게 40~49세에서만 지역간 균형발전이 가장 높았다. ㅇㅇㅇ기자 어떻게 조사했나 이번 조사는 경기일보의 의뢰로 조원씨앤아이가 2019년 12월28일(土)부터 30일(月)까지 사흘간, 인천광역시 거주 만19세 이상 남녀를 대상으로 ARS 여론조사(유선전화 RDD 12%+통신사 제공 휴대전화 가상번호 88% 방식, 성,연령,지역별 비례할당무작위추출)를 실시한 결과이며,표본수는 805명(총 통화시도 17,366명, 응답률 4.6%), 표본오차는 95% 신뢰수준에 ±3.5%p임. 그 밖의 사항은 중앙선거여론조사심의위원회 홈페이지 참조 ※오차보정방법 : [림가중] 성별, 연령별, 지역별 가중값 부여(2019년 11월말 행정안전부 발표 주민등록인구기준)
	지난 25일 서울 성동구에서 23년째 PC방을 운영하고 있는 이모씨(47)는 올해 추석 계획을 묻는 기자의 지문에 "가게를 지키는 일"이라고 답했다. 코로나19로 적자가 너무 심해져 하루라도 가게를 비울 수 없다는 것이다.(관련 기사 ☞ "나라도 돈 벌겠다" 중2 아들말에…PC방 사장님 3일째 집에 못갔다)
	이후 2020대한민국지속가능혁신리더대상 조직위 운영 사무국으로 이메일 또는 우편(서울시 중구 청계천로 11(서린동, 청계한국빌딩 16층))을 통해 6월 30(화)(오후 6시까지 도착분에 한함)까지 제출하면 된다. 응모 자격은 정부 상훈 관련법에 부합하는 지자체·기관·법인 및 단체·개인으로 접수된 신청서류는 반환되지 않는다. 평가는 1차 서류심사, 2차 실사를 포함한 심층심사, 3차 최종평가를 거쳐 최종 수상자를 선정한다. 자세한 내용은 아래 개요를 참고하시기 바랍니다. 대한민국을 이끄는 혁신리더들의 많은 참여 바랍니다. [2020 대한민국지속가능혁신리더대상 개요] 주 최 : 2020 대한민국지속가능혁신리더대상 조직위원회 주 관 : 머니투데이, 더리더 접수마감 : 2020년 6월 30(화)

	접수문의 : 02-724-0952(머니투데이 더리더) 접 수 처 : 이메일(awards@mt.co.kr) 시상일시 : 2020년 7월 중 시상장소 : 여의도 켄싱턴호텔 신청대상 : 정치·사회·경제·교육·체육·문화·예술·환경 등 각 분야의 지속적인 혁신 공로가 인정되는 지자체 및 우수 기관, 단체, 개인리더 등 신청양식 : 더리더 홈페이지 우측 상단 배너 클릭, 기사하단 신청서 다운로드에서 클릭 후 내려받기 가능
	특히 생체리듬으로 알려진 '써카디안(circadian) 리듬'은 간헐적 단식에서 아주 중요한 요소다. 써카디안 리듬과 중추시계, 말초시계가 일치돼야 건강한 일상이 가능하기 때문. 햇볕이 비출 때 일어나고 일정한 시간에 건강한 음식을 취하며 해가 지면 잠자리에 드는 '원시 인류'의 생활을 따라야 한다고 저자는 강조한다. ◇호르메시스와 간헐적 단식=박용우 지음. 블루페가수스 펴냄. 276쪽/1만5000원.
	주목받은 신인에게 주는 '넥스트 리더'는 위클리, 크래비티, 엔하이픈에게 돌아갔다. {IMG:2}다음은 '2020 TMA' 수상자(작) 명단. ▲ 대상 : 방탄소년단 ▲ 리스너스 초이스 : 방탄소년단 ▲ 월드와이드 아이콘 : 세븐틴, 방탄소년단 ▲ TMA 인기상 : 슈퍼주니어 ▲ 올해의 아티스트 : 마마무&화사, 강다니엘, 방탄소년단, 갓세븐, 트와이스, 뉴이스트, 아이즈원, 몬스타엑스, 세븐틴, 슈퍼주니어 ▲ 글로벌 핫티스트 : 스트레이 키즈, (여자)아이들, 에이티즈, 더보이즈 ▲ 베스트 퍼포머 : 있지, 투모로우바이투게더, 제시 ▲ 넥스트 리더 : 위클리, 크래비티, 엔하이픈 ▲ 팬앤스타 최다 득표상(가수) : 슈퍼주니어 ▲ 팬앤스타 최다 득표상(개인) : 황치열 ▲ 팬앤스타 초이스상(가수) : 슈퍼주니어 ▲ 팬앤스타 초이스상(개인) : 황치열
	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. 위의 내용은 단순 정보이지만 문장으로 볼 수 있기에 사용 한다.
	유족으로는 딸 이희경씨, 동생 은화(전 이화여대 교수)·효숙·성숙씨, 올케 이부자씨가 있다. 여성단체들은 여성장으로 고인을 배웅하기로 했다. 빈소는 창원경상대병원 장례식장 VIP 1호실에 마련됐다. (055)214-1910. 김정화 기자 clean@seoul.co.kr
	경찰은 유서 내용 등을 토대로 A 소방사가 극단적 선택을 한 것으로 보고 유족 등을 상대로 정확한 사망원인을 조사하고 있다.

	※ 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. 이보희 기자 boh2@seoul.co.kr
	■이부영은 누구인가 이부영 전 열린우리당 의장은 1980년대를 대표하는 재야 민주투사이자 정치 원로다. 동아일보 해직 언론인 출신으로 민주화 투쟁을 하다 수차례 옥고를 치렀다. 1990년에 3당 합당에 반대해 만든 민주당을 통해 정계에 입문한 뒤 14~16대 서울 강동갑에서 3선을 했다. 1995년 당시 김대중 총재가 이끄는 새정치국민회의에 합류하지 않고 통합민주당에 남아 있다가 합당 후 한나라당에서 원내총무, 부총재 등을 지냈다. 2004년 17대 총선에서 과반 의석인 152석을 차지했던 열린우리당 의장을 맡았다. 2015년 정계를 은퇴했고, 지난해부터는 자유언론실천재단 이사장으로서 올바른 언론 환경 조성에 노력하고 있다. ▲1942년 서울 출생 ▲서울대 정치학과 ▲동아일보 기자 ▲14~16대 국회의원 ▲한나라당 부총재 ▲열린우리당 의장 ▲동아시아평화국제회의 조직위원장 ▲자유언론실천재단 이사장
	몸매관리 비법으로 밀크어트를 소개한 그녀는 자신만의 다이어트 비법인 건강음료를 공개해 화제가 되고 있다. 손쉽게 만들 수 있는 오영주 표 밀크어트 건강음료 3가지를 소개한다. ■ 아보카도 스무디 <재료> 우유 200ml, 아보카도 1/2개, 바나나 1개 <만드는 방법> 아보카도를 반으로 갈라 씨와 껍질을 제거하고, 우유, 아보카도, 바나나 등 모든 재료를 믹서에 넣고 갈아주면 완성이다. ■ 고구마라떼 <재료> 우유 300ml, 삶은 고구마 1개 <만드는 방법> 삶은 고구마는 껍질을 벗긴 뒤 우유와 함께 믹서에 넣고 갈아준다. 만약 고구마라떼를 마실 때 목 넘김을 부드럽게 하고 싶다면 고구마를 잘게 잘라 믹서에 넣으면 된다. 고구마를 대신해 블루베리, 바나나, 딸기 등 과일로 대체 가능하며, 기호에 따라 꿀이나 시럽으로 당도를 조절한다.
기사와 상관 없는 광고 혹은 반복되는 문장, 오류의 경우	아래 기사는 본 기사와 상관 없이 다른 기사의 내용이 오류로 잘못 들어간 경우이다. 본 기사와 상관 없는 내용은 삭제한다. ========== 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.

	한편 질병관리본부 중앙방역대책본부는 4일 오전 0시 기준 코로나19 확진자가 5328명이라고 밝혔다. 전날 오전 0시와 비교하면 516명이 늘었다. 사망자는 전날 하루 사이에 4명이 추가돼 총 32명이다. 격리 해제된 확진자는 7명이 늘어 41명이다. [출처: 서울신문에서 제공하는 기사입니다.] https://en.seoul.co.kr/news/newsView.php?id=20200304500092#csidx4dab120876716f7a9506745d61f1391

<부록3>
말뭉치 종류별 구축 예시

❍ 각 말뭉치 종류별 비교 예시(전체 기사가 아닌 내용 일부 발췌)

신문기사말뭉치	김 위원장은 이 대표를 만나 축하 인사를 건내며 “원만하게 정치를 풀어갈 수 있도록 노력해달라”고 요청했고, 이 대표 역시 “늘 지도해 주셨듯 이번에는 더 많은 지도를 해달라”고 답했다. 이 대표는 기자 시절 김종인 위원장과 취재원과 기자 사이의 관계로 만난 바 있다. 이 대표는 이날 오전 tbs 라디오 ‘김어준의 뉴스공장’에 출연해 “전당대회 바로 다음날 전화를 드려 신고를 하고 김 위원장이 추진하는 일이 잘하는 것 같다. 잘 실현되게 돕겠다”고 말한 바 있다.
문장 말뭉치	<s>김 위원장은 이 대표를 만나 축하 인사를 건내며 “원만하게 정치를 풀어갈 수 있도록 노력해달라”고 요청했고, 이 대표 역시 “늘 지도해주셨듯 이번에는 더 많은 지도를 해달라”고 답했다.</s> <s>이 대표는 기자 시절 김종인 위원장과 취재원과 기자 사이의 관계로 만난 바 있다.</s> <s>이 대표는 이날 오전 tbs 라디오 ‘김어준의 뉴스공장’에 출연해 “전당대회 바로 다음날 전화를 드려 신고를 하고 김 위원장이 추진하는 일이 잘하는 것 같다. 잘 실현되게 돕겠다”고 말한 바 있다.</s>
문장 교정 말뭉치	<s>김 위원장은 이 대표를 만나 축하 인사를 건네며 “원만하게 정치를 풀어갈 수 있도록 노력해달라”고 요청했고, 이 대표 역시 “늘 지도해주셨듯 이번에는 더 많은 지도를 해달라”고 답했다.</s> <s>이 대표는 기자 시절 김종인 위원장과 취재원과 기자 사이의 관계로 만난 바 있다.</s> <s>이 대표는 이날 오전 tbs 라디오 ‘김어준의 뉴스공장’에 출연해 “전당대회 바로 다음날 전화를 드려 신고를 하고 김 위원장이 추진하는 일이 잘하는 것 같다. 잘 실현되게 돕겠다”고 말한 바 있다.</s>
내용	문장말뭉치 문단 분할. 피인용문 내 문단분할은 진행하지 않음. 맞춤법 수정

신문기사말뭉치	심사위원장을 맡은 송하엽 교수는 “당선작은 상부차로를 축소하고 선형공원을 제시하며 지하보도와 입체적인 연결을 제시하는 안이다. 지하에 자연광을 도입하며 균일하게 만든 아치구조 아래 길다란 책 서고를 만든 점이 인상적이며 실제 동선으로 사용되는 점도 시민친화적”이라고 총평했다.
문장 말뭉치	<s>심사위원장을 맡은 송하엽 교수는 “당선작은 상부차로를 축소하고 선형공원을 제시하며 지하보도와 입체적인 연결을 제시하는 안이다. 지하에 자연광을 도입하며 균일하게 만든 아치구조 아래 길다란 책 서고를 만든 점이 인상적이며 실제 동선으로 사용되는 점도 시민친화적”이라고 총평했다.</s>
문장 교정 말뭉치	<s>심사위원장을 맡은 송하엽 교수는 “당선작은 상부차로를 축소하고 선형공원을 제시하며 지하보도와 입체적인 연결을 제시하는 안이다. 지하에 자연광을 도입하며 균일하게 만든 아치구조 아래 기다란 책 서고를 만든 점이 인상적이며 실제 동선으로 사용되는 점도 시민친화적”이라고 총평했다.</s>
내용	문장말뭉치 문단 분할. 피인용문 내 문단분할은 진행하지 않음. 맞춤법 수정

신문기사말뭉치	JTBC 새 월화드라마 ‘야식남녀’에 셰프 정일우는 고단한 하루 끝에 팬시리 허기가 지는 날, 사람들이 찾는 심야식당 ‘비스트로(bistro)’에 시청자들을 초대했다. 셰프 박진성(정일우)의 ‘비스트로’는 시계바늘이 밤10시를 가리키면 작은 간판에 불이 들어오며 골목을 따스하게 비춘다. 야식을 먹는다는 건, 어쩌면 진짜로 배가 고파서라기 보다, 마음이 헛헛해서일지도 모른다.
문장 말뭉치	<s>JTBC 새 월화드라마 ‘야식남녀’에 셰프 정일우는 고단한 하루 끝에 팬시리 허기가 지는 날, 사람들이 찾는 심야식당 ‘비스트로(bistro)’에 시청자들을 초대했다.</s> <s>셰프 박진성(정일우)의 ‘비스트로’는 시계바늘이 밤10시를 가리키면 작은 간판에 불이 들어오며 골목을 따스하게 비춘다.</s> <s>야식을 먹는다는 건, 어쩌면 진짜로 배가 고파서라기 보다, 마음이 헛헛해서일지도 모른다.</s>
문장 교정 말뭉치	<s>JTBC 새 월화드라마 ‘야식남녀’에 셰프 정일우는 고단한 하루 끝에 팬스레 허기가 지는 날, 사람들이 찾는 심야식당 ‘비스트로(bistro)’에 시청자들을 초대했다.</s> <s>셰프 박진성(정일우)의 ‘비스트로’는 시곗바늘이 밤 10시를 가리키면 작은 간판에 불이 들어오며 골목을 따스하게 비춘다.</s> <s>야식을 먹는다는 건, 어쩌면 진짜로 배가 고파서라기보다, 마음이 헛헛해서일지도 모른다.</s>
내용	문장말뭉치 문단 분할. 맞춤법, 띄어쓰기 수정

신문기사말뭉치	지엔티파마는 혈관 재개통 치료를 받은 뇌졸중 환자에서 장애를 개선하고 부작용을 줄이는 효과로 올해 초 미국 특허청에 우선권 특허를 신청했다. 지엔티파마는 혈전제거수술이나 혈전용해제로 재개통 치료를 받은 뇌졸중 환자에서 넬로넴다즈의 약효와 안전성이 확인됨에 따라 뇌세포보호약물들이 뇌졸중 치료의 새로운 장을 열 것으로 기대하고 있다.
문장 말뭉치	<s>지엔티파마는 혈관 재개통 치료를 받은 뇌졸중 환자에서 장애를 개선하고 부작용을 줄이는 효과로 올해 초 미국 특허청에 우선권 특허를 신청했다.</s> <s>지엔티파마는 혈전제거수술이나 혈전용해제로 재개통 치료를 받은 뇌졸중 환자에서 넬로넴다즈의 약효와 안전성이 확인됨에 따라 뇌세포보호약물들이 뇌졸중 치료의 새로운 장을 열 것으로 기대하고 있다.</s>
문장 교정 말뭉치	<s>지엔티파마는 혈관 재개통 치료를 받은 뇌졸중 환자에서 장애를 개선하고 부작용을 줄이는 효과로 올해 초 미국 특허청에 우선권 특허를 신청했다.</s> <s>지엔티파마는 혈전제거수술이나 혈전용해제로 재개통 치료를 받은 뇌졸중 환자에서 넬로넴다즈의 약효와 안전성이 확인됨에 따라 뇌세포보호약물들이 뇌졸중 치료의 새로운 장을 열 것으로 기대하고 있다.</s>
내용	문장말뭉치 문단 분할. 맞춤법 수정

신문기사말뭉치	1년 가까이 당을 이끌게 됐지만 김 내정자 앞에 놓인 과제도 적지 않다. 내년 재·보궐 선거에서 패배한다면 그 책임을 고스란히 짊어지게 된다. 물론 그 때까지 임기여서 어짜피 물러날 예정이지만, 선거 승리 땐 임기 연장 문제가 자연스레 거론될 가능성이 크다.
문장 말뭉치	<s>1년 가까이 당을 이끌게 됐지만 김 내정자 앞에 놓인 과제도 적지 않다.</s> <s>내년 재·보궐 선거에서 패배한다면 그 책임을 고스란히 짊어지게 된다.</s> <s>물론 그 때까지 임기여서 어짜피 물러날 예정이지만, 선거 승리 땐 임기 연장 문제가 자연스레 거론될 가능성이 크다.</s>
문장 교정 말뭉치	<s>1년 가까이 당을 이끌게 됐지만 김 내정자 앞에 놓인 과제도 적지 않다.</s> <s>내년 재·보궐 선거에서 패배한다면 그 책임을 고스란히 짊어지게 된다.</s> <s>물론 그 때까지 임기여서 어차피 물러날 예정이지만, 선거 승리 땐 임기 연장 문제가 자연스레 거론될 가능성이 크다.</s>
내용	문장말뭉치 문단 분할. 맞춤법 수정

신문기사말뭉치	노동당 기관지 노동신문은 26일 “개성시에서 악성 비루스(바이러스)에 감염된 것으로 의심되는 월남 도주자가 3년 만에 불법적으로 분계선을 넘어 지난 7월19일 귀향하는 비상사건이 발생하였다”면서 “당 중앙위원회 정치국은 개성시에 치명적이며 파괴적인 재앙을 초래할 수 있는 위험이 조성된 것과 관련하여 25일 당 중앙위원회 본부청사에서 비상확대회의를 긴급소집하였다”고 보도했다.
문장 말뭉치	<s>노동당 기관지 노동신문은 26일 “개성시에서 악성 비루스(바이러스)에 감염된 것으로 의심되는 월남 도주자가 3년 만에 불법적으로 분계선을 넘어 지난 7월19일 귀향하는 비상사건이 발생하였다”면서 “당 중앙위원회 정치국은 개성시에 치명적이며 파괴적인 재앙을 초래할 수 있는 위험이 조성된 것과 관련하여 25일 당 중앙위원회 본부청사에서 비상확대회의를 긴급소집하였다”고 보도했다.</s>
문장 교정 말뭉치	<s>노동당 기관지 노동신문은 26일 “개성시에서 악성 비루스(바이러스)에 감염된 것으로 의심되는 월남 도주자가 3년 만에 불법적으로 분계선을 넘어 지난 7월 19일 귀향하는 비상사건이 발생하였다”면서 “당 중앙위원회 정치국은 개성시에 치명적이며 파괴적인 재앙을 초래할 수 있는 위험이 조성된 것과 관련하여 25일 당 중앙위원회 본부청사에서 비상확대회의를 긴급소집하였다”고 보도했다.</s>
내용	문장말뭉치 문단 분할. 문장전각/반각문자 3 치환. 띄어쓰기 수정

신문기사말뭉치	사전증여신탁의 운용 상품으로는 상장지수펀드(ETF)를 활용해 지수, 채권, 금을 포함한 대체자산 등에 분산 투자하는 자산배분형 상품으로, '콴텍'의 위험관리 기술력을 탑재해 다른 자산배분형 상품 대비 안정성에 중점을 두어 장기 투자에 적합하게 설계됐다. 콴텍은 금융위원회가 주관하는 로보어드바이저 테스트베드에서 역대 최다 알고리즘을 보유한 업체로 금융권과의 협업을 확대하고 있으며 **항후에는** 손님이 직접 금 현물, ETF 등을 직접 운용 지시 가능하도록 운용의 폭을 넓힐 예정이다.
문장 말뭉치	<s>사전증여신탁의 운용 상품으로는 상장지수펀드(ETF)를 활용해 지수, 채권, 금을 포함한 대체자산 등에 분산 투자하는 자산배분형 상품으로, '콴텍'의 위험관리 기술력을 탑재해 다른 자산배분형 상품 대비 안정성에 중점을 두어 장기 투자에 적합하게 설계됐다.</s> <s>콴텍은 금융위원회가 주관하는 로보어드바이저 테스트베드에서 역대 최다 알고리즘을 보유한 업체로 금융권과의 협업을 확대하고 있으며 **항후에는** 손님이 직접 금 현물, ETF 등을 직접 운용 지시 가능하도록 운용의 폭을 넓힐 예정이다.</s>
문장 교정 말뭉치	<se>사전증여신탁 운용 상품으로는 ETF를 활용하여 지수·채권·금을 포함한 대체자산 등에 분산 투자하는 자산배분형 상품으로, '콴텍'의 위험관리 기술력을 탑재하여 타 자산배분형 상품 대비 안정성에 중점을 두어 장기 투자에 적합하게 설계됐다.</s> <se>콴텍은 금융위원회가 주관하는 로보어드바이저 테스트베드에서 역대 최다 알고리즘을 보유한 업체로 금융권과의 협업을 확대 중으로 **향후에는** 손님이 직접 금 현물, ETF 등을 직접 운용 지시 가능하도록 운용의 폭을 넓힐 예정이다.</s>
내용	문장말뭉치 문단 분할. 맞춤법 수정

신문기사말뭉치	농산물 생산자와 판매자, 소비자 모두가 필요한 정보를 저장하고 투명하게 공유할 수 있다. 과기부는 “정보공유와 증빙, 검수 작업이 간소화됐고 학교급식에 적용하거나 온라인 판매점포를 여는 등 서비스 창출이 가능할 것”이라고강조했다. 자세한 내용은 이달 안으로 한국정보화진흥원(NIA) 홈페이지(www.nia.or.kr)를 통해 확인할 수 있다. 송경희 과기부 소프트웨어정책관은 “민간소프트웨어시장 확대를 위해 공공 부문이 먼저 민간의 서비스 개발 수요를 선제적으로 반영해 플랫폼을 구축했다”며 “이를 기반으로 민간 서비스 창출과 연계될 수 있도록 정책적으로 지원하겠다”고말했다.
문장 말뭉치	<s>농산물 생산자와 판매자, 소비자 모두가 필요한 정보를 저장하고 투명하게 공유할 수 있다.</s> <s>과기부는 “정보공유와 증빙, 검수 작업이 간소화됐고 학교급식에 적용하거나 온라인 판매점포를 여는 등 서비스 창출이 가능할 것”이라고강조했다.</s> <s>자세한 내용은 이달 안으로 한국정보화진흥원(NIA) 홈페이지(www.nia.or.kr)를 통해 확인할 수 있다.</s> <s>송경희 과기부 소프트웨어정책관은 “민간소프트웨어시장 확대를 위해 공공 부문이 먼저 민간의 서비스 개발 수요를 선제적으로 반영해 플랫폼을 구축했다”며 “이를 기반으로 민간 서비스 창출과 연계될 수 있도록 정책적으로 지원하겠다”고말했다.</s>
문장 교정 말뭉치	<s>농산물 생산자와 판매자, 소비자 모두가 필요한 정보를 저장하고 투명하게 공유할 수 있다.</se><se>과기부는 “정보공유와 증빙, 검수 작업이 간소화됐고 학교급식에 적용하거나 온라인 판매점포를 여는 등 서비스 창출이 가능할 것”이라고 강조했다.</s> <s>자세한 내용은 이달 안으로 한국정보화진흥원(NIA) 홈페이지(www.nia.or.kr)를 통해 확인할 수 있다.</s> <s>송경희 과기부 소프트웨어정책관은 “민간소프트웨어시장 확대를 위해 공공 부문이 먼저 민간의 서비스 개발 수요를 선제적으로 반영해 플랫폼을 구축했다”며 “이를 기반으로 민간 서비스 창출과 연계될 수 있도록 정책적으로 지원하겠다”고 말했다.</s>
내용	문장말뭉치 문단 분할. 띄어쓰기 수정

〈Abstract〉

Collection and refinement of the original text of newspaper articles data 2021

It has been three years since this project has been carried out to address copyright issues through the collection of original texts of newspaper articles from various fields. This project is also related to preparing for the fourth industrial revolution by establishing sophisticated corpora to improve artificial intelligence (AI) technology and to apply the data to academic research.

Although data is crucial when it comes to training AI, it is difficult for individuals companies and even the academia to secure large amounts of corpora for training. Against this backdrop, the National Institute of Korean Language (NIKL) has collected newspaper articles to establish and provide corpora for the public to use.

In 2020, not only did the NIKL collect original texts of newspaper articles to build corpora that reflected the contemporary Korean language, but it also secured the right to access the corpora for the development of industrial and academic technology and research.

The project can be broken down into the following work scopes: the collection of data from original texts of newspaper articles (more than 10 million words per month), the resolution of copyright-related issues through copyright agreements, the removal and refining of duplicate articles, the establishment of raw corpora from newspaper articles, and the list-up of metadata for each article.

For media selection, a total of 35 media channels were selected after thorough discussion with two organizations—the Korea Press Foundation and Chosun Ilbo. The contract and appendices were written and copyright issues were addressed by notarizing the terms and conditions to the two organizations.

As for building raw corpora, it was suggested that three sets of copora be built to expand useage: corpora of newspaper articles using the existing method of collection, corpora of paragraphs divided into sentences for AI training and corpora of proofread sentences.

The minimum unit of corpora for newspaper articles that have been released so far is paragraphs. However, most AIs are trained by using sentences. This is especially true for morphological analysis and machine translation where the base unit is sentences. This is why a set of corpora was formed by dividing each

paragraph into sentences. Moreover, spelling and spacing errors in a text interfere with the AI training which is why the corpora of proofread sentences was additionally established to reduce major errors that may interfere with the AI's learning process.

From 35 participating media channels, a total of 3,013,829 articles and 561,739,209 words were secured.

The characteristics of each medium were distinguished and any omitted data was identified. During this process, it was found that one particular medium had omitted letters from the raw data of the Korea Press Foundation.

During the the first stage of the data refinement process, similar articles, data that was more or less than a specific number of words and data with many errors were sifted and removed from the group.

During the second stage of the data refinement process, people read the articles and deleted several parts by hand including parts that were not considered as sentences or unnecessary to the body, and articles that could not be included due to the possibility of copyright issues. After this filtering process, punctuation used in the articles including double quotation marks and single quotation marks, and the Chinese characters (F900-FAFF) in the CJK Compatibility Ideographs were all standardized.

In the process of building the corpora for sentences, paragraphs in the corpora from newspaper articles were divided into sentences based on the location of end marks. The end marks inside quotations were not divided. Thus, one paragraph was split into an average of approximately 1.7 sentences. Moreover, punctuation marks such as interpuncts and spaces were standardized as well.

The corpora of proofread sentences are data comprised of sentences that went through a refining process of obvious spacing and spelling errors. Most of the errors that were corrected were ones that appeared frequently throughout all of the data. It is projected that this data will be useful for training AI if it is coupled with the corpora comprised of divided sentences.

A total of 730,017 articles amounting to 203,585,743 words were finally selected.

This data was classified into nine different topics by the NIKL. A list of metadata requested by the NIKL including the number of words, dates, titles of the articles, etc. was also prepared.

The final corpus file was sent in the format of a JSON file.

Thanks to the establishment of this large-scale corpora that is free of copyright

issues and available to the public, it is projected that the corpora will be applicable in various fields such as the development of AI technology and academic research.

Key words: corpus from newspaper, artificial intelligence,
newspaper articles, AI training data, contemporary Korean

<기획·연구>

국립국어원 이승재 언어정보과장

국립국어원 이현주 학예연구관

국립국어원 황용주 학예연구관

<사업 참여자>

사업 책임자 윤종웅((주)윤즈정보개발 소장)

사업 참여자 강성준((주)윤즈정보개발 연구원)

김보희((주)윤즈정보개발 연구원)

김하은((주)윤즈정보개발 연구원)

남가윤((주)윤즈정보개발 연구원)

박지영((주)윤즈정보개발 연구원)

서경찬((주)윤즈정보개발 책임연구원)

안소연((주)윤즈정보개발 연구원)

윤여민((주)윤즈정보개발 연구원)

이승철((주)윤즈정보개발 수석연구원)

이재용((주)윤즈정보개발 연구원)

지의선((주)윤즈정보개발 연구원)

최원수((주)윤즈정보개발 연구원)

발행인: 국립국어원장

발행처: 국립국어원

서울시 강서구 금낭화로 154

전화 02-2669-9775, 전송 02-2669-9727

인쇄일: 2021년 10월 7일

발행일: 2021년 10월 7일

인 쇄: 다큐팩토리

※ “이 책은 국립국어원의 용역비로 수행한 ‘2021년 신문 기사 원문 자료 수집 및 정제’ 사업의 결과물을 발간한 것입니다.”