국립국어원 2019-01-05

발 간 등 록 번 호
11-1371028-000762-01

웹 말뭉치 구축

사업 책임자 | 이영희

국립국어원

국립국어원 2019-01-05

발 간 등 록 번 호
11-1371028-000762-01

웹 말뭉치 구축

사업 책임자
이 영 희

국립국어원

제 출 문

국립국어원장 귀하

국립국어원과 체결한 연구용역 계약에 따라 '웹 말뭉치 구축'에 관한 연구 보고서를 작성하여 제출합니다.

■ 사업기간: 2019년 5월 27일 ~ 2019년 11월 23일

2019년 11월 23일

사업 책임자: 이 영 희((주)메트릭스코퍼레이션)

사업 수행자 (주)메트릭스코퍼레이션
사업 책임자 이영희
사업 참여자 김수진, 김홍건, 정종서, 김예슬, 박수정,
이상훈, 신성호, 신현주, 채윤철, 하지영, 윤지은

〈사업 수행자〉 ㈜메트릭스코퍼레이션

사업 책임자	이영희((주)메트릭스코퍼레이션)
사업 참여자	김수진((주)메트릭스코퍼레이션)
	김홍건((주)메트릭스코퍼레이션)
	정종서((주)메트릭스코퍼레이션)
	김예슬((주)메트릭스코퍼레이션)
	박수정((주)메트릭스코퍼레이션)
	이상훈((주)메트릭스코퍼레이션)
	신성호((주)메트릭스코퍼레이션)
	신현주((주)메트릭스코퍼레이션)
	채윤철((주)메트릭스코퍼레이션)
	하지영((주)메트릭스코퍼레이션)
	윤지은((주)메트릭스코퍼레이션)

<국문 초록>

웹 말뭉치 구축

본 사업은 4차 산업혁명 대비 기반 기술 개발, 인공 지능 기술 개발 및 활용을 위한 대규모 말뭉치 구축을 목적으로 국어 자원의 활용도와 가치를 제고하기 위해 누리 소통망(SNS), 블로그, 게시판 등에서 실제로 사용된 웹 언어 자료를 모아 컴퓨터가 읽을 수 있는 형태로 분석한 말뭉치를 국가적으로 구축하여 우리말 인공 지능 개발과 국어 연구 등에 공공 자료로 활용하는 데 그 목적이 있다. 이에 따른 주요 과업의 결과를 요약하면 다음과 같다.

첫째, 웹 말뭉치 구축 사업에 참여할 게시자를 모집하며 주요 웹 게시물 작성자와 온라인 패널, 일반 웹 게시자를 대상으로 활발한 홍보 활동을 진행하였고, 참여자 대상으로 저작권 이용 허락 계약을 체결하는 과정에서 웹 말뭉치 구축에 대한 인지도와 말뭉치 구축 사업의 필요성에 대한 관심을 확보하였다.

둘째, 웹 원문 자료로 국내에서 활발히 사용하고 있는 누리 소통망, 블로그, 게시판, 리뷰 등 다양한 사이트에서 작성된 여러 가지 유형의 게시물을 수집하였다. 수집한 웹 원문 자료는 누리 소통망 2,000,000건, 블로그 10,000건, 게시판 10,000건, 리뷰 100,000건으로 전체 2,120,000건의 웹 원문 자료를 수집하고 웹 말뭉치를 구축하였다. 특히, 매체별로 다양한 사이트에서 웹 원문 자료를 수집함으로써 사이트 특성에 따른 다양한 웹 언어 자료를 확보하였다.

셋째, 과업에서 규정한 웹 말뭉치 구축 지침에 따라 웹 원시 말뭉치를 구축함으로써 민간에서 변환 및 호환이 용이한 공공재로서의 말뭉치를 구축하였다. 웹 원문 자료에 포함된 정보를 기계적으로 수집해 변형이나 왜곡 없이 사이트에 게시된 형태 그대로 웹 말뭉치로 구축함으로써, 실제 웹 게시물 작성자들이 사용하는 언어를 연구하는 자료로 활용될 수 있도록 하였다.

주요어: 웹 말뭉치, 웹 말뭉치 수집, 원시 웹 말뭉치

차 례

제1장 서론

제2장 참여자 모집 및 이용 허락 계약 체결

제3장 웹 원시 말뭉치 구축

차 례

제4장 결론

표 차례

그림 차례

제 1 장

서 론

1. 사업 목적

본 사업은 인공 지능 산업 발전을 위한 대규모 고품질 우리말 자원 수요 증대를 위해 추진한 2019년 국어 말뭉치 구축 사업 중 원문 수집 및 저작권 처리에 해당하는 사업으로, 대규모 웹 언어 자료를 수집하는 '웹 말뭉치 구축' 사업이다. '웹 말뭉치 구축' 사업은 누리 소통망(SNS), 블로그, 게시판 등에서 실제로 사용된 웹 언어 자료를 모아 컴퓨터가 읽을 수 있는 형태로 분석한 말뭉치를 국가적으로 구축하여 우리말 인공 지능 개발과 국어 연구 등에 공공 자료로 활용할 수 있도록 하기 위해 추진한 사업으로, 4차 산업혁명 대비 기반 기술 개발, 인공 지능 기술 개발 및 활용을 위한 대규모 말뭉치 구축으로 국어 자원의 활용도와 가치를 제고하는 데 그 필요성이 있다.

본 사업의 추진 목표는 다음과 같다.

□ 웹 말뭉치 구축 추진 목표

- 자연어처리 시스템 개발 등의 활용을 위한 웹 말뭉치 구축
- 민간에서의 자유로운 활용을 위한 공공재로서의 말뭉치 구축
- 단계별 품질 점검을 통한 품질 향상

2. 사업 수행 범위

본 사업은 위와 같은 사업의 필요성과 목적에 따라 웹 말뭉치 구축을 진행한다. 말뭉치란 컴퓨터가 읽을 수 있는 형태로 입력하고 분석한 대규모 언어 빅데이터로, 말뭉치는 쓰임에 따라 종류가 다양하다. 특히, 원시 말뭉치는 다른 정보 없이 단순하게 입력된 자료로 모든 분석 말뭉치를 만들기 위해 필요한 자료이다. 본 사업의 주요 내용은 웹 원문 자료를 수집해 원시 말뭉치를 구축하는 것으로, 구체적인 사업 수행 범위는 다음과 같다.

□ 웹 말뭉치 구축 추진 내용

- 웹 원문 자료(누리 소통망(SNS: 트위터, 페이스북, 인스타그램 등), 블로그, 게시판, 리뷰 등) 수집
  - 누리 소통망(SNS) : 200만 발화(등록된 게시물 건수) 이상
  - 블로그 : 1만 페이지(등록된 게시물 건수) 이상

- 게시판 : 1만 페이지(등록된 게시물 건수) 이상
- 리뷰 : 10만 리뷰(등록된 게시물 건수) 이상

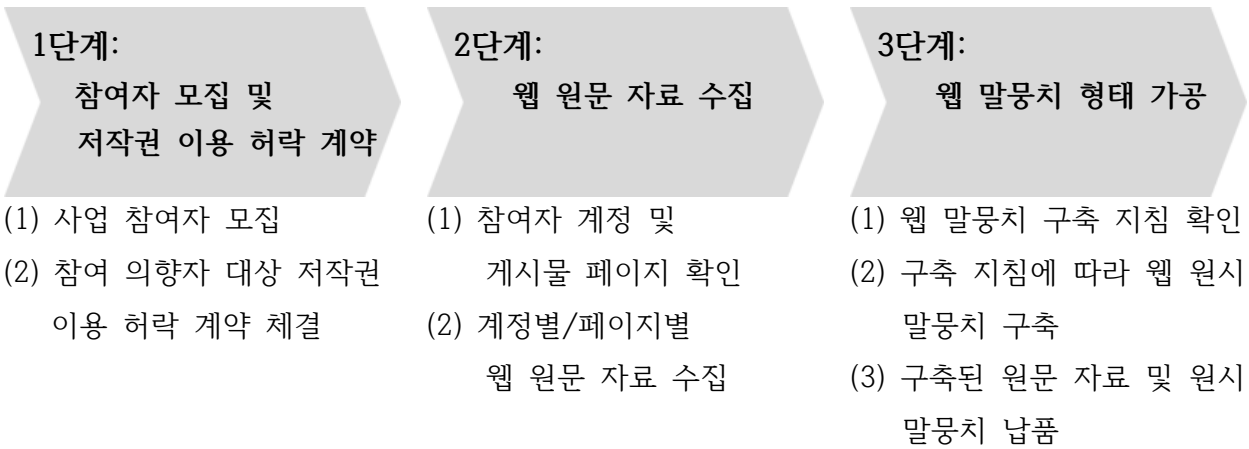
• 웹 원시 말뭉치 구축
- 수집된 웹 원문 자료에 대해 표지를 부착하여 원시 말뭉치 형태로 가공 (단, 띄어쓰기 교정 등 추가 작업은 포함하지 않음)
- 수집 날짜, 웹 주소 등 구축 대상 자료에 대한 메타 정보를 포함하여 구축

• 웹 말뭉치 저작권 이용 허락(이용 동의) 계약 체결
- 수집된 원문 자료를 대상으로 원문 자료 저작권자와 저작권 이용 허락 또는 이용 동의 계약 체결
- 저작권 이용 허락(이용 동의) 대상 권리는 원시 말뭉치의 저장, 복제, 전송, 배포, 2차적 저작물 작성권을 포함하여 체결

3. 사업 수행 절차

본 사업의 수행은 1단계로 참여자 모집 및 저작권 이용 허락 계약을 진행하고, 2단계로 계약이 완료된 참여자의 계정을 확인해 웹 원문 자료를 수집하며, 마지막으로 3단계는 수집된 원문 자료를 웹 말뭉치 형태로 가공하는 절차로 진행되었다. 각 단계별 절차 및 주요 수행 내용은 다음과 같다.

<그림 1> '웹 말뭉치 구축' 수행 절차



4. 기대 효과

‘웹 말뭉치 구축’ 사업을 통한 기대 효과는 다음과 같다.

□ 웹 말뭉치 구축 기대 효과

- 민간에서 활용 가능한 국가 공공재로서의 말뭉치 확대 구축 및 국어 자원의 활용도와 가치 향상에 기여
- 4차 산업혁명 대비 기반 기술 개발 및 인공 지능 기술 개발, 활용을 위한 대규모 말뭉치 구축으로 국어 자원의 활용도와 가치 제고
- 민간 공유를 통해 언어 인공 지능 등 관련 산업 활용을 위한 기반을 마련하고 국어 및 국어문화 연구, 국어정책 수립의 기초 자료로 활용

본 사업과 함께 추진되고 있는 국어 말뭉치 구축 사업을 통해 인공 지능 스피커, 대화형 로봇, 로봇 개인 비서 등 한국어 인공 지능의 성능을 향상시킬 것으로 기대되며, 향후 4차 산업혁명 시대의 인공 지능 서비스 개발 및 기술 혁신을 위한 중요 자료가 될 전망이다.

제 2 장

참여자 모집 및 이용 허락 계약 체결

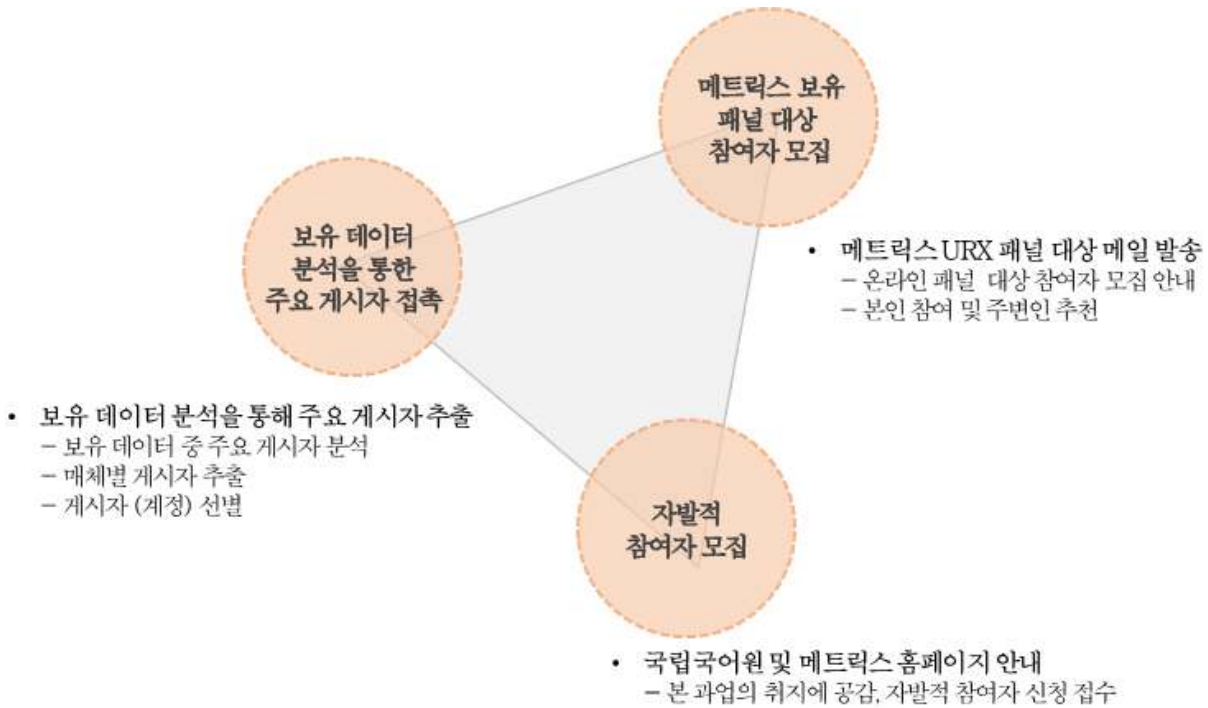
1. 참여자 모집 및 선정

본 사업에 참여하기 위해서는 참여자 본인이 누리 소통망, 블로그, 게시판 등에 직접 작성한 웹 게시물을 최소 1건 이상 보유해야 하며, 참여 대상이 되는 웹 게시물에 대한 저작권 이용 허락 계약을 체결할 의향이 있어야 한다는 두 가지 사항을 충족하여야 한다. 따라서 참여자 본인이 직접 작성한 웹 게시물이 있는지 확인이 필요하며, 사업 참여 의향을 확인하는 것이 필수적이다. 한정된 사업 기간 내에 본 사업 목적에 적합한 참여자 모집을 완료하기 위하여 다각도로 참여자를 모집하는 방안이 필요하였다. 이에 적합한 참여자 모집을 위해 다음과 같은 세 가지 모집 방법을 통하여 참여자 모집을 진행하였다.

1.1. 참여 대상 모집 방법

'웹 말뭉치 구축' 의 참여자는 다음과 같은 세 가지 방법으로 모집을 진행하였다.

<그림 2> 참여자 모집 방법



1) 보유 데이터 분석을 통한 주요 게시자 모집

첫 번째 게시자 모집 방법은 실제 웹 게시글 작성 여부를 확인하고 적합한 대상자에게 참여를 권유하는 방식으로, 사업 수행 기관의 내부 보유 데이터를 활용하여 웹 게시글을 확인하여 적합한 게시자를 찾는 방법을 사용하였다. 내부 축적된 웹 게시물 약 2억 건 중, 2011년에서 2018년에 작성된 게시물 119,365,000건을 분석하여, 1차로 약 12,000개의 작성자 아이디를 추출하였으며, 2차로 약 6,000개의 작성자 계정 및 아이디를 추출하였다. 추출 기준은 웹 게시물이 600건 이상이며, 2018년까지 활발한 게시물 작성 활동을 유지한 계정 및 아이디를 중심으로 추출하였다. 작성자 계정 및 아이디 추출 후, 판매처나 제조사 등에서 홍보용 활동을 목적으로 작성된 계정을 제외하고, 웹 게시물 작성 활동이 활발하고 게시글 내용이 우수한 약 7,300개의 작성자 계정 및 아이디를 선별하였다. 보유 데이터 분석을 통한 참여자 선정 절차는 다음과 같다.

<표 1> 보유 데이터 분석을 통한 참여자 선정 방법

1	**내부 보유 데이터 분석**

- 내부 보유 웹 게시물 데이터 중,
  2011.1.1.~2018.12.31. 기간 내 작성 게시물 총 119,365,000건 분석

2	**기준에 적합한 작성자 아이디 추출**

- 웹 게시물 600건 이상 및 2018년도 웹 게시물이 있는 계정/아이디 추출
  - 1차: 12,000개 추출
  - 2차: 6,000개 추출

3	**참여 대상 아이디 선정**

- 판매처 및 회사 홍보용 등 부적합 계정/아이디 제거,
  최종 7,300개 작성자 계정/아이디 선별

2) 사업 수행 기관의 온라인 패널 대상 참여자 모집

두 번째 방식은 사업 수행 기관의 온라인 패널인 URX(메트릭스코퍼레이션 리서치 패널 브랜드로, 'YOUR eXperience'의 약자) 회원을 이용하는 방식이다. URX 패널은 2019년 5월 기준, 약 1,225,000명이 회원으로 가입되어 있으며, 상시 온라인 설문조사 응답 협약이 되어 있어, 적극적인 참여 유도가 용이하였다. URX 패널 현황은 다음과 같다.

<그림 3> 메트릭스코퍼레이션 온라인 패널 URX 현황

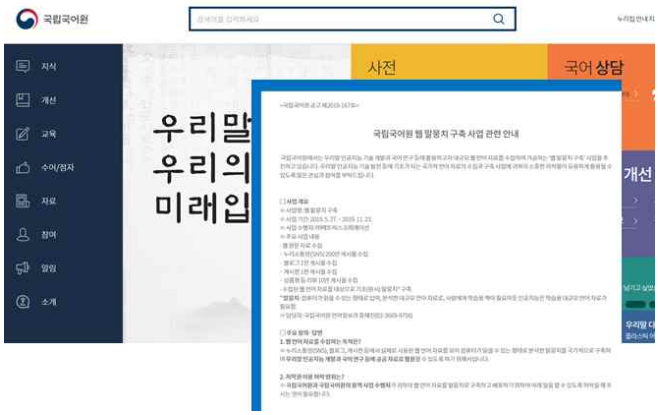
구분	구성비	패널수(명)
계	100.0	1,225,000
서울	33.8	412,825
인천	5.8	71,051
경기	20.7	253,575
부산	7.5	91,873
대구	4.8	58,802
광주	3.3	40,425
대전	3.3	40,427
울산	1.5	18,374
강원	2.1	25,725
경북	3.4	41,656
경남	4.2	51,451
전북	2.6	31,858
전남	2.2	26,956
충북	2.2	26,959
충남	1.9	23,275
제주	0.6	7,353

그러나, 패널 회원 본인이 직접 참여하는 것만으로는 한정된 일정 내에 충분한 모집이 이루어지기 어렵기 때문에, URX 패널 본인의 참여 권유에 그치지 않고, 주변에 있는 참여 조건 대상자를 추천하고 참여를 독려하는 방식으로 진행하였다. 즉, 패널 회원 본인이 웹 게시물 작성자인 경우 본인이 직접 참여하였으며, 패널 회원 중 국립국어원의 '웹 말뭉치 구축' 사업의 취지에 공감하여 참여자 모집 활동을 추진할 의향이 있는 패널은 참여자 모집 요원으로 활동하며 주변에서 참여 조건 대상자를 찾아 참여를 권유하거나 참여 홍보 활동을 수행하여 참여자 모집을 진행하였다.

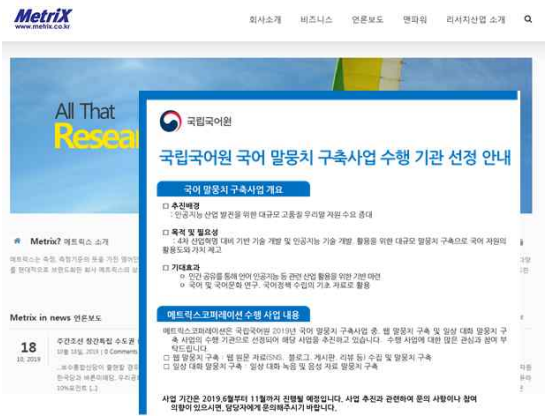
3) 자발적 참여자 모집

세 번째로는 국립국어원 홈페이지(www.korean.go.kr) 및 사업 수행 기관인 메트릭스코퍼레이션 홈페이지(www.metrix.co.kr)의 팝업창(pop-up)에 제시된 '웹 말뭉치 구축' 사업 참여 안내를 통한 자발적인 관심 및 참여로 참여자를 모집하였다. 국립국어원 홈페이지에는 다양한 국어 관련 자료 이용자 및 연구자 등이 많이 방문하였을 것으로 예상되며, 홈페이지 방문자의 경우 국립국어원에 대한 긍정적인 인식과 기대감이 형성되어 있어 자발적인 참여로 연결이 용이하게 이루어졌을 것으로 예상된다. 메트릭스코퍼레이션 홈페이지의 경우, 다양한 설문 조사 및 국가 통계 조사를 담당하는 면접원들의 접속이 상시로 이루어지고 있어, 과업에 대한 인지도를 확보하기 용이하였다. 다만, 사업 착수 후 약 1개월 이후에 홈페이지를 통한 안내 팝업창이 제시되어, 앞선 두 가지 참여자 모집 방법 대비 이를 통한 참여자 수는 많지 않으나, 본 사업의 취지와 목적에 대한 공감을 통한 자발적인 참여라는 점에서 의미가 매우 크다고 할 수 있다.

<그림 4> 국립국어원 홈페이지 내 홍보 화면



<그림 5> 메트릭스코퍼레이션 홈페이지 내 홍보 화면



1.2. 참여자 접촉 및 참여 안내 과정

참여자 모집 대상자 중 패널 회원과 자발적 참여자를 제외하고, 보유 데이터 분석을 통한 주요 게시자는 작성자 아이디 및 계정, 프로필상 제시된 일부 이메일 주소를 제외하고 접촉이 가능한 다른 연락처가 확보되지 않은 상태이다. 따라서, 작성자 아이디와 계정을 이용해 쪽지 혹은 이메일을 발송한 후, 회신이 오는 경우에만 이후 참여 절차를 진행할 수 있는 어려움이 있었다. 참여 거절에 대한 답변이 회신되는 경우를 제외하고, 2회 이상 쪽지 및 이메일을 발송하여 본 사업에 대한 안내가 충분히 이루어지도록 진행하였다. 참여자 접촉 및 참여 안내를 진행한 절차는 다음과 같다.

<그림 6> 참여자 접촉 및 참여 안내 방법

1 **사업 참여 안내 이메일 또는 쪽지 발송**

- 이메일 또는 쪽지 발송
  - 데이터 분석을 통해 추출한 아이디 및 계정 대상자 이메일 및 쪽지 발송
  - 패널 회원은 패널 대상 등록된 이메일을 통해 안내 메일 발송
- 사업의 목적 및 취지 안내

2 **참여 의향 회신 이메일 및 쪽지 수신**

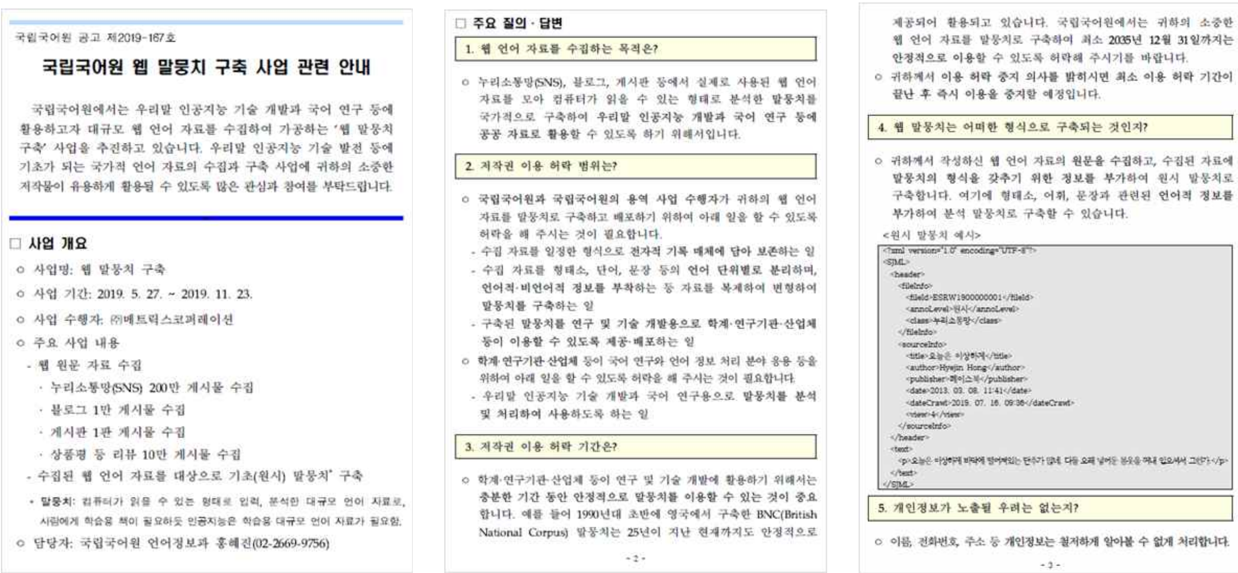
- 참여 의향자가 직접 발송한 이메일 및 쪽지 수신
  - 사업 참여 안내 메일 또는 쪽지에 대한 회신
  - 홈페이지에 제시된 담당자 이메일 주소로 자발적 참여자가 발송한 이메일 수신
- 회신자 아이디 및 계정 확인

3 **최종 참여 의향자 대상 참여 방법 상세 안내 및 참여자 개별 연락처 확인**

- 참여 방법에 대한 상세한 안내 및 설명
- 저작권 이용 허락 계약 및 사례비 지급을 위한 개별 연락처 확인

사업 참여 안내 이메일 및 쪽지 발송 후, 참여 의향에 대한 회신을 하는 참여자를 대상으로 개별 접촉을 통해 사업의 목적 및 취지에 대해 상세히 설명하고, 참여 방법을 안내하였다. 사업의 목적 및 취지에 대한 안내는 국립국어원에서 제공한 '국립국어원 웹 말뭉치 구축 사업 관련 안내' 문서를 이용하였다. 안내 문서에는 사업의 목적 및 사업 개요, 주요 질의·답변에 대한 내용이 포함되어 있다.

<그림 7> 국립국어원 제공 '웹 말뭉치 구축 사업 관련 안내'



국립국어원 공고 제2019-167호

국립국어원 웹 말뭉치 구축 사업 관련 안내

국립국어원에서는 우리말 인공지능 기술 개발과 국어 연구 등에 활용하고자 대규모 웹 언어 자료를 수집하여 가공하는 '웹 말뭉치 구축' 사업을 추진하고 있습니다. 우리말 인공지능 기술 발전 등에 기초가 되는 국가적 언어 자료의 수집과 구축 사업에 귀하의 소중한 저작물이 유용하게 활용될 수 있도록 많은 관심과 참여를 부탁드립니다.

□ **사업 개요**

- ○ 사업명: 웹 말뭉치 구축
- ○ 사업 기간: 2019. 5. 27. ~ 2019. 11. 23.
- ○ 사업 수행자: ㈜메트릭스코퍼레이션
- ○ 주요 사업 내용
  - 웹 원문 자료 수집
    - · 누리소통망(SNS) 200만 게시물 수집
    - · 블로그 1만 게시물 수집
    - · 게시판 1만 게시물 수집
    - · 상품평 등 리뷰 10만 게시물 수집
  - 수집된 웹 언어 자료를 대상으로 기초(원시) 말뭉치* 구축

* 말뭉치: 컴퓨터가 읽을 수 있는 형태로 입력, 분석한 대규모 언어 자료로, 사람에게 학습용 책이 필요하듯 인공지능은 학습용 대규모 언어 자료가 필요함.

- ○ 담당자: 국립국어원 언어정보과 홍혜진(02-2669-9756)

□ **주요 질의 · 답변**

1. 웹 언어 자료를 수집하는 목적은?

○ 누리소통망(SNS), 블로그, 게시판 등에서 실제로 사용된 웹 언어 자료를 모아 컴퓨터가 읽을 수 있는 형태로 분석한 말뭉치를 국가적으로 구축하여 우리말 인공지능 개발과 국어 연구 등에 공공 자료로 활용할 수 있도록 하기 위해서입니다.

2. 저작권 이용 허락 범위는?

○ 국립국어원과 국립국어원의 용역 사업 수행자가 귀하의 웹 언어 자료를 말뭉치로 구축하고 배포하기 위하여 아래 일을 할 수 있도록 허락을 해 주시는 것이 필요합니다.
- 수집 자료를 일정한 형식으로 전자적 기록 매체에 담아 보존하는 일
- 수집 자료를 형태소, 단어, 문장 등의 언어 단위별로 분리하며, 언어적·비언어적 정보를 부착하는 등 자료를 복제하여 변형하여 말뭉치를 구축하는 일
- 구축된 말뭉치를 연구 및 기술 개발용으로 학계·연구기관·산업체 등이 이용할 수 있도록 제공·배포하는 일

○ 학계·연구기관·산업체 등이 국어 연구와 언어 정보 처리 분야 응용 등을 위하여 아래 일을 할 수 있도록 허락을 해 주시는 것이 필요합니다.
- 우리말 인공지능 기술 개발과 국어 연구용으로 말뭉치를 분석 및 처리하여 사용하도록 하는 일

3. 저작권 이용 허락 기간은?

○ 학계·연구기관·산업체 등이 연구 및 기술 개발에 활용하기 위해서는 충분한 기간 동안 안정적으로 말뭉치를 이용할 수 있는 것이 중요합니다. 예를 들어 1990년대 초반에 영국에서 구축한 BNC(British National Corpus) 말뭉치는 25년이 지난 현재까지도 안정적으로

- 2 -

제공되어 활용되고 있습니다. 국립국어원에서는 귀하의 소중한 웹 언어 자료를 말뭉치로 구축하여 최소 2035년 12월 31일까지는 안정적으로 이용할 수 있도록 허락해 주시기를 바랍니다.

○ 귀하께서 이용 허락 중지 의사를 밝히시면 최소 이용 허락 기간이 끝난 후 즉시 이용을 중지할 예정입니다.

4. 웹 말뭉치는 어떠한 형식으로 구축되는 것인지?

○ 귀하께서 작성하신 웹 언어 자료의 원문을 수집하고, 수집된 자료에 말뭉치의 형식을 갖추기 위한 정보를 부가하여 원시 말뭉치로 구축합니다. 여기에 형태소, 어휘, 문장과 관련된 언어적 정보를 부가하여 분석 말뭉치로 구축할 수 있습니다.

<원시 말뭉치 예시>

```
<?xml version="1.0" encoding="UTF-8"?>
<SJML>
  <header>
    <fileInfo>
      <fileId>ESRW1900000001</fileId>
      <annoLevel>원시</annoLevel>
      <class>누리소통망</class>
    </fileInfo>
    <sourceInfo>
      <title>오늘은 이상하게</title>
      <author>Hyejin Hong</author>
      <publisher>페이스북</publisher>
      <date>2013. 03. 08. 11:41</date>
      <dateCrawl>2019. 07. 16. 09:36</dateCrawl>
      <view>4</view>
    </sourceInfo>
  </header>
  <text>
    <p>오늘은 이상하게 바닥에 떨어져있는 단추가 많네. 다들 오래 넣어둔 봄옷을 꺼내 입으셔서 그런가?</p>
  </text>
</SJML>
```

5. 개인정보가 노출될 우려는 없는지?

○ 이름, 전화번호, 주소 등 개인정보는 철저하게 알아볼 수 없게 처리합니다.

- 3 -

상세한 사업의 목적과 취지 안내 이후, 참여 의향을 최종적으로 확인하고 구체적인 웹 말뭉치 구축 사업 참여 방법을 안내하였다. 참여를 위해서는 본인이 작성한 게시물의 매체와 계정을 확인하여 계정을 특정하는 온라인 주소(URL) 확보가 필요하다. 또한, 저작권 이용 허락 계약 체결과 저작권 이용 허락에 대한 사례비 지급을 위해 저작권자 본인의 연락처가 반드시 필요하므로, 본인 명의 휴대폰 번호와 본 사업에 참여하고자 하는 게시물이 포함된 온라인 주소(URL) 두 가지를 참여자 신청 시 필수 사항에 포함시켰다. 즉, 본인의 성명, 본인 명의의 휴대폰 번호, 본 사업에 참여하고자 하는 본인이 작성한 게시물이 포함된 온라인 주소(URL) 세 가지 사항이 필수 요소이며, 이외에 성별이나 주소, 연령 등의 정보는 별도로 확인하거나 확보하지 않았다. 참여 대상 온라인 주소(URL)를 직접 확인하여 본인이 작성한 게시물이 최소 1건 이상 있는 것이 확인되면, 저작권 이용 허락 계약 체결을 위한 안내 메시지를 본인 명의의 휴대폰 번호로 발송하였다. 메시지 안내를 통한 저작권 이용 허락 계약 체결 방법은 다음 장의 '2. 저작권 이용 허락 계약 체결'에서 상세히 설명하도록 한다.

참여자 선정 및 모집을 통해 확보한 참여자 수는 총 2,869명이다. 이 중, 저작권 이용

허락 계약 체결 과정, 본인 작성 여부 확인 및 계정 확인 과정에서 중도 탈락하거나 참여를 거절한 622명을 제외하고 최종 2,247명의 참여자 모집을 완료하였다.

2. 저작권 이용 허락 계약 체결

2.1. 저작권 이용 허락 계약의 내용

본 사업 참여자는 게시물을 작성한 원문 자료 저작권자로, 참여자와의 웹 게시물에 대한 저작권 이용 허락 계약 체결이 필요하다. 특히, 본 사업의 경우, 웹 게시물을 수집해 원시 말뭉치로 구축하는 일 뿐만 아니라 복제·변형물을 연구 및 기술 개발용으로 학계·연구기관·산업체 등이 이용할 수 있도록 제공·배포가 가능하도록 저작권자로부터 저작권 이용 허락 계약을 체결하는 과제가 포함되어 있다. 향후 발생할 수 있는 법률적 분쟁을 최소화하고 민간 활용도를 제고하기 위해 저작권 이용 허락 계약 체결은 본 사업에서 필수적인 추진 과제이다.

저작권 이용 허락 계약서 양식 및 내용은 국립국어원에서 법률 검토를 거친 후 수행기관에 제공하였으며, 수행 기관과의 세부 요건에 대한 협의를 거쳐 최종적으로 확정하였다. 저작권 이용 허락 계약 내용에 대한 검토는 사업 수행 기관인 메트릭스코퍼레이션 법률 자문팀을 통해 재검토를 완료하였다. 확정된 저작권 이용 허락 계약서를 이용하여 수집된 원문 자료를 대상으로 원문 자료 저작권자와 저작권 이용 허락 계약을 체결하였으며, 계약서의 세부 내용은 다음과 같다.

1) 저작권 이용 허락 대상 권리의 내용

저작권 이용 허락 계약서에 제시된 대상 권리의 내용은 복제권, 전송권, 배포권, 2차적 저작물 작성권을 포함하며, 계약서상 제시된 내용은 다음과 같다.

제2조 (계약의 대상)

본 계약의 이용 허락 대상이 되는 권리는 아래의 저작물(이하 “대상 저작물”)에 대한 저작재산권 중 당사자가 합의한 권리로 한다.

종별 : 어문저작물

권리 : 복제권, 전송권, 배포권, 2차적 저작물 작성권

저작권 이용 허락 대상 권리의 내용은 구체적인 설명이 필요한 부분이므로, 계약서상에 다음과 같은 세부 권리의 내용을 포함시켰다.

※ 저작권 이용 허락 대상 권리의 내용

1. 국립국어원 및 국립국어원이 발주한 용역 사업의 수행자가 대상 저작물을 일정한 형식으로 전자적 기록 매체에 담아 보존하는 일
2. 국립국어원 및 국립국어원이 발주한 용역 사업의 수행자가 자모, 음절, 어휘, 어절, 구절, 문장 및 텍스트 단위의 국어 연구와 언어 정보 처리 분야에 응용하기 위해 대상 저작물을 복제·변형(목차·머리말·도표·그림·각주 등의 편집 및 삭제, 언어 단위별 분리, 언어적·비언어적 정보 부착 등)하는 일
3. 국립국어원 및 국립국어원이 발주한 용역 사업의 수행자가 대상 저작물 및 그 복제·변형물을 연구 및 기술 개발용으로 학계·연구기관·산업체 등이 이용할 수 있도록 제공·배포하는 일
4. 대상 저작물 및 그 복제·변형물을 제공·배포받은 학계·연구기관·산업체 등이 국어 연구와 언어 정보 처리 분야 응용을 위하여 대상 저작물 및 그 복제·변형물을 분석 및 처리하여 사용하는 것을 허락하는 일

2) 대상 저작물 이용 허락 기간

민간에서 활용 가능한 국가 공공재로서의 말뭉치 확대 구축이 필요하므로, 저작권 이용 허락 기간은 영구적 또는 준영구적이어야 할 필요성이 있었다. 따라서 대상 저작물의 이용 허락 기간은 계약 체결일부터 2035. 12. 31.까지로 하며, 권리자가 이용 허락을 중지하고자 하는 의사를 밝히지 아니하면 이용 허락이 5년 단위로 자동 갱신이 되는 내용을 계약서상에 명시하였다.

제3조 (이용 허락 기간)

대상 저작물의 이용 허락 기간은 계약 체결일부터 2035. 12. 31.까지로 하며, 권리자가 이용 허락을 중지하고자 하는 의사를 밝히지 아니하면 이용 허락이 5년 단위로 자동 갱신된다. 권리자가 이용 허락 중지 의사를 밝히면 그 의사 내용에 따라 이용 허락을 중지하여야 하며, 그렇지 아니하면 이용 허락 내용이 유지된다.

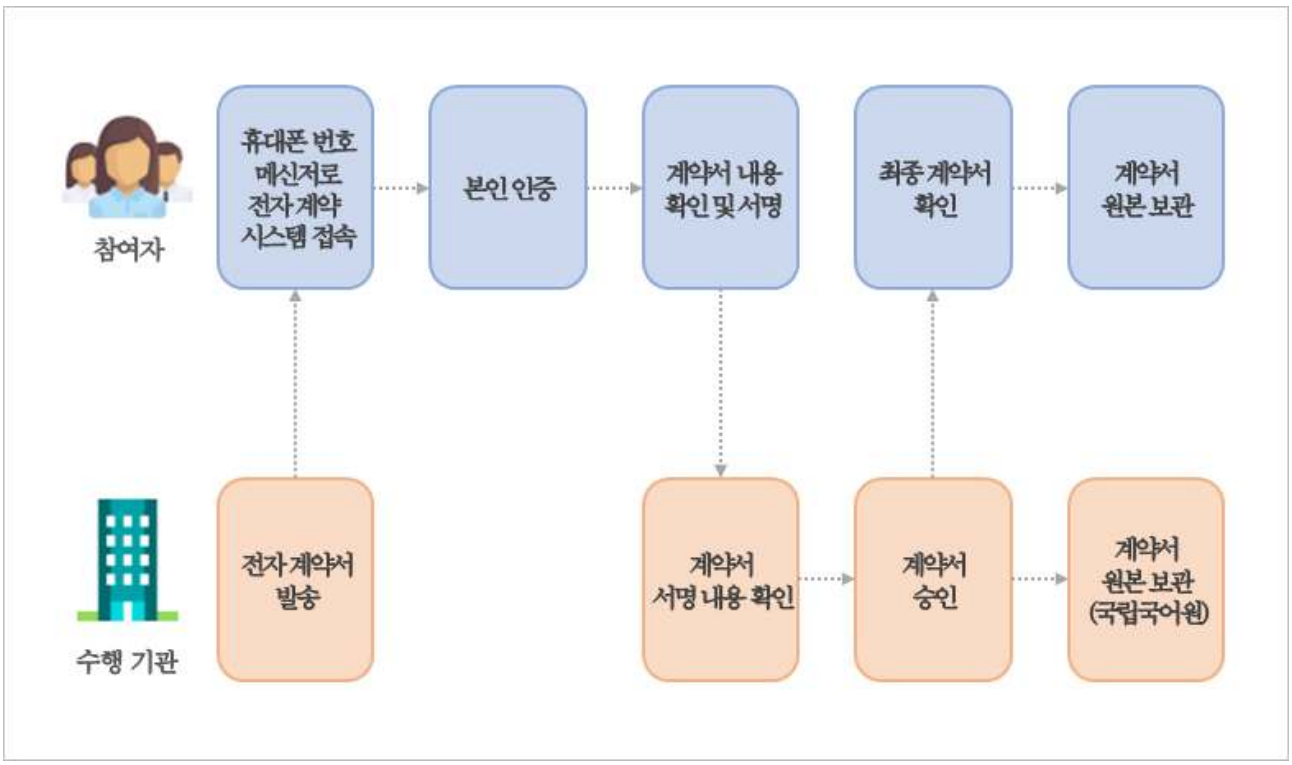
2.2 저작권 이용 허락 계약 체결

저작권 이용 허락 계약을 체결하기 위해서는 웹 원문 자료의 저작권자를 만나 계약을 체결하고, 상호 간 계약서 각 1부씩을 보관하는 것이 원칙이다. 그러나 웹 원문 자료의 저작권자를 직접 찾아가 계약을 체결하기 위해서는 많은 시간과 비용이 소요되며, 저작권자의 시간적인 부담이 크기 때문에 번거로운 과정이므로 참여자 모집 시 제약 요인이 될 수 있다. 따라서 본 사업 수행 시 저작권 이용 허락 계약 체결은 전자계약 시스템을 이용해 진행하였다.

전자 계약은 일반적인 문서를 통한 계약의 개념과 동일하나, 다만 그 방식이 전자문서를 통해서 이루어지는 데에 차이가 있다. 「전자서명법」 제3조 제2항, 제3항, 그리고 「전자문서 및 전자거래 기본법」 제4조 제1항에 따라, 전자서명은 당사자 간의 약정에 따른 서명, 서명날인 또는 기명날인으로서의 효력을 가진다. 즉, 전자문서의 형식적 효력(데이터 메시지의 문서성)을 인정함에 따라 데이터 메시지에 의한 계약도 법적 효력을 갖게 된다.

따라서, 최근 디지털 데이터 확보의 중요성 및 전자문서 사용 환경 변화의 추세에 따라 전자 계약 방식으로 저작권 이용 허락 계약 체결을 진행하였다.

<그림 8> 저작권 이용 허락 전자 계약 절차



전자 계약 시스템은 IT 보안 전문 기관인 한국정보인증에서 블록체인 기술을 이용해 개발한 ‘signOK’ 시스템을 이용하였다. 전자 계약 절차는 본인 명의의 휴대폰 번호로 전자 계약 서명 요청 메시지를 발송하면, 저작권자가 이를 수신하여 시스템에서 본인 인증 절차를 완료한 후 계약서 내용을 확인하고 계약서에 서명하고, 서명에 대한 내용을 관리자가 확인 및 승인하면 계약이 완료된다. 완료된 계약서는 계약 당사자인 저작권자 본인과 국립국어원에서 PDF 파일 형태로 각 1부씩 보관하였다.

웹 게시물 저작권 이용 허락 계약을 완료한 참여자에게는 소정의 사례금을 지급하였으며, 최종적으로 저작권 이용 허락 계약을 완료한 참여자는 2,065명이다.

제 3 장

웹 원시
말뭉치 구축

1. 웹 원문 자료 수집

1.1. 수집 대상 매체 확인

웹 말뭉치 구축을 위해 대상 매체에 게시된 웹 원문 자료를 수집해야 하며, 대상 매체별 수집 개발이 필수적이다. 본 사업에 해당하는 대상 매체는 누리 소통망, 블로그, 게시판, 리뷰 네 가지이다. 본 사업의 범위에 포함된 네 가지 매체를 대상으로 수집 사이트 검토 작업을 진행하였다.

1) 누리 소통망 및 블로그

국내 웹 게시자가 많이 이용하는 누리 소통망은 트위터(www.twitter.com), 페이스북(www.facebook.com), 인스타그램(www.instagram.com)이 대표적이며, 블로그의 경우는 네이버 블로그(blog.naver.com), 다음 블로그(blog.daum.net), 티스토리(www.tistory.com) 세 개의 사이트가 대표적인 사이트로, 실제 참여자 모집 시 누리 소통망과 블로그는 위의 6개 사이트 중심으로 진행되었다.

<그림 9> 누리 소통망 및 블로그 매체

2) 게시판

게시판의 경우 누리 소통망, 블로그와 달리 국내에 다양한 사이트에 이용자가 분포되어 있다. 대부분 네이버 카페(cafe.naver.com), 다음 카페(cafe.daum.net) 이용자가 가장 많은 비중을 차지하고 있으나, 클리앙(www.clien.net), 엠엘비파크(mlbpark.donga.com), 뽐뿌(www.ppomppu.co.kr), 디시인사이드(www.dcinside.com), 이글루스(www.egloos.com), 오

르비(orbi.kr), 웃긴대학(web.humoruniv.com) 등의 사이트 게시판 역시 국내 사용자의 웹 게시물이 많다. 웹 원문 자료 수집을 위해서는 누리 소통망, 블로그의 주요 6개 사이트 외에 참여자가 웹 게시물을 작성한 다수의 게시판 사이트에 대한 사전 수집 개발 및 점검이 필요하였다.

<그림 10> 게시판 매체

3) 리뷰

웹 원문 자료 대상 매체 중 누리 소통망, 블로그, 게시판과는 달리, 리뷰의 경우는 수집 대상 매체 특성에 따른 구분이 아닌 게시글 내용의 성격에 따라 분류가 필요하므로, 누리 소통망, 블로그, 게시판 매체를 통해 수집된 웹 원문 자료 중, 리뷰 기준에 해당하는 게시물 선별을 통한 구분이 필요하다. 웹 게시글 중 리뷰에 해당하는 게시물은 본문에 상품이나 서비스에 대한 평가 및 이용 후기가 포함된 글을 의미하며, 넓은 범위에서 보면 상품 또는 서비스를 이용한 느낌과 만족도에 대해 언급을 하는 게시물 역시 리뷰에 해당하는 범위로 해석할 수 있다. 그러나 본 사업에서는 리뷰 선별을 위한 본문의 내용의 세부 분석을 과업의 범위로 지정하지 않았으며, 사업 수행 일정과 비용의 제한이 있어 웹 게시글의 본문 내용을 면밀히 파악하여 리뷰로 규정하는데 한계가 있었다. 따라서 본 사업에서는 웹 게시글의 제목을 기준으로 명확하게 리뷰로 규정할 수 있는 게시글을 리뷰로 정의하기로 하였다. 본 사업에서 적용한 리뷰에 대한 기준 및 기준은 다음과 같다.

<표 2> 리뷰 선정 기준 및 예시

리뷰 선정 기준	선정 기준에 따른 실제 게시물 예시
1) 상품평	상품평 좋은걸로 찾아봤어요
2) 후기	레몬디톡스 구매후기
3) 리뷰	스포츠선크림 리뷰
4) 개봉기	갤럭시S10 개봉기
5) 사용기	스노우피크 `폴딩 토치` 구입 사용기
6) 언박싱	갤럭시노트 7 사전예약 언박싱
7) 설치기	유아놀이매트 설치기
8) 맛집 방문	벌교 꼬막정식 꼬막 맛집 방문 후기
9) 여행기	강촌여행기 이모저모
10) 서평	서평 : 엄마표 영어 이제 시작합니다
11) 답사기	돌잔치 전문 업체 답사기
12) 추천	인사동 맛집 추천 차(茶)이야기
13) (제품)선택	여행용 향수로 향기좋은 고체향수 선택
14) 가볼만한 곳	경주 가볼만한 곳 보문단지

1.2. 수집 대상 매체 접근 방법

수집 대상 매체는 리뷰를 제외한 누리 소통망, 블로그, 게시판 매체를 대상으로 접근 방법을 검토하였다. 사업 수행 기관은 웹 게시물 수집을 위한 자체 개발 수집기를 보유하고 있으며, 국내 대부분의 사이트에 대한 수집기 개발이 완료되어 있어 추가적인 개발 과정은 필요하지 않았다. 그러나 본 사업 특성상 매체별 접근 방식이 상이하여 수집 접근 방식별로 수집기 시험 및 점검 작업 수행이 필요하였다. 일반적으로 웹 게시물 수집 시 수집 키워드를 적용하여 키워드 검색 조건에 따라 검색된 결과물을 수집하는 방법으로 진행하나, 본 사업 특성상 수집 키워드가 존재하지 않으며, 작성자의 계정 또는 아이디 기반으로 수집이 이루어져야 하는 특징이 있었다. 또한, 게시판의 경우는 단순히 작성자 계정이나 아이디가 아닌, 직접 작성한 게시물 단위별 수집이 필요해 매체별 차별화된 수집 방식으로 접근이 필요하였다. 구체적인 매체별 접근 방법은 다음과 같다.

<표 3> 매체별 수집 접근 방법

매체 구분	매체별 수집 접근 방법
누리 소통망	• 참여자 개인 계정별 수집 - 트위터, 인스타그램, 페이스북의 개인 계정별 게시물 중 참여자 본인 작성 게시물 수집 진행 * 개인 계정 주소 예시: twitter.com/younghee  
블로그	• 블로그 주소별 수집 - 개인 블로그 주소 내 게시된 전체 게시물 수집 진행 * 개인 블로그 주소 예시: blog.naver.com/PostList.nhn?blogId=bluesea 
게시판	• 게시판 내 참여자 게시물 단위별 수집 - 게시판 내 참여자 아이디로 작성된 게시물 페이지별 수집 진행  

1) 누리 소통망 수집 접근 방법

개인 계정별 웹 페이지가 존재하는 누리 소통망의 경우, 참여자 본인의 계정 주소를 기반으로 해당 계정에 게시된 게시물을 수집하되, 타인이 작성한 댓글 등은 제외하고 수집을 진행하는 것이 필요하다. 따라서 참여 신청 시 참여자 본인 소유의 계정 주소를 받아 해당 계정 내 참여자 본인이 작성한 모든 글을 대상으로 수집하였다.

2) 블로그 수집 접근 방법

블로그는 본인이 자유롭게 자신의 글을 올릴 수 있는 웹 사이트로, 블로그 내 여러 개의 게시글이 있는 항목으로 구성되어 있는 경우가 있다. 참여자 본인의 웹 블로그 주소를 기반으로 블로그 내 모든 항목에 작성된 게시물을 수집하는 방식으로 수집 진행이 필요하다. 따라서, 참여 신청 시 참여자 본인이 개설한 블로그 주소를 받아 해당 주소 내 항목에 있는 모든 글을 수집하였다.

3) 게시판 수집 접근 방법

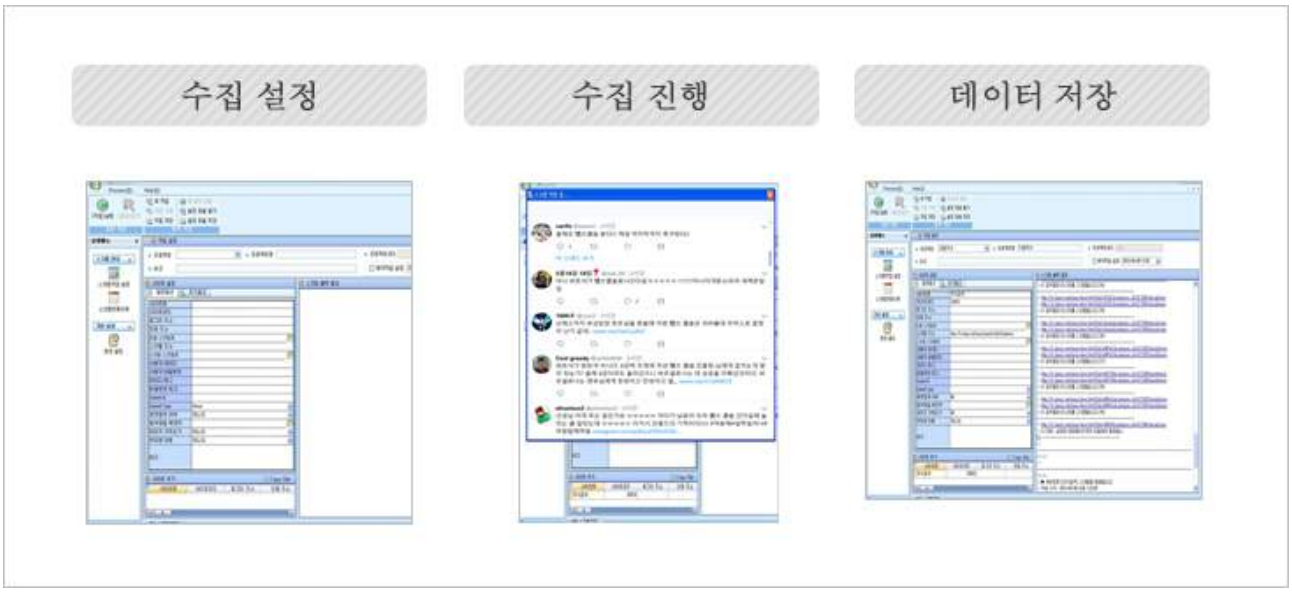
자신의 계정 또는 웹 사이트에 본인이 작성한 글을 올리는 누리 소통망이나 블로그와는 달리, 온라인 게시판은 불특정 다수가 다양한 주제에 대한 생각을 논하는 포럼 형식으로 운영되므로, 특정 게시판 내에는 본인의 글 외에 다수가 작성한 게시물이 포함되어 있다. 따라서, 특정 계정이나 게시판 주소를 기준으로 수집이 불가능하며, 참여자 본인이 작성한 게시물 페이지 단위별 접근이 필수적이다. 따라서, 게시판은 누리 소통망 및 블로그와는 달리 개별 게시물 단위별 수집을 진행해 많은 시간과 노력이 소요되었다. 게시판 수집을 위해서는 게시자 본인이 게시물을 작성한 게시판의 사이트 주소와 사이트 내 게시판 명칭, 사이트 내 게시판에서 작성한 게시물의 개별 웹 주소(URL)가 필요하였다.

1.3. 웹 원문 자료 수집

웹 원문 자료의 수집은 사업 수행 기관의 자체 개발 수집기인 'Buzz Crawler'를 이용하여 수집하였다. 일반적으로 웹 게시물 수집 시 수집 키워드 설정을 통해 검색된 결과물을 수집하는 방식으로 진행하나, 본 사업의 경우 계정 단위 또는 게시물 개별 페이지 단위로 수집이 진행되어야 하므로, 별도의 수집 키워드 설정 과정 없이 사이트

내 계정별 웹 주소(URL) 단위별로 수집을 설정해 게시물을 수집하는 방식으로 진행되었다.

<그림 11> 메트릭스코퍼레이션 자체 개발 버즈 수집 엔진: Buzz Crawler

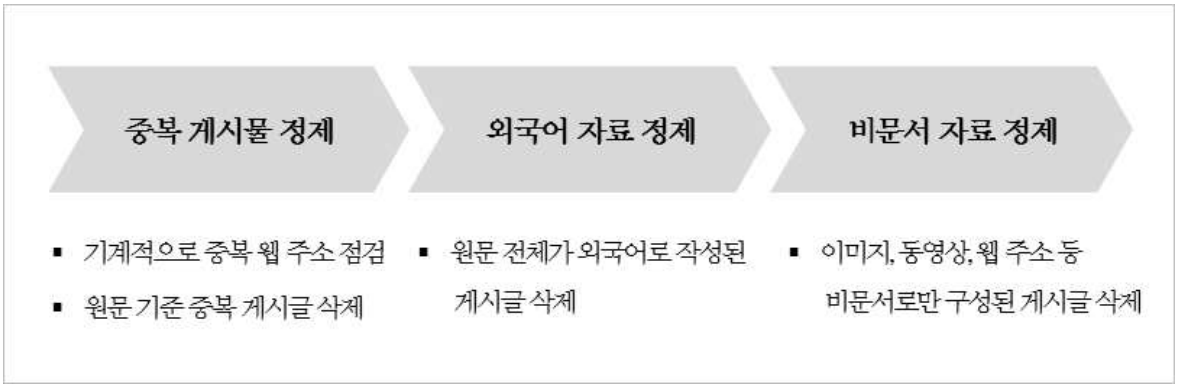


1.4. 웹 원문 자료 정제

1) 비적합 게시글 정제

웹 원문 자료의 수집이 완료되면, 검수 시스템을 이용하여 비적합 게시글을 정제하는 작업을 수행하였다. 정제는 사업 수행 기관 자체 개발 검수 시스템인 'Buzz Rechecker' 시스템을 이용하여 중복 게시글, 전문 외국어로 작성된 게시물, 이미지 또는 동영상이나 웹 주소 등 비문서로만 구성된 웹 원문 자료는 삭제하였다.

<그림 12> 비적합 게시글 정제 과정



2) 게시자별 수집 한도 설정

본 사업의 목적에 따라 다양한 자료의 수집 및 균형적인 말뭉치 구성을 위해 매체별 참여자 계정의 수집 한도를 제한할 필요가 있었다. 사업 착수 전 참여자 계정별 수집 한도를 최대 6,000건으로 제한하기로 규정한 바 있다. 따라서, 웹 원문 자료 수집 완료 후, 계정별 수집된 게시물을 집계하여 6,000건이 초과되는 계정은 초과 건수부터 웹 원문 자료 삭제 작업을 진행하였다.

1.5. 웹 원문 자료 수집 결과

참여자의 게시물을 매체별로 수집한 결과, 총 2,878,606건이 수집되었다. 자료 정제 기준에 따라 비적합 게시물과 게시자별 수집 한도 초과 게시물을 정제한 결과, 총 758,606건을 삭제하였으며 최종 2,120,000건의 웹 원문 자료를 확보하였다.

2. 웹 말뭉치 구축

2.1. 웹 말뭉치 구축 지침

웹 말뭉치 구축은 국립국어원에서 규정한 '웹 말뭉치 구축 지침'에 따라 구축을 진행하였다. 구축 지침은 국립국어원에서 제시한 지침에 따랐으며, 말뭉치 구축 사업 전반적으로 기본적인 틀이 갖추어져 있었으므로 별도의 추가나 조정이 필요하지 않았다. 웹 말뭉치 구축 지침은 크게 파일 형식 및 개요에 대한 지침과 헤더 및 마크업 지침으로 구성되어 있다.

1) 파일명 부여 지침

파일명 부여 방식은 총 14자리를 기준으로, 각 자리별 부여 지침에 따라 파일명을 부여하도록 되어 있다. 첫째 자리는 웹 말뭉치를 의미하는 내용으로 변동이 필요 없으며, 매체 및 장르 구분과 말뭉치 유형에 따라 둘째 자리부터 넷째 자리는 파일의 내용에 따라 지침에 맞추어 부여된다. 각 자릿수별 부여 지침은 다음과 같다.

<표 4> 파일명 부여 방식(지침)

자릿수	내용	부여 지침	지침 기준
1	구분	E	웹 말뭉치
2	매체 및 장르 대분류	S	누리 소통망
		B	블로그
		P	게시글
		R	리뷰
3, 4	말뭉치 유형 구분	OR	원문 자료
		RW	원시 말뭉치
5, 6	구축년도	19	
7 ~ 14	일련번호	########	번호

2) 인코딩 방식

인코딩 방식은 구축 지침에 제시된 규정에 따라 다른 인코딩과의 왕복 변환이 간단하고 하위 호환성이 보장되는 UTF-8로 진행하였다.

3) 헤더와 마크업 부착 지침

헤더와 마크업은 SJML의 기본 구조 요소에 따라 부착하였다. SJML은 파일의 메타 정보가 담겨 있는 'header'와 텍스트 정보를 포함하는 'text'로 구성된다. 'header' 에 포함되는 정보는 파일의 정보에 포함되는 'fileInfo' 정보와 게시물의 기본 정보에 해당하는 'sourceInfo', 게시자의 인구통계학적 특성 정보에 해당하는 'profileInfo'로 구성되어 있다. 'fileInfo'는 파일의 고유 식별자로 말뭉치 파일명을 의미하며, 주석 및 샘플링 방식, 장르 구분에 대한 정보를 부착한다. 'sourceInfo'는 제목, 게시자, 수집 사이트, 게시 날짜, 수집 날짜, 웹 주소(URL), 조회 수 정보로 수집 시 해당 정보를 포함해 수집해야만 기록이 가능한 정보이다. 웹 원시 자료 수집 시 수집과 동시에 사이트명, 제목, 본문, 작성자, 작성일, 웹 주소 등 메타 정보를 포함하여 수집하는 시스템을 이용하여 해당 정보는 별도의 확인 작업을 거치지 않고 자동 축적된 정보를 이용하였다. 'profileInfo' 정보는 필수적인 항목에 해당되는 정보는 아니며, 직업, 출생

지, 주 성장지, 학업 등은 별도의 정보 확인이 필요해 본 사업에서는 해당 정보를 별도로 확보하지 않았다. 다만, 저작권 이용 허락 계약 시 본인 확인을 위해 성별 및 연령 정보가 확보되어 있어, 해당 정보를 포함시켰다. 헤더와 마크업 지침은 다음과 같다.

<표 5> 헤더와 마크업 지침

요소	메타 정보		내용
<header>	<fileInfo>	<fileId>	파일의 고유 식별자(말뭉치 파일명)
		<annoLevel>	주석 수준: 원시
		<sampling>	표본 수집 방식
		<class>	구축 계획에 따른 장르 분류: 누리 소통망, 블로그, 게시글, 리뷰
	<sourceInfo>	<title>	제목
		<author>	게시자
		<publisher>	수집 사이트
		<date>	게시 날짜
		<dateCrawl>	수집 날짜
		<url>	URL 주소
		<view>	조회수
	<profileInfo>	<personId> 외	게시자 성별(sex)
			게시자 연령(age)
<text>	<p>		단락 경계 정보 또는 줄 바꿈 경계 정보

요소에 따른 메타 정보의 세부 내용은 다음과 같다. ‘header’ 요소의 ’fileInfo’의 메타 정보는 본 사업에서 규정한 지침에 따라 정의된 정보로 구성되어 있다. fileId는 원시 말뭉치 파일명 부여 지침에 따라 설정된 파일명이며, ‘annoLevel’은 파일이 원시 말뭉치에 해당하므로 일괄적으로 ‘원시’ 정보로 기재하였다. ‘sampling’은 표본 수집 방식으로, 본 사업에서는 게시자를 모집하여 게시자가 직접 작성한 게시글 중 무작위 추출을 통해 게시물을 선별하였으므로, ’게시자 모집 후 무작위 추출‘로 기재하였다.

'class'는 구축 계획에 따른 장르로, 본 사업에서는 누리 소통망, 블로그, 게시판, 리뷰의 네 가지 유형에 대한 정보이다.

'sourceInfo'는 앞서 설명한 바와 같이, 게시물의 기본 정보에 해당하는 내용으로, 웹 게시물 수집 시 수집과 함께 추출되는 정보이다. 'title'은 "나의 결혼식 후기", "한끼 뚝딱 해결하기 좋은 마약계란 만들기" 등과 같은 웹 게시물의 제목에 해당하는 정보이다. 'author'는 게시자명에 해당되는 정보로, 게시물 작성 시 작성자의 닉네임, 별명 등의 정보이다. 'publisher'는 수집 사이트 정보로, 'fileInfo'의 'class'와 유사하나, 구체적인 사이트 정보에 대한 내용이다. 즉, 누리 소통망의 경우, '트위터', '인스타그램', '페이스북' 등의 구체적인 사이트 정보가 포함되며, 블로그의 경우는 '네이버 블로그', '다음 블로그', 게시판은 '네이버 카페', '다음 카페' 등 구체적인 사이트 정보를 입력하였다. 'date' 및 'dateCrawl'은 각각 게시물을 작성한 게시 날짜와 게시물을 실제 수집한 수집 날짜를 의미한다. 'url' 주소는 게시물 고유의 웹 주소로, "twitter.com/D3LLM0Nt/status/1160118499471048706"와 같은 유형으로 구성되어 있다. 'view'는 해당 게시물의 조회 수 정보이다. 게시물에 따라 0 또는 2534 등 게시물을 조회한 정보가 입력된다. 'profileInfo'는 앞서 설명한 바와 같이 저작권 이용 허락 계약서에 기재된 성별 및 연령 정보를 입력하였다.

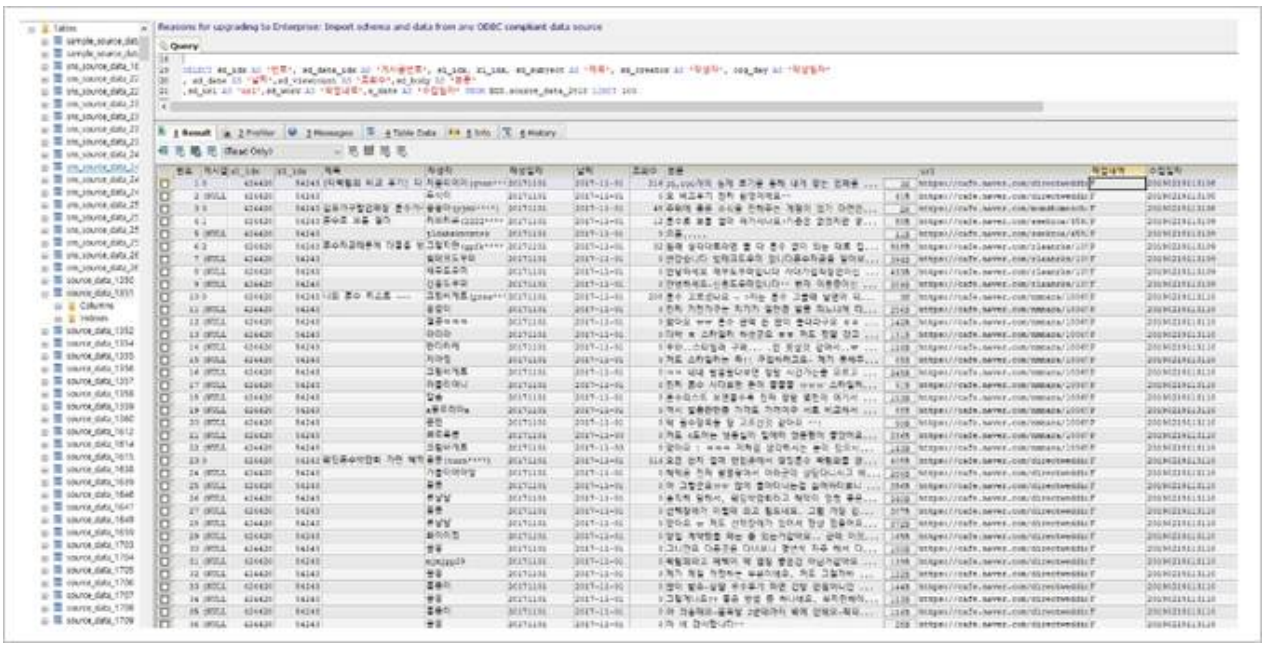
'text' 정보는 게시글 본문에 해당하는 내용으로, 게시자가 작성한 글의 내용을 입력하며, 온라인상 제시된 정보상 줄바꿈 경계 정보에 따라 <p> 태그를 부착하였다. 문장 경계 정보에 따른 <s> 태그 부착 및 단락 경계 정보에 따른 <p> 태그 부착은 온라인상 게시글을 작성하는 특성상 현실적으로 적용하기 어려웠다. 왜냐하면, 단락 또는 문장에 따라 줄바꿈을 하기 보다는 게시글 중 주요 내용을 부각시키거나 습관적으로 단락이 아닌 주요 문구 중심으로 줄바꿈을 사용하는 경향이 있어 문장 경계 및 단락 경계 정보가 실제 단락 또는 문장의 의미와 다르게 사용되었기 때문이다. 따라서, 실제 'text'는 게시글 내 줄바꿈 정보를 기반으로 <p> 태그만을 부착하는 방식으로 진행되었다.

2.2. 헤더와 마크업 부착

앞서 명시한 바와 같이, 웹 원시 자료 수집 시 수집과 동시에 사이트명, 제목, 본문, 작성자, 작성일, 웹 주소 등 메타 정보를 포함하여 수집하는 시스템을 이용하였으므로 해당 정보는 별도의 확인 작업을 거치지 않고 자동 축적된 정보를 이용함으로써 부착 과정에서 오류 및 오기의 가능성을 최소화하였다. 다만, 추가 요청 정보인 'profileInfo'

정보에 해당하는 게시자 성별 및 연령은 저작권 이용 허락 계약서에 기재된 정보를 수기로 입력하였으며, 정보 입력의 정확성 확인을 위해 정보 입력 후 입력 오류 검증 확인 작업을 진행하였다. 확보된 정보는 통합 파일 서버 내 자동 자료 입출력 도구를 이용해 자동으로 헤더 부착 및 메타 정보를 xml 형태 및 SJML 형태로 변환하였다.

<그림 13> 자동 헤더 및 마크업 자료 입출력 화면



2.3. 웹 말뭉치 구축 결과

웹 말뭉치 구축 지침에 따라 파일명과 헤더 및 마크업을 부착한 실제 결과물은 다음과 같다.

1) 원문 자료 파일명 부여 결과

원문 자료는 내용의 확인 및 분석이 용이한 엑셀 파일 형식으로 구성하였다. 파일명은 지침에 따라 매체별로 구분하여 부여하였으며, 실무에서 원문 자료 파일 확인의 용이성과 납품 회차별 구분을 위해 1개 파일에서 최대 3개 파일로 분할하였다. 최종 원문 자료 파일명은 다음과 같이 부여되었다.

<표 6> 원문 자료 파일명 부여 결과

매체	파일명
누리 소통망(S)	ESOR1910000001
	ESOR1910000002
	ESOR1910000003
블로그(B)	EBOR1920000001
	EBOR1920000002
게시판(P)	EPOR1930000001
리뷰(R)	EROR1940000001
	EROR1940000002

2) 원시 말뭉치 파일명 부여 결과

원시 말뭉치 파일은 SJML 파일로 구성하였으며, 파일명은 지침에 다라 매체별로 구분하여 부여하였다. SJML 파일의 수는 최종 목표 원문 자료의 수와 동일하다.

<표 7> 원시 말뭉치 파일명 부여 결과

매체	파일명	파일 수(건)
누리 소통망(S)	ESRW1910000001 ~ ESRW1912000000	2,000,000
블로그(B)	EBRW1920000001 ~ EBRW1920010000	10,000
게시판(P)	EPRW1930000001 ~ EPRW1930010000	10,000
리뷰(R)	ERRW1940000001 ~ ERRW1940100000	100,000

3) 원시 말뭉치 구축 결과

실제 SJML 원시 말뭉치 파일을 구축한 결과는 매체별로 다음과 같다.

<그림 14> 누리 소통망 원시 말뭉치 구축 결과

```
<?xml version='1.0' encoding='UTF-8'?>
<SJML>
  <header>
```

```
    <fileInfo>
      <fileId>ESRW1910635175</fileId>
      <annoLevel>원시</annoLevel>
      <sampling>게시자 모집 후 무작위 추출</sampling>
      <class>누리소통망</class>
    </fileInfo>
    <sourceInfo>
      <title>근데 내가 전후 분량을 안봐서 그런건지 그냥 재미가 없는건지 모르겠네</title>
      <author>서줄무늬</author>
      <publisher>twitter</publisher>
      <date>2019-08-08 02:09:47</date>
      <dateCrawl>2019-08-14 13:51:03</dateCrawl>
      <url>https://twitter.com/JGWiegler/status/1159149649271218177</url>
      <view>0</view>
    </sourceInfo>
    <profileInfo>
      <personId sex="" age="29"></personId>
    </profileInfo>
  </header>
  <text>
    <p>근데 내가 전후 분량을 안봐서 그런건지 그냥 재미가 없는건지 모르겠네</p>
  </text>
</SJML>
```

<그림 15> 블로그 원시 말뭉치 구축 결과

```
<?xml version='1.0' encoding='UTF-8'?>
<SJML>
  <header>
    <fileInfo>
      <fileId>EBRW1920005235</fileId>
      <annoLevel>원시</annoLevel>
      <sampling>게시자 모집 후 무작위 추출</sampling>
      <class>블로그</class>
    </fileInfo>
    <sourceInfo>
      <title>***</title>
      <author>하리</author>
```

```
        <publisher>네이버 블로그</publisher>
        <date>2018-07-18 23:57:39</date>
        <dateCrawl>2019-08-02 21:38:23</dateCrawl>

<url>http://blog.naver.com/PostView.nhn?blogId=nba1030&logNo=221321836559</url>
        <view>0</view>
      </sourceInfo>
      <profileInfo>
        <personId sex="F" age="25"></personId>
      </profileInfo>
    </header>
    <text>
      <p>여름에 해볼 작업 출발지점(기초 아이디어)_피지컬 컴퓨팅과 결합</p>
      <p>2인용 벤치X 1인용 2+@개</p>
      <p>musical bench.ino</p>
      <p>내 컴퓨터 저장</p>
      <p>네이버 클라우드 저장</p>
      <p>+ RGB LED 조명</p>
      <p>스크래치 종이는 단순하고 예전에 해봤으니 응용해서 접목 가능할듯...</p>
    </text>
  </SJML>
```

<그림 16> 게시판 원시 말뭉치 구축 결과

```
<?xml version='1.0' encoding='UTF-8'?>
<SJML>
  <header>
    <fileInfo>
      <fileId>EPRW1930003236</fileId>
      <annoLevel>원시</annoLevel>
      <sampling>게시자 모집 후 무작위 추출</sampling>
      <class>게시글</class>
    </fileInfo>
    <sourceInfo>
      <title>요새 할 짓 없어서 에세이도 읽고 있는데</title>
      <author>쩝쩝접</author>
      <publisher>orbi.kr</publisher>
      <date>2019-07-25 03:27:56</date>
```

<dateCrawl>2019-08-02 13:52:55</dateCrawl>
<url>https://www.orbi.kr/00023785384/</url>
<view>179</view>
</sourceInfo>
<profileInfo>
<personId sex="M" age="25"></personId>
</profileInfo>
</header>
<text>
김웅 검사가 쓴
검사내전 책을 읽던 중에
P.206
``우리나라에서는 적지 않은 사람들이 경쟁 위주의 교육을 학교폭력의 원인으로 지목한다. 또 대부분의 학교에서는 조심스럽지만 일관되게 피해자의 유별난 성격도 한몫을 했다고 항변한다. 그러나 학교폭력의 원인을 일방적으로 사회에 돌리고, 피해자의 탓으로 모는 것은 비과학적인 무지의 소산이다.``
``사회과학자들은 이미 이 부분에 대해서 수많은 연구를 해왔다. 연구 결과 청소년 폭력의 원인에 대해서 크게 두 가지 대표적인 학설이 정립되어 있다. 로버트 애그뉴(Robert Agnew)가 주장한 `일반긴장이론(General Strain Theory)`, 고트프레드슨과 허쉬(Gottfredson and Hirschi)가 주장한 `범죄의 일반이론(General Theory of Crime)`이 그것이다.``
P.207
``흔히 범죄나 청소년 범죄를 사회 탓으로 돌린다. 경쟁 위주의 입시 등으로 원인을 돌리는 것은 여러모로 편리하고 저항도 덜 받는다. 모두에게 책임을 돌리게 되면 아무도 책임을 지지 않아도 되기 때문이다. 구조적인 문제라는 피상적인 말잔치로 포장되는 것이다. 하지만 일반이론은 이를 정면으로 반박하고 있다. 그 때문에 처음 발표된 1990년대부터 지속적으로 비난을 받아왔다. 하지만 1993년 그래스믹(Grasmick)의 연구에서부터 2005년 맥도날드(McDonald)의 연구에 이르기까지, 이를 반박하기 위해 실시된 여러 조사들에서 오히려 일반이론이 옳다는 것이 입증되고 있다. 따라서 학교폭력의 원인을 경쟁이나 사회 탓으로만 돌리는 것은 비과학적이라고 말할 수 있다.``
이 부분 보고서 많이 놀랐음
사회과학 쪽 공부 따로 해보고는 싶네
</text>
</SJML>

<그림 17> 리뷰 원시 말뭉치 구축 결과

```
<?xml version='1.0' encoding='UTF-8'?>
<SJML>
  <header>
```

```
      <fileInfo>
        <fileId>ERRW1940098344</fileId>
        <annoLevel>원시</annoLevel>
        <sampling>게시자 모집 후 무작위 추출</sampling>
        <class>리뷰</class>
      </fileInfo>
      <sourceInfo>
        <title>명동/남산맛집 1978테라스 _ 연말모임장소 추천!</title>
        <author>빵소</author>
        <publisher>네이버 블로그</publisher>
        <date>2018-12-22 18:16:06</date>
        <dateCrawl>2019-08-27 18:47:12</dateCrawl>

<url>http://blog.naver.com/PostView.nhn?blogId=sojeong1224&logNo=221424969069</url>
        <view>0</view>
      </sourceInfo>
      <profileInfo>
        <personId sex="F" age="23"></personId>
      </profileInfo>
    </header>
  <text>
      <p>남산맛집으로 오래전부터 유명한 #1978테라스</p>
      <p>남산돈까스맛집 바로 옆!</p>
      <p>밖에도 메뉴판이 있어서 좋더라구요+_+</p>
      <p>건물이 엄청 넓어요 지하도 있고 이층도 있고..!</p>
      <p>저는 일층 창가자리로~</p>
      <p>먼저 에이드들이 나오고</p>
      <p>점심시간때라 그런지 손님들이 참 많았어요</p>
      <p>랍스타빅스타라는 랍스타오일파스타에요</p>
      <p>새우,올리브,홍합 등 해산물이 들어간오일파스타!</p>
      <p>랍스타등껍질은 토핑용이라 아쉽던 ㅎ..ㅎ</p>
      <p>그래도 살이 들어있는데 엄청 맛있어요!</p>
      <p>바로 요것! 정말 맛이 없을 수 없던 ㅠㅠ</p>
      <p>파스타도 간이 잘 맞고 맛있더라구요!</p>
      <p>그리고 제일 좋았던 베리베리가든피자:)</p>
      <p>블루베리+리코타치즈 조합은 당연 사랑이에여 ><</p>
      <p>화덕피자라 도우가 얇은데 쫄깃하니 너무 맛있던!</p>
      <p>빵순이라 그런지 파스타보다 피자가 더 손이 가더라구요 흐흐</p>
```

```
      <p>요렇게 한상차림 배불리 잘 먹었습니다 ㅎㅎ</p>
      <p>단짠단짠 느낌이라 물리지 않고 잘 먹었던!</p>
      <p>사장님이 너무 잘먹어서 보기 좋다며 덕담까지 해주셨어요 호호.. 가족들끼리 외식하기에도 좋고 남산 앞이니 데이트하기도 좋은 장소같아요!</p>
      <p>1978테라스</p>
      <p>서울특별시 중구 예장동 838</p>
      <p>모바일에서 작성된 글입니다.</p>
      <p>블로그앱에서 보기</p>
      <p>블로그앱 설치 URL을</p>
      <p>네이버앱 알림으로 전송했습니다.</p>
      <p>알림이 오지 않는다면,</p>
      <p>네이버앱을 최신버전으로 업데이트 하거나,</p>
      <p>로그아웃상태인지 확인해주세요</p>
      <p>다시 보내기</p>
      <p>확인</p>
      <p>닫기</p>
    </text>
  </SJML>
```

제 4 장

결 론

1. 사업 요약

본 사업은 인공 지능 산업 발전을 위한 대규모 고품질 우리말 자원 수요 증대를 위해 추진한 대규모 웹 언어 자료를 수집하는 사업으로, 누리 소통망(SNS), 블로그, 게시판 등에서 실제로 사용된 웹 언어 자료를 모아 말뭉치로 구축하였다.

1.1. 참여자 선정 및 모집

본 사업에 참여하기 위해서는 참여자 본인이 누리 소통망, 블로그, 게시판 등에 직접 작성한 웹 게시물을 최소 1건 이상 보유해야 하며, 참여 대상이 되는 웹 게시물에 대한 저작권 이용 허락 계약을 체결할 의향이 있어야 한다는 두 가지 사항을 충족하여야 한다. 따라서 참여자 본인이 직접 작성한 웹 게시물이 있는지 확인이 필요하며, 사업 참여 의향을 확인하는 것이 필수적이다. 한정된 사업 기간 내에 본 사업 목적에 적합한 참여자 모집을 완료하기 위하여 다각도로 참여자를 모집하는 방안이 필요하였다. 이에 적합한 참여자 모집을 위해 다음과 같은 세 가지 모집 방법을 통하여 참여자 모집을 진행하였다.

첫 번째, 사업 수행 기관의 보유 데이터를 활용하여 웹 게시글을 확인, 적합한 게시자를 찾는 방법이다.

두 번째, 사업 수행 기관의 온라인 패널인 URX 패널을 이용해 참여를 유도하는 방법으로, 패널 회원 본인이 직접 참여하는 방법 외에 주변의 참여 조건 대상자를 찾아 추천 및 독려하는 방법을 함께 이용하였다.

세 번째, 국립국어원 홈페이지 및 사업 수행 기관인 메트릭스코퍼레이션 홈페이지의 팝업창 안내를 통해 자발적인 참여를 유도하는 방법이다.

1.2. 저작권 이용 허락 계약 체결

본 사업의 참여자는 게시물을 작성한 원문 자료 저작권자로, 참여자와의 웹 게시물에 대한 저작권 이용 허락 계약 체결이 필요하다. 직접 대면하지 않고 다수의 참여자를 대상으로 신속하고 용이하게 계약 체결을 진행하기 위해, 전자 계약 시스템을 이용하여 저작권 이용 허락 계약을 진행하였다. 웹 게시물 저작권 이용 허락 계약을 완료한 참여자는 2,065명이며, 최종 참여자 대상으로 소정의 사례금을 지급하였다.

1.3. 웹 원문 자료 수집

본 사업에 해당하는 수집 대상 매체는 누리 소통망, 블로그, 게시판, 리뷰 네 가지 매체이다. 누리 소통망의 경우, 트위터, 인스타그램, 페이스북 등의 사이트로, 참여자의 계정을 대상으로 참여자가 직접 작성한 게시물을 수집하는 방식으로 진행하였다. 블로그는 네이버 블로그, 다음 블로그, 티스토리 등의 사이트로, 참여자가 운영하는 블로그의 주소 내 게시된 게시물을 수집하는 방식으로 진행하였다. 게시판의 경우, 누리 소통망이나 블로그와는 달리 다양한 커뮤니티 사이트에서 게시글이 발생하고 있으며, 특정 계정이나 주소 대상이 아닌, 해당 사이트 내 참여자의 게시글이 있는 게시판 내 게시물 단위별로 수집을 진행였으며 다른 매체 대비 많은 시간과 노력이 소요되었다.

웹 원문 자료 수집 결과로 총 2,878,606건이 수집되었으며, 비적합 게시글 정제 및 참여자별 수집 한도 제한 기준에 따라 758,606건을 삭제하여, 최종적으로 2,120,000건의 웹 원문 자료를 확보하였다.

1.4. 웹 말뭉치 구축

웹 말뭉치 구축은 구축 지침에 따라 파일명 및 헤더와 마크업 부착을 진행하였다. 파일명은 매체 및 장르와 말뭉치 유형, 일련번호 정보를 포함하여 총 14자리로 부여하였으며, 인코딩 방식은 UTF-8으로 진행하였다. 헤더와 마크업 부착은 'header' 요소 내 'fileInfo'와 'sourceInfo'에 해당되는 11가지 메타 정보에 대한 태그 부착과 'text' 요소로 구성된다. 최종 원문 자료 파일은 누리 소통망 3개, 블로그 2개, 게시판 1개, 리뷰 2개의 총 8개 파일로 구성하였으며, 원시 말뭉치 파일은 누리 소통망 2,000,000개, 블로그 10,000개, 게시판 10,000개, 리뷰 100,000개 파일로 구축을 완료하였다.

2. 사업의 의의 및 기대 효과

'웹 말뭉치 구축' 사업을 통한 기대 효과는 다음과 같다.

□ 웹 말뭉치 구축 기대 효과

- 민간에서 활용 가능한 국가 공공재로서의 말뭉치 확대 구축 및 국어 자원의 활용도와 가치 향상에 기여

- 4차 산업혁명 대비 기반 기술 개발 및 인공 지능 기술 개발, 활용을 위한 대규모 말뭉치 구축으로 국어 자원의 활용도와 가치 제고
- 민간 공유를 통해 언어 인공 지능 등 관련 산업 활용을 위한 기반을 마련하고 국어 및 국어문화 연구, 국어정책 수립의 기초 자료로 활용

본 사업과 함께 추진되고 있는 국어 말뭉치 구축 사업을 통해 인공 지능 스피커, 대화형 로봇, 로봇 개인 비서 등 한국어 인공 지능의 성능을 향상시킬 것으로 기대되며, 향후 4차 산업혁명 시대의 인공 지능 서비스 개발 및 기술 혁신을 위한 중요 자료가 될 전망이다.

3. 사업 추진 관련 제언

본 사업은 웹 게시자로부터 직접 작성한 게시물을 대상으로 사업 참여 의향을 이끌어 내야하며, 게시물에 대한 저작권 이용 허락 계약을 체결해야 하는 어려움이 있는 과제이다. 2,000명 이상의 다수 참여자를 모집하고 참여자로부터 저작권 이용 허락 계약을 체결하기 위해서는 참여 대상자의 사업 수행에 대한 인지와 과업의 필요성에 대한 공감이 필수적이다. 또한, 다양한 사이트를 대상으로 2,120,000건의 웹 게시물을 안정적으로 수집할 수 있는 전문적인 기술력이 필요하며, 과업의 원활한 수행 및 유동적인 대응을 위해 자체적인 수집 기술을 보유하고 있는가가 중요한 요소이다. 본 과업 추진 결과를 바탕으로 하여 원활한 과업 추진을 위해 다음과 같은 사항을 제언하고자 한다.

3.1. 사업에 대한 사전 안내

본 사업 진행 시, 대규모의 참여자를 모집하는 과정에서 수행 기관에 대한 확인 및 사업 내용에 대한 구체적인 사실 확인 문의가 다수 발생하였다. 주로 문의가 발생했던 내용은 웹 언어 자료를 수집하는 목적과 실제 어떤 범위 및 방식으로 구축이 되는지에 대한 문의가 많았으며, 저작권의 이용 허락 범위와 허락 기간 외에 개인 정보 노출 우려에 대한 문의도 발생하였다. 참여자 모집 과정에서 수행 기관 및 담당자 연락처를 명시하였음에도 불구하고 참여 의향자가 보다 정확한 사실 확인을 위해 수행 기관이 아닌 국립국어원으로 직접 문의를 하는 사례가 많아, 참여자 모집이 본격적으로 시작

된 후 약 1개월이 지난 시점에서 국립국어원에서 제공하는 공문 및 ‘웹 말뭉치 구축 사업 안내문’이 배포되었다. 국립국어원에서 직접 제공한 공문과 사업 안내문은 참여자들의 신뢰성 확보와 사업의 정확한 내용 확인에 매우 유용하게 활용되었다. 따라서 참여자 모집 이전에 사업 안내문 사전 안내 및 홍보를 충분하게 진행한 후에 본격적인 참여자 모집이 이루어지는 것이 필요하다.

3.2. 충분한 예산 확보

본 사업의 취지에 대한 자세한 설명이 이루어지는 경우, 대부분의 참여자가 사업의 목적과 의미에 대해 공감하였다. 국어에 대한 의미와 중요성에 대한 깊은 국민적 공감대가 형성되어 있기 때문이라고 해석된다. 그러나 사업에 대한 구체적인 목적 및 취지 설명과 활용도에 대한 충분한 설명이 선행되어야만 가능하며, 참여자가 작성한 다수의 게시물에 대해 저작권 이용 허락을 구하기까지 많은 노력과 시간이 필요하다. 참여자 모집을 위해 방대한 데이터를 분석하고, 선정된 대상자에게 본 사업에 대해 설명을 진행하기까지 대규모 인력이 투입되었으며, 많은 절차와 시간이 소요되었다.

또한, 원시 말뭉치 구축을 위한 자료 수집 및 처리 과정상 일반적인 온라인 게시물 분석을 위한 수집과는 달리 많은 수집 시간과 수집 서버의 부담이 필요하였다. 자료 수집 시 줄바꿈 정보 등 국어 분석에 필요한 정보가 누락 없이 수집이 되어야 하며, 온라인상 육안으로는 보이지 않으나 웹 페이지 정보에 숨어 있어 기계적 수집 시 함께 수집되는 정보인, 기타 정보나 표 및 그림 등의 정보를 제외하고 수집을 진행해야 한다. 따라서 수집의 속도와 서버 부담을 고려하지 않고 필요한 정보와 불필요한 정보를 선택적으로 수집해야 하는 어려움이 있었다.

실제로 많은 비용이 투입된 부분은 참여자 모집과 수집 과정이었으며, 참여자 모집 및 자료 처리를 위한 충분한 예산 확보가 필요한 부분이다.

3.3. 참여자 부담 요소 검토

본 사업 추진 중 가장 많은 노력이 필요했던 부분은 저작권 이용 허락에 대한 계약 체결이다. 참여자 입장에서는 다소 생소한 저작권에 대한 이용 허락 계약을 체결해야 한다는 부담이 있었으며 특히, 이용 허락 기간 및 중지 방법에 대한 문의가 가장 많이 발생하였다. 저작권 이용 허락 계약서 제3조에 “대상 저작물의 이용 허락 기간은 … 2035. 12. 31.까지로 하며, 권리자가 이용 허락을 중지하고자 하는 의사를 밝히지 아니

하면 이용 허락이 5년 단위로 자동 갱신된다. 권리자가 이용 허락 중지 의사를 밝히면 그 의사 내용에 따라 이용 허락을 중지하여야 하며, 그렇지 아니하면 이용 허락 내용이 유지된다"는 내용이 제시되어 있다. 웹 게시물 내용 특성상 일상, 육아, 연애 등 개인 사생활에 대한 내용이 포함되어 있는 경우가 많아, 이용 허락 기간에 대해 부담을 토로하는 참여자들의 문의가 발생하였다. 물론, 이용 허락 중지 의사를 밝히는 경우 자동 갱신이 되는 것을 중지할 수 있으나, 국립국어원에 직접 연락하여 이용 허락 중지 의사를 밝히기 용이하지 않은 점을 부담으로 느끼고 있었다. 본 사업에서는 저작권에 대한 준영구적인 이용 허락을 목적으로 하고 있으므로 이용 허락 기간에 대한 변경이나 조정은 현실적으로 어려울 것으로 보이나, 참여자가 준영구적인 이용 허락에 대한 부담을 느끼지 않도록 하기 위하여, 민간에서 활용을 위한 배포 시 참여자의 정보를 추적할 수 있는 게시자명, 웹 주소(URL) 등의 정보를 제외하는 등 세심한 배려가 필요하다.

<부록1> 국가 언어 자원(말뭉치) 구축 및 활용 저작권 이용 허락 계약서

국가 언어 자원(말뭉치) 구축 및 활용 저작권 이용 허락 계약서

저작자 및 저작권 이용허락자 _____(이하 “권리자”이라 함)와 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 저작재산권 이용허락과 관련하여 권리자와 이용자 사이의 권리관계를 명확히 하는 것을 목적으로 한다.

제2조 (계약의 대상)

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

저작물:

저작자:

종별: ☑ 어문저작물

권리: ☑ 복제권, ☑ 전송권, ☑ 배포권, ☑ 2차적저작물작성권

※ 저작권 이용허락 대상 권리의 내용

1. 국립국어원 및 국립국어원이 발주한 용역 사업의 수행자가 대상저작물을 일정한 형식으로 전자적 기록 매체에 담아 보존하는 일
2. 국립국어원 및 국립국어원이 발주한 용역 사업의 수행자가 자모, 음절, 어휘, 어절, 구절, 문장 및 텍스트 단위의 국어 연구와 언어 정보 처리 분야에 응용하기 위해 대상저작물을 복제·변형(목차·머리말·도표·그림·각주 등의 편집 및 삭제, 언어 단위별 분리, 언어적·비언어적 정보 부착 등)하는 일
3. 국립국어원 및 국립국어원이 발주한 용역 사업의 수행자가 대상저작물 및 그 복제·변형물을 연구 및 기술 개발용으로 학계·연구기관·산업체 등이 이용할 수 있도

록 제공·배포하는 일

4. 대상저작물 및 그 복제·변형물을 제공·배포받은 학계·연구기관·산업체 등이 국어 연구와 언어 정보 처리 분야 응용을 위하여 대상저작물 및 그 복제·변형물을 분석 및 처리하여 사용하는 것을 허락하는 일

제3조 (이용허락 기간)

대상저작물의 이용허락 기간은 계약 체결일부터 2035. 12. 31.까지로 하며, 권리자가 이용허락을 중지하고자 하는 의사를 밝히지 아니하면 이용허락이 5년 단위로 자동 갱신된다. 권리자가 이용허락 중지 의사를 밝히면 그 의사 내용에 따라 이용허락을 중지하여야 하며, 그렇지 아니하면 이용허락 내용이 유지된다.

제4조 (권리자의 의무)

(1) 권리자는 이용자에게 대상저작물에 관하여 본 계약서 제2조에 따른 저작재산권을 이용할 권리를 제3조의 기간 동안 비독점적으로 허락한다.

(2) 권리자는 이용자에게 계약 체결일로부터 10일 이내에 대상저작물의 이용을 위해 필요한 상당한 자료를 인도하여야 한다. 다만, 대상저작물이 한국저작권위원회에 등록되어 있지 않은 경우 이용자가 요청하면 이용허락자는 대상저작물의 저작재산권을 등록한 후 위 의무를 이행한다.

(3) 권리자는 대상저작물에 제3자의 이용허락권, 질권 등이 존재하는 경우, 이용자에게 그 사실을 사전에 알려야 한다.

(4) 권리자는 대상저작물의 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 대상저작물을 제3조의 이용허락 기간 동안 제2조의 이용 허락을 받은 범위 내에서 비독점적으로 자유롭게 이용할 수 있다.

(2) 이용료는 설정하지 아니한다.

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

(4) 이용자는 대상저작물의 이용함에 있어서 저작인격권을 침해하지 아니한다. 다만, 제2조에 따른 목적에 한하여 제2조에 따른 변형을 할 수 있으며, 대상저작물의 본질적인 내용을 변경하지 않는 범위 내에서 권리자에게 그 사실을 사전에 고지한 후 사소한 수정 및 편집을 할 수 있다.

제6조 (확인 및 보증)

(1) 권리자는 이용자에게 다음 각 호의 사항을 확인하고 보증한다.

1. 대상저작물의 저작권이용허락을 체결하는 데 필요한 권리 및 권한을 적법하게 보유하고 있다는 것
2. 대상저작물의 내용이 제3자의 저작권, 상표권, 인격권을 비롯한 일체의 권리를 침해하지 아니한다는 것
3. 대상저작물에 대하여 이용자에게 사전에 알린 제3자의 권리 외에는 이용자의 이용을 제한할 수 있는 부담이 더 이상 존재하지 아니한다는 것

(2) 이용자는 권리자에게 다음 각호의 사항을 확인하고 보증한다.

1. 대상저작물에 적용된 이용허락 조건에 의해서만 대상저작물 재이용을 허락할 것
2. 대상저작물을 제3자의 명예권을 비롯한 인격적 권리를 침해하는 방식으로 이용하지 아니할 것

제7조 (계약내용의 변경)

본 계약 내용 중 일부를 변경할 필요가 있는 경우에는 권리자와 이용자의 서면합의에 의하여 변경할 수 있으며, 그 서면합의에서 달리 정함이 없는 한, 변경된 사항은 그 다음 날부터 효력을 가진다.

제8조 (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 정

하여 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.

제9조 (손해배상)

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

제10조 (비용의 부담)

계약 체결에 따른 비용은 이용자가 전부 부담한다.

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 내용 및 대상저작물의 내용을, 상대방의 서면에 의한 승낙 없이 제3자에게 공개하여서는 아니 된다.

제13조 (기타부속합의)

(1) 권리자와 이용자는 본 계약의 내용을 보충하거나, 이 계약에서 정하지 아니한 사항을 규정하기 위하여 부속합의서를 작성할 수 있다.

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

제14조 (계약의 해석 및 보완)

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

제15조 (계약 효력 발생일)

본 계약의 효력은 계약 체결일로부터 발생한다.

2019 년 _____월 _____일

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

<부록2> 웹 말뭉치 구축 지침

웹 말뭉치 구축 지침

1. 파일 형식 및 개요

1.1. 파일명 부여 방식

첫째 자리 :구분	둘째 자리 :매체 및 장르 대분류	셋째, 넷째 자리 :말뭉치 유형 구분	다섯째, 여섯째 자리 :구축년도	8자리 일련번호
E: 웹 말뭉치	S: 누리소통망 B: 블로그 P: 게시글 R: 리뷰	OR: 원문 자료 RW: 원시 말뭉치	19	########

- 예시

· ESOR1900000001.txt 누리소통망 원문 자료 1번째 파일(파일 형식은 원문 자료에 따라 변경 가능)

· ESRW1900000001.sjml 누리소통망 원시 말뭉치 1번째 파일

· EBRW1900000011.sjml 블로그 원시 말뭉치 11번째 파일

1.2. 인코딩

- UTF-8

2. 헤더와 마크업

2.1. SJML의 기본 구조

가. SJML의 요소　　사이트, 제목, 날짜, URL

<fileInfo>	<fileId>	파일의 고유 식별자(말뭉치 파일명)
	<annoLevel>	주석 수준: 원시
	<sampling>	표본 수집 방식
	<class>	구축 계획에 따른 장르 분류: 누리소통망, 블로그, 게시글, 리뷰
<sourceInfo>	<title>	제목
	<author>	게시자
	<publisher>	수집 사이트
	<date>	게시 날짜
	<dateCrawl>	수집 날짜
	<url>	URL 주소
	<view>	조회수
<profileInfo> *필수 항목 아님	<personId> 외	게시자 성별(sex) 게시자 연령(age)

(1) <header>: 파일의 메타 정보를 담는 요소

(2) <text>: 텍스트 정보 포함

- <p>: 단락 경계 정보 또는 줄 바꿈 경계 정보

나. SJML의 형식

- SJML의 형식은 XML의 형식을 따른다.
- 형식 선언: 문서의 첫 번째 행에 아래와 같이 문서 형식을 선언한다.

```
<?xml version="1.0" encoding="UTF-8"?>
```

- 들여쓰기: 들여쓰기(소프트탭: 스페이스 2개)를 통해 요소의 계층을 시각화한다.

```
<SJML>
  <header>
    <fileInfo>
      <fileId>ESRW1900000001</fileId>
      ....
    </fileInfo>
  </header>
  <text>
   ....
  </text>
</SJML>
```

2.2. 마크업된 웹 말뭉치 파일 예시

```
<?xml version="1.0" encoding="UTF-8"?>
<SJML>
  <header>
    <fileInfo>
      <fileId>ESRW1900000001</fileId>
      <annoLevel>원시</annoLevel>
      <sampling>패널 대상 게시자 모집 후 무작위 추출</sampling>
      <class>누리소통망</class>
    </fileInfo>
    <sourceInfo>
      <title>TWEET : “사람들은 캐나다에 여자 프로축구 팀이 한 팀이라도 있는 것처럼 ‘여자팀에서 뛰라’고 얘기한다. 그런데 우리 나라에는 프로 팀이 한 팀도 없다. 어떻게 공정한 기회가 주어진다는 건</title>
      <author>미깡맨</author>
      <publisher>트위터</publisher>
      <date>2018. 05. 04. 10:40:09</date>
      <dateCrawl>2019. 06. 25. 16:33:20</dateCrawl>
      <url>https://twitter.com/raul07duff11/status/992217284218245122</url>
```

```
            <view>0</view>
          </sourceInfo>
        </header>
        <text>
          <p>“사람들은 캐나다에 여자 프로축구 팀이 한 팀이라도 있는 것처럼 ‘여자팀에서 뛰라’고 얘기한다. 그런데 우리 나라에는 프로 팀이 한 팀도 없다. 어떻게 공정한 기회가 주어진다는 건가? 그녀를 뛰게 하라”</p>
    </text>
</SJML>
```

<Abstract>

Web Corpus Construction

The purpose of this project is to build nationally a corpus that is analyzed in a computer-readable form by collecting web language data from social networking sites(SNS), blogs and web-bulletin boards to enhance the utilization and value of Korean language resources for the purpose of developing infrastructure technologies for the fourth Industrial Revolution, and developing and utilizing artificial intelligence technologies. The results of the following major tasks are summarized as follows.

First, we recruited publishers to participate in the web corpus construction project and carried out active promotional activities for major web post creators, online panels and general web publishers. And in the process of entering into a copyright permit agreement with participants, we have secured awareness of Web corpus construction and interest in the need for the corpus construction project.

Second, various types of postings were collected from various sites, such as SNS, blogs, web bulletin boards, and reviews, which are actively used in Korea as original web documents. The collected web original data collected was 2,000,000 SNS, 10,000 blogs, 10,000 web bulletin boards, and 100,000 reviews, which collected a total of 2,120,000 original web documents and built a web corpus. In particular, various web language data according to site characteristics were obtained by collecting original web text data from various sites by media.

Third, in accordance with the Web corpus deployment guidelines stipulated

in the project, we established the corpus as a public good that is easy to convert and compatible in the private sector. By mechanically collecting the information contained in the original web document and establishing it as a web corpus in the form posted on the site without any distortion or distortion, it could be used as a material to study the language used by actual web postwriters.

Keywords : web corpus, web corpus construction, inartificial web corpus

사업 책임자 이영희((주)메트릭스코퍼레이션 인터넷사업부 이사)
사업 참여자 김수진((주)메트릭스코퍼레이션 인터넷사업부 팀장)
김홍건((주)메트릭스코퍼레이션 인터넷사업부 부장)
정종서((주)메트릭스코퍼레이션 인터넷사업부 대리)
김예슬((주)메트릭스코퍼레이션 인터넷사업부 주임)
박수정((주)메트릭스코퍼레이션 인터넷사업부 연구원)
이상훈((주)메트릭스코퍼레이션 연구부 차장)
신성호((주)메트릭스코퍼레이션 연구부 과장)
신현주((주)메트릭스코퍼레이션 개발운영팀 팀장)
채윤철((주)메트릭스코퍼레이션 전산팀 팀장)
하지영((주)메트릭스코퍼레이션 실사팀 팀장)
윤지은((주)메트릭스코퍼레이션 실사팀 과장)
담당 연구원 이승재(국립국어원 언어정보과장)
홍혜진(국립국어원 언어정보과 학예연구관)

발행인: 국립국어원장
발행처: 국립국어원
서울시 강서구 금낭화로 154
전화 02-2669-9775, 전송 02-2669-9757
인쇄일: 2019년 11월 23일
발행일: 2019년 11월 23일
인 쇄: (주)타라그래픽스

※ 이 책은 국립국어원의 용역비로 수행한 '웹 말뭉치 구축' 사업의 결과물을 발간한 것입니다.