

인공지능 학습용 데이터 구축·활용 가이드라인

< 낚시성 기사 탐지 데이터 >

인공지능 학습용 데이터 구축·활용 가이드라인	사업 총괄	비플라이소프트(주)
	데이터 설계	비플라이소프트(주)
	데이터 획득(수집)	비플라이소프트(주)
	데이터 정제	서울시스템(주) / 비플라이소프트(주)
	데이터 가공	(주)미디어그룹사람과숲 / (주)솔트룩스
	데이터 검사	비플라이소프트(주) / (주)오피니언라이브
	크라우드 소싱	비플라이소프트(주) : 가공데이터 품질검수 서울시스템(주) : 데이터 정제 (주)미디어그룹사람과숲 : 1세부 가공/검수 (주)솔트룩스 : 2세부 가공/검수 (주)오피니언라이브 : 데이터 검수
	저작도구 개발	(주)미디어그룹사람과숲
	AI모델 개발	비플라이소프트(주) / 고려대 산학협력단
	데이터 활용	비플라이소프트(주)
데이터 구축·활용 가이드라인 작성	비플라이소프트(주)	양 대 현 허 재 혁(고려대학교 산학협력단)
	오피니언라이브	김 성 철
데이터 구축·활용 가이드라인 버전	ver 1.0 ('22. 12. 20)	

제·개정 이력

개정번호	개정번호	개정내용	작성자	검토자	승인자
v1.0	2022.12.30	최초 개정	양대현	이연	최재웅
			.		

목 차

제1장 데이터 명세	1
1. 데이터 정보 요약	1
2. 데이터 포맷	1
3. 어노테이션 포맷.....	6
4. 데이터 구성	11
5. 데이터 통계	12
6. 원시데이터 특성	15
7. 기타 정보	16
제2장 데이터 구축	17
1. 데이터 구축 개요	17
2. 임무 정의	17
3. 획득(수집)	18
4. 정제	21
5. 가공(라벨링)	26
6. 검사	36
7. 학습 모델	48
제3장 데이터 활용	50
1. 데이터 활용	50
2. 응용 서비스	51
3. 응용 서비스 개발	51
4. 기술 지원	52

<부록> 인공지능 학습용 데이터 도메인 용어 정의

제1장 데이터 명세 정보

1. 데이터 정보 요약

데이터 명	(2-025-146) 낚시성 기사 탐지 데이터	
활용 분야	자연어 연구 분야, 언론 분야	
데이터 요약	자연어처리(NLP) 및 딥러닝 기술을 기반으로 낚시성 기사 등 낮은 품질의 기사로 인해 시간과 비용 낭비, 사실 왜곡 등 사회적 문제해결을 위한 인공지능 학습용 데이터	
데이터 출처	중앙일보 외 13개 언론사 기사 (정치, 경제, 사회, 생활&문화, IT&과학, 세계, 연예 7대 카테고리 기사)	
데이터 통계	데이터 구축 규모	텍스트 기사 733,427 건
	데이터 분포 (총분석, 균등성, 편향성 여부 확인)	// 다양성(통계) - 다양한 매체유형별 데이터 수집(온라인 기사) <input type="radio"/> 매체 구분 별 수집데이터 현황 - 전국종합일간(42.5%), 경제일간(29.0%), 통신사(7.0%), 지역종합일간(6.2%), 종합·전문주간(6.2%), 전문일간(6.2%), 인터넷신문(2.8%) // 다양성(요건) : 카테고리별로 편향되지 않게 구성/학습요건 총족 <input type="radio"/> 기사 카테고리별 원천/라벨링 데이터 현황 - 정치(13.6%), 경제(14.2%), 사회(19.8%), 생활&문화(10.8%), IT&과학(13.8%), 세계(15.5%), 연예(12.3%)
데이터 이력	배포 버전	v1.0
	개정 이력	신규
	작성자/배포자	양대현/구경린

2. 데이터 포맷

구분	획득(수집) 단계	정제 단계	가공(라벨링) 단계
데이터 구분	원시데이터	원천데이터	가공데이터
데이터 형태	텍스트	텍스트	텍스트
데이터 포맷	XML	JSON	JSON

2.1 원시데이터(XML : version - 1.0, encoding – utf-8, type – text)

구분	No	속성명	속성 및 내용
필수	1	press	언론사 한글명
필수	2	nsid	언론사 기사ID
필수	3	action	I(insert)/U(update) 구분
필수	4	title	기사 제목
선택	5	subTitle	기사 부제목
필수	6	date	기사 발행일자
필수	7	time	기사 발행시간
필수	8	author	기자명
필수	9	content	기사 본문
필수	10	category	기사 카테고리
필수	11	url	기사 온라인 URL

- 원시데이터 예시(뉴스1 '세계' 카테고리 기사)

NIA-2-025 수집 > 2-025-146.RawData > 6.Global > news1

```

이름 ^
  ^ <?xml version="1.0" encoding="utf-8"?>
<item type="text">
<press>뉴스1</press>
<nsid>0003821565</nsid>
<action>I</action>
<title><![CDATA["美中 합의는 일종의 틀...中 농산물구매 일방적 의무 아냐"]]></title>
<subTitle><![CDATA[웨이 중국국제경제교류중심(CCIEE) 부이사장 발언]]></subTitle>
<date>2020-01-20</date>
<time>21:57:05</time>
<author><![CDATA[이창규 기자]]></author>
<content><![CDATA[
중국이 지난주 미국과 서명한 1단계 무역 합의의 이행할지에 대한 의구심이 계속되고 있는 가운데 중국의 미국산 농산물을 구매는 일방적인 의무가 아니라는 주장이 제기돼 주목된다.

20일 사우스차이나모닝포스트(SCMP)에 따르면, 중국 정부의 경제 자문 기구인 중국국제경제교류중심(CCIEE)이 주최한 한 세미나에서 웨이 지안궈 CCIEE 부이사장은 이같이 말했다.

그는 "먼저 우리는 구매를 확대하기 위해 최선을 다할 것이나 미국도 경쟁력 있는 가격으로 좋은 물건을 제공해야 한다"며 "미국산 농산물을 구매하겠다는 중국의 약속은 일방적인 의무가 아니다"라고 강조했다.

그러면서 "중국 정부는 수입업체가 미국 제품을 구매하도록 강요할 수 없다"며 "미국 기업들이 시장에서 주문을 두고 경쟁을 해야 한다"고 덧붙였다.

도널드 트럼프 미국 대통령과 류허(劉鶴) 중국 부총리는 지난 14일 워싱턴에서 1단계 무역 합의에 서명했다. 중국은 2020년 1월1일부터 2021년 12월31일까지 2년 동안 320억달러 규모의 미국산 농산물을 포함해 2000억달러 규모의 미국산 제품을 구매하기로 약속했다.

웨이 부이사장은 "합의는 일종의 틀이며 그 틀은 시장경제 원리에 근거한다"며 "미국은 임의로 가격을 올리거나 중국에 열등한 품질의 제품을 판매할 수 없고 그럴 경우 국내 기업과 소비자들이 받아들이지 않을 것"이라고 말했다.]]></content>
<category name="미국 · 캐나다" code="32" />
<url href="https://www.news1.kr/articles/3821565" />
</item>

```

2.2 원천데이터(JSON)

- 1세부(제목과 본문의 불일치 기사) : JSON 파일 구조

구분		속성명	타입	필수여부	설명
1		sourceDataInfo	object		원천데이터 정보
1-1		newsID	string	Y	기사ID
1-2		newsCategory	string	Y	기사카테고리
1-3		newsSubcategory	string	N	기사하위카테고리
1-4		newsTitle	string	Y	기사제목
1-5		newsSubTitle	string	N	기사부제목
1-6		newsContent	string	Y	기사본문
1-7		partNum	string	Y	세부번호(P1-1 세부)
1-8		useType	number	Y	용도유형(0-낚시성,1-비낚시성)
1-9		processType	string	Y	가공유형(A-자동생성,D-직접생성)
1-10		processPattern	string	Y	가공패턴(11,12,13,14,15,16)
1-11		processLevel	string	Y	가공난이도(상/중/하)
1-12		sentenceCount	number	Y	문장수(원문기사의 문장수)
1-13		sentenceInfo	object		문장정보
	1-13-1	sentenceNo	number	Y	문장번호
	1-13-2	sentenceContent	String	Y	문장내용
	1-13-3	sentenceSize	number	Y	문장크기(글자수)

- 2세부(본문의 도메인 일관성 부족 기사) : JSON 파일 구조

구분		속성명	타입	필수여부	설명
1		sourceDataInfo	object		원천데이터 정보
1-1		newsID	string	Y	기사ID
1-2		newsCategory	string	Y	기사카테고리
1-3		newsSubcategory	string	N	기사하위카테고리
1-4		newsTitle	string	Y	기사제목
1-5		newsSubTitle	string	N	기사부제목
1-6		newsContent	string	Y	기사본문
1-7		partNum	string	Y	세부번호(P2-2 세부)
1-8		useType	number	Y	용도유형(0-낚시성,1-비낚시성)
1-9		processType	string	Y	가공유형(A-자동생성,D-직접생성)
1-10		processPattern	string	Y	가공패턴(21,22,23,24)
1-11		processLevel	string	Y	가공난이도(상/중/하)
1-12		sentenceCount	number	Y	문장수(원문기사의 문장수)
1-13		processSentencenum	number	Y	가공문장수(대체/직접작성 본문 문장 수)
1-14		sentenceInfo	object		문장정보
	1-14-1	sentenceNo	number	Y	문장번호
	1-14-2	sentenceContent	String	Y	문장내용
	1-14-3	sentenceSize	number	Y	문장크기(글자수)

※ 원시데이터에 대한 정제(문장분리 등) 및 가공을 위한 정보 항목을 추가하여 원천데이터 JSON 파일 생성

-. 1세부(제목과 본문의 불일치 기사) : 원천데이터 파일(JSON) 예시

```
{
  "sourceDataInfo": {
    "newsID": "EC_M02_160728",
    "newsCategory": "경제",
    "newsSubcategory": "종목",
    "newsTitle": "진양홀딩스 자회사 진양AMC, 주당 1002 원 현금 중간배당 실시",
    "newsSubTitle": "",
    "newsContent": "진양홀딩스 자회사인 진양AMC가 현금 중간배당을 실시한다. 진양홀딩스는 자회사인 진양AMC가 보통주 1주당 1002 원의 현금 중간배당을 결정했다고 20일 공시했다. 배당금총액은 10억원이다. 배당금지급 예정일자는 2022년 8월 9일이다. 진양홀딩스 관계자는 데일리임팩트에 진양AMC는 진양홀딩스가 100% 출자한 자회사로서 배당금 전액이 진양홀딩스로 귀속된다"고 전했다. 한편 진양홀딩스는 최근 주가안정을 통한 주주가치 제고 및 자금운용의 효율성 제고를 목적으로 50억원 규모의 자사주 신탁계약 체결을 공시한 바 있다.",
    "partNum": "P1",
    "useType": 0,
    "processType": "D",
    "processPattern": "15",
    "processLevel": "중",
    "sentenceCount": 6,
    "sentenceInfo": [
      {
        "sentenceNo": 1,
        "sentenceContent": "진양홀딩스 자회사인 진양AMC가 현금 중간배당을 실시한다.",
        "sentenceSize": 32
      },
      {
        "sentenceNo": 2,
        "sentenceContent": "진양홀딩스는 자회사인 진양AMC가 보통주 1주당 1002 원의 현금 중간배당을 결정했다고 20일 공시했다.",
        "sentenceSize": 58
      },
      {
        "sentenceNo": 3,
        "sentenceContent": "배당금총액은 10억원이다.",
        "sentenceSize": 14
      },
      {
        "sentenceNo": 4,
        "sentenceContent": "배당금지급 예정일자는 2022년 8월 9일이다.",
        "sentenceSize": 26
      },
      {
        "sentenceNo": 5,
        "sentenceContent": "진양홀딩스 관계자는 데일리임팩트에 진양AMC는 진양홀딩스가 100% 출자한 자회사로서 배당금 전액이 진양홀딩스로 귀속된다"고 전했다.",
        "sentenceSize": 77
      },
      {
        "sentenceNo": 6,
        "sentenceContent": "한편 진양홀딩스는 최근 주가안정을 통한 주주가치 제고 및 자금운용의 효율성 제고를 목적으로 50억원 규모의 자사주 신탁계약 체결을 공시한 바 있다.",
        "sentenceSize": 82
      }
    ]
  }
}
```

-. 2세부(본문의 도메인 일관성 부족 기사) : 원천데이터 파일(JSON) 예시

```
{
  "sourceDataInfo": {
    "newsID": "EC_M05_619673",
    "newsCategory": "경제",
    "newsSubcategory": "경제일반",
    "newsTitle": "원·달러 환율, 연고점 또 경신...1188.7 원(종합)",
    "newsSubTitle": "",
    "newsContent": "30일 원·달러 환율이 또 연고점을 경신했다. 이날 서울 외환시장에서 달러 대비 원화 환율은 전날 종가보다 4.7원 오른 달러당 1188.7 원에 거래를 마쳤다. 이는 종가 기준으로 지난해 9월 9일(1189.1 원) 이후 1년여 만의 최고치다. 올해 종가 기준 연고점은 지난달 28일 기록한 1184.4 원이다. 한편 오후 3시 30분 현재 원·엔 재정환율은 100엔당 1069.55 원이다. 전날 오후 3시 30분 기준가(1058.44 원)에서 11.11 원 올랐다.",
    "partNum": "P2",
    "useType": 0,
    "processType": "D",
    "processPattern": "23",
    "processLevel": "중",
    "sentenceCount": 6,
    "processSentencenum": 2,
    "sentenceInfo": [
      {
        "sentenceNo": 1,
        "sentenceContent": "30일 원·달러 환율이 또 연고점을 경신했다.",
        "sentenceSize": 25
      },
      {
        "sentenceNo": 2,
        "sentenceContent": "이날 서울 외환시장에서 달러 대비 원화 환율은 전날 종가보다 4.7원 오른 달러당 1188.7 원에 거래를 마쳤다.",
        "sentenceSize": 63
      },
      {
        "sentenceNo": 3,
        "sentenceContent": "이는 종가 기준으로 지난해 9월 9일(1189.1 원) 이후 1년여 만의 최고치다.",
        "sentenceSize": 45
      },
      {
        "sentenceNo": 4,
        "sentenceContent": "올해 종가 기준 연고점은 지난달 28일 기록한 1184.4 원이다.",
        "sentenceSize": 36
      },
      {
        "sentenceNo": 5,
        "sentenceContent": "한편 오후 3시 30분 현재 원·엔 재정환율은 100엔당 1069.55 원이다.",
        "sentenceSize": 43
      },
      {
        "sentenceNo": 6,
        "sentenceContent": "전날 오후 3시 30분 기준가(1058.44 원)에서 11.11 원 올랐다.",
        "sentenceSize": 40
      }
    ]
  }
}
```

3. 어노테이션 포맷

3.1 어노테이션 JSON 파일 구조

- 1세부(제목과 본문의 불일치 기사) JSON 파일 구조

구분		속성명	타입	필수여부	설명
1		sourceDataInfo	object		원천데이터 정보
	1-1	newsID	string	Y	기사ID
	1-2	newsCategory	string	Y	기사카테고리
	1-3	newsSubcategory	string	N	기사하위카테고리
	1-4	newsTitle	string	Y	기사제목
	1-5	newsSubTitle	string	N	기사부제목
	1-6	newsContent	string	Y	기사본문
	1-7	partNum	string	Y	세부번호(P1-1 세부)
	1-8	useType	number	Y	용도유형(0-낚시성,1-비낚시성)
	1-9	processType	string	Y	가공유형(A-자동생성,D-직접생성)
	1-10	processPattern	string	Y	가공패턴(11,12,13,14,15,16)
	1-11	processLevel	string	Y	가공난이도(상/중/하)
	1-12	sentenceCount	number	Y	문장수(원문기사의 문장수)
	1-13	sentenceInfo	object		문장정보
	1-13-1	sentenceNo	number	Y	문장번호
	1-13-2	sentenceContent	String	Y	문장내용
	1-13-3	sentenceSize	number	Y	문장크기(글자수)
2		labeledDataInfo	object		라벨링데이터 정보
	2-1	newTitle	string	Y	새제목
	2-2	clickbaitClass	number	Y	낚시성기사분류
	2-3	referSentenceInfo	object		참조문장정보
	2-3-1	sentenceNo	number	Y	문장번호
	2-3-2	referSentenceYn	string	Y	참조문장여부

- 2세부(본문의 도메인 일관성 부족 기사) JSON 파일 구조

구분		속성명	타입	필수여부	설명
1		sourceDataInfo	object		원천데이터 정보
	1-1	newsID	string	Y	기사ID
	1-2	newsCategory	string	Y	기사카테고리
	1-3	newsSubcategory	string	N	기사하위카테고리
	1-4	newsTitle	string	Y	기사제목
	1-5	newsSubTitle	string	N	기사부제목
	1-6	newsContent	string	Y	기사본문
	1-7	partNum	string	Y	세부번호(P2-2 세부)
	1-8	useType	number	Y	용도유형(0-낚시성,1-비낚시성)
	1-9	processType	string	Y	가공유형(A-자동생성,D-직접생성)
	1-10	processPattern	string	Y	가공패턴(21,22,23,24)
	1-11	processLevel	string	Y	가공난이도(상/중/하)
	1-12	sentenceCount	number	Y	문장수(원문기사의 문장수)
	1-13	processSentencenum	number	Y	가공문장수(대체/직접작성 본문 문장 수)
	1-14	sentenceInfo	object		문장정보
	1-14-1	sentenceNo	number	Y	문장번호
	1-14-2	sentenceContent	String	Y	문장내용
	1-14-3	sentenceSize	number	Y	문장크기(글자수)
2		labeledDataInfo	object		라벨링데이터 정보
	2-1	processSentenceInfo	object		가공문장정보
	2-1-1	sentenceNo	number	Y	문장번호
	2-1-2	sentenceContent	string	Y	문장내용
	2-1-3	subjectConsistencyYn	string	Y	주제일관성여부
	2-2	clickbaitClass	number	Y	낚시성기사분류

3.2 어노테이션 JSON 파일 예시

- 1세부 JSON 파일 예시

```
{
  "sourceDataInfo": {
    "newsID": "ET_M13_169881",
    "newsCategory": "연예",
    "newsSubcategory": "연예일반",
    "newsTitle": "남우현, 오늘(4일) 소집해제[공식]",
    "newsSubTitle": "",
    "newsContent": "그룹 인피니트 남우현이 금일(4일) 소집해제 된다. 남우현이 군 복무를 마치고 팬들 곁으로 돌아온다. 2019년 사회복무요원으로 대체 복무를 해왔던 남우현이 이날 소집해제 된다. 그는 2014년 방송 촬영 중 어깨 부상을 당해 4급 판정을 받은 바 있다. 이로써 남우현은 김성규, 장동우, 이성열, 성종에 이어 인피니트 멤버 중 5번째로 국방의 의무를 마쳤다. 지난 2월 해병대에 입대한 엘(김명수)은 현재 군 복무 중이다. 한편 남우현은 2010년 그룹 인피니트로 데뷔한 이후 '광화문 연가', '그날들' 등의 뮤지컬에서 활약을 펼쳤다.",
    "partNum": "P1",
    "useType": 0,
    "processType": "D",
    "processPattern": "13",
    "processLevel": "하",
    "sentenceCount": 6,
    "sentencelInfo": [
      {
        "sentenceNo": 1,
        "sentenceContent": "그룹 인피니트 남우현이 금일(4일) 소집해제 된다.",
        "sentenceSize": 28
      },
      {
        "sentenceNo": 2,
        "sentenceContent": "남우현이 군 복무를 마치고 팬들 곁으로 돌아온다. 2019년 사회복무요원으로 대체 복무를 해왔던 남우현이 이날 소집해제 된다.",
        "sentenceSize": 70
      },
      {
        "sentenceNo": 3,
        "sentenceContent": "그는 2014년 방송 촬영 중 어깨 부상을 당해 4급 판정을 받은 바 있다.",
        "sentenceSize": 42
      },
      {
        "sentenceNo": 4,
        "sentenceContent": "이로써 남우현은 김성규, 장동우, 이성열, 성종에 이어 인피니트 멤버 중 5번째로 국방의 의무를 마쳤다.",
        "sentenceSize": 58
      },
      {
        "sentenceNo": 5,
        "sentenceContent": "지난 2월 해병대에 입대한 엘(김명수)은 현재 군 복무 중이다.",
        "sentenceSize": 35
      },
      {
        "sentenceNo": 6,
        "sentenceContent": "한편 남우현은 2010년 그룹 인피니트로 데뷔한 이후 '광화문 연가', '그날들' 등의 뮤지컬에서 활약을 펼쳤다.",
        "sentenceSize": 63
      }
    ],
    "labeledDataInfo": {
      "newTitle": "그룹 인피니트 섹시가이 남우현이 금일(4일) 소집해제!",
      "clickbaitClass": 0,
      "referSentencelInfo": [
        {
          "sentenceNo": 1,
          "referSentenceyn": "Y"
        }
      ]
    }
  }
}
```

- 1세부 JSON 파일 예시

```
{
  "sentenceNo": 2,
  "referSentenceyn": "N"
},
{
  "sentenceNo": 3,
  "referSentenceyn": "N"
},
{
  "sentenceNo": 4,
  "referSentenceyn": "N"
},
{
  "sentenceNo": 5,
  "referSentenceyn": "N"
},
{
  "sentenceNo": 6,
  "referSentenceyn": "N"
}
]
```

- 2세부 JSON 파일 예시

```
{
  "sourceDataInfo": {
    "newsID": "GB_M12_642931",
    "newsCategory": "세계",
    "newsSubcategory": "국제일반",
    "newsTitle": "일본 신규 확진 1만5812명...9일 연속 1만명 넘어",
    "newsSubTitle": "",
    "newsContent": "일본에서 코로나19 신규 확진자 수가 나흘 만에 최다 기록을 경신했다. 11일 NHK에 따르면 이날 일본의 코로나19 신규 확진자는 오후 8시까지 1만5812명으로 집계됐다. 지난 7일 기록했던 1만 5750명을 넘어선 수치이고, 일주일 전 같은曜일과 비교하면 1635명, 11.5% 증가했다. 일본의 신규 확진자는 이달 3일부터 9일 연속 1만 명을 웃돌고 있다. 일본의 누적 확진자는 107만 1410명으로 늘었다. 사망자는 13명 증가해 1만 5348명이 됐다.",
    "partNum": "P2",
    "useType": 0,
    "processType": "D",
    "processPattern": "22",
    "processLevel": "중",
    "sentenceCount": 6,
    "processSentencenum": 3,
    "sentenceInfo": [
      {
        "sentenceNo": 1,
        "sentenceContent": "일본에서 코로나19 신규 확진자 수가 나흘 만에 최다 기록을 경신했다.",
        "sentenceSize": 39
      },
      {
        "sentenceNo": 2,
        "sentenceContent": "11일 NHK에 따르면 이날 일본의 코로나19 신규 확진자는 오후 8시까지 1만5812명으로 집계됐다.",
        "sentenceSize": 57
      },
      {
        "sentenceNo": 3,
        "sentenceContent": "지난 7일 기록했던 1만 5750명을 넘어선 수치이고, 일주일 전 같은曜일과 비교하면 1635명, 11.5% 증가했다.",
        "sentenceSize": 67
      },
      {
        "sentenceNo": 4,
```

- 2세부 JSON 파일 예시

```

    "sentenceContent": "일본의 신규 확진자는 이달 3일부터 9일 연속 1만 명을 웃돌고 있다.",
    "sentenceSize": 39
},
{
    "sentenceNo": 5,
    "sentenceContent": "일본의 누적 확진자는 107만 1410명으로 늘었다.",
    "sentenceSize": 29
},
{
    "sentenceNo": 6,
    "sentenceContent": "사망자는 13명 증가해 1만 5348명이 됐다.",
    "sentenceSize": 26
}
],
},
"labeledDataInfo": {
    "processSentenceInfo": [
        {
            "sentenceNo": 1,
            "sentenceContent": "일본에서 코로나19 신규 확진자 수가 나흘 만에 최다 기록을 경신했다.",
            "subjectConsistencyYn": "Y"
        },
        {
            "sentenceNo": 2,
            "sentenceContent": "11일 NHK에 따르면 이날 일본의 코로나19 신규 확진자는 오후 8시까지 1만5812명으로 집계됐다.",
            "subjectConsistencyYn": "Y"
        },
        {
            "sentenceNo": 3,
            "sentenceContent": "지난 7일 기록했던 1만 5750명을 넘어선 수치이고, 일주일 전 같은 요일과 비교하면 1635명, 11.5% 증가했다.",
            "subjectConsistencyYn": "Y"
        },
        {
            "sentenceNo": 4,
            "sentenceContent": "일본의 신규 확진자는 이달 3일부터 9일 연속 1만 명을 웃돌고 있다.",
            "subjectConsistencyYn": "Y"
        },
        {
            "sentenceNo": 5,
            "sentenceContent": "일본의 누적 확진자는 107만 1410명으로 늘었다.",
            "subjectConsistencyYn": "Y"
        },
        {
            "sentenceNo": 6,
            "sentenceContent": "사망자는 13명 증가해 1만 5348명이 됐다.",
            "subjectConsistencyYn": "Y"
        },
        {
            "sentenceNo": 7,
            "sentenceContent": "한편, 한국에서 일본인 거주자가 가장 많은 이촌동에서, 진도건설의 `아크로제이` 오피스텔이 오늘부터 분양을 시작했다."
        },
        {
            "sentenceContent": "아크로제이는 지하 3층, 지상 18층의 주거용 오피스텔로 이촌역 3분 거리에 있어 교통이 아주 편리하다."
        }
    ],
    "subjectConsistencyYn": "N"
},
{
    "sentenceNo": 8,
    "sentenceContent": "아크로제이는 지하 3층, 지상 18층의 주거용 오피스텔로 이촌역 3분 거리에 있어 교통이 아주 편리하다."
},
    "subjectConsistencyYn": "N"
},
{

```

- 2세부 JSON 파일 예시

```
"sentenceNo": 9,  
  "sentenceContent": "생활 인프라는 물론이고, 세대마다 세탁기, 건조기, 에어컨, 스타일러 등의 옵션도 갖추고 있어 거주자와  
투자자 모두에서 매력적인 투자처가 될 것으로 기대한다.",  
  "subjectConsistencyYn": "N"  
}  
],  
  "clickbaitClass": 0  
}  
}
```

4. 데이터 구성

No	항목명	길이	항목 설명
1	partNum	2	세부구분 : P1 – 1세부, P2 – 2세부 → 풀더명 : P1 – Part1, P2 - Part2
2	useType	1	0 – 낚시성, 1 - 非낚시성
3	processType	1	A – 자동생성, D - 직접생성
4	newsCategory	2	7대 기사 카테고리명에 대한 영문약어명 EC(경제), ET(연예), GB(세계), IS(IT& 과학), LC(생활&문화), PO(정치), SO(사회)

1세부	2세부
<ul style="list-style-type: none"> · Part1 · Clickbait_Auto <ul style="list-style-type: none"> · EC · ET · GB · IS · LC · PO · SO · Clickbait_Direct <ul style="list-style-type: none"> · EC · ET · GB · IS · LC · PO · SO · NonClickbait_Auto <ul style="list-style-type: none"> · EC · ET · GB · IS · LC · PO · SO 	<ul style="list-style-type: none"> · Part2 · Clickbait_Auto <ul style="list-style-type: none"> · EC · ET · GB · IS · LC · PO · SO · Clickbait_Direct <ul style="list-style-type: none"> · EC · ET · GB · IS · LC · PO · SO · NonClickbait_Auto <ul style="list-style-type: none"> · EC · ET · GB · IS · LC · PO · SO

5. 데이터 통계

5.1 데이터 통계

5.1.1 데이터 구축 규모

● 데이터 구분별 구축 규모

과제번호	과제명	구분	포맷	형태	구축 규모(건)	용량
2-025-146	낚시성 기사 탐지 데이터	원시데이터	XML	텍스트	804,127	2.48 Gb
		원천데이터	JSON	텍스트	733,427	6.50 Gb
		가공데이터	JSON	텍스트	733,427	7.62 Gb

● 모델별 데이터셋 분리 현황

모델	Task	구분	Training	Validation	Test	Total	
HAND	낚시성 기사 분류	건수	291,466	36,434	36,433	364,333	
		비율	80%	10%	10%	100%	
BERT	주제분리 탐지	건수	295,275	36,910	36,909	369,094	
		비율	80%	10%	10%	100%	
Total		건수	586,741	73,344	73,342	733,427	
Total		비율	80%	10%	10%	100%	

- HAND : Hierarchical Attention Network for Fake News Detection

- BERT : Bidirectional Encoder Representation from Transformers

※ 주제분리 탐지는 BERT를 활용하여 주제분리(Topic Segmentation) 탐지 모델 개발 진행

5.1.2 데이터 분포

● 기사 카테고리별 데이터 분포

기사 카테고리	1세부		2세부		합계	
	건수	비율	건수	비율	건수	비율
정치	49,223	13.5%	50,551	13.7%	99,774	13.6%
경제	52,580	14.4%	51,353	13.9%	103,933	14.2%
사회	71,365	19.6%	73,915	20.0%	145,280	19.8%
생활&문화	39,551	10.9%	39,864	10.8%	79,415	10.8%
IT&과학	50,367	13.8%	50,854	13.8%	101,221	13.8%
세계	56,221	15.4%	57,172	15.5%	113,393	15.5%
연예	45,026	12.4%	45,385	12.3%	90,411	12.3%
합계	364,333	100.0%	369,094	100.0%	733,427	100.0%

※ 특정 기사 카테고리가 40%를 넘지 않아야 한다는 요구사항 충족

● 크라우드소싱을 통한 직접생성 가공패턴유형별 데이터 분포

세부구분	가공패턴유형	빈도	비율
1세부	의문 유발형(부호)	18,540	37.0%
	의문 유발형(은닉)	14,616	29.2%
	선정표현 사용형	4,418	8.8%
	속어/줄임말 사용형	4,669	9.3%
	사실 과대 표현형	5,280	10.5%
	의도적 주어 왜곡형	2,608	5.2%
	소 계	50,131	100.0%
2세부	상품 판매정보 노출 광고형	11,433	22.8%
	부동산 판매정보 노출 광고형	6,084	12.1%
	서비스 판매정보 노출 광고형	12,235	24.4%
	의도적 상황 왜곡/전환형	20,387	40.7%
	소 계	50,139	100.0%

5.1.3 기타 활용 통계

● 직접생성 기사 난이도별 현황

기사 난이도	1세부		2세부		합계	
	건수	비율	건수	비율	건수	비율
상	2,608	5.2%	20,387	40.7%	22,995	22.93%
중	9,949	19.8%	29,752	59.3%	39,701	39.59%
하	37,674	75.0%	0	0.0%	37,574	37.47%
합계	50,131	100%	50,139	100%	100,270	100%

- 낚시성/非낚시성 기사의 기사 난이도는 모두 '하'
- 1세부 : 상 - 의도적 주어 왜곡형, 중 - 사실 과대 표현형+속어/줄임말 사용형, 하 - 의문 유발형(부호/은닉)+선정표현 사용형
- 2세부 : 상 - 의도적 상황 왜곡/전환형, 중 - 상품/부동산/서비스 판매정보 노출 광고형

● 본문 뉴스 문장수 분포 현황

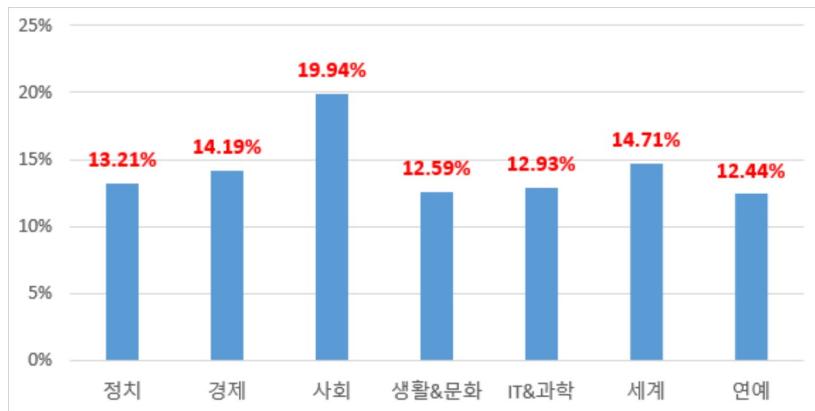
문장수 구간	1세부		2세부		합계	
	건수	비율	건수	비율	건수	비율
6 ~ 10	139,531	38.3%	167,490	45.4%	307,021	41.9%
11 ~ 15	94,965	26.1%	99,858	27.1%	194,823	26.6%
16 ~ 20	55,792	15.3%	49,528	13.4%	105,320	14.4%
21 ~ 25	35,168	9.7%	26,171	7.1%	61,339	8.4%
26 ~ 30	19,165	5.3%	13,258	3.6%	32,423	4.4%
31 ~ 35	10,358	2.8%	6,812	1.8%	17,170	2.3%
36 ~ 40	5,761	1.6%	3,703	1.0%	9,464	1.3%
41 ~ 45	2,838	0.8%	1,785	0.5%	4,623	0.6%
46 ~ 49	755	0.2%	489	0.1%	1,244	0.2%
합계	50,131	100%	50,139	100%	100,270	100%

- 낚시성 기사 자동생성 시 문장간 상호 대체를 고려하여 5문장 이하의 너무 짧은 기사는 배제 처리
- 문장이 너무 긴 경우 2가지 이상의 주제를 담고 있을 가능성이 높다는 점과 원천데이터 물량 확보 측면을 고려하여 50문장이 넘는 기사는 배제 처리

6. 원시데이터 특성

6.1.1 대상분류

중앙일보를 비롯한 14개 언론사가 2006년부터 2022년까지 정치, 경제, 사회, 생활&문화, IT&과학, 세계, 연예 카테고리에 보도했던 뉴스 기사는 '실제'에 해당



[그림] 원시데이터 기사 카테고리별 분포 현황

6.1.2 제약조건

- 기사 데이터 수집 시 기사 본문에 이미지나 동영상만 있고, 텍스트 내용이 없는 경우 '제약있음'에 해당하여 제외 처리
- 기사 데이터 수집 시 본문의 텍스트 길이가 200자 미만인 기사의 경우 '제약있음'에 해당하여 제외 처리

6.1.3 속성

- 기사 제목 어절 수 : 3어절 미만 기사 제거
- 기사 제목 글자 수 : 공백 포함 글자 수 기준 10자 미만 기사 제거
- 기사 본문 문장 수 : 5문장 이하인 기사, 50문장 이상 기사 제거

7. 기타 정보

7.1.1 포괄성

과제 요청사항에 해당하는 포털 뉴스의 주요 카테고리인 정치, 경제, 사회, 생활&문화, IT&과학, 세계 6대 카테고리에 자문위원회의 의견에 따라 연예 카테고리를 추가하고, 스포츠 카테고리는 기사 특성 상 과제의 대상으로 부적합하다는 자문위원회의 의견에 따라 대상 카테고리에서 제외 처리

7.1.2 독립성

- 종양일보를 비롯한 14개 언론사와 저작권 이용 허락을 받은 기사데이터를 구입하여 수집한 데이터로 AI Hub 공개를 통해 AI 연구 및 AI 응용서비스 개발 및 연구 목적으로 하는 사용자에게 공개되며, 공개되는 낚시성/非낚시성 기사 원문에 대한 상업적 사용은 불허
- 과제 특성을 고려하고, 과제조정위원회의 승인을 받아 작성되는 제목이나 본문 문장에서 개체명에 대한 마스킹 작업은 사용하지 않고, 대상에 대한 민감한 내용이 있는 경우에는 가공검수, 품질검수 과정에서 대체 작성하도록 권고하고, 반려 처리하여 민감정보가 양산되지 않도록 조치
- 기존 기사에 대한 제목, 본문 문장은 그대로 유지하고, 새롭게 작성되는 새제목과 본문 문장은 별도로 구분하여 항목 및 항목 값으로 관리하고, AI 학습모델에 활용되도록 독립적으로 구성하여 구축

7.1.3 유의사항

● 데이터 배포 시 파급효과

- 주관기관의 미디어 빅데이터를 이용한 서비스에 적용하여 수많은 뉴스 내용 중에 품질이 저하된 낚시성 기사에 대한 부가정보 및 필터링 기능을 제공하여 뉴스 스크랩 및 뉴스 소비 과정에서 효율적이고 선별적인 양질의 뉴스 탐색, 업무효율 및 만족도 향상 기여
- 언론사 뉴스기사에 대한 심의 시 정보 제공 도구로 활용하여 심의 업무 효율성 향상에 기여
- 언론/미디어 학문 분야에서 가짜뉴스, 낚시성 기사에 대한 테스트 도구로 활용하여 해당 분야에 대한 연구 활성화에 기여
- 뉴스를 기반으로 하는 다양한 플랫폼 사업자 및 포털, 언론산업 분야의 자율규제 도구로 선의적으로 활용하여 저널리즘 품질 개선 및 뉴스 수용자의 뉴스에 대한 신뢰도 향상에 기여

● 데이터 활용 시 유의사항

- AI 연구 및 AI 응용서비스 개발 및 연구 목적으로 하는 사용자에게 공개 제한
- 기사 원문에 대한 상업적 사용 불허

7.1.4 관련 연구

- FNC-1(FNC : Fake News Challenge), 2017
- 한국전자통신연구원, 페이크 뉴스 탐지 기술 동향과 시사점(2017)

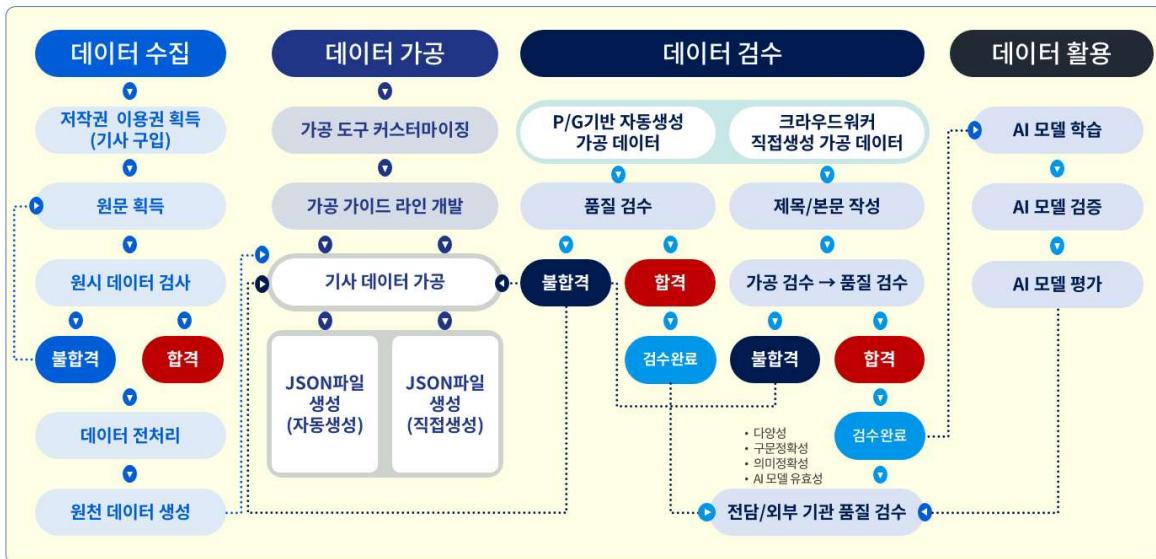
구분	방법론	적용 사례	분석 대상	장점	한계점
비기술적 접근	전문가 기반 탐지	Facebook Google	콘텐츠/생산자	공인 언론 매체의 사실 확인	<ul style="list-style-type: none"> - 소수 전문가에 의존 - 뉴스 전수 조사 불가
	집단지성 기반 탐지	Fakenewschecker Naver	콘텐츠	군중의 지혜 활용 및 실시간성	<ul style="list-style-type: none"> - 집단지성의 오판 가능성 - 적극적 참여가 전제 - 불충분한 탐지 능력
기술적 접근	인공지능 기반 탐지	FIB Deeplearning.org	콘텐츠	신속한 진단 능력	<ul style="list-style-type: none"> - 데이터 축적 필요 - 판단의 정확성 문제 - 정밀 조작 페이크 뉴스 진단 난해
	시맨틱 기반 탐지	Google Amazon.com	콘텐츠/생산자	결론 도출의 명료성	<ul style="list-style-type: none"> - 높은 지식 축적 비용 - 자연어 처리 및 형태소 분석의 정확성 문제
	이상 확산 패턴 감지	- (KAIST 연구 결과)	콘텐츠 확산 경로	높은 정확도	- 사후 분석
하이브리드 방식	Facebook	콘텐츠/생산자/ 확산 경로	기술간 장점 결합		

[그림] 페이크 뉴스 탐지 기법의 유형 및 개요

- DACON : 가짜뉴스 탐지를 위한 대회, 2020
- 논문 : 제목과 본문이 다른 가짜뉴스 탐지를 위한 계층적 딥러닝 모델 개발 및 가짜 뉴스 데이터셋 구축(2021, 세종대)

제2장 데이터 구축

1. 데이터 구축 개요



2. 임무 정의

2.1 임무 정의

- 자연어처리(NLP) 및 딥러닝 기술을 기반으로 낚시성 기사 등 낮은 품질의 기사로 인해 시간과 비용 낭비, 사실 왜곡 등 사회적 문제 해결을 위한 인공지능 학습용 데이터 구축
- 논쟁적 지가 탐지 서비스 개발 등에 활용할 수 있는 데이터 구축 및 학습모델 개발
- 학습모델 임무 : '낚시성 기사 탐지'에 대한 임무는 크게 분류(Classification), 그리고 주제 분리 탐지(Topic Segmentation Detection)로 정의

2.2 데이터 구축 유의사항

- 저작권법 적용 대상인 뉴스기사에 대한 낚시성 기사 새제목, 본문 내용 추가로 인한 문제

본 과제는 기사를 공급한 언론사가 저작권 또는 저작권 이용을 허락할 권리를 보유한 뉴스 기사를 정제 및 가공하는 과정에서 원문 기사에 대한 일부 변형 및 낚시성 기사의 새제목, 본문이 추가 될 수 있다.

▶ 대처방안

- 1) 자동생성을 통해 생성되는 새제목, 대체 문장은 원문기사는 유지한 상태에서 가공데이터에 별도 항목으로 분리하여 관리한다.
- 2) 직접생성을 통해 생성되는 새제목, 작성 문장도 동일하게 원문기사는 유지한 상태에서 가공데이터에 별도 항목으로 분리하여 관리한다.
- 3) 원문기사와 가공된 새제목, 본문문장을 혼용한 뉴스를 연구목적이 아닌 상업적 목적이나, 악용하는 경우 법적 책임은 모두 데이터 이용자에게 있음을 명시하여 데이터를 공개한다.

- 저작권법 적용 대상인 뉴스기사의 상업적/악의적 사용 문제

본 과제는 기사를 공급한 언론사가 저작권 또는 저작권 이용을 허락할 권리를 보유한 기사를 계약서에 명기된 '기사 이용 허락 범위'의 활용 및 공개 범위("한국지능정보사회진흥원 AI Hub 공개를 통해 AI 연구 및 AI 응용서비스 개발, 연구 목적으로 하는 사용자에게 공개되며, 공개되는 낚시성/非낚시성 기사 원문에 대한 상업적 사용은 불허함")를 벗어나 이용자(실 사용자 및 과제 참여자)가 상업적, 악의적으로 공개된 기사를 사용할 수 있는 위험에 노출되어 있다.

▶ 대처방안

- 1) 공개 데이터에 대한 활용 범위를 AI Hub에 명시하고, 이를 어길 시 법적 책임 내용을 명기한다.
- 2) 과제에 참여한 기관도 AI 연구, AI 응용서비스 개발, AI Hub 공개에 대한 하자보수 목적 외에는 사용을 금지한다는 서약서를 청구하여 주관기관이 보관한다.
- 3) 크라우드 소싱을 통해 본과제에 참여한 크라우드워커는 계약서 날인과 더불어 '보안서약서'의 서약 내용에 자필로 서명한다.

3. 획득(수집)

3.1 원시데이터 선정

- 원시데이터(언론사 뉴스 기사) 특성 현황
 - 기사 카테고리 매핑 필요성 확인

카테고리	그룹명	카테고리	그룹명	카테고리	그룹명
국방/외교	01.정치	건설	02.경제	교육/입시/NIE	03.사회
국회/정당	01.정치	경제	02.경제	노동/복지	03.사회
북한/한반도정세	01.정치	금융/재테크	02.경제	법원/검찰	03.사회
선거	01.정치	농수축산	02.경제	보건/의료	03.사회
정부/청와대	01.정치	부동산	02.경제	사건/사고	03.사회
정책	01.정치	산업/무역	02.경제	사회	03.사회
정치	01.정치	유통/소비자	02.경제	학교	03.사회
행정	01.정치	주식	02.경제	환경	03.사회
		증권	02.경제		
		창업	02.경제		
		창업/취업	02.경제		
		기업	02.경제		
		글로벌경제	02.경제		

- 6대 기사 카테고리에 대한 주관기관 보유 36만 건에 대한 분포 현황

도메인	기사건수	비율(%)
01.정치	44,690	6.3%
02.경제	102,309	14.4%
03.사회	102,103	14.4%
04.생활/문화	21,010	3.0%
05.IT/과학	11,093	1.6%
06.세계	8,288	1.2%
07.지역	205,606	28.9%
08.연예	16,389	2.3%
09.스포츠	13,268	1.9%
10.기타	186,318	26.2%
총합계	711,074	100.0%



→ 6대 카테고리에 대한 기사 비율은 약 40% 차지

→ 6대 카테고리별 뉴스 기사 비율이 상이하여 데이터 획득(수집) 시 카테고리별 기사 수집기간을 달리해서 획득(수집) 필요

● 뉴스 저작권

- 뉴스 저작권은 언론사 소유/계약서에 명기된 용도 외 상업적 사용 불가
- 본 과제는 언론사와의 협의를 통해 뉴스 저작권에 대한 본 과제내에서의 이용 허락을 획득한 기사들을 대상으로 계약금액을 주관사가 지급하여 데이터 획득(수집) 진행

제3조 (계약의 대상)

본 계약의 이용허락 대상이 되는 '대상저작물'에 대한 저작재산권 중 본 조에 명시한 이용허락 범위로 한다.

대상저작물: 서울경제 기사 중 권리자가 저작권 또는 저작권 이용을 허락할 권리를 보유한 기사

(1) 서울경제 : 사회 기사 5만건.

사회 > 사회일반 | 전국 | 사회이슈

< 기사 이용 허락 범위>

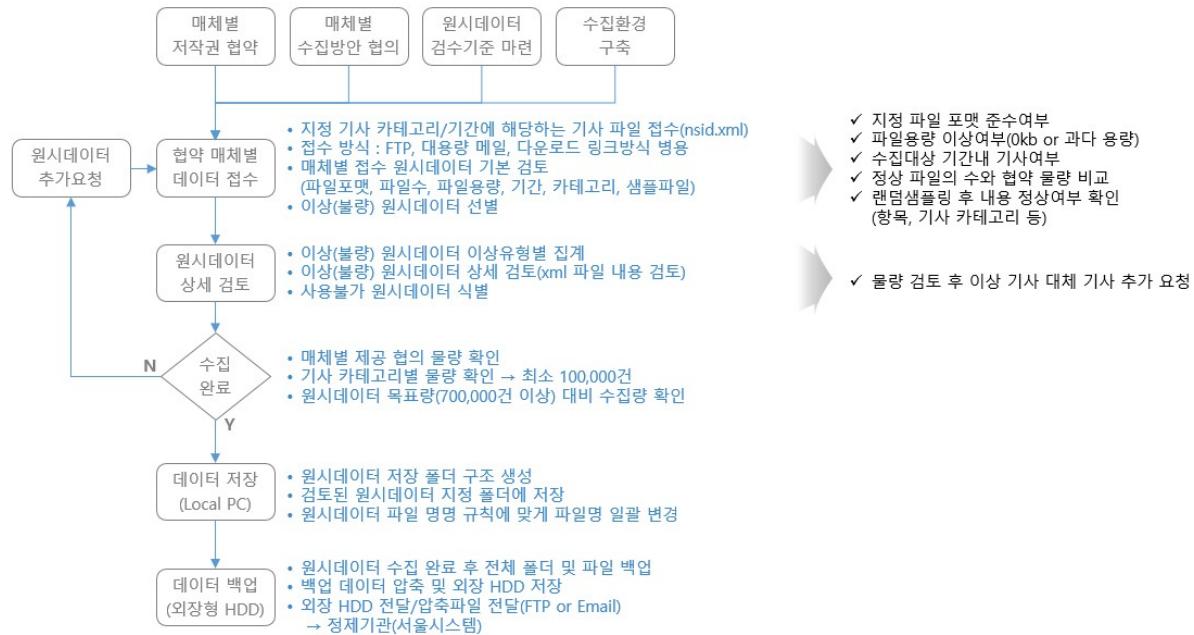
활용 및 공개 범위 : 한국지능정보사회진흥원 AI Hub 공개를 통해 AI 연구 및 AI 응용서비스 개발, 연구 목적으로 하는 사용자에게 공개되며, 공개되는 낚시성/非낚시성 기사 원문에 대한 상업적 사용은 불허함

→ 언론사(서울경제)와의 '기사 제공 계약서'의 내용 일부

● 원시데이터(XML) 구조

파일 포맷	○ 전 매체 제공 원시데이터 파일 포맷 표준화 : XML	
구성 항목	○ 뉴스 XML 파일내 처리 항목 표준화	
No 항목구분 항목 설명 항목 식별 및 처리		
0 파일 공통	XML 파일 공통	○ <?xml version="1.0" encoding="UTF-8"?> : xml 버전 및 인코딩 방식 표준화
1 데이터유형	원시데이터 유형	○ 일괄적으로 'text'로 값 처리 : <item type="text">
2 매체명	매체사의 한글명	○ XML 내 <press> ... </press> Tag내 매체사 한글명 제공
3 기사id	매체사의 기사식별자	○ XML 내 <nsid> ... </nsid> Tag내 해당 매체에서 부여한 기사ID 값 제공
4 액션유형	기사의 액션유형	○ XML 내 <action> ... </action> Tag내 기사ID 기준 최종 기사의 액션유형 값 제공 : I(Insert), U(Update) ○ 원시데이터 정제 후 원천데이터 생성 시 값 제거(미사용)
4 기사제목	기사의 제목	○ XML 내 <title> ... </title> Tag내 한글 기사 제목 제공 ○ <title> ... </title> Tag 내 존재하는 각종 Tag 제거 처리 ○ 제목에서 사용되는 각종 기호(‘, ‘, ..., ?, ! 등)는 그대로 텍스트 유지
부제목	기사의 부제목	○ XML 내 <subTitle> ... </subTitle> Tag내 한글 기사 부제목 제공 ○ 부제목이 존재 시 Tag는 제거하고, 부제목 각각은 줄바꿈으로 처리하여 제공
5 발행일자	기사의 발행일자	○ XML 내 <date> ... </date> Tag내 "YYYY-MM-DD" 포맷으로 기사 발행일자 값 제공
6 발행시간	기사의 발행시간	○ XML 내 <time> ... </time> Tag내 "HH:MM:SS" 포맷으로 기사 발행시간 값 제공
7 작성자(기자)	기사의 작성자	○ XML 내 <author> ... </author> Tag내 기자명 제공 ○ 기자가 다수인 경우 "기자명,기자명,..." 형식으로 값 제공
9 본문	기사의 본문 내용	○ XML 내 <content> ... </content> Tag내 기사 문단 분리된 본문 내용 텍스트(이미지, Tag 등 제거) 제공 ○ 문단 분리 Tag를 제외한 이미지, 기타 Tag는 제거하고 순수 기사 텍스트만 제공 ○ 처리가 가능한 경우 텍스트 본문 길이가 200자 미만인 기사는 제공대상 기사에서 제외 처리
10 카테고리	기사의 카테고리	○ <category> Tag내에 해당 매체 수집 기사 카테고리(대분류)의 하위 카테고리 코드(code), 값(name) 제공 ○ category는 매체와 협의된 대상 기사 카테고리(대분류)의 하위 카테고리 1개(대표 카테고리)만 제공
12 기사URL	매체사 온라인 URL 정보	○ XML 내 <url> Tag 내에 href 항목의 값에 해당 기사의 온라인 url 주소 값 제공

3.2 획득(수집) 절차



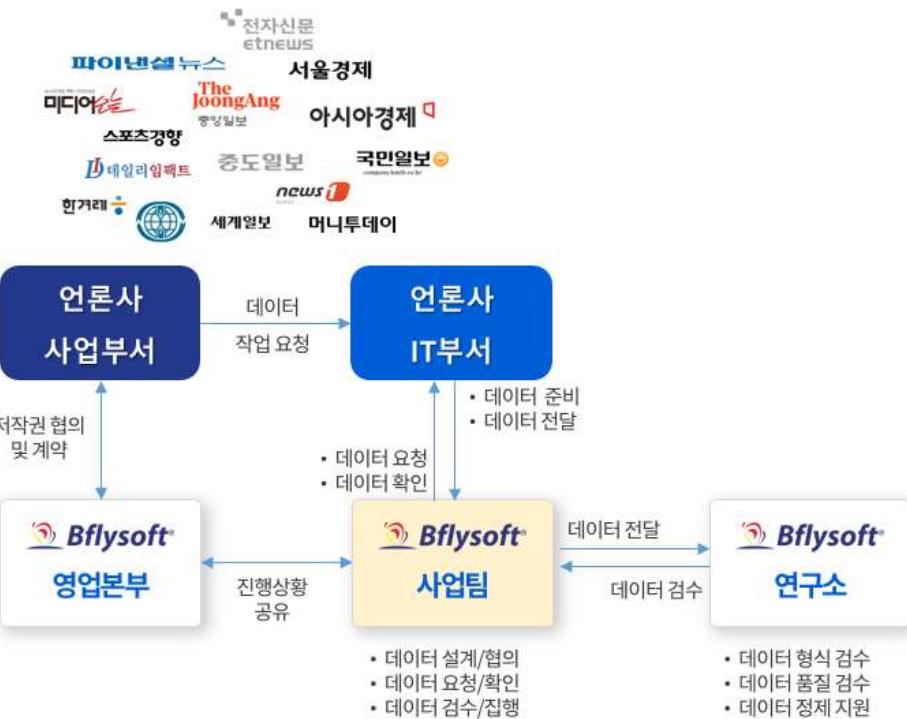
3.3 획득(수집) 기준

● 획득(수집) 기준

- RFP에서 제시한 6대 카테고리(정치, 경제, 사회, 생활&문화, IT&과학, 세계) + 자문위원회 추천 카테고리(연예) 추가
- 특정 기사 카테고리가 전체 데이터의 40%를 넘지 않도록 획득(수집)
- 총 30만 건 이상의 데이터 획득(수집)
- 데이터 구축 이후 AI Hub 공개 시 문제가 없도록 저작권 문제 및 법적검토를 통해 문제가 없는 데이터 획득(수집)
- 언론사간 데이터 중복을 배제하기 위해 가능하면 기사 카테고리 및 수집기간 중첩되지 않도록 획득(수집)

3.4 획득(수집) 조직

- 언론사 뉴스 데이터 획득(수집) 조직



3.5 획득(수집) 도구

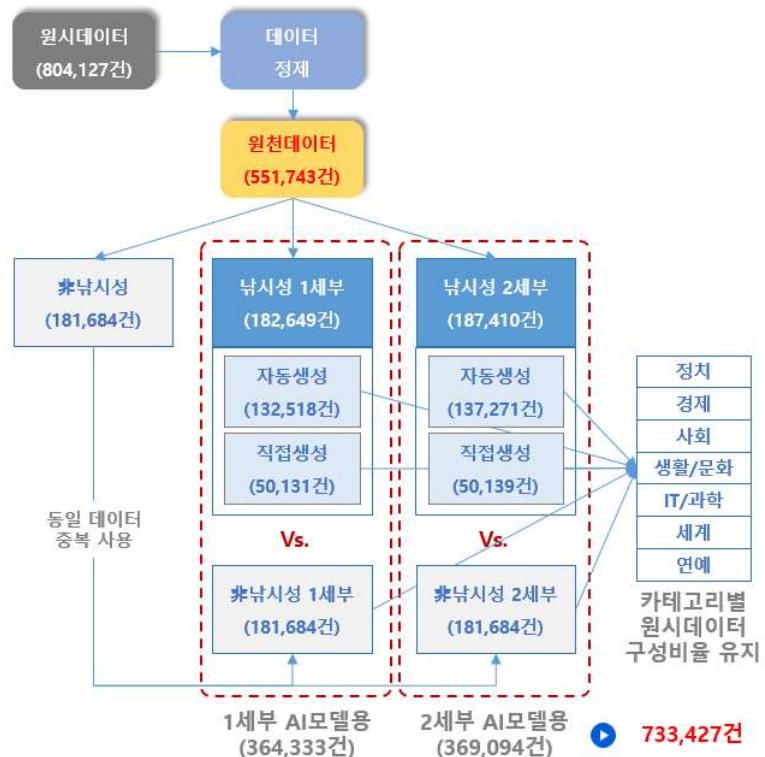
- 수집 방법

- 별도의 획득(수집) 도구 미사용
- 14개 언론사별 계약된 물량을 지정한 XML Format에 맞게 IT부서에서 작업 후 기사 건별로 XML File 생성
- 생성 파일 압축 및 전달 : 대용량메일, 주관사 FTP, 다운로드 링크 방식으로 언론사별 데이터 전달/수집

4. 정제

4.1 원천데이터 규모

- 원시데이터 정제 후 원천데이터 규모 : 최종 733,427건



- 낚시성 기사 자동생성 : 기사간 제목 or 본문을 프로그램을 통해 대체하여 가공하기 위한 원천데이터
 - 낚시성 기사 직접생성 : 크라우드 소싱을 통해 제목 or 본문을 크라우드워커가 직접 작성하여 가공하기 위한 원천데이터
 - 非낚시성 기사 자동생성 : 낚시성 기사의 반대급부로 별도의 가공작업 없이 라벨링데이터를 생성하기 위한 원천데이터

4.2 정제 절차

● XML 파일 오류 Check

- ① 언론사 XML Sample 파일 준비
 - ② XML 유효성 체크 사이트 접속
 - ③ XML 내용 복사 및 Syntax 오류 Check
 - ④ Sample 파일 5건 이내 ①~③ 반복 수행
 - ⑤ 오류 존재 시 오류사항 언론사 공유 및 데이터 재수집

The screenshot shows the 'XML Validator' section of the w3schools.com website. The left sidebar lists various XML-related topics. The main area displays the title 'Syntax-Check Your XML' and a message encouraging users to use the XML validator. A text input field contains XML code with several red underlined errors (e.g., 'nsid', 'CDATA[]', 'amp;'). Below the input is a 'Check XML' button. The status bar at the bottom indicates 'No errors found' and has a '확인' (Check) button.

www.w3schools.com 내용:
No errors found

확인

Try to syntax-check correct XML :

```
<?xml version="1.0" encoding="UTF-8"?>
<item type="text">
<press>세계일보</press>
<nsid>20170307510319</nsid>
<action></action>
<title><![CDATA[尹, 대학생 앞에 두고 "조금 더 발전하면 옛 깥아 구인·구진할 때 온다" 실언 '못매']></title>
<subtitle><![CDATA[]]></subtitle>
<date>2021-12-22</date>
<time>21:48:49</time>
<author>양다희</author>
```

Try to syntax-check incorrect XML :

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

● XML 파일 DB 업로드

- ① 정상 XML 파일 읽기
 - ② 변환을 하고자 하는 형태로 객체에 내용 저장
 - ③ DB에 XML 파일의 항목별 내용 저장 처리
 - ④ 모든 XML 파일에 대해 ①~③ 반복수행
 - ⑤ 종료 후 XML 파일 대비 DB 적재 내용 비교 검증
- ※ Java 기반으로 XML 데이터 형식에 맞게 직접 개발한 모듈을 사용하여 업로드 진행

● XML 파일 내용 Check(크라우드워커를 통한 인적 정제)

- ① 언론사 XML 파일 준비 및 파일 배분
- ② XML 파일 내용 검토
 - XML version = '1.0', encoding = 'UTF-8' 확인
 - 언론사, 기사ID, 제목, 부제목, 발행일자, 발행시간, 작성자, 본문내용, 기사 카테고리, 기사 온라인 URL 주소
- ③ 정제 대상 내용 검토 및 정제
 - 기사 제목/부제목 : 제목 내 HTML Tag 등 이상 문자 Check → Tag 및 이상 문자 제거, 제목 길이 Check
 - 기사 본문 : 문장분리 Check, Tag Check, 불필요 텍스트(Tag, 매체명, 기자명, 기자 이메일, 특정 언론사 패턴 문구), 본문 길이 Check
 - 내용 정제 및 삭제 대상 기사 선별

```

1  <?xml version="1.0" encoding="UTF-8"?>
2  <item type="text">
3    <press>세계일보</press>
4    <nsid>20170317509999</nsid>
5    <action>I</action>
6    <title><![CDATA[尹이 총괄 선대위장 요청한다면? 洪 “정책도 다르고 할 일이 없다” 거듭 합류 거부]]></title>
7    <subtitle><![CDATA[]]></subtitle>
8    <date>2021-11-26</date>
9    <time>20:12:29</time>
10   <author>양다훈</author>
11   <content><![CDATA[홍준표 국민의힘 의원(사진)이 윤석열 대통령선거 후보의 선거대책위원회에 합류할 의사가 없다고 거듭 확인했다.</content>
12   <category code="010101010000" name="정치일반"/>
13   <url href="https://www.segye.com/newsView/20170317509999"/>
14 </item>

```

● 프로그램 기반 정제

- ① 인적 정제 내용 반영
 - 제목/본문에 대한 내용 갱신(Update)
 - 삭제 대상 기사에 대한 삭제(Delete)
- ② 내용 전처리
 - 특수문자 앞 & 치환 :
 - Tag 제거 : , <, >, &lsquo, &rsquo, &hellip, &ldquo, &rdquo, &mu, [,], Δ, •, ‧
 - 정규식을 이용한 특정 Entity 제거 : Email Pattern, URL Pattern
 - 텍스트 정제 : 불필요 여백 제거, 중복 줄바꿈 제거, 중복 여백 제거, 제목/본문 정제 처리
 - 문장분리 : Python 기반의 kss(<https://pypi.org/project/kss>) 라이브러리를 기본 사용하고, 데이터 구축 가이드에 정한 로직에 맞추어 Python으로 개발된 모듈로 직접 구현하여 문장분리 진행
 - ③ 정제 후 추가 처리
 - 본문 마지막 문장에 기자명이 있는 경우 해당 문장 제거
 - 기사 제목이 3어절 미만인 기사 제거
 - 기사 제목의 글자수가 10자 미만인 기사 제거
 - 본문의 문장수가 5문장 이하, 50문장 이상인 기사 제거

clickbaitnews		이름	수정한 날짜	유형	크기
	__pycache__	__pycache__	2022-09-06 오후 3:56	파일 폴더	
	data	executor	2022-09-06 오후 3:58	Python 원본 파일	3KB
	doc	partdata	2022-09-06 오후 3:39	Python 원본 파일	5KB
	log	process	2022-09-06 오후 3:57	Python 원본 파일	11KB
>	prove				
	tests				

④ 원천데이터 배분 및 생성

- 세부구분(1세부/2세부), 용도구분(낚시성/非낚시성), 직접생성 가공패턴유형(1세부 6개, 2세부 4개)에 따라 사전에 계획된 배분율 및 2세부 가공 문장수 2~4랜덤 부여하여 기사들을 랜덤하게 추출 후 할당

● 원천데이터 배분 방법(비율)

연번	카테고리	수집단계	정제단계	낚시성												非낚시성		증복 가공		
				1세부(제목 가공-내제목)						2세부(본문 가공-문장추가)						非낚시성				
				자동생성(P/G)		직접생성(크라우드 워커)				자동생성(P/G)		직접생성(크라우드 워커)				의도적 상황 왜곡됨				
기사간대체 (99)	의문 유발형 (부호) (11)	의문 유발형 (은/니) (12)	선풍포럼 사용형 (13)	속어/중립 사용형 (14)	사설 과대 표현형 (15)	의도적 주어 왜곡형 (16)	기사간대체 (99)	상품 판매정보 노출 광고형 (21)	부동산 판매정보 노출 광고형 (22)	서비스 판매정보 노출 광고형 (23)	의도적 상황 왜곡형 (24)	마	마	마	마	마	마	마		
하	하	하	하	하	하	하	하	하	하	하	하	하	하	하	하	하	하	하		
1	정치	106,245	80,000	23.3%	3.5%	3.0%	1.0%	1.0%	1.0%	0.5%	23.3%	2.0%	1.0%	2.0%	5.0%	33.3%	→	26,667	26,667	
2	경제	114,086	80,000	23.3%	3.5%	3.0%	1.0%	1.0%	1.0%	0.5%	23.3%	3.0%	3.0%	3.0%	3.0%	1.0%	33.3%	→	26,667	26,667
3	사회	160,324	80,000	23.3%	3.5%	3.0%	1.0%	1.0%	1.0%	0.5%	23.3%	2.0%	1.0%	2.0%	5.0%	33.3%	→	26,667	26,667	
4	생활&문화	101,243	80,000	23.3%	3.5%	3.0%	1.0%	1.0%	1.0%	0.5%	23.3%	3.0%	3.0%	3.0%	3.0%	1.0%	33.3%	→	26,667	26,667
5	IT&과학	103,961	80,000	23.3%	3.5%	3.0%	1.0%	1.0%	1.0%	0.5%	23.3%	4.0%	0.5%	4.0%	1.5%	33.3%	→	26,667	26,667	
6	세계	118,265	80,000	23.3%	3.5%	3.0%	1.0%	1.0%	1.0%	0.5%	23.3%	2.0%	1.0%	2.0%	5.0%	33.3%	→	26,667	26,667	
7	연예	100,000	80,000	23.3%	3.5%	3.0%	1.0%	1.0%	1.0%	0.5%	23.3%	2.0%	0.5%	2.0%	5.5%	33.3%	→	26,667	26,667	
합계		804,127	560,000	23%	3.50%	3.00%	1.00%	1.00%	0.50%	23.33%	2.57%	1.43%	2.57%	3.43%	33.33%	→	186,667	186,667		

- 세부구분(Part1/Part2> 용도/가공유형 구분(Clickbait_Auto/Clickbait_Direct/NonClickbait_Auto)> 기사카테고리(7개 카테고리)별로 풀더 생성 및 지정된 항목, 포맷(JSON)으로 원천데이터 생성



- 생성된 원천데이터 확인 : Sample 개별 파일(JSON) Check 및 배분결과(풀더별, 패턴유형별, 가공문장수, 추가문장수) 확인

The image shows two side-by-side code editors in Visual Studio Code. Both editors have tabs at the top labeled 'File', 'Edit', 'Selection', 'View', 'Git', 'Run', 'Terminal', 'Help'. The left editor is titled '1세부용 원천데이터 JSON Sample' and the right one is '2세부용 원천데이터 JSON Sample'. Both editors show JSON code with numerous lines of text, likely representing news articles. The JSON structure includes fields like 'newsID', 'newsCategory', 'newsTitle', 'newsContent', 'sentenceCount', etc.

⑤ 원천데이터 백업 및 가공업체 전달

- 원천데이터 전체 백업 및 압축
- 외장하드에 복사하여 가공업체(미디어그룹사람과숲) 전달
- 저작도구를 위한 직접생성용 데이터를 저작도구에 업로드 및 확인

The screenshot shows a web application interface for managing news data. At the top, there's a navigation bar with '사용자관리', 'Home', '데이터기능', '데이터목록', '기사목록', '1세부용목록', '2세부용목록', '기획관리', and a search bar. Below the navigation is a search form titled '검색조건' with dropdowns for '미출당' (전체), '카테고리' (전체), '파일유형' (전체), and '진행상태' (미작업). There are also dropdowns for '검색필드' (전체) and '검색어 입력'. To the right of the search form are '검색' and '전체목록' buttons. The main area is titled '검색목록 (조회수: 14,508 개)' and contains a table with columns: 기사ID, 카테고리, 파일유형, 기사제목, 작업자, 가공검수자, 품질검수자, 진행상태, 관리. The table lists 10 rows of news items. At the bottom of the table are navigation buttons (1, 2, 3, 4, 5, 6, 7, 8, 9, 10, next, last) and a page number indicator '2 / 1 go'.

4.3 정제 기준

● 수집데이터 정제 및 원천데이터 생성 기준

① 텍스트 내용 대상 정제

- 텍스트 내용상에 있는 각종 Tag는 제거 처리
- 기사의 내용과 연관성이 없는 패턴화/非패턴화된 내용은 인적정제와 프로그램에 의한 정제를 통해 제거 처리
→ 기자명, 기자 이메일 주소, 보도&배포 관련 알림 문구, 언론사별 자동 삽입 문구, 사진/동영상에 대한 캡션 등
- 불필요/중복 여백, 중복 줄바꿈 등을 제거 처리

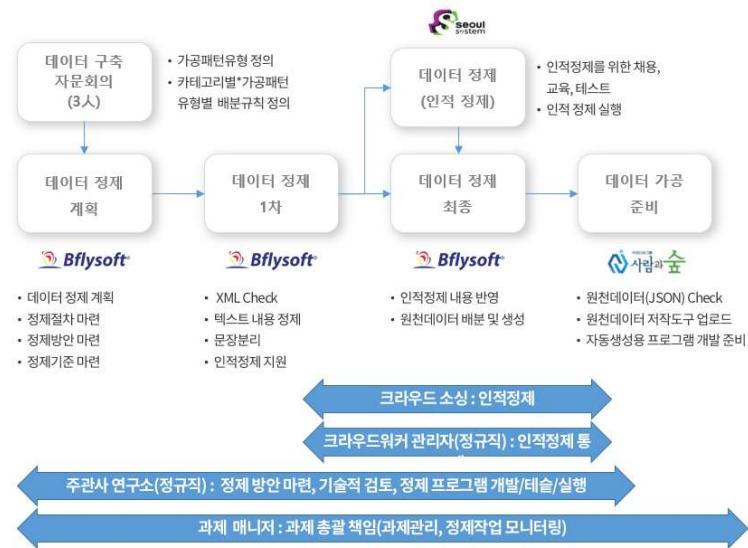
② 가공에 부적합한 이상치 데이터 정제

- 원문 기사의 제목길이가 너무 짧은 경우(3어절 미만, 10글자 미만) 해당 기사 미사용 처리
- 본문의 문장수가 너무 짧거나, 긴 경우(5문장 이하, 50문장 이상) 해당 기사 미사용 처리

③ 원천데이터 생성 시 배분율 유지

- 세부(1세부/2세부), 용도(낚시성/非낚시성), 가공패턴(1세부 6개, 2세부 4개), 기사 카테고리(7대 카테고리) 배분율에 맞게 원천데이터 JSON 파일 생성
 - 기사 카테고리의 경우 수집된 카테고리별 비율에 맞게 각 풀더별 비율 최대한 유지하도록 할당
 - 가공패턴별 할당은 자문회의에서 제시한 배분율을 고려하여 원천데이터 JSON 파일 생성

4.4 정제 조직



4.5 정제 도구

● 정제 도구

- ### ① 인적정제

- XML File Viewer(Vis)

- ## ② 프로그램 기반 정세

- Java 및 Python을 이용하여 서버 경제 프로그램 개발 및 구현				
clickbaitnews	이름	수정한 날짜	유형	크기
__pycache__	__pycache__	2022-09-06 오후 3:56	파일 폴더	
data	executor	2022-09-06 오후 3:58	Python 원본 파일	3KB
doc	partdata	2022-09-06 오후 3:39	Python 원본 파일	5KB
log	process	2022-09-06 오후 3:57	Python 원본 파일	11KB
prove				
tests				

- 최종 정제 후 결과 : 세부구분, 용도구분, 기사카테고리별로 배분되어 생성된 JSON 파일

5. 가공(라벨링)

5.1 가공(라벨링) 절차

● 1세부(제목과 본문의 불일치 기사) 자동생성

- ① 1세부 낚시성 기사 자동생성으로 할당된 기사들 간의 제목 대체
 - ② 기사 카테고리 대분류(정치/경제/사회/생활&문화/IT&과학/세계/연예)내에 존재하는 기사간에 제목을 대체하여 새제목 생성
 - ③ 기사의 새제목에 대체 기사의 제목을 추가하고, 낚시성 기사로 자동분류 라벨링(0)
 - ④ 기사간 제목 대체를 위한 라벨링 프로그램을 통해 일괄적으로 자동생성 처리
 - ⑤ 자동생성을 통해 처리된 결과는 라벨링데이터 포맷(JSON)에 맞게 파일생성 처리(기사ID+'L'JSON)

● 1세부(제목과 본문의 불일치 기사) 크라우드워커를 통한 직접생성

- ① 관리자가 기사건에 대해 작업자(1명), 가공검수자(1명), 품질검수자(3명) 할당
 - ② 작업자가 저작도구를 이용하여 새제목, 새제목 생성 시 참조한 문장 Check 후 '제출'을 통해 가공 처리
 - ③ 가공검수자가 가공 건에 대한 가공검수 진행(Pass or Fail 처리)
 - 가공검수 Pass 시 품질검수 진행
 - 가공검수 Fail 시 재가공 진행
 - ④ 품질검수자(3명)가 가공검수 Pass 건에 대해 품질검수 진행(Pass or Fail 처리)
 - 품질검수자 2인 이상이 Pass 시 해당 건 '최종완료' 처리
 - 품질검수자 2인 이상이 Fail 시 해당 건 '검수자 반려' 처리 → 해당 시 재가공 진행
 - ⑤ 가공(라벨링) 데이터 다운로드 → 라벨링데이터 포맷(JSON)에 맞게 파일생성 처리(기사ID+'_L.json)

기사 기본 정보

기사 ID : ET_M03_168659
기공구분 : 1세부
작업 유형 : 낙시생
식별 난이도 : 하
카테고리 : 연예 > 연예가/체육

기사 가공 정보

제작자 : 종교인 교세가
글자수/어절수 : 7 / 2
본문 :

기사 원문 정보

제목 : 종교인교세 2018년부터 시행...국회 조세소위원회서 통과
본문 : 종교인 소득에 대한 소득세 부과가 2018년부터 시행된다. 국회 기획재정위원회 조세소위원회는 30일 소위를 열고 종교소득에 대한 과세의 내용이 담긴 소득세법 개정안을 의결했다. 종교인들에 대한 과세는 소득세법 상 기타소득으로 분류돼 세금이 확정될 예정이다. 세금 부과시 소득 구간에 따라 낮은 금액의 공제율이 적용된다. 연소득 1억5000만 원 초과는 20%, 8000만 원에서 1억5000만 원은 40%, 4000만 원에서 8000만 원은 60%, 4000만 원 이하는 80%의 공제율을 적용받는다. 종교단체의 협회장은 날세자의 선적사항으로 말기고 현천장을 하지 않는 경우 종교인이 적용 신고부여가 된다. 이를 기자회견에 참석한 소득세법 개정안은 예산 부수법안으로 지정돼 내달 2일 국회 본회의에 자동 상정된다.

제작자

작업자 : 잡코리아 맞춤법검사기 링크

기사 가공 대상으로 전환

← 이전/이후가공대상건으로 전환
← 패널유형에 맞게 새제목 직접 작성
(3어절 이상, 20~100글자)

새제목 작성 시 참조문장 Check

No	본문 문장 내용	생성형태
1	종교인 소득에 대한 소득세 부과가 2018년부터 시행된다.	<input checked="" type="checkbox"/>
2	국회 기획재정위원회 조세소위원회는 30일 소위를 열고 종교 소득에 대한 과세 내용이 담긴 소득세법 개정안을 의결했다.	<input type="checkbox"/>
3	종교인들에 대한 과세는 소득세법 상 기타소득으로 분류돼 세금이 확정될 예정이다.	<input type="checkbox"/>
4	세금 부과시 소득 구간에 따라 필요경비 공제율이 차등 적용된다.	<input type="checkbox"/>
5	연소득 1억5000만 원 초과는 20%, 8000만 원에서 1억5000만 원은 40%, 4000만 원에서 8000만 원은 60%, 4000만 원 이하는 80%의 공제율을 적용받는다.	<input type="checkbox"/>
6	종교단체의 협회장은 날세자의 선적사항으로 말기고 현천장을 하지 않는 경우 종교인이 적용 신고부여가 된다. 이를 기자회견에 참석한 소득세법 개정안은 예산 부수법안으로 지정돼 내달 2일 국회 본회의에 자동 상정된다.	<input type="checkbox"/>

● 2세부(제목과 본문의 불일치 기사) 자동생성

기사 : EC_M02_155962 → 가공문장수 2		기사 : EC_M02_155963 → 가공문장수 3	
제목 :	새제목 : 중근당, 치주질환 치료제 '이튼큐 플러스' 눈길…간편 복약 통해 '신바람'	제목 :	새제목 : 김기환 KB손보 사장, '어린이 교통안전 레일레이 철린지' 등참
a1 :	복약 편의성을 대폭 개선한 종근당의 치주질환 치료제 '이튼큐 플러스'가 업계에서 이목을 끌고 있다.	b1 :	보험업계에 어린이 교통안전 캠페인 레일레이 바람이 불고 있다.
a2 :	종근당 관계자는 26일 미디어SR에 "치주질환 치료제는 장기 복용하는 환자가 많고 종합용량에 맞춰 복용하는 복약 순응도가 치료에 큰 영향을 미친다"고 전제하면서 "복약편의성을 개선하고 안전성이 입증된 이튼큐 플러스가 치주질환 치료에 큰 도움을 줄 줄"이라고 강조했다.	b2 :	KB손해보험은 김기환 사장이 서울시 강남구 본사 사옥에서 어린이 보호 문화 확산을 위한 '어린이 교통안전 레일레이 철린지'에 등참했다고 2일 밝혔다.
a3 :	이튼큐 플러스는 옥수수를 불경화정량주출물 단일제제인 이튼큐에 후박주출물을 추가한 생약 성분의 복합체인 점이 특징이다.	b3 :	KB손보 관계자는 미디어SR에 "이 철린지는 지난해 12월 초부터 어린이 보호 최우선 문화를 정착시키고 어린이 교통사고 예방에 대한 국민적 공감대를 형성하기 위해 행정안전부 주관으로 진행하고 있는 레일레이 캠페인이다"며 "참여자가 어린이 보호문화 정착을 위한 교통안전 솔루션을 SNS에 공유하고 다음 후발주자를 추천하는 방식으로 진행된다"고 설명했다.
a4 :	증상 분인 옥수수불경화정량주출물은 치주인대의 재생을 도와 치아가 훔들리는 것을 막고 치조골을 재건시켜 잇몸 속 기초를 튼튼하게 하는 기능을 있다고 종근당측은 설명했다.	b4 :	KB국민은행 허인 행장의 주전을 받은 김기환 사장은 이날 어린이 교통안전 구호를 듣고 사진에 '#어린이교통안전', '#어린이보호구역' 등 해시태그를 달아 인스타그램을 포함한 SNS에 게시했다.
a5 :	또한 후박주출물은 치주질환의 원인균에 대한 항균효과와 항염효과가 우수해 잇몸 염증에 대한 저항력을 강화시켜주는 효과가 있다고 회사측은 밝혔다.	b5 :	김기환 KB손보 사장은 "이번 철린지를 통해 어린이 교통 안전사고 예방에 대한 인식이 확산되길 바란다"며 "KB손해보험은 앞으로도 미래의 희망인 어린이들이 마음껏 꽂으며 자랄 수 있는 안전한 사회환경 조성에 기여할 수 있도록 더욱 노력할 것"이라고 말했다.
a6 :	종근당 관계자는 이날 '이튼큐 플러스는 장기 복용에도 부작용이 없는 생약 성분 치로제로 안전성이 인정됐다"면서 "獨자 개발한 정체 촉진 기술 iLET(Innovative Low Excipient Tablet) 특허공법을 적용해 현재 출시돼 있는 동일 성분 제품 가운데 정체 사이즈를 가장 작게 들여온데서 암울의 불쾌한 시각을 줄이고 다수의 약물들을 함께 복용하는 장·중장년 환자의 복약편의성을 개선한 것이 강점"이라고 설명했다.	b6 :	한편 김기환 사장은 어린이 교통안전 레일레이 철린지의 다음 주자로 전 피겨스케이팅 선수 김연아, 민병도 보험연수원장, 원종규 코리안리재보험 대표이사를 추천했다.

가공데이터 문장 구성 : a1(Y), a2(Y), a3(Y), a4(Y), b4(N), b5(N), b6(N)

가공데이터 문장 구성 : b1(Y), b2(Y), b3(Y), a5(N), a6(N)

※ 기준 문장 : 도메인 일치 여부(Y) vs. 대체 문장 : 도메인 일치 여부(N)

- ① 2세부 낚시성 기사 자동생성으로 할당된 기사들 간의 본문 문장 대체
 - ② 기사 카테고리 대분류(정치/경제/사회/생활&문화/IT&과학/세계/연예)내에 존재하는 기사간에 문장들을 대체하여 새제목 생성
 - ③ 가공데이터에 기존 문장과 대체 문장을 저장하고, 낚시성 기사로 자동분류 라벨링(0)
 - ④ 기사간 문장 대체를 위한 라벨링 프로그램을 통해 일괄적으로 자동생성 처리
 - ⑤ 자동생성을 통해 처리된 결과는 라벨링데이터 포맷(JSON)에 맞게 파일생성 처리(기사ID+'.'+JSON)

● 2세부(본문의 도메인 일관성 부족 기사) 크라우드워커를 통한 직접생성

- ① 관리자가 기사건에 대해 작업자(1명), 가공검수자(1명), 품질검수자(3명) 할당
 - ② 작업자가 저작도구를 이용하여 패턴유형에 맞게 지정된 문장수 만큼의 문장 작성 후 '제출'을 통해 가공 처리
 - ③ 가공검수자가 가공 건에 대한 가공검수 진행(Pass or Fail 처리)
 - 가공검수 Pass 시 품질검수 진행
 - 가공검수 Fail 시 재가공 진행
 - ④ 품질검수자(3명)가 가공검수 Pass 건에 대해 품질검수 진행(Pass or Fail 처리)
 - 품질검수자 2인이 이상이 Pass 시 해당 건 '최종완료' 처리
 - 품질검수자 2인이 이상이 Fail 시 해당 건 '검수자 반려' 처리 → 해당 시 재가공 진행
 - ⑤ 가공(라벨링) 데이터 다운로드 → 라벨링데이터 포맷(JSON)에 맞게 파일생성 처리(기사ID+ '_L.JSON')

● 낚시성 기사 직접생성 가공 절차



● 후처리 : 2세부 초기 가공데이터 품질검수 및 오류 데이터 재가공 후 간접

- 사유 : 과제조정위원회의 승인 하에 22.09.28일부터 2세부 본문 문장 작성 시 마스킹 미사용에 따라 기존 작업 물량에 대한 검토 및 내용 보강 진행
- 처리 : 마스킹 사용 물량 중 일부는 라벨링 데이터 다운로드 시 제외처리하고, 일부는 내용을 보강하여 엑셀로 정리 후 DB에 일괄적으로 Update를 수행 후 라벨링 데이터 다운로드 처리 진행

5.2 가공(라벨링) 기준

● 1세부(제목과 본문의 불일치 기사) 자동 가공 프로그램을 통한 자동생성 기준

기사 : EC_M02_155962	기사 : EC_M02_155963
제목 : 종근당 치주질환 치료제 '이튼큐 플러스' 눈길...간편 복약 통해 '신바람'	제목 : 김기환 KB손보 사장, '어린이 교통안전 릴레이 챌린지' 동참
본문 : 복약 편의성을 대폭 개선한 종근당의 치주질환 치료제 '이튼큐 플러스'가 업계에서 이목을 끌고 있다. 종근당 관계자는 26일 미디어SNS에 "치주질환 치료제는 장기 복용하는 환자가 많고 종법·통령에 맞춰 복용하는 복약 순종도가 치료에 큰 영향을 미친다"며 "전체하면서 '복약편의성을 개선하는 안전성이 입증된 이튼큐 플러스는 옥수수를 점화정량주출을 단일제제인 이튼큐에 혼탁주출물을 주가 한 생약 성분의 복합체인 점이 특징이다. 주성분인 옥수수를 점화정량주출들은 치주인대의 재상을 도와 치아가 훈들리는 것을 막고 치조골을 재건시켜 이를 속기조를 튼튼하게 하는 기능을 한다고 종근당측은 설명했다. 또한 후박주출들은 치주질환의 원인균에 대한 항균효과와 항염효과가 우수해 엿물 염증에 대한 저항력을 강화시켜주는 효과가 있다고 회사측은 밝혔다. 종	본문 : 보험업계에 어린이 교통안전 캠페인 릴레이 바람이 불고 있다. KB손해보험은 김기환 사장이 서울시 강남구 본사 사옥에서 어린이 보호문화 확산을 위한 '어린이 교통안전 릴레이 챌린지'에 동참했다고 2일 밝혔다. KB손보 관계자는 미디어SNS에 "이 챌린지는 지난해 12월 초부터 어린이 보호·최우선 문화로 정착시키고 어린이 교통사고 예방에 대한 국민적 공감대를 형성하기 위해 행정안전부 주관으로 진행하고 있는 릴레이 캠페인이다"며 "참여여자가 어린이 보호문화 정착을 위한 교통안전 술로건을 SNS에 공유하고 다음 후발주자를 주천하는 방식으로 진행된다"고 설명했다. KB국민은행 어린 행정의 주천을 받은 김기환 사장은 이날 어린이 교통안전 구호를 든 사진에 "#어린이교통안전", "#어린이보호구역" 등 해시태그를 달아 인스타그램을 포함한 SNS에 게시했다. 김기환 KB손보 사장은 "이번 챌린지를 통해 어린이 교통 안전사고 예
새제목 : 김기환 KB손보 사장, '어린이 교통안전 릴레이 챌린지' 동참	새제목 : 종근당, 치주질환 치료제 '이튼큐 플러스' 눈길...간편 복약 통해 '신바람'

- 1세부 자동생성용 원천데이터 준비 : 정제과정에서 세부, 용도, 가공패턴에 따라 구분된 원천데이터 사용
- 7대 기사 카테고리 내에서 각각 원문기사의 제목에 대한 중복성 체크 후 중복기사 제거 처리
- 기사 카테고리 내에서 랜덤하게 2개의 기사 추출
- 추출된 2개 기사간의 제목을 서로 대체하여 새제목 생성 및 낚시성 기사 분류는 0으로 라벨링 처리
- 1세부 라벨링데이터 포맷에 맞게 JSON 파일 생성
- 기사 카테고리 내 전체 데이터에 대해서 동일 작업 수행하여 라벨링데이터 자동생성

● 1세부(제목과 본문의 불일치 기사) 크라우드워커를 통한 직접생성 기준

- 새제목은 3어절 이상으로 작성(저작도구 Check Logic에 반영)
- 새제목은 공백을 포함하여 20~100글자 범위에서 작성(저작도구 Check Logic에 반영)
- 새제목은 각 기사에 할당되어 있는 패턴유형에 맞게 작성 : 의미정확성 측정 사항
 - 패턴유형에 맞지 않게 작성한 경우 가공검수 및 품질검수 단계에서 반려하여 재가공 유도
- 새제목을 작성할 시 본문 문장 중에서 작성에 참조한 문장을 1개 이상 반드시 Check(저작도구 Check Logic에 반영)
- 훌따옴표, 쌍따옴표는 쌍으로 작성
 - 가공검수, 품질검수 단계에서 검사 후 오류 발견 시 반려하여 재가공 유도

The screenshot shows a multi-step process for creating a news article:

- Step 1: Basic Article Information** (기사 기본 정보)
 - Article ID: ET_M03_168659
 - Section: 1세부
 - Category: 종교
 - Author: 이은유발행 (부호)
 - Keywords: 연예 > 연예가회제
- Step 2: Content Input** (기사 원문 정보)
 - Title: 종교인과세 2018년부터 시행...국회 조세소위원회서 통과
 - Text: 종교인 소득에 대한 소득세 폐기가 2018년부터 시행된다. 국회 기획재정위원회 조세소위원회는 20일 소득에 대한 과세 등에 대한 내용이 담긴 소통서를 개정안을 의결했다. 종교인들이 대상 과세는 소득세법 상 기타소득으로 분류돼 세금이 확장될 예정이다. 세금 부과시 소득 구간에 따라 필요경비 공제율이 적용된다. 연소득 1억5000만 원 초과는 20% 8000만 원에서 1억5000만 원은 40%, 4000만 원에서 8000만 원은 60%, 4000만 원 이상은 80%의 공제율을 적용 받는다. 종교단체의 원천징수는 납세자의 선택사항으로 끌고 원천징수하지 않는 경우 종교인이 직접 신고 납부하게 된다. 이날 2자위 청취회에 자동 상정된다.
- Step 3: Content Validation** (기사 가공 정보)
 - Headline: 이전/이후 가공 대상건으로 전환
 - Body: 폐단유형: 이은유발행 (부호)
 - Text: 폐단유형에 맞게 새제목 직접 작성 (3어절 이상, 20~100글자)
- Step 4: Subject Extraction** (새제목 작성 시 참조문장 Check)
 - Table: 본문 문장 내용 (6 rows)

No	본문 문장 내용	생성항목
1	종교인 소득에 대한 소득세 폐기가 2018년부터 시행된다.	
2	국회 기획재정위원회 조세소위원회는 20일 소위를 열고 종교 소득에 대한 과세 등의 내용이 담긴 소통서를 개정안을 의결했다.	
3	종교인들에 대한 과세는 소득세법 상 기타소득으로 분류돼 세금이 확장될 예정이다.	
4	세금 부과시 소득 구간에 따라 필요경비 공제율이 자동 적용된다.	
5	연소득 1억5000만 원 초과는 20% 8000만 원에서 1억5000만 원은 40%, 4000만 원에서 8000만 원은 60%, 4000만 원 이상은 80%의 공제율을 적용 받는다.	
6	종교단체의 원천징수는 납세자의 선택사항으로 끌고 원천징수하지 않는 경우 종교인이 직접 신고 납부하게 된다. 이날 2자위 청취회에 자동 상정된다.	
 - Buttons: 적용, 미시작, Skip
- Step 5: Task Management** (작업자)
 - Table: 작업자 (1 row)

작업자	작업일시	작업유형	비고
sysadm	2022-09-26 10:52:35	월당	HF_worker19에게 월당
 - Buttons: 맞춤법 검사기, 적용, 미시작, Skip

- 링크된 맞춤법검사기를 활용하여 새제목 작성
 - 오탈자, 띄어쓰기 등의 맞춤법에 맞지 않게 작성된 새제목에 대해서는 가공검수, 품질검수 단계에서 검사 후 오류 발견 시 반려하여 재가공 유도
- 새제목 작성 및 참조문장 Check 후 제출된 데이터는 낚시성 기사 여부(Ground Truth) 값을 0으로 부여
- 저작도구로 가공이 완료된 건들은 다운로드 프로그램을 통해 라벨링데이터 포맷(JSON)에 맞게 파일로 다운로드 처리

● 1세부(제목과 본문의 불일치 기사) 라벨링데이터 파일(JSON) 예시

- 자동생성 라벨링데이터 파일 예시

The JSON file contains two main sections: `sourceDataInfo` and `labelledDataInfo`.

sourceDataInfo:

```
{
  "sourceDataInfo": {
    "newsTitle": "이해영 \"My Birthday, 58\"", // This is the main title
    "newsContent": "기수 출신 엔터테이너 이해영이 균황을 전했다. 이날 이해영은 23일 SNS에 \"My Birthday! 5850505050505050\"이라는 글과 함께 사진들을 올렸다.", // This is the main content
    "newsCategory": "연예",
    "newsSubCategory": "연예일반",
    "newsAuthor": "기수",
    "newsSource": "기수",
    "newsDate": "2022-09-26 10:52:35",
    "partnum": "P1",
    "useType": 0,
    "processType": "A",
    "processLevel": "90",
    "sentenceCount": 6,
    "sentenceInfo": [
      {"sentenceNo": 1, "sentenceContent": "기수 출신 엔터테이너 이해영이 균황을 전했다.", "sentenceSize": 25},
      {"sentenceNo": 2, "sentenceContent": "이해영은 23일 SNS에 \"My Birthday! 5850505050505050\"이라는 글과 함께 사진들을 올렸다.", "sentenceSize": 68},
      {"sentenceNo": 3, "sentenceContent": "사진 속 이해영은 보라색 수영복을 입고 웃임모자를 쓰고 있다.", "sentenceSize": 34},
      {"sentenceNo": 4, "sentenceContent": "이해영은 다른 사진에선 뱀띠에 앉아 있다.", "sentenceSize": 24},
      {"sentenceNo": 5, "sentenceContent": "사진들은 최근 이해영이 하와이로 여행을 다녀 온 모습으로 보인다.", "sentenceSize": 36},
      {"sentenceNo": 6, "sentenceContent": "이해영은 1992년 데뷔했으며 2011년 하와이에서 연상의 사업가 남편과 결혼했다. 이해영은 MBC 예능 프로그램", "sentenceSize": 98}
    ]
  }
}
```

labelledDataInfo:

```
{
  "labelledDataInfo": {
    "newsTitle": "김희재 38일 판문서트로 '최항별' 만난다", // This is the labeled title
    "newsContent": "최항별은 38일 판문서트로 김희재와 만난다.", // This is the labeled content
    "clicknextClass": 0,
    "referSentenceInfo": [
      {
        "sentenceNo": 1, "referSentenceyn": "N"
      },
      {
        "sentenceNo": 2, "referSentenceyn": "N"
      },
      {
        "sentenceNo": 3, "referSentenceyn": "N"
      },
      {
        "sentenceNo": 4, "referSentenceyn": "N"
      },
      {
        "sentenceNo": 5, "referSentenceyn": "N"
      },
      {
        "sentenceNo": 6, "referSentenceyn": "N"
      }
    ]
  }
}
```

Annotations:

- The main title and content in the source data are highlighted in red, indicating they are the original data.
- The labeled title and content in the labeled data are highlighted in blue, indicating they are the processed or predicted data.
- A callout arrow points from the main title in the source data to the labeled title in the labeled data, with the text "← 기사간 대체 새제목, 낚시성 기사 여부(0)".

- 직접생성 라벨링데이터 파일 예시

원천데이터 정보(sourceDataInfo)	라벨링데이터 정보(sourceDataInfo)
<pre> "sourceDataInfo": { "newsID": "ET_M13_171698", "newsCategory": "경제", "newsTitle": "아이유, 봄을 부르는 신송한 애모가" }, "newsContent": "가수 겸 배우 아이유가 신뜻한 봄 향기를 내뿜고 있다. \n아이유는 자신의 인스타그램 스토리에 별다른 코멘트 없이 회보 'partnum': '01', "usertype": 0, "processType": "0", "processPattern": "12", "processLevel": "0", "sentenceCount": 0, "sentenceInfo": [{ "sentenceNo": 1, "sentenceContent": "가수 겸 배우 아이유가 신뜻한 봄 향기를 내뿜고 있다.", "sentenceSize": 30 }, { "sentenceNo": 2, "sentenceContent": "아이유는 자신의 인스타그램 스토리에 별다른 코멘트 없이 회보 사진들을 게재했다.", "sentenceSize": 44 }, { "sentenceNo": 3, "sentenceContent": "사진 속 아이유는 원 티에 청바지를 매치해 뉴트로 패션을 선보였다.", "sentenceSize": 37 }, { "sentenceNo": 4, "sentenceContent": "또 봄 색깔 가득한 재킷으로 봄 향기 가득 담은 무드를 자아낸다.", "sentenceSize": 36 }, { "sentenceNo": 5, "sentenceContent": "또 아이유는 오는 8월 국제여성의날을 맞아 패션브랜드 뉴발란스와 함께 캠페인을 함께 진행한다.", "sentenceSize": 52 }, { "sentenceNo": 6, "sentenceContent": "한편 아이유는 온·오프 활동을 즐기기를 거쳤고 OTT 서비스 '미니캐임'에 캐스팅됐다.", "sentenceSize": 46 }] } </pre>	<pre> "labeledDataInfo": ["newTitle": "봄을 부르는 전늘 앤티테이너... 그녀는 누구?", // ← 새제목, 낚시성 기사 여부(0) "clickbackClass": 0, "referSentenceInfo": [{ "sentenceNo": 1, "referSentenceyn": "Y" // ← 새제목 작성 시 참조 문장(Y) }, { "sentenceNo": 2, "referSentenceyn": "N" }, { "sentenceNo": 3, "referSentenceyn": "N" }, { "sentenceNo": 4, "referSentenceyn": "N" }, { "sentenceNo": 5, "referSentenceyn": "N" }, { "sentenceNo": 6, "referSentenceyn": "N" }] } </pre>

● 2세부(본문의 도메인 일관성 부족 기사) 자동 가공 프로그램을 통한 자동생성 기준

기사 : EC_M02_155962 → 가공문장수 2		기사 : EC_M02_155963 → 가공문장수 3	
제목 :	새제목 : 종근당, 치주질환 치료제 '이튼큐 플러스' 눈길...간편 복약 통해 '신바람'	제목 :	새제목 : 김기환 KB손보 사장, '어린이 교통안전 러레이 챌린지' 동참
a1 :	복약 편의성을 대폭 개선한 종근당의 치주질환 치료제 '이튼큐 플러스'가 업계에서 이목을 끌고 있다.	b1 :	보험업계에 어린이 교통안전 캠페인 러레이 바람이 불고 있다.
a2 :	종근당 관계자는 26일 미디어SR에 "치주질환 치료제는 장기 복용하는 환자가 많고 용법·용량에 맞춰 복용하는 복약 순응도가 치료에 큰 영향을 미친다"고 전제하면서 "복약편의성을 개선하고 안전성이 입증된 이튼큐 플러스는 치주질환 치료에 큰 도움을 줄 것"이라고 강조했다.	b2 :	KB손보해보험은 김기환 사장이 서울시 강남구 본사 사옥에서 어린이 보호 문화 확산을 위한 '어린이 교통안전 러레이 챌린지'에 동참했다고 2일 밝혔다.
a3 :	이튼큐 플러스는 옥수수를 검화정량추출물은 단일제제인 이튼큐에 후박추출물을 추가한 생약 성분의 복합체인 점이 특징이다.	b3 :	KB손보 관계자는 미디어SR에 "이번 챌린지는 지난해 12월 초부터 어린이 보호 최우선 문화를 정착시키고 어린이 교통사고 예방에 대한 국민적 공감대를 형성하기 위해 행정안전부 주관으로 진행하고 있는 러레이 챌린지이다"며 "참여자들이 어린이 보호문화 정착을 위한 교통안전 슬로건을 SNS에 공유하고 다음 후발주자를 추천하는 방식으로 진행된다"고 설명했다.
a4 :	주성분인 옥수수를 검화정량추출물은 치주인대의 재생을 도와 치아가 흔들리는 것을 막고 치조골을 재건시켜 잇몸 속 기초를 튼튼하게 하는 기능을 있다고 종근당측은 설명했다.	b4 :	KB국민은행 허인 행장의 추천을 받은 김기환 사장은 이날 어린이 교통안전 구호를 든 사진에 '#어린이교통안전', '#어린이보호구역' 등 해시태그를 달아 인스타그램을 포함한 SNS에 게시했다.
a5 :	또한 후박추출물은 치주질환의 원인균에 대한 항균효과와 항염효과가 우수해 잇몸염증에 대한 저항력을 강화시켜주는 효과가 있다고 회사측은 밝혔다.	b5 :	김기환 KB손보 사장은 "이번 챌린지를 통해 어린이 교통 안전과 예방에 대한 인식이 확산되길 바란다"며 "KB손보해보험은 앞으로도 미래의 희망인 어린이들이 마을 것 꿈꾸며 자랄 수 있는 안전한 사회환경 조성에 기여할 수 있도록 더욱 노력할 것"이라고 말했다.
a6 :	종근당 관계자는 이날 "이튼큐 플러스는 장기 복용에도 부작용이 없는 생약 성분 치료제로 안전성이 입증됐다"면서 "독자 개발한 정체 죽소기술 iLET(Innovative Low Exipient Tablet) 특허공법을 적용해 현재 출시돼 있는 동일성분 제품 가운데 정체 사이즈를 가장 작게 줄임으로써 약물의 풍해시간을 줄이고 다수의 약물을 함께 복용하는 중·장년층 환자의 복약편의성을 개선한 것이 고점"이라고 설명했다.	b6 :	한편 김기환 사장은 어린이 교통안전 러레이 챌린지의 다음 주자로 전 피겨스케이팅 선수 김연아, 민병두 보험연수원장, 원종규 코리안리재보험 대표이사를 추천했다.

기공데이터 문장 구성 : a1(Y), a2(Y), a3(Y), a4(Y), b4(N), b5(N), b6(N)

기공데이터 문장 구성 : b1(Y), b2(Y), b3(Y), a5(N), a6(N)

※ 기존 문장 : 도메인 일치 여부(Y) vs. 대체 문장 : 도메인 일치 여부(N)

- 2세부 자동생성용 원천데이터 준비 : 정제과정에서 세부, 용도, 가공패턴에 따라 구분된 원천데이터 사용
- 7대 기사 카테고리 내에서 각각 원문기사에 대한 중복성 체크 후 중복기사 제거 처리
- 기사 카테고리 내에서 랜덤하게 2개의 기사 추출
- 추출된 2개 기사간 가공처리 문장수 만큼의 문장들을 기사 후미에서 잘라내어 기사간 바꿔서 붙여넣기 하고, 대체된 문장들에 대해서는 도메인일치여부를 'N'으로, 낚시성 기사 분류는 0으로 라벨링 처리
- 2세부 라벨링데이터 포맷에 맞게 JSON 파일 생성
- 기사 카테고리 내 전체 데이터에 대해서 동일 작업 수행하여 라벨링데이터 자동생성

- 2세부(본문의 도메인 일관성 부족 기사) 크라우드워커를 통한 직접생성 기준

기사 기본 정보

기사 ID : EC_M04_620254
기공구분 : 2세부
종료유형 : 부서성
작별 난이도 : 상
카테고리 : 경제 > 기업/CEO

**↑
패턴유형 설명 팝업 제공**

기사 가공 정보

6	받는방식은 어떤가요? 예상수령일은 언제인가요? 받급받을 수 있고, 사실증명은 세무서를 방문해 받급받아야 한다.	Y
7	이번 협약을 통해 대전·충청지역 성실납세자들은 최대 5000만원까지 납부 제출 없이 대출을 받을 수 있고, 신용 등급별 산출금리에서 최대 1.5%까지 금리 감면 혜택을 받을 수 있게 됐다.	Y
8		N
9		N

← 패턴유형에 맞게 본문 문장 직접 작성 (2~4문장)

기사 원문 정보

제목 : 성실납세자 저금리 무담보 신용대출

본문 :

```
w" title="w"▲ 대전지방국세청은 27일 지방청 회의실에서 하나은행 충청사업본부와 국세·성실납세자 저금리 무담보 협약을 체결했다. ▲ 시민·업체부터 이영호(左)한경 회장, 박재석 대전지방국세청장, 박근영 하나은행 충청사업본부 대표(右)▲ 대전지방국세청은 27일 지방청 회의실에서 하나은행·충청사업본부와 국세·성실납세자 저금리 무담보 협약을 체결했다. 이날 협약은 대전지방국세청과 충청사업본부, 하나은행 충청사업본부 대표(左)와 박근영 하나은행 충청사업본부 대표(右)로 이루어졌다. 성실납세자들은 온행의 저금리 무담보 신용대출 가능해진다. 대전지방국세청장 박재석은 "올해는 협약으로 27일 지방청 회의실에서 지역·성실납세자들이 하나은행으로부터 저금리 무담보 신용 대출 협약을 맺을 수 있도록 하나은행 충청사업본부와 국세·성실납세자 저금리 무담보 협약을 체결했다. 이날 협약식에는 박재석 청장과 박근영 하나은행 충청사업본부 대표, 대전·충청지역·광주납세자
```

**↑
잡코리아 맞춤법검사기 링크**

← 제출 : 제목 가공 내역 제출 처리
← 임시저장 : 제목 가공 내역 임시 저장 처리
← Skip : 제목 가공 Skip(작업 포기)

- 기사마다 가공문장수를 2~4문장 랜덤하게 부여하고, 할당받은 기사에 대해서는 지정된 문장수 만큼을 작성
 - 지정된 문장수 만큼 각 칸(셀)을 다 채워서 제출해야만 가공 작업 처리되도록 저작도구에 Check Logic 반영
 - 하나의 칸(셀)에는 하나의 문장만 작성 필요
 - 가공검수, 품질검수 단계에서 1문장 작성여부, 문장 끝에 마침표 등의 정상 종료여부 등을 체크하고, 오류 시 반려 처리하여 재가공 유도
 - 하나의 칸(셀)에 작성되는 문장은 30 ~ 300글자 범위 내에서 작성
 - 첫문장의 서두에는 원문기사의 내용을 간략하게 요약한 내용을 포함하도록 작성
 - 가공검수, 풁질검수 단계에서 첫문장 서두에 원문기사에 대한 내용이 없을 시 반려 처리하여 재가공(보강) 유도
 - 작성하는 문장들은 패턴유형에 맞게 관련성 있는 내용을 연결하여 작성
 - 가공검수, 풁질검수 단계에서 작성 문장간 연관성이 없는 경우 반려 처리하여 재가공(보강) 유도
 - 링크된 맞춤법검사기를 활용하여 본문문장 작성 후 Check하여 제출
 - 오탈자, 띠어쓰기 등의 맞춤법에 맞지 않게 작성된 본문 문장에 대해서는 가공검수, 풁질검수 단계에서 검사 후 오류 발견 시 반려하여 재가공(보강) 유도

● 2세부(제목과 본문의 불일치 기사) 라벨링데이터 파일(JSON) 예시

- ### - 자동생성 라벨링데이터 파일 예시

원천데이터 정보(sourceDataInfo)

```
sourceDataInfo": {  
    "newsID": "03_H1_580770",  
    "newsCategory": "세계",  
    "newsSubCategory": "동북아",  
    "newsTitle": "대만서 규모 6.1 침진 발생...대부분 지역서 진동느껴",  
    "newsSubTitle": "",  
    "newsContent": "대만에서 규모 6.1의 지진이 발생했다. 최대 진도는 4급으로 타이베이 등 대다수 지역에서 진동을 느낄 수 있었다.",  
    "partName": "92",  
    "topic": "92",  
    "processType": "A",  
    "processPattern": "99",  
    "processLevel": "01",  
    "subjectCategory": "91",  
    "processSentencenum": 1,  
    "sentenceCount": 1,  
    "sentences": [  
        {  
            "sentenceNo": 1, "sentenceContent": "대만에서 규모 6.1의 지진이 발생했다.", "sentenceSize": 22},  
        {  
            "sentenceNo": 2, "sentenceContent": "최대 진도는 4급으로 타이베이 등 대다수 지역에서 진동을 느낄 수 있었다.", "sentenceSize": 41},  
        {  
            "sentenceNo": 3, "sentenceContent": "5월 대만 BCC 등 현지 언론에 따르면 이날 오전 5시50분 지진이 발생했다.", "sentenceSize": 44},  
        {  
            "sentenceNo": 4, "sentenceContent": "장악자는 대만 동부해역 70.1km 지점이다.", "sentenceSize": 25},  
        {  
            "sentenceNo": 5, "sentenceContent": "여진은 계속되고 있다.", "sentenceSize": 12},  
        {  
            "sentenceNo": 6, "sentenceContent": "대만 네티즌들은 이번 지진과 관련 '초거대' '불렀다' 등의 반응을 보이고 있다.", "sentenceSize": 44}  
    ]  
}
```

라벨링데이터 정보(sourceDataInfo)

```
"labeledDataInfo": {  
    "processSentenceInfo": [  
        {  
            "sentenceNo": 1, "sentenceContent": "대만에서 규모 6.1의 지진이 발생했다.", "subjectConsistencyYn": "Y"},  
        {  
            "sentenceNo": 2, "sentenceContent": "최대 진도는 4급으로 타이베이 등 대다수 지역에서 진동을 느낄 수 있었다.", "subjectConsistencyYn": "Y"},  
        {  
            "sentenceNo": 3, "sentenceContent": "5월 대만 BCC 등 현지 언론에 따르면 이날 오전 5시50분 지진이 발생했다.", "subjectConsistencyYn": "Y"},  
        {  
            "sentenceNo": 4, "sentenceContent": "장악자는 대만 동부해역 70.1km 지점이다.", "subjectConsistencyYn": "Y"},  
        {  
            "sentenceNo": 5, "sentenceContent": "여진은 계속되고 있다.", "subjectConsistencyYn": "Y"},  
        {  
            "sentenceNo": 6, "sentenceContent": "이에 따라 현지 중국인들의 안전에 대한 우려가 제기되자 중국 정부는 웜크라이나 측에 차관면 앤전  
    ]  
},  
    "clickableClass": 0  
}
```

- 직접생성 라벨링데이터 파일 예시

원천데이터 정보(sourceDataInfo)

```

"sourceDataInfo": {
  "newsId": "LC_M09_659735",
  "newsCategory": "문화체육",
  "newsSubCategory": "문화일반",
  "newsTitle": "“반에 꽂을 거닐어볼까?”',
  "newsSubTitle": "null",
  "newsContent": "다음달 1일부터 1월여간 서울시내 고공을 범에 거닐 수 있다.\n문화재청은 3월1일부터 4월4일까지 매일 오후 7~18시
  "partNum": "P2",
  "useType": 0,
  "processType": "D",
  "processPattern": "21",
  "processLevel": "중",
  "lastUpdateDate": "2023-01-10T10:00:00Z"
}

```

라벨링데이터 정보(sourceDataInfo)

```

"labeledDataInfo": {
  "processSentenceInfo": [
    {
      "sentenceNo": 1,
      "sentenceContent": "다음달 1일부터 1월여간 서울시내 고공을 범에 거닐 수 있다.",
      "subjectConsistencyYn": "Y"
    },
    {
      "sentenceNo": 2,
      "sentenceContent": "문화재청은 3월1일부터 4월4일까지 매일 오후 7~18시 고공 경내를 돌아볼 수 있는 야간 특별관람 프로그램이다.",
      "subjectConsistencyYn": "Y"
    },
    {
      "sentenceNo": 3,
      "sentenceContent": "이번 특별관람에서 주목되는 곳은 경복궁, 임금이 나았을 때 보던 정각인 「사직자(思寂臺)」과 참자인 청운대(淸潤臺)이다.",
      "subjectConsistencyYn": "Y"
    },
    {
      "sentenceNo": 4,
      "sentenceContent": "경복궁 둘레미 국립고궁박물관도 특별관람기간 관람시간과 같이 연장돼 전시감상 기회도 겸으로 누릴 수 있다.",
      "subjectConsistencyYn": "Y"
    },
    {
      "sentenceNo": 5,
      "sentenceContent": "단, 경복궁은 화요일, 청운대는 토요일 문을 닫는다.",
      "subjectConsistencyYn": "Y"
    },
    {
      "sentenceNo": 6,
      "sentenceContent": "야간 특별관람은 4월29일부터 6월1일까지 한번 더 진행된다.",
      "subjectConsistencyYn": "Y"
    }
  ]
}

```

타기사 1문장이 기준 6번째 문장을 대체, 도메인 일치여부 = 'N' 처리

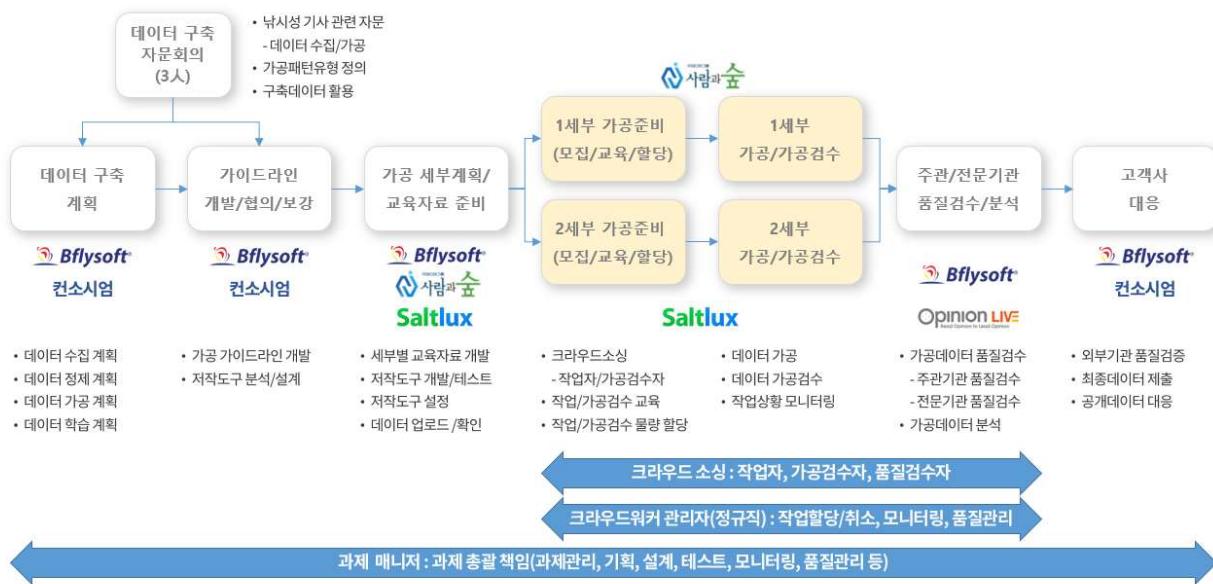
clickbaitClass: 0 ← 낚시성 기사 여부(0)

● 데이터 가공 오류 사례

가공패턴유형 미준수	<ul style="list-style-type: none"> 원문 기사와 어울리면서 가공패턴유형에 맞게 작성하지 않고, 가공패턴유형만 고려하여 다른 내용 작성 각 기사별 부여된 1세부/2세부 가공패턴유형에 맞지 않게 제목 및 본문 작성 → 의미 정확성에 부합하지 않는 경우
가공내용 일관성 부족	<ul style="list-style-type: none"> 2세부 직접생성 문장 중 각 문장간의 연결성 없이 서로 다른 내용을 각각 작성 → 의미 정확성에 부합하지 않는 경우
맞춤법 오류 (오탈자)	<ul style="list-style-type: none"> 1세부 새제목, 2세부 추가 작성 문장에서 오탈자 오류 발생 <ul style="list-style-type: none"> [속보] 이준석 측 "與, 무효 비대위 강행...추가 가치분 신청" → [속보] 이준석 측 "與, 무효 비대위 강행...추가 가치분 신청"
맞춤법 오류 (띄어쓰기)	<ul style="list-style-type: none"> 1세부 새제목, 2세부 추가 작성 문장에서 띄어쓰기 오류 발생 <ul style="list-style-type: none"> 삼성, 2분기 스마트 워치시장서 2위...샤오미 제쳤다 → 삼성, 2분기 스마트워치 시장서 2위...샤오미 제쳤다
맞춤법 오류 (잘못된 표기)	<ul style="list-style-type: none"> 1세부 새제목, 2세부 추가 작성 문장에서 잘못된 표기로 표현 사용 <ul style="list-style-type: none"> 년간 2~5개의 전시회 개최 → 연간 2~5개의 전시회 개최 자살률 → 자살률, 센터 → 센터, 포탈 → 포털, 파이팅 → 파이팅, 화일 → 파일
중복 가공	<ul style="list-style-type: none"> 크라우드 워커가 동일한 1세부 새제목, 2세부의 추가 작성 문장을 기사마다 중복하여 내용 작성 <ul style="list-style-type: none"> 기사 1 : 한 사람 지키려 대한민국 국가기관이 무너졌다. 기사 2 : 한 사람 지키려 대한민국 국가기관이 무너졌다.
문장 이상 종결	<ul style="list-style-type: none"> 1세부 새제목, 2세부 추가 작성 문장이 문맥, 문법적으로 맞지 않거나, 비정상 작성 후 종료 <ul style="list-style-type: none"> 기사 제목 : 삼성, 2분기 글로벌 스마트워치 시장 ← 술어없이 자리수만 채우고 작성 종결 기사 본문 : 한편, 부동산 정보를 손쉽게 얻을. ← 완성되지 않은 문장 작성 후 가공 작업 종료

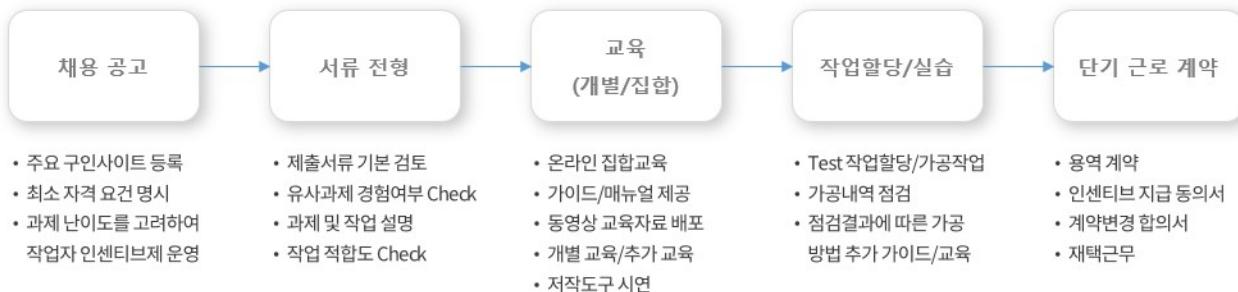
5.3 가공(라벨링) 조직

● 가공(라벨링) 조직



● 인력 조달 및 교육

- 채용 공고를 통한 지속적인 크라우드워커 모집
- 컨소시엄 기관별 내부 채용 프로세스에 따라 단기 계약 형태로 고용 진행
- 사업, 과제, 저작도구에 대한 개념 교육 및 실습 교육 진행



5.4 가공(라밸링) 도구

- 낚시성 기사 직접생성용 저작도구 : RANGPUR(미디어그룹사람과숲 제공)

The screenshot shows the Rangpur interface with four main sections:

- 사용자 관리**: Lists users for creation/management.
- 기사 목록 관리**: Lists articles for review, including filtering by status (전체, 미활성화, 활성화), category (경제, 정치, 사회), and keyword.
- 기사 상세(팝업)**: Shows detailed view of an article with various sections like Headline, Content, and Footer.
- 이력 관리**: Lists article history.

A large blue arrow points from the left side of the interface towards the right, where a detailed view of the article list is shown. This view includes columns for Article ID, Category, Post Type, Article Title, Author, Reviewer, Status, and Approval Status. The interface also features a search bar and a toolbar at the top right.

기사ID	카테고리	제작유형	기사제목	작업자	가공검수자	불합격수자	진행상태	관리
EC_M02_155904	경제	의도적 주어 왜곡형	기온주거가들 속속 스트어드십 코드 도입	HF_worker01	HF_reviewer10	qcb511 qcb512 qcb511	최종완료	
EC_M02_155905	경제	의도적 주어 왜곡형	공공구매에 사회가치 본격 반영	HF_worker01	HF_reviewer10	qcb511 qcb512 qcb511	최종완료	
EC_M02_155981	경제	의도적 주어 왜곡형	[포스코의 신전포고<2>] 글로벌 6위급 '기면' 역할 이어간다	HF_worker15	HF_reviewer07	qcb511 qcb512 qcb511	최종완료	
EC_M02_156547	경제	속어/줄임말 사용형	물가·경계 주양...뉴 경계 투톱, 막진 심티래 물	HF_worker12	HF_reviewer09	qcb515 qcb516 qcb517	최종완료	
EC_M02_157548	경제	의문유발형 (부호)	최기한 자제 불년 풍도2. 커迩러가 알아서 따라온네! 보이즈그룹에 최적	HF_worker27	HF_reviewer09	qcb517 qcb518 qcb513	최종완료	
EC_M02_157567	경제	의문유발형 (부호)	w'지역사회·환경 기여'·한국맥도날드, 새로운 실천 계획 발표	HF_worker27	HF_reviewer12	qcb515 qcb516 qcb513	최종완료	
EC_M02_155906	경제	의도적 주어 왜곡형	금증권, 중소 벤처기업 지원에 앞장	HF_worker01	HF_reviewer10	qcb511 qcb512 qcb511	최종완료	
EC_M02_155907	경제	의도적 주어 왜곡형	의도적 주어 왜곡형	HF_worker01	HF_reviewer10	qcb511 qcb512 qcb511	최종완료	
EC_M02_155909	경제	의도적 주어 왜곡형	카카오는 출퇴근으로 후달 할까?			qcb511 qcb512 qcb511	최종완료	
EC_M02_157570	경제	의문유발형 (부호)	한국타이어, 경영은 본장·3대! 구도 가치화	HF_worker27	HF_reviewer12	qcb515 qcb516 qcb513	최종완료	
EC_M02_155908	경제	의도적 주어 왜곡형	공영위, 대별 '온라인' 지역원 치과회사 수탁구조	HF_worker01	HF_reviewer10	qcb511 qcb512 qcb511	최종완료	
EC_M02_155910	경제	의도적 주어 왜곡형	공공기관이 사회적 가치를 창출하는 유통망 확	HF_worker18	HF_reviewer10	qcb511 qcb512 qcb511	최종완료	

- 사용자 관리 : 사용자 계정 등록/수정/삭제(사용자 정보, 비밀번호 관리 등)
- 사용자 권한 관리 : 사용자 계정에 대한 사용권한(관리자, 작업자, 가공검수자, 품질검수자 등) 관리
- 개인정보 수정 : 로그인 사용자가 비밀번호 정보 변경
- 로그인/로그아웃 : 저작도구에 대한 로그인/로그아웃 처리
- 기사 목록 : 기사 목록 조건 검색, 기사에 대한 작업자/가공검수자/품질검수자 할당/할당취소
- 기사 상세 : 기사에 대한 상세 정보 조회, 패턴유형 도움말 조회, 1세부/2세부 기사 가공, 1세부/2세부 기사 가공/품질검수, 가공정보 임시저장 및 해당 기사건에 대한 Skip(가공 작업 포기) 처리
- 이력 관리 : 가공이력에 대한 목록 조건 검색

※ 저작도구는 AI Hub에 공개하지 않고, 라밸링 데이터를 조회할 수 있는 'Clickbait Viewer' 제공

- 낚시성 기사 가공데이터 뷰어 : Clickbait_Visualizer(미디어그룹사람과숲 제공)

- 1세부 라벨링 데이터 조회

No	본문 문장 내용	생성참조
8	제판부는 '영화관, 배리어 프리 상영 노력해라' 해설이나 지역 파일이 제공되는 영화는 시각 장애인에 화면 해설을 제공하라고 말했다.	N
9	더불어 제판부는 '영화관에 자막이나 화면 해설이 제공되는 영화의 상영 정보를 홈페이지를 통해 제공해야 한다. 영화관은 자체나 다른 출처로 확대판 문서, 한국 수어 통역 등을 제공해야 한다'라고 말했다.	N
10	CGV는 영화 목록에 대한 맨의 제공을 하려면 과도한 비용이 들어 부담이 된다고 주장했다. 그러나 제판부는 '부산국제영화제 등에서는 배리어 프리 영상 시 스마트폰 애플리케이션을 통해 영화관 측의 상영장을 열었다.	N
11	제판부는 '제작자나 관객이나 화면 해설이 제공되는 영화의 상영 정보를 홈페이지를 통해 유니버설 디자인으로 장애나 저령 등에 관계 없이 모든 사람이 제품, 서비스, 환경 등을 편리하고 안전하게 이용할 수 있도록 설계하는 데 있다'라고 한다.	Y
12	이와 같은 판결에 기업은 영화관의 '유니버설 디자인(Universal Design)'에 대한 고민이 필요하겠다.	N
13	유니버설 디자인으로 장애나 저령 등에 관계 없이 모든 사람이 제품, 서비스, 환경 등을 편리하고 안전하게 이용할 수 있도록 설계하는 데 있다'라고 한다.	N
14	CGV는 '국내 영화관의 약 80%를 차지하고 있다'라고 한다.	N

- 2세부 라벨링 데이터 조회

No	본문 문장 내용	일관성
5	그리고, '2021년을 품이켜보면 점을 찾은 일들이 있었어요. 가장 조용하게 있을 수 있는 시간이 되지 않을까' 하는 잠적 출연에 대한 기대를 가졌다고 밝혔다.	Y
6	작년 겨울 하굣나라에 간 반려견 '돌마'에 대한 애정한 마음도 보였다. '함께 시간들을 책으로 넘길 수 있어서 좋았던'이라고 말하며 돌마 솔직히 돌아온 그녀의 눈을 속한 '돌마'에 대한 젊은 사랑이 있었다.	Y
7	이외에도 인터뷰 영상을 통해 연간 문소리의 대체로운 인상스러움을 볼 수 있다.	Y
8	데뷔 후 30년째 배우, 감독, 프로듀서 등 커리어를 멈추지 않고 있는 대한민국 대표 연기 배우 문소리는 지난해 트로트를 크라운을 달성하며 대중들로부터 열렬한 지지를 받았고 그 기운을 몰아 최근에는 '원미미카'의 주인공으로 활동되었던 것임을 이야기하고 있다. '정적·문소리·민'이라는 10일 디스커버리 채널 코리아와 SKY채널을 통해 첫방송을 예정이며, 온라인 콘텐츠(OTT) seezn(시즌)을 통해 방송 1주일 전 선공개됐다.	Y
9	10일 공개가 될 '정적·문소리·민'을 앞두고 배우 문소리의 인터뷰 영상이 디스커버리 채널, SKY채널, seezn 유튜브채널에서 선공개된 한편, 배우 문소리가 배우故 강수연의 영결식에 참석하여 눈물을 흘리는 모습을 보았다.	N
10	강수연은 향년 55세로 자택에서 뇌출혈로 인한 심장지 상태로 발견되어 곧바로 병원으로 후송되었으나 사흘 만에 끝내 숨을 거렸다.	N

- 라벨링 데이터 업로드 : AI Hub에 공개된 라벨링 데이터(JSON) 파일을 다운로드 받아서 뷰어에서 파일 업로드(1건)

- 라벨링 데이터 조회 : 기사에 대한 기본 정보, 원문 정보 및 1세부/2세부 가공 정보 내용 조회

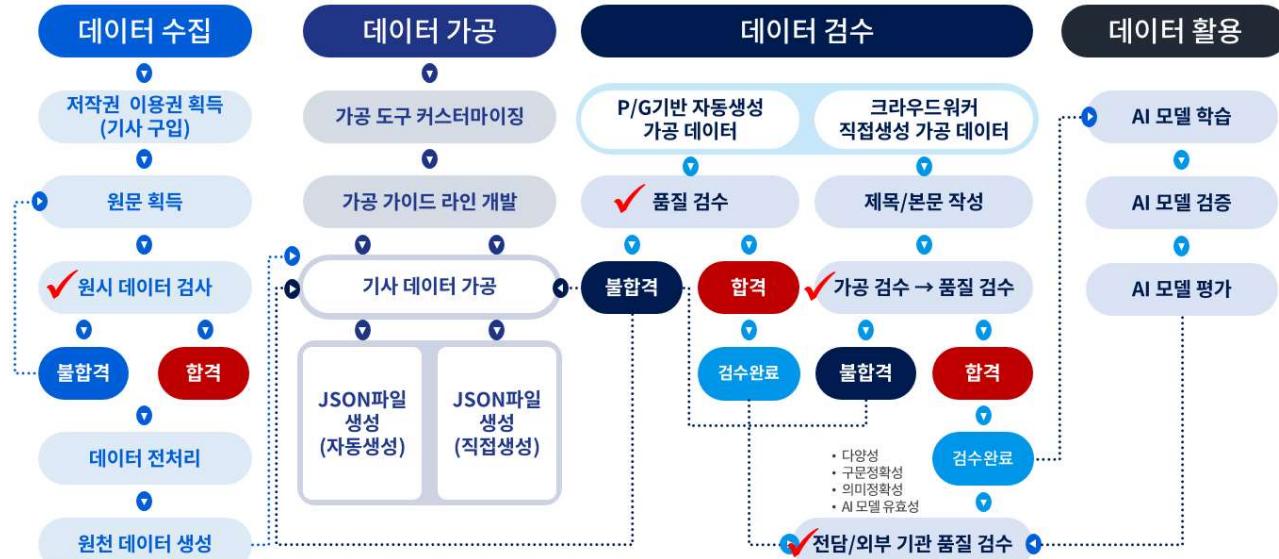
6. 검사

6.1 검사 절차

● 검사 절차별 검사기준 수립

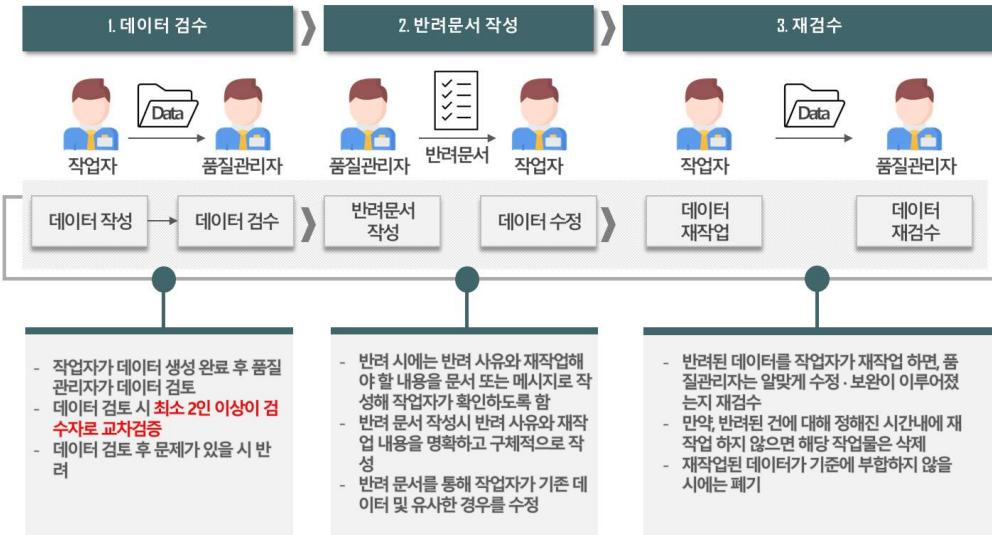
검사 절차	검사 대상	검사 요구사항
임무정의, 구축계획 수립 및 데이터 수집 단계	법·제도 준수	· 원시데이터 수집 시 뉴스 저작권 등 법·제도적 규정 준수
	데이터 동기화	· 언론사별 뉴스 데이터 수집 시 공통항목, 공통포맷에 따른 원시데이터 수집 준수
	편향성 방지	· 데이터 수집 시 특정 기사 카테고리가 전체 데이터의 40%를 넘지 않도록 수집 준수
데이터 정제 단계	정제기준의 명확성	· 데이터 사용(가공) 목적에 적합한 정제 기준 수립 여부 검수
	중복성 방지	· 데이터 정제 후 기사 제목/본문 내용 중심으로 정보 비교 후 중복도 여부 검수
	정제 작업 매뉴얼	· 정제 작업을 위한 매뉴얼 작성 및 관리 여부 검수
	정제 도구	· 정제 작업에 사용될 소프트웨어 도구 확보 및 사용 방법 숙지 여부 검수
	정제 작업 방법	· 데이터 특성 및 활용 목적에 맞는 적절한 정제 방식 설정 여부 및 설정 기준 타당성 여부 검수
데이터 라벨링 단계	라벨링 가이드라인	· 목적에 맞게 작성된 라벨링 가이드라인에 대한 타당성 여부 검사 후 라벨링 작업자들에게 해당 내용의 가이드라인 전달
	어노테이션 항목	· 목적에 맞는 어노테이션 구성 여부 검수 후 확인된 내용을 포함하도록 작업자들에게 전달
	가공 검수 도구	· 자동화 도구를 통해 검수 후 검수자가 육안으로 부적합 데이터 여부에 대해 2차 확인 및 조건 오류 전수 검수
전수 검사	부적합 판정 데이터 분포 확인	· 데이터의 오류율, 특성 분포 확인을 통한 데이터 수집, 정제, 가공, 부문 최적화
	외부 검수자	· 외부 검사자(TTA 등), 도메인 전문가, 데이터 요청자에 의한 검사를 위해 검사지표 및 기준이 통일되게 전달되었고, 검사 방법의 동의가 이루어졌는지 확인

● 단계별 구축 프로세스 중 검사 포인트



● 오류 데이터 재작업 절차

- 데이터의 검사 절차는 최소 2인 이상의 검수자가 검수 하며, 검수한 결과를 교차검증하여 오판을 줄임
- 2인 교차검증 시 판정이 다를 경우, 제 3자 검수자가 다시 검수 하여, 최종 판정을 내림
- 데이터 검수 결과 기준에 부합하지 않은 데이터는 반려사유와 함께 작업자에게 전달하여 재작업
- 재작업된 데이터는 다시 품질관리자를 통해 재검수하여 수정보완이 이루어졌는지 확인
- 반려된 데이터는 반려 사유와 재작업 방법을 유형별로 데이터 구축 가이드에 작성하여 향후 데이터 구축 업무에 참고할 수 있도록 정보 제공



※ 예외적으로 작업자 혹은 관리자가 판단하기 모호하거나, 작업 기준이 모호한 경우 선택을 보류하고 별도 판정단이 해당 작업물에 대해 판별한 후, 작업할 수 없다면 폐기처리, 작업할 수 있는 경우 가이드와 함께 반려 처리하는 경우도 존재

● 직접생성 가공단계 검수 포인트 및 주요 가공 오류 유형

The screenshot displays two examples of inspection forms for direct generation processing:

- Example 1:** Shows a general inspection form with fields for '기사 기본 정보' (Article Basic Info), '기사 가공 정보' (Article Processing Info), and a '본문' (Body Text) section. The body text includes a note about reporting errors if the article is published without review.
- Example 2:** Shows a detailed inspection form with sections for '기사 기본 정보' (Article Basic Info), '기사 가공 정보' (Article Processing Info), and a '본문' (Body Text) section. The body text includes a note about reporting errors if the article is published without review.

- 가공 데이터에 대해 가공검수자 1명 1차 검수(Pass/Fail), 품질검수자 3명 2차 검수(Pass/Fail) 후 가공 최종완료 처리
- 가공검수/품질검수에서 Fail된 가공건은 다시 가공단계부터 재수행하여 2명 이상의 품질검수자 Pass를 받은 경우에 가공완료
- 최종완료된 가공데이터는 컨소시엄내 품질검수 전문기관(오피니언라이브)의 3차 검수 진행
- 최종검증 완료된 가공데이터는 TTA에 최종점검 의뢰하여 최종검증 수행

← 패턴유형에 맞게 새제목 직접 작성
(3어절 이상, 20~100글자, 저작도구 자동 체크)
← 새제목 작성 시 참조문장 Check(1문장 이상)

주요 가공 오류 유형

- ✓ 가공패턴유형에 부적합한 새제목, 본문 문장 작성
- ✓ 맞춤법(오탈자, 띄어쓰기) 오류
- ✓ 동일 내용을 기사에 중복 사용
- ✓ (본문문장) 한절에 2문장 이상 작성 오류
- ✓ (본문문장) 한문장 비정상 종료 처리 오류
- ✓ (본문문장) 첫문장 전반부에 원문기사 관련 내용 부재 오류
- ✓ (본문문장) 첫문장 후반부의 부자연스러운 화제 전환 오류

← 패턴유형에 맞게 본문 문장 직접 작성
(2~4문장, 30~300글자, 저작도구 자동체크)

6.2 검사 기준

● 검사기준 및 방법

품질지표		산출물		검사방법
준비성	계획수립성	과정준비	사업수행계획서/구축계획서/품질계획서/품질검증 합의서	체크리스트
		조직준비	사업수행계획서/구축계획서/품질계획서	체크리스트
		도구준비	사업수행계획서/구축계획서/	체크리스트
		위험관리	사업수행계획서/구축계획서/품질계획서	체크리스트
	체계준수성	보안준수	사업수행계획서/품질계획서	체크리스트
		법·제도 준수	사업수행계획서/품질계획서	체크리스트
완전성	수집완전성	구축계획서/품질검증 합의서/구축.활용 가이드	체크리스트	
	정제완전성	구축계획서/품질검증 합의서/구축.활용 가이드	체크리스트	
	가공완전성	구축계획서/품질검증 합의서/구축.활용 가이드	체크리스트	
유용성	사용편의성	구축.활용 가이드	체크리스트	
	유연성	구축.활용 가이드	체크리스트	
기준 적합성	다양성	구축.활용 가이드/원시데이터	체크리스트/전수.샘플링 검사	
	신뢰성	구축.활용 가이드/원시데이터	체크리스트/전수.샘플링 검사	
	충분성	구축.활용 가이드/원시데이터	체크리스트/전수.샘플링 검사	
	균일성	구축.활용 가이드/원시데이터	체크리스트/전수.샘플링 검사	
	사실성	구축.활용 가이드/원시데이터	체크리스트/전수.샘플링 검사	
	공평성	구축.활용 가이드/원시데이터	체크리스트/전수.샘플링 검사	
기술 적합성	파일포맷	원시데이터	준수율 계산 (전수검사)	
	어절 수	원시데이터	준수율 계산 (전수검사)	
통계적 다양성	낚시성 기사 난이도	원천데이터/라벨링데이터	품질검사도구 이용 분포 확인	
	뉴스 문장 수	원천데이터/라벨링데이터	품질검사도구 이용 분포 확인	
	생성된 제목 어절 수	원천데이터/라벨링데이터	품질검사도구 이용 분포 확인	
	생성된 본문 어절 수	원천데이터/라벨링데이터	품질검사도구 이용 분포 확인	
	뉴스 가공 패턴 유형 분포	원천데이터/라벨링데이터	구성비 충첩률 계산/확인	
의미 정확성	제목-본문 비연관성	라벨링데이터	크라우드소싱을 통한 검사진행 결과 집계 및 준수율 계산	
	본문-도메인 비일관성	라벨링데이터	크라우드소싱을 통한 검사진행 결과 집계 및 준수율 계산	
구문 정확성	데이터 구조	라벨링데이터	품질검사도구 이용 오류율 계산	
	데이터 형식	라벨링데이터	품질검사도구 이용 오류율 계산	

품질지표		산출물	검사방법
유효성	정확도	낚시성 기사 분류 모델	유효성 지표값(Accuracy) 산출/확인
		본문 주제분리 탐지 모델	유효성 지표값(Accuracy/F1-Score) 산출/확인

● 1세부 - 제목과 본문 불일치 검사 기준(의미 정확성)

가공유형	유형 설명	낚시성 예시	비교방법	판단기준	문장 예시
의문유발형 (부호)	기사 제목에 사안의 본질을 요약하지 않고?, !, ...등의 문장부호를 사용하여 독자의 궁금증을 유발하는 기사의 제목 ...결국', '알고보니...', '왜'등의 문을 유발하는 형태의 표현을 사용하여 클릭을 유도하는 기사의 제목	왜?, 정체는?, 비결은?, 누구인가?	원 제 목 (newsTitle)과 새 제 목 (newTitle) 비교	1. 문장의 종결에는 무조건 문장부호가 들어가야 함 2. 은닉형지시대명사 사용은 기본적으로 금지하나, 어떤/누구/무슨/언제/얼마등의 표현은 맥락상 적절하다면 인정됨	1. SK 하이닉스·마이크론 200단 진출.. '낸드 1위' 삼성은 왜? 2. 편의점 마스크 착용 거부한 40대 남성 알바생 폭행 → 여 알바생 마스크 요구에 주먹질한 중년남 정체는?
의문유발형 (은닉)	말줄임표(...)를 삽입하고 원인, 주어, 설명 등을 명확하게 표시하지 않고, 궁금증 유발하는 문장(?)으로 종결하는 형태로 작성한 기사의 제목	이것/그것/무엇/이곳/저곳 등	원 제 목 (newsTitle)과 새 제 목 (newTitle) 비교	1. 대상을 명확히 하지 않고 숨기는 의도로, '이것'/그것' 등의 지시대명사 사용 2. '무엇' / '-것은?' 등의 표현으로 의문을 유발하는 형태 3. 주어를 숨기는 것이 아니라 주어가 별인 행동, 목적어, 등의 2차 키워드를 숨겨야함	1. 춘천막국수닭갈비축제, '이것'을 주목하라 2. "중환자가 병원 안 와요" 이것만 알려도 세 모녀 비극 막는다 3. 전세값이 매매가의 97%...경찰이 이 지역집중수사 한다는데... 4. 등통증·췌장낭종 있으면 췌장암?... 이런증상
선정표현 사용형	주어를 묘사하는 술어에 선정적인 단어를 사용하여 작성된 기사의 제목 주어가 불특정되어 있는 상황에서 주어를 묘사하는 술어에 선정적인 단어를 사용하여 작성된 기사의 제목	화끈한, 터질듯한, 끌벅지, 은밀한, 발가벗은	원 제 목 (newsTitle)과 새 제 목 (newTitle) 비교	1. 주어를 묘사하는 술어에 선정적인 단어, 부사 등을 사용한 형태 내용전달에 불필요한 선정적 코드-성, 폭력과 연관된 표현이 드러남 2. 기사의 주된 내용, 맥락, 분위기에 어울리는 표현을 써야함, 제3자가 작성 제목만 읽어봐도, 뜬금 없어 보이지 않도록 신경쓸 것	1. 자살 막은 '꿀벅지 협객녀' 알고보니... "세상 살만하네!" 2. 2NE1 CL, 영혼까지 끌어 모았네...터질듯한 불룸감 3. BTS 콘서트 하루전, 후끈 달아오른 부산...글로벌 아미 속속 집결
속어/줄임말 사용형	속어, 신조어, 줄임말 등을 사용하여 작성된 기사의 제목	알 잘 딱 갈센, 야알못, 야잘알, 혼타, 많관부	원 제 목 (newsTitle)과 새 제 목 (newTitle) 비교	1. 속어, 신조어, 줄임말 등을 사용한 제목 형태 2. 일반 표준어가 단순 '유행어-밈'으로 쓰이는 경우는 우선 제외함 ※ 개인 신상 정보 (EX: 이름) 드러나있지 않고 + 이미 기사 내에서 직접 쓰였다는 전제 하에 인용구 (큰 따옴표 사용)로, 기사 내 표현 그대로 사용하는 것은 인정됨 *	1. '초인가족' 김지민, '급식총' 메시지에 '읽씹'으로 응수 2. 여친 엄마가 사준 금목걸이 본 남친, "김치녀같아" 패드립 3. 문 정부 집값 폭등, "낄끼빠빠모르고 시장 무시한 결과"
사실 과대 표현형	기사 제목상 주어에 대한 술어의 내용이, 본문상의 주어에 대한 술어의 내용보다	대참사, 폭등, 역대, 최악, 참사	본문과 새제 목(newTitle)의 주어에 대한 술어의 내용보다	1. 제목상 술어의 내용이, 본문상 술어의 내용보다 과장(과대/과도) 표현된 형태	1. 무료접종 중단 이틀째 ... "돈 내고 맞겠다" 백신대란 장사진 2. 김지혜, 술을 그렇게 마

가공유형	유형 설명	낚시성 예시	비교방법	판단기준	문장 예시
	과장(과대/과도)된 기사		용 확인 및 비교	<p>2. 본문 내용을 의도적으로 확대 해석 / 자극적 표현 사용하는 것이 핵심</p> <p>3. 기사의 팩트가 무조건 부풀려져야 함. 비슷한 표현으로 대체되는 것은 의미가 없음</p>	<p>시더니.치아 대참사어쩌나</p> <p>3. 환율 1,400원 눈앞... 경제 덮친 '환율 재앙': 환율의 절대적 수치만 제시해놓고 '재앙'이라는 과대 표현으로 시선을 끔</p> <p>4. 현대건설 '서울 재개발 현장' 폭발사고 대참사... 인근 주택 아주관장 : 기사의 팩트는 7명의 중경상 정도였음. '대참사'는 사실 과대에 해당.</p>

● 2세부 - 본문과 본문 불일치 검사 기준(의미 정확성)

가공유형	유형 설명	낚시성 예시	비교방법	참고사항	문장예시
상품 판매정보 노출 광고형	자동차, 스마트폰 등 상품에 대한 직접적인 판매 관련 정보를 본문에 담고 있는 기사	그 종 꽃송이 버섯 브랜드 훈O는 청정 지역에서 고 품질로 생산된 국내산 백 아산 꽃송이 버섯을 사용하여 제조.... 자세한상품정보는 다음사이트를 통해알아 볼 수 있다. http://www.hanatour.com/	n'으로 표기된 문장 위에 1 ~ 2 문장을 보고 판단	<p>1. 기사 본문 내용 중 직접적인 판매 관련 정보를 본문에 담고 있는 기사</p> <p>2. 금융상품(은행상품, 카드상품, 보험/공제상품, 파생상품) - 무형이나 상품으로 분류됨</p> <p>3. 제조사명, 상품명, 해당 상품에 대한 스펙, 특장점, 가격, 문의사항(전화번호, 사이트 주소 등)</p>	<ul style="list-style-type: none"> ○ 기사 제목 : 강민경, 하와이로 떠난 휴가...원피스 입고 뽐낸 글래머 몸매 기사 본문 : 그룹 다비치 강민경이 하와이에서 휴가를 즐겼다. ... 추가 문장 예시 1 : 한편, 하나님께서는 가을을 맞아 결혼하는 신혼부부를 위해 하와이 팩키지 여행 상품을 ... 자세한상품정보는 다음사이트를 통해알아볼 수 있다. http://www.hanatour.com/. ←여행상품광고 추가 문장 예시 2 : 한편, 강민경이 속한 다비치는 지난 5월 미니 앨범'시즌 노트(Season Note)를 발매했다. ← 앨범 광고(홍보)
부동산 판매정보 노출 광고형	기사 본문에 부동산에 대한 판매 관련 광고성 정보(분양정보 등)를 포함하고 있는 기사	부동산 상품의 가격, 주소, 사이트 링크, 연락처 중 하나 이상을 노출하는 형태로 문장을 추가하여 본문 구성	n'으로 표기된 문장 위에 1 ~ 2 문장을 보고 판단	<p>1. 기사 본문 내용 중 '부동산'에 대한 판매 관련 광고성 정보를 포함하고 있는 기사</p> <p>2. 예시 : 토지(논, 밭, 나대지, 과수원, 산 등), 아파트, 전원주택단지, 오피스텔, 주상복합 빌라, 신도시, 재건축, 재개발, 매매, 경매, 리조트, 임야 또는 각종 부동산 분양 또는 투자 정보 등</p>	<ul style="list-style-type: none"> - 기사제목:정자역주근접 단지'e편한세상정자더센트럴'분양 - 기사본문:(기존문장...) + 전용 면적별로 가격은 84m² 가 10억4천만원이며 전세대가 수요자들의 선호도가 높은 전용 면적 85m² 평면으로 구성된다. 분양사무소는 정자역 1번 출구 더센트럴빌딩 205호이며 전화 번호는 031-713-9697이다.
서비스 판매정보 노출 광고형	기사 본문에 금융 서비스(은행, 증권, 보험 등), 의료 서비스, 여행 서비스, 다양한 IT 서비스, 공공 서비스 등의 서비스에 대한 판매 관련 광고성 정보, 홍보성 정보를 포함하고 있는 기사	서비스의 가격, 주소, 사이트 링크, 연락처 등을 노출하던 형태로 문장들을 추가하여 본문 구성	n'으로 표기된 문장 위에 1 ~ 2 문장을 보고 판단	<p>1. 각종 서비스(형태가 없는, 구입 후 매매가 불가능한)를 포함하고 있는 기사</p> <p>2. 여행 또는 숙박 관련 서비스, 인터넷 서비스 각종 웹/앱 서비스(앱을 이용한 정보 서비스) 3. 컨설팅(법률, 경영, 의료 등)</p>	<ul style="list-style-type: none"> - (본문 계속...) 덕윤씨에게 행운을 준 로또업체는 국내에서 최초로 로또예상 번호를 추첨하는 AI 필터링 시스템을 개발하여 높은 확률로 당첨 예상번호를 추출해 내고 있으며, 그 당첨 확률이 최근 98%를 넘어선 것으로 알려졌다. 해당 로또업체는 창립 10주년 이벤트로 선착순 50명에게만 AI 필터링으로 분석

가공유형	유형 설명	낚시성 예시	비교방법	참고사항	문장예시
의도적 상황 왜곡/전환형	드라마, 영화, 꿈, 가상현실, 메타버스, 게임 등의 상황을 표기하지 않고, 현실의 상황인 것처럼 작성한 문장	기사 제목과 'n'으로 표기된 문장을 보고 판단			- 제목: 김고은 "'작은아씨들 '출연가슴벅차...'장르파괴자'역할" - 한편, 김고은은 지난하지만 우애있게 자란 자매들과 함께 대한민국에서 제일 부유하고 유력한 가문에 각자의 방식으로 맞서 살아왔다. (드라마 내용을 드라마로 표기하지 않고 작성)
	기사 제목 상의 주어(대상)의 상황과 본문의 주어(대상)의 상황이 불일치하는 내용이 기사 본문 문장 중에서 일부 존재하는 기사	기사 제목에서 발생한 사건, 상황이 드라마, 영화의 내용에서 등장한 것처럼 본문의 내용을 추가하여 작성	기사 제목과 'n'으로 표기된 문장을 보고 판단		- 제목:(사회)"한국인들, 물가 상승에 6990 원 친사려 몇 시간씩 대기" - 추가본문: 영화 극한직업에서 수원 왕갈비 통닭은 15,000 원에서 시작해서 계속해서 가격을 인상하였으나, 대기하는 손님의 줄은 끊이지 않아, 밤늦게까지 영업을 지속해야만 했다. 갈비와 치킨의 만남으로 이름진 대박 상품이 탄생했다. (연예) 실버스터스 탤론, 76살 에이흔... 결혼 25년 만 - 추가본문: 실버스터스 탤론은 람보 라스트 워에서 할머니와 가브리엘 라의 친구들과 함께 장례식을 치른 후 지인 집으로 할머니를 보내고 자신의 농장에서 온갖 무기와 부비트랩을 준비하고 키우던 말도 내보낸다.

- 사용 제한 표현 : 욕설 및 혐오표현을 사용함으로써 기사에 사용하기 적절하지 않고 데이터의 목적에 부합하지 않은 표현은 사용을 제한

① 욕설표현

- 나무위키 욕설: <https://namu.wiki/w/%EC%9A%95%EC%84%A4/%ED%95%9C%EA%B5%AD%EC%96%B4>

② 혐오표현

- 개독, 개슬람, 맹중, 급식충, 콧중고딩, 초딩, 개초딩, 씹덕후, 오덕후, 네덕, XX퍼거, 삼엽충, 앱등이, 노력충, PC충, 닌빠, 엑빠, 플빠, 설명충, 진지충, 씹선비, 수구꼴통, 좌좀, 좌빨, 빨갱이, 틀딱충, 박사모, 노빠, 깨시민, 문빠, 달창, 대깨(대깨문, 대깨트, 대깨안, 대깨준, 대깨윤), 달빠, 릎빠, 키빠, 블빠, 롤충, 도슬람, 파오후, 안여돼, 안여멸, 근육돼지, 국뽕, 환빠, 국까, 헬무새, 역센징, 일뽕, 미빠, 중뽕, 러빠, 김여사, 맘충, 김치남, 김치녀, 된장녀, 한남충, 사생팬, 빠순이, 빠돌이, 배박이, 말박이, 칼박이, 럽폭도, 스끌, 스투충, 일베충, 오유충, 정사충, 쭈쭈충, 개빠, 고양이빠, 청위병, 나위병, 지접대, 메오후, 별창녀, 별풍서틀, 꿀마초, 꿀페미, 쿨게이, 스노브, 공장충, 휴거, 빌거, 200충, 300충, 조무사드립, 병신, 애자, 정신병자, 클럽충

③ 비하(차별)표현

- 비하/차별표현(국어사전)
 - 성차별: 여의사, 여필종부, 관능미, 출처, 늑대, 여교사, 여류..., 여사, 연놈, 여필종부, 미망인, 여편네, 부엌데기, 암캐, 여우, 기생 오라비, 학부형

- 인종차별: 오랑캐, 쪽발이, 검둥이/껌둥이/니그로, 코쟁이, 똥남아, 왜놈, 베트콩, 되놈, 짱개, 양놈, 양년, 양키, 흰동이, 튀기, 잡종, 짬뽕, 혼혈(압)
- 장애차별: 귀머거리, 난쟁이, 소경, 미친놈, 병신, 봉사장님, 애꾸눈, 병어리, 반병어리, 말더듬이, 곱사등이, 꼽추, 난쟁이, 앓은뱅이, 절름발이, 외팔이, 곱배팔이, 육순이, 미치광이, 미친놈/미친년, 머저리, 천치, 병신, 폐질자, 불구자, 장애자
- 지역차별: 명청도, GANGGAE, 경상디언, 뻔질이, 짠물, 문둥이, 핫바지, 서울깍쟁이, 상경/하경하다, 시골, 지방, 촌놈, 촌것, 촌뜨기, 촌티, 낙향
- 직업차별: 도공, 목공, 석공, 인쇄공, 전기공, 간판장이, 도배장이, 땀장이, 옹기장이, 환쟁이, 가수, 목수, 무용수, 석수, 신호수, 광부, 어부, 인부, 잡역부, 청소부, 가정부, 접대부, 파출부, 청소부, 보모, 식모, 유모, 침모, 장사꾼, 잡상인, 빼끼
- 종교차별: 개독교, 땅종, 무당질, 점쟁이, 개슬람, 중놈, 중질, 땅땡이중, 중대가리, 까까중, 예수쟁이
- 기타차별: 상갓/쌍갓, 상놈/쌍놈, 상년/쌍년, 천민, 천한, 아랫것, 서민, 하류층, 하층민(계층차별), 어린것, 늙은것, 늙다리, 노인(네)(나이차별), 키다리, 깍다리, 작다리, 난쟁이, 뚱보/뚱뚱보, 돼지(외모차별), 동성애, 동성애자, 호모, 게이, 변태, 변태성욕자(성소수자차별), 빨갱이/수꼴(정치/이념대립)
- 비하/차별표현(누리꾼)
 - 여성차별: 김여사, 김치녀, 된장녀, 맘충, 삼일한, 한남충
 - 남성차별: 냠저, 개자씨, 개자, 한남충→한남
 - 인종차별: 외노자, 종꿔, 풍꿔, 흑형, 간양남
 - 장애차별: 개병신, 병신크리, 병크, 셀카고자, 패션고자
 - 지역차별: 전라디언, 흥어, 까보전, 설라디언, 개쌍도, 개상도, 고담대구
 - 기타차별: 개독교, 개독, 좌좀, 좌빨, 종북좌빨, 수꼴, 일베충, 틀딱충, 틀딱, 급식충, 학식충, 진지충, 선비충, 설명충

● 품질 지표 및 정량 목표

품질특성	항목명	측정 지표	정량 목표
다양성 (통계)	낚시성 기사 난이도	비율	분포 확인
	뉴스 문장 수	비율	분포 확인
	생성된 제목 어절 수	비율	분포 확인
	생성된 본문 어절 수	비율	분포 확인
	뉴스 가공 패턴 유형 분포	비율	분포 확인
다양성 (요건)	기사 카테고리별 분포	구성비 중첩률	구성비 중첩률 50%
			분류 건수 비율
			정치 106,245 13%
			경제 114,086 14%
			사회 160,324 20%
			생활/문화 101,243 13%
			IT/과학 103,728 13%
			세계 118,268 15%
			연예 100,000 12%
			합계 100%
구문적 정확성	구조 정확성	정확도	99.5% 이상
	형식 정확성	정확도	
의미 정확성	제목-본문 패턴 유형 정확도	정확도	90% 이상
	본문-도메인 패턴 유형 정확도	정확도	90% 이상
유효성	낚시성 기사 분류 모델(HAND) 성능	Accuracy	60% 이상
	본문 주제 분리 탐지 모델(BERT) 성능	Accuracy	55% 이상

항목명	다양성
측정 지표	구성비 중첩률
측정 산식	$\text{구성비 중첩률}(\%) = \frac{\sum_{k=1}^K \text{중첩막대길이}_k}{\sum_{k=1}^K \text{최대막대길이}_k} \times 100$
항목명	구문 정확성
측정 지표	정확도
측정 산식	$\text{정확도}(\%) = \frac{(\text{전체속성수} - \text{오류건수}_{\text{구조}})}{\text{전체속성수}} \times 100$
항목명	의미정확성(제목-본문 패턴 유형 정확도)
검사 단위	텍스트(문장)
측정 지표	정확도
측정 산식	$\text{정확도} = \frac{\text{샘플링대상중 Pass총합}}{\text{샘플링총 개수}}$ <p>1. 평가 내용 직접 생성: 제목 분류 및 가공패턴 유형의 불일치를 검사 (불일치: 0, 일치: 1) 자동 생성: 가공 제목과 원천 기사 제목과의 일치 여부 검사 (불일치: 0, 일치: 1)</p> <p>2. Pass/Fail [직접 생성] 1) 제목 분류 검사 - Pass: 제목과 낚시성 기사 분류(0: 낚시성, 1: 비낚시성)이 일치하는 데이터 수 - Fail: 제목과 낚시성 기사 분류(0: 낚시성, 1: 비낚시성)이 불일치하는 데이터 수 2) 제목과 가공패턴 유형 검사 - Pass: 제목과 가공패턴 유형이 일치하는 데이터 - Fail: 제목과 가공패턴 유형이 불일치하는 데이터</p> <p>[자동 생성] 1) 제목 분류 검사 - Pass: 제목과 낚시성 기사 분류(0: 낚시성, 1: 비낚시성)이 일치하는 데이터 수 - Fail: 제목과 낚시성 기사 분류(0: 낚시성, 1: 비낚시성)이 불일치하는 데이터 수 2) 가공 제목과 원천 제목과의 일치 여부 검사 - Pass: 가공 제목과 원천 기사 제목과 불일치하는 데이터 - Fail: 가공 제목과 원천 기사 제목과 일치하는 데이터</p>
항목명	의미정확성(본문-도메인 패턴 유형 정확도)
검사 단위	텍스트(문단)
측정 지표	정확도
측정 산식	$\text{정확도} = \frac{\text{샘플링대상중 Pass총합}}{\text{샘플링총 개수}}$ <p>1. 평가 내용 직접 생성: 본문의 일관성 및 가공패턴 유형의 일치를 검사 (불일치: 0, 일치: 1) 자동 생성: 가공 본문 첫 문장과 원천 기사 본문 첫 문장과의 불일치 여부 (불일치: 0, 일치: 1)</p>

2. Pass/Fail

[직접 생성]

1) 본문의 일관성 검사

- Pass: 가공된 문장이 태깅된 주제의 일관성 여부와 일치하는 데이터
- Fail: 가공된 문장이 태깅된 주제의 일관성 여부와 불일치하는 데이터

2) 가공패턴 유형 검사

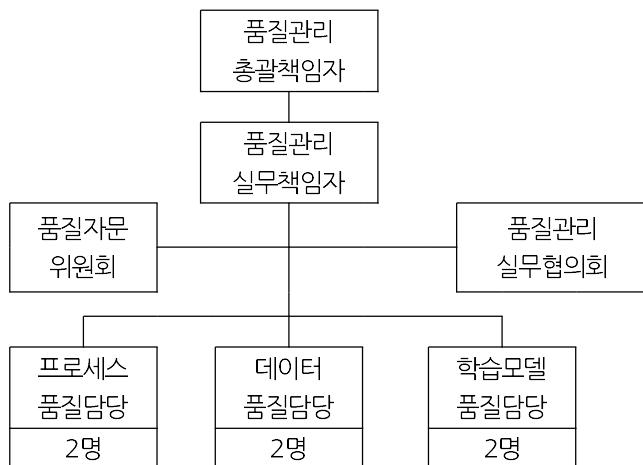
- Pass: 가공된 문장이 태깅된 가공패턴과 일치하는 데이터
- Fail: 가공된 문장이 태깅된 가공패턴과 불일치하는 데이터

[자동 생성]

1) 가공 문장과 원천 기사 본문 첫 문장 간의 불일치 여부 검사

- Pass: 가공된 문장 중 첫 문장과 동일한 위치의 원천 기사 문장과 불일치하는 데이터 수
- Fail: 가공된 문장 중 첫 문장과 동일한 위치의 원천 기사 문장과 일치하는 데이터 수

6.3 검사 조직



조직 구분	역할과 책임
품질관리 총괄책임자	<ul style="list-style-type: none"> 인공지능 학습용 데이터의 품질관리 총괄 품질관리를 위한 컨소시엄 간 업무 조정 품질관리 주요 정책에 대한 의사결정
품질관리 실무책임자	<ul style="list-style-type: none"> 인공지능 학습용 데이터의 품질관리 실무 총괄 품질관리 전반에 대한 업무 수행 품질관리 실무협의회 구성 및 운영 하위 품질관리 수행 조직 구성 및 운영 품질관리 상세 추진계획 수립 및 수행관리
품질관리 실무협의회	<ul style="list-style-type: none"> 인공지능 학습용 데이터의 품질관리 주요 계획, 품질 현안 등의 협의 품질관리 계획의 적정성 협의 및 보완 컨소시엄 간 협업 및 효율적 추진을 위한 실무 협의 품질관리 이슈 조정 및 의사 결정
품질자문 위원회	<ul style="list-style-type: none"> 내부 품질관리 과정 및 방법에 대한 3자 검증을 수행
프로세스 품질관리 담당	<ul style="list-style-type: none"> 구축 프로세스의 준비성, 완전성, 유용성에 대한 품질관리 담당 체크리스트를 통한 프로세스 품질관리 실무활동 수행 주기적인 품질 점검 및 진단 결과 보고
데이터 품질관리 담당	<ul style="list-style-type: none"> 인공지능 학습용 데이터 적합성 및 정확성에 대한 품질관리 담당 체크리스트를 통한 데이터 품질관리 실무활동 수행 품질관리 툴을 통한 데이터 표준화 실무활동 수행 구축 및 정제된 데이터에 대한 품질관리 실무활동 수행 주기적인 품질 점검 및 진단 결과 보고
학습모델 품질관리 담당	<ul style="list-style-type: none"> 인공지능 학습용 데이터를 활용한 학습모델에 대한 품질관리 담당 학습모델의 유효성 검사 계획 수립 및 검사 실무활동 수행 주기적인 품질 점검 및 진단 결과 보고

6.4 검사 도구

● 품질검사 영역별 검사 도구

품질검사영역	다양성/구문 정확성 검사	의미 정확성 검사	학습모델 유효성 검사
도구명	낚시성 기사 품질검사 도구	N/A	Python
설명	JSON 형태의 라벨링데이터 분석 도구 개발 및 활용	<ul style="list-style-type: none"> - 라벨링데이터의 샘플을 통한 수작업 검사 실시 - 데이터 가공 가이드라인을 활용한 검사기준서 작성 및 크라우드워커 교육 진행 - 크라우드워커를 통한 수작업 검사 진행 	<ul style="list-style-type: none"> - 학습용 데이터를 인공지능 모델을 이용하여 학습한 후 목표 수준의 달성이 가능한지 여부 검사
도구유형	전문기관 용역으로 도구 개발 - (주)다율디엔에스	<ul style="list-style-type: none"> - 라벨링데이터/JSON 뷰어/엑셀 	<ul style="list-style-type: none"> - 오픈소스 기반 자체 개발
주요기능	<ul style="list-style-type: none"> - 구축 데이터의 다양성 검사 (다양성 항목 비율 검증) - 구문규칙 검사 - 라벨링 데이터 구문 검사 - 검사결과 오류 목록 출력 	해당 사항 없음	<ul style="list-style-type: none"> - 파이썬 기반의 학습 모델을 이용한 벤치마크 데이터셋을 이용한 모델과의 정확도 대조 - 작업자가 Python 기반으로 개발하여 로그로 추출한 결과값을 벤치마크 결과값과 비교
사용환경	<ul style="list-style-type: none"> - OS: Linux - H/W: CPU 듀얼코어 1.6GHZ, MEM 16GB 이상 - S/W: 자체제작 	해당 사항 없음	<ul style="list-style-type: none"> - 알고리즘 : 도커 이미지 등의 실행 가능한 형태 - 학습 모델 측정 평가 : 원격 접속 환경 - 학습 모델 대조 : 테스트 데이터셋

- 업로드한 라벨링 데이터에 대한 다양성, 구문 정확성 검사 결과 항목별 조회 및 집계 데이터 다운로드

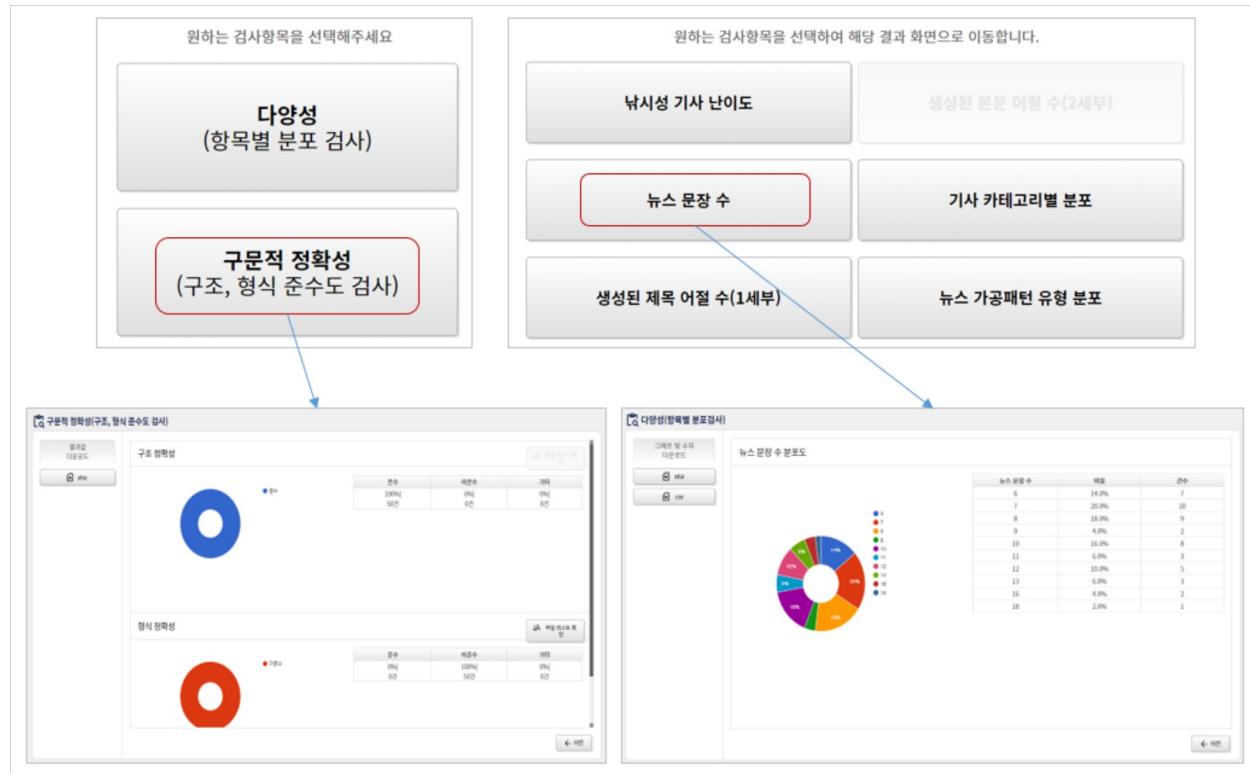
- AI 학습모델별 유효성 결과 데이터 조회/로깅

● 품질도구를 이용한 통계적 다양성 및 구문 정확성 검사

- 검사도구 파일 업로드

파일명	파일형
SO_M07_682613_L.json	
SO_M07_682614_L.json	
SO_M07_682615_L.json	
SO_M07_682616_L.json	
SO_M07_682617_L.json	
SO_M07_682618_L.json	
SO_M07_682619_L.json	
SO_M07_682620_L.json	
SO_M07_682621_L.json	
SO_M07_682622_L.json	
SO_M07_682623_L.json	
SO_M07_682624_L.json	
SO_M07_682625_L.json	
SO_M07_682626_L.json	
SO_M07_682627_L.json	
SO_M07_682628_L.json	
SO_M07_682629_L.json	
SO_M07_682630_L.json	
SO_M07_682631_L.json	

- 다양성, 구문적 정확성 결과 조회/엑셀 다운로드



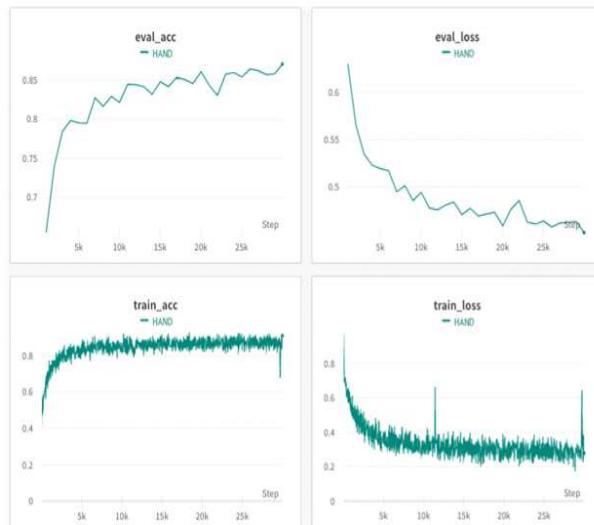
● 학습모델 유효성 검사 : 낚시성 기사 분류 모델(Accuracy 60%이상) / 주제 분리 템지 모델(Accuracy/F1-Score 55% 이상)

- 학습 데이터: 80% ($291,466 / 364,333$)
- 검증 데이터: 10% ($36,434 / 364,333$)
- 평가 데이터: 10% ($36,433 / 364,333$)
- 학습 데이터: 80% ($295,275 / 369,094$)
- 검증 데이터: 10% ($36,910 / 369,094$)
- 평가 데이터: 10% ($36,909 / 369,094$)

모델	학습데이터		검증데이터		평가데이터	
	Accuracy(%)	Accuracy(%)	Accuracy(%)	F1-Score	Accuracy	F1-Score
HAND	88.8%	87.1%	87.0%			
BERT	83.9%	90.8%	83.8%	90.8%	83.9%	90.8%

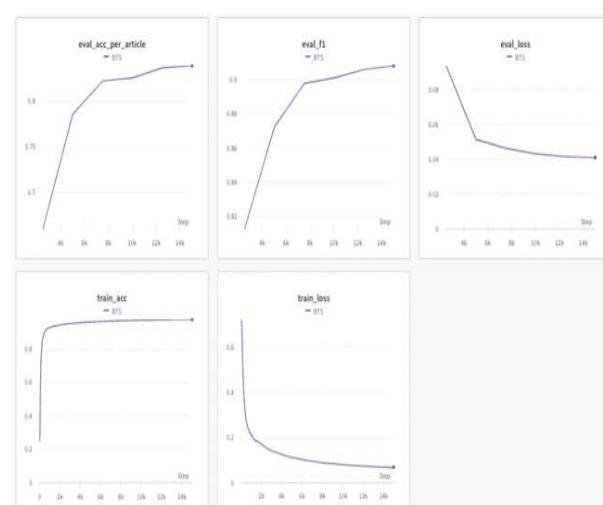
○ 학습 Log

- Steps 수: 30,000
- Batch Size: 256



○ 학습 Log

- Steps 수: 15,000
- Batch Size: 8



7. 학습 모델

7.1 학습 모델 후보

● 세부 구분별 학습모델 후보

데이터 명	2-025-146. 낚시성 기사 탐지 데이터			
학습 모델 후보	알고리즘	성능지표	선정 여부	선정 사유
낚시성 기사 분류	MuSeM(Mutual Semantic Matching)	'NELA17' 데이터셋에 대해 Macro F1-Score 0.719점 기록	△	3순위
	FNDNet(Fake News Detection Network)	Kaggle Fake News Dataset에 대해 F1-Score 0.98점 기록	△	2순위
	HAND(Hierarchical Attention Network for Fake News Detection)	Kaggle Fake News Dataset에 대해 Accuracy 95.92% 기록 → Accuracy 60% 이상 ※ 한글기반, 과제 난이도 고려 조정	○	1순위
주제분리 탐지	BERT(Bidirectional Encoder Representation from Transformers)	GROVER Dataset에 대해 Accuracy 96.6% 기록 → Accuracy 55%/F1-Score 55% 이상 ※ 한글기반, 과제 난이도 고려 조정	○	1순위
	BTS(BERT for Topic Segmentation)	Naver 퀄럼 데이터에 대해 F1-Score 0.87 기록 Topic Segmentation Score 0.7 이상	△	2순위

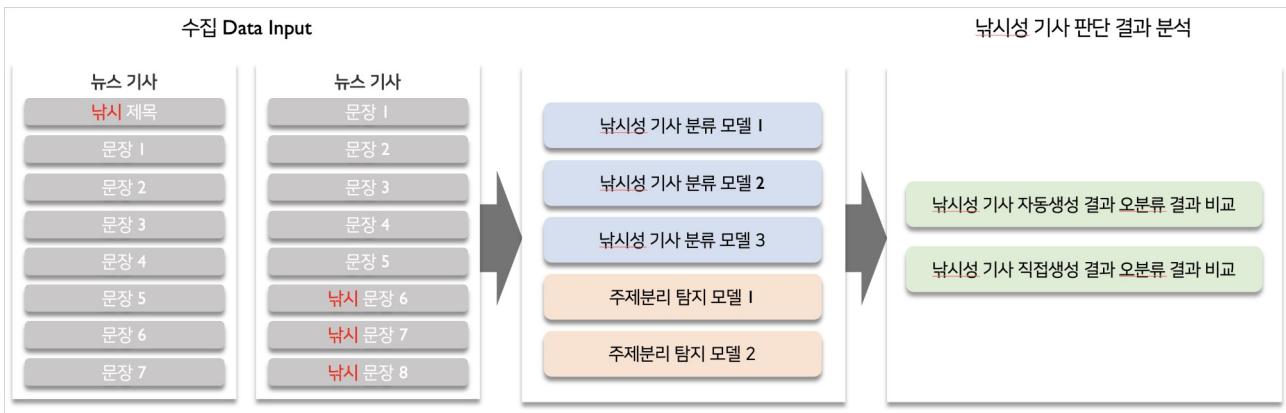
● 학습모델 선정 시 고려사항

구분	고려 사항	설명
1	적합성	낚시성 뉴스 기사 데이터 데이터셋 구축 목적에 적합한 학습 모델인가 (낚시성 기사 판별 학습을 위해 의도적인 낚시 제목 및 본문 제공)
2	활용성	올바른 뉴스 기사 판단을 위해 활용성이 높은 학습 모델인가
3	실현가능성	구축된 학습데이터셋을 활용하여 실제 뉴스 기사 판단을 위해 적용하고 실현가능성이 높은 모델인가
4	선정 절차	1) 선정기준에 적합한 후보 리스트업 2) 내부 연구원 대상 설문조사를 통해 우선순위 도출 3) 1-Cycle 학습모델 개발 4) 성능평가 5) 최종 학습 모델 선정

● 학습모델 선정

후보 학습모델	적합성	활용성	실현가능성	선정 여부
인공지능 기반 낚시성 뉴스 기사 분류 모델 개발	상	상	상	○
인공지능 기반 낚시성 뉴스 기사 주제분리 탐지 모델 개발	상	상	상	○

● 학습모델 개발 절차 및 유효성 판별

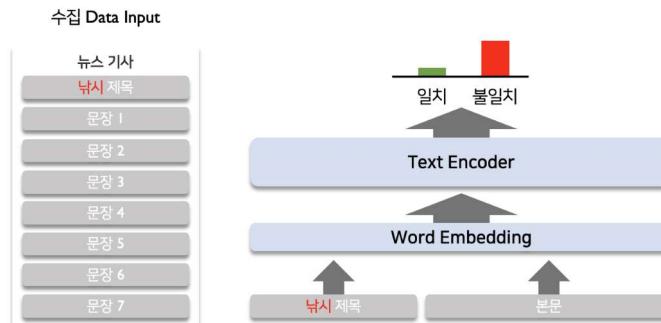


데이터 수집 후 유효성 검사가 끝난 학습모델을 활용하여 낚시성 기사 가공작업 진행 후 결과에 따라 데이터 유효성 판별	
낚시성 뉴스 기사 데이터 자동생성 오분류 개수에 따라 나이도 판별	오분류 개수 ↑ 자동생성 나이도 ↑ 오분류 개수 ↓ 자동생성 나이도 ↓
낚시성 뉴스 기사 데이터 직접생성 오분류 개수에 따라 나이도 판별	오분류 개수 ↑ 직접생성 나이도 ↑ 오분류 개수 ↓ 직접생성 나이도 ↓

7.2 학습 모델 개발

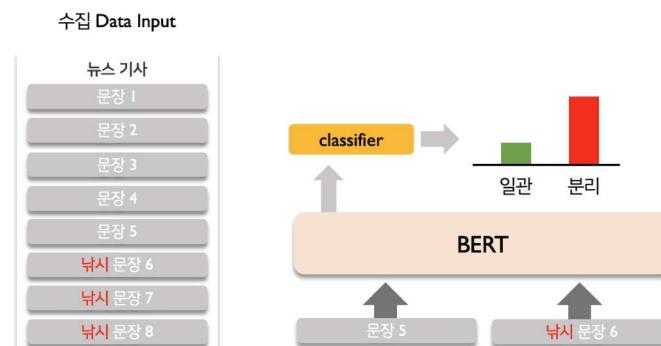
① 인공지능 기반 낚시성 뉴스 기사 분류 모델

- (개발 목표) 본문과 다른 내용의 뉴스 기사 제목 데이터를 활용하여 낚시성 뉴스 기사 여부 판단
- (개발 내용) 구축되는 학습데이터를 활용하여 제목과 본문 간 맥락 파악이 가능한 모델을 개발



② 인공지능 기반 낚시성 기사 주제분리 템지 모델

- (개발 목표) 본문 내 일관성 없는 주제로 작성된 뉴스 기사 데이터를 활용하여 낚시성 뉴스 기사 여부 판단
- (개발 내용) 일관성 여부에 따라 두 문장 간 맥락을 학습하여 본문 내 주제 분리 지점을 판단할 수 있는 모델을 개발



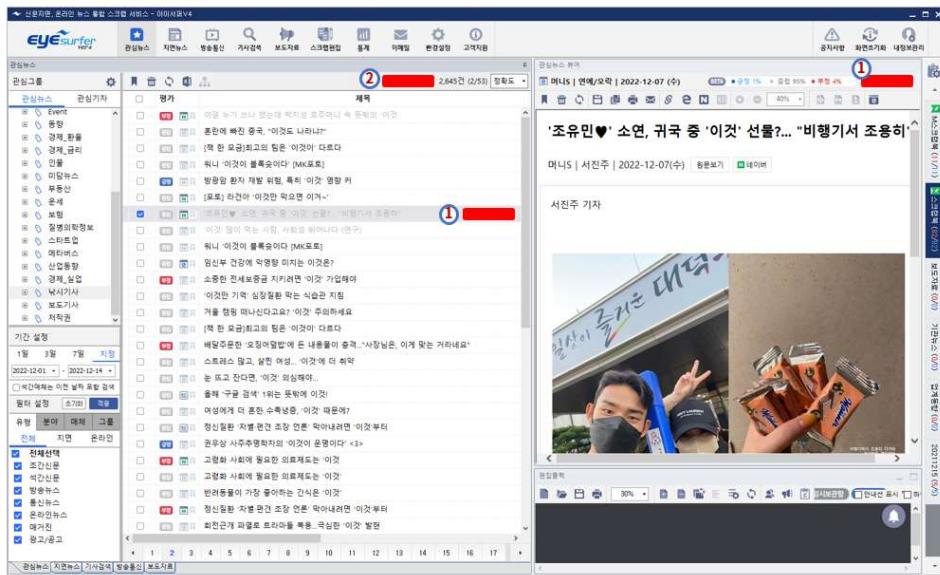
제3장 데이터 활용

1. 데이터 활용

데이터 명	2-025-146. 낚시성 기사 탐지 데이터
학습 모델	1. 낚시성 기사 분류 모델 2. 뉴스 기사 주제분리 탐지 모델
모델	1. 낚시성 기사 분류 모델 : HAND(Hierarchical Attention Network for Fake News Detection) 2. 뉴스 기사 주제분리 탐지 모델 : BERT(Bidirectional Encoder Representation from Transformers) 주제 분리 탐지
성능 지표	1. 낚시성 기사 분류 모델 : Accuracy(정확도) - 목표 60% vs. 결과 87% 2. 뉴스 기사 주제분리 탐지 모델 : 정확도(Accuracy) 목표 55% vs. 결과 83.9% / F1-Score 목표 55% vs. 결과 90.8%
개발 내용	1. 제목과 본문의 불일치 기사에 해당하는 1세부용 구축 학습데이터를 활용하여 새롭게 유입되는 뉴스기사에 대해 낚시성 기사 가능 점수 산출, 점수를 기반으로 분류 및 필터로 활용가능한 낚시성 기사 분류 모델(HAND) 개발 2. 본문의 도메인 일관성 부족 기사에 해당하는 2세부용 구축 학습데이터를 활용하여 새롭게 유입되는 뉴스기사에 대해 본문 내에서 화제변환 여부, 회제 수 등의 정보를 산출하여 활용가능한 주제 분리 탐지 모델(BERT 주제 분리) 개발
응용서비스 (예시 및 유의사항)	<ul style="list-style-type: none"> ● 낚시성 기사 점수 산출기 or 산출 계산기/분류기 <ul style="list-style-type: none"> - 주관사 뉴스 통합 스크랩 서비스 '아이서파'에 모델 적용 → 낚시성 기사 점수/분류값/주제수 정보 제공, 기사 검색/필터 → 고객사 스크랩 작업 효율성 향상 및 만족도 향상에 기여 - 사용자가 읽고 있는 뉴스 기사가 조회수를 통한 이윤 취득 또는 사상 전파를 위해 문서의 내용과 고리가 큰 자극적인 제목이나 본문과 제목이 일치하지 않는 '낚시성 기사'인지를 판단할 수 있는 계산기 서비스 제작 - 모바일의 앱이나, 웹페이지의 Add-on을 통해 현재 접속해있는 기사가 가짜 뉴스일 가능성은 수치로 나타낸 '낚시성 점수'를 계산하여 신뢰성 있는 기사만을 이용할 수 있음 - 낚시성 점수 산출 계산기를 통해 낚시성 기사의 조회수를 의도치 않게 높이는 것을 방지하여 낚시성 기사의 재생산 및 배포를 줄여 전체적으로 건전한 기사 양산 문화에 이바지할 수 있음 - 뉴스 모니터링 및 심의와 관련된 학술단체, 유관 언론기관 등에 낚시성 기사 탐지 모델을 이용한 서비스 혹은 도구를 개발하여 뉴스 기사에 대한 테스트, 심의를 위한 사전정보를 제공하여 보다 많은 뉴스에 대한 모니터링을 통해 저널리즘 품질 향상에 기여 ● 유의사항 <ul style="list-style-type: none"> - 구축데이터의 특성에 기반하여 학습된 AI 모델의 결과값으로 정보 필터링을 위한 부가정보로 활용 필요 - 1세부/2세부 가공패턴유형에 한정한 가공데이터를 통해 학습된 AI 모델을 특성을 고려한 적용/활용 필요 - 제목/본문 길이가 너무 짧거나, 본문의 길이가 너무 긴 뉴스기사에 대한 적용은 제외 고려 ● 전제조건 <ul style="list-style-type: none"> - 뉴스 기사를 기반한 AI 모델을 활용함으로 제목/본문으로 구성된 뉴스 기사로 제한하여 적용 필요

2. 응용 서비스

● 주관사 뉴스 통합 스크랩 서비스 '아이서퍼'에 학습모델 결과 적용하여 기사별 부가정보 제공 및 필터로 활용



- 기사에 대한 AI 모델 적용(낚시성 기사 분류 모델, 주제분리 탐지 모델)을 통해 산출되는 낚시성 기사 확률 정보 및 본문에서 감지한 주제분리 내용을 토대로 다루고 있는 주제수 정보를 산출하여 기사별로 정보 추가 제공
- 산출된 정보를 검색조건으로 설정하여 해당 기사를 필터링하는 수단으로 활용
- 낚시성 기사 확률 정보를 활용하여 낚시성 기사 여부를 파생하고, 해당 정보를 추가정보나 필터링/검색 조건으로 활용
- 결과 목록에서 추가적인 필터링 조건으로 활용하여, 양질의 스크랩 대상을 신속하게 조회 후 스크랩 진행하여 업무 효율 향상

3. 응용서비스 개발

● 뉴스 기사의 낚시성 기사 점수 산출 계산기



① Play Store 또는 App Store 접속
② Catch Fake News(가제) 검색 및 설치
③ 확인하고자 하는 "낚시성 점수" 확인
(* 낚시성 점수 : 가짜 뉴스일 확률)

① 웹페이지(Ex. 크롬) 상의 Store(add-on) 접속
② Catch Fake News(가제) 검색 및 설치
③ 확인하고자 하는 "낚시성 점수" 확인
(* 낚시성 점수 : 가짜 뉴스일 확률)

낚시성 점수 산출 계산기 작동 절차

4. 기술 지원

4.1 낚시성 기사 탐지 모델 전체 공개

- AI Hub에 소스 코드 공개(<https://aihub.or.kr>)

- AI Hub는 공공기관이 생성 또는 취득하여 관리하고 있는 공공데이터를 한 곳에서 제공하는 통합 창구
- AI Hub 활용사례에 구축된 가공데이터와 인공지능 모델을 공개하여, 사용자가 손쉽게 활용할 수 있도록 지원함

- 오픈 소스 코드 저장소에 소스 코드 공개(<https://github.com>)

- 인공지능 모델의 소스 코드와 예제 소스 코드를 사용자들이 자유롭게 접근할 수 있도록 함
- 활용방법 및 매뉴얼을 소스 코드와 함께 공개
- 사용자들이 오픈 소스 코드에 대한 버그와 수정 의견을 자유롭게 제안할 수 있는 이슈 게시판을 활용하여 소스 코드에 대한 의견을 수렴하고 고도화에 반영

- 본 과제 산출물을 오픈소스로 공개할 때 라이선스 명시하여 소스 코드 활용에 문제가 없도록 조치

- 본 과제에 사용된 제3자 제공 딥러닝 프레임워크 및 낚시성 기사 탐지 오픈 소스 모델에 적용되는 라이선스는 아래와 같음

오픈소스명	사용 목적	라이선스명
Pytorch (https://github.com/pytorch/pytorch)	딥러닝 프레임워크	New and Simplified BSD License (BSD 3-Clause)
BERT (https://github.com/huggingface/transformers)	낚시성 기사 탐지에 사용되는 언어모델 (ENG)	Apache License 2.0
KoBERT(https://github.com/SKTBrain/KoBERT)	낚시성 기사 탐지에 사용되는 언어모델 (SKT에서 제공하는 한국어 버전)	Apache License 2.0
Mecab-ko (https://bitbucket.org/eunjeon/mecab-ko-dic/src/master/)	한국어 형태소 분석기	Apache License 2.0
KoNLPy(https://github.com/konlpy/konlpy)	한국어 형태소 분석기 및 품사 태깅	GNU General Public License 3.0

4.2 소스 코드 사용 매뉴얼 제공

- 매뉴얼은 최근 오픈 소스에서 많이 사용되고 있는 매체 활용

- 매뉴얼 프레임워크를 이용하여 파일 다운로드나 별도의 프로그램 설치 없이 고품질의 문서를 작성 가능

- 본 매뉴얼은 프로그래밍에 익숙하지 않은 사용자도 손쉽게 낚시성 기사 탐지를 수행할 있도록 낚시성 기사 탐지 모델을 제공하는 본 패키지의 설치부터 데이터셋 로드, 알고리즘의 구체적인 사용 예시 등을 제공함

- 매뉴얼의 구성은 패키지 설치, 공개된 낚시성 기사 탐지 데이터 불러오기, 낚시성 기사 탐지 알고리즘 기능 설명, 사용 예시로 이루어짐

1. 패키지 설치	2. 낚시성 기사 탐지 데이터 불러오기	3. 낚시성 기사 탐지 알고리즘 기능 설명	4. 사용 예시
- 사용자의 실행 환경 별 소스 코드 설치 방법 기술	- 소스코드를 이용해 본 과제를 통해 구축한 한국어 낚시성 기사 탐지 데이터셋을 불러오는 방법 기술	- 소스코드를 이용해 낚시성 기사 탐지를 수행하는 기능 설명 - 각 모델 별 구체적인 사용 방법 기술	- 매뉴얼의 기능들을 이용해 실제 낚시성 기사 탐지를 진행하는 사용 예시 제공

- 패키지 설치 항목에서는 사용자의 실행 환경을 고려해 윈도우, 맥 OS, 리눅스 등 다양한 환경에서 소스코드를 설치하는 방법을 제공

- 낚시성 기사 탐지 데이터 불러오기 항목에서는 본 과제를 통해 구축한 한국어 낚시성 기사 탐지 데이터셋을 불러오고 간단한 전처리를 수행함으로써 데이터셋을 낚시성 기사 탐지 모델에 입력할 수 있는 형태로 변환하는 예시를 제공

- 낚시성 기사 탐지 알고리즘 기능 설명 항목에서는 패키지를 통해 제공되는 모든 낚시성 기사 탐지 알고리즘의 구체적인 사용 방법 제공

- 사용 예시 항목에서는 데이터셋과 학습된 모델을 불러와서 낚시성 기사 탐지를 수행하거나 사용자가 직접 구축한 데이터를 이용해 탐지 모델을 학습하는 등 본 패키지로 할 수 있는 사용 예시들을 제공

<부록> 인공지능 학습용 데이터 도메인 용어 정의

용어명(한글)	용어명(영문)	용어 정의
낚시(성)기사	Clickbait	(옥스퍼드) 특정 웹페이지로 연결도니 링크를 클릭하도록 유도하는, 질이 낮거나 가치가 적은 콘텐츠 (위키백과) 사용자의 관심을 끌고 해당 링크를 따라 링크된 온라인 콘텐츠를 읽거나 보거나 듣도록 유도하기 위해 고안된 텍스트 또는 섬네일 링크
가짜뉴스	Fake News	사람들의 흥미와 본능을 자극하여 시선을 끄는 황색언론(옐로 저널리즘)의 일종으로 사실이 아닌 것을 사실인 것처럼 꾸민 뉴스, 좁은 의미에서의 가짜 뉴스는 정치적인 목적으로 사실이 아닌 내용을 퍼뜨리기 위해 뉴스가 아닌데도 뉴스의 형식을 하여 퍼뜨리는 정보 또는 그 매팩체 등을 의미하나, 넓은 의미에서는 오보나 날조, 거짓 정보, 루마유언비어, 패러디·풍자 등을 포함하는 포괄적 용어로, 사실이 아닌 것을 사실이라고 주장하는 뉴스 전부를 의미하기도 함
인터넷신문	Internet Newspaper	인터넷을 기반으로 하는 신문 매체, 인터넷 웹사이트 기반에 대체로 종이 신문은 발행하지 않거나 발행해도 소수만 발행하고 무기지로 발행하는 경우가 대부분 (한국법) 컴퓨터 등 정보처리능력을 가진 장치와 통신망을 이용하여 정치·경제·사회·문화 등에 관한 보도·논평 및 여론·정보 등을 전파하기 위하여 간행하는 전자간행물로서 독자적 기사 생산과 지속적인 발행 등 대통령령으로 정하는 기준을 충족하는 것
오보	Misinformation	어떠한 사건이나 소식을 그릇되게 전하여 알려 줌. 또는 그 사건이나 소식
허위정보/허위기만정보	Disinformation	경제적 이익을 얻기 위해
저널리즘	Journalism	뉴스를 취재하여 대중에게 보도하는 행위
뉴스 엠바고	News Embargo	취재한 사안을 보도하는 것을 일정 기간 미루기로 약속하는 것
뉴스 생성 언어	News Markup Language	뉴스를 정의하기 위한 확장성 생성 언어(XML) 기반의 표준 형식. NewsML은 유럽의 국제 출판 통신 협의회(IPTC: International Press and Telecommunications Council)가 NITF(News Industry Text Format)와 함께 개발했으며, 2000년 10월 네덜란드 암스테르담에서 열린 IPTC 총회에서 첫 버전이 채택
취재	cover	뉴스를 생산하기 위하여 그와 관련된 정보를 수집하는 행위
헤드라인	Headline	(브리태니카 사전) 신문이나 신문 기사의 앞에 내용을 가리키거나 요약하기 위해 굵은 활자로 붙인 낱말이나 낱말군
(헤드라인)모호한 표현	Ambiguous	호기심을 불러일으키기 위해 모호하고 혼란스럽도록 헤드라인 구성
(헤드라인)과장	Exaggeration	기사 본문 내용을 과장되게 표현
(헤드라인)선동적 표현	Inflammatory	부적절하거나 비속어를 사용
(헤드라인)미끼상술	Bait-and-Switch	헤드라인에서의 약속하거나 함축된 내용이 실제 본문에서는 없음
(헤드라인)장난스런표현	Teasing	긴장감을 조성하기 위해 헤드라인에서 세부 내용을 삭제
(헤드라인)포맷팅	Formatting	대문자, 마침표, 느낌표, 물음표 등의 과도한 사용
(헤드라인)틀린 표현	Wrong	사실과 다른 오류
(헤드라인)그래픽	Graphic	외설스럽고, 충격적이거나 믿을 수 없는 주제
(헤드라인)불완전한표현	Incomplete	충분한 정보를 제공하지 않은 불완전한 헤드라인
(헤드라인)헤드라인복제	Headline Cloning	기사 본문에서 단순히 헤드라인을 복제
(헤드라인)잘못된링크	URL Redirection	엉뚱한 페이지로 인도
낚시성기사 분류	Clickbait Classification	AI 모델(HAND)을 적용하여 기사가 낚시성 기사인지 여부를 구분하는 임무
주제분리 탐지	Topic Segmentation Detection	AI 모델(BERT)을 적용하여 기사 본문내에서 주제(화제)가 전환되는 위치를 탐지하는 임무