대한민국 대전환 한국판뉴딜

2020년 인공지능 학습용 데이터 교육영상 “문서요약 텍스트” 소개 및 활용

2021.03

비플라이소프트㈜
이현미

과학기술정보통신부 NIA 한국지능정보사회진흥원

2021.03

INDEX

01 문서요약 텍스트 데이터 소개

문서요약 텍스트 데이터의 구축 목적

다양한 주제의 한국어 원문으로부터 추출요약문과 생성요약문을
도출해낼 수 있도록 인공지능을 훈련하기 위한 데이터셋

활용 분야

뉴스기사 요약, 법률문서 요약, 사업보고서 요약 등
한국어 문서의 핵심내용을 신속하고 정확하게 파악할 수 있는 AI 요약기술 개발

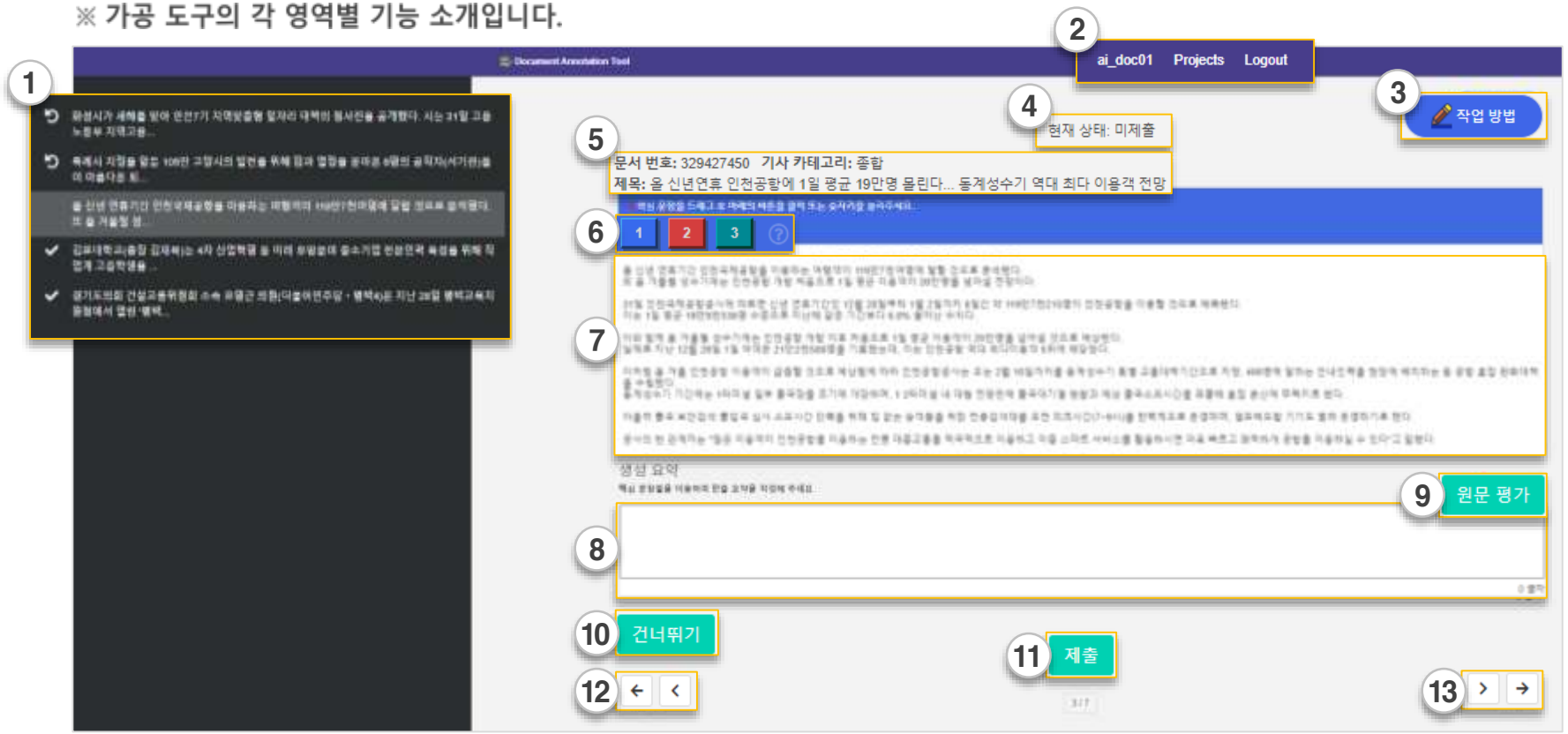
01 문서요약 텍스트 데이터 소개

데이터의 종류 및 규모

데이터의 종류	신문	기고문	잡지	법원 판결문
원문형태	PDF 원문 추출 뉴스 텍스트	PDF 원문 추출 칼럼 및 오피니언 텍스트	PDF 원문 추출 웹진 기사 텍스트	법원 판결문 텍스트
데이터셋 규모	원문 및 추출/생성 요약문 각 30만 건	원문 및 추출/생성 요약문 각 6만 건	원문 및 추출/생성 요약문 각 1만 건	원문 및 추출/생성 요약문 각 3만 건
중요성	기본요약 알고리즘 확보	사실관계가 아닌 개인의견 요약 형태	장문 요약 및 문법적 다양성 확보	핵심 사실관계 추출
원문확보 채널	개별 언론사 저작권 협의 완료			공개된 법원 판결문

02 가공 작업 환경

※ 가공 도구의 각 영역별 기능 소개입니다.



1. 사용자 작업 내역 목록
2. 사용자 메뉴(작업정보, 로그아웃 등)
3. 작업 가이드 라인
4. 문서 상태 : 승인 / 미제출 / 반려
5. 원문 상태 정보 표기
6. 주요 문장 선택 아이콘 : 단축키 제공(숫자키1~3)
7. 기사 본문 표시
8. 기사 원문 평가
9. 생성 요약 작성 영역
10. '건너뛰기 ' 버튼
11. '제출 ' 버튼
12. 본인에게 할당된 문서 좌측 이동 버튼
13. 본인에게 할당된 문서 우측 이동 버튼

02 가공 작업 환경

추출요약 생성

기사 원문을 전체적으로 읽습니다. 그리고, 기사 원문을 대표할 만한 문장을 3개 선정하여 라벨링 작업을 합니다.
주요 문장 내 텍스트를 드래그 후 표시된 아이콘을 클릭하거나 단축키(숫자 1~3)를 사용하여 표기합니다.

1. **주요 키워드를 더블 클릭합니다.**

2. **1번 아이콘을 클릭 또는 숫자 1번 키를 누르면, 해당 문장이 모두 블록으로 변경됨을 확인할 수 있습니다.**

주요 Q&A

1. **선택한 블록을 변경하려면?**

-> 블록처리된 문장 우측 끝에 ⓧ 아이콘이 보입니다. 해당 아이콘을 클릭하면 선택한 블록이 해제 됩니다.

2. **3개의 문장 순서를 변경하려면?**

-> ⓧ 아이콘을 클릭하여 선택했던 블록을 모두 해제 한 후, 처음부터 다시 라벨링합니다.

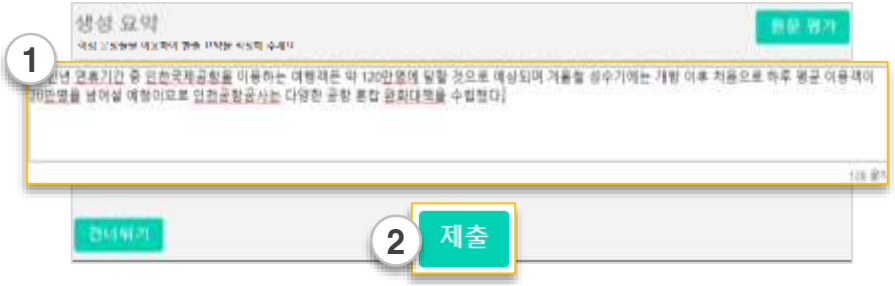
02 가공 작업 환경

원문 평가

추출요약 완료 후, 우측 하단에 보이는 '원문평가' 버튼을 클릭합니다.
기사에 대한 평가를 진행 후 '제출' 버튼을 클릭.

생성요약 입력 후 제출하기

추출요약 완료 후, 3개의 추출문장을 1개의 자연스러운 문장으로 재구성하여 '생성요약 ' 입력 영역에 작성합니다.



1. **생성요약 작정** ▶ 2. **'제출' 버튼 클릭**

03 검수 작업 환경

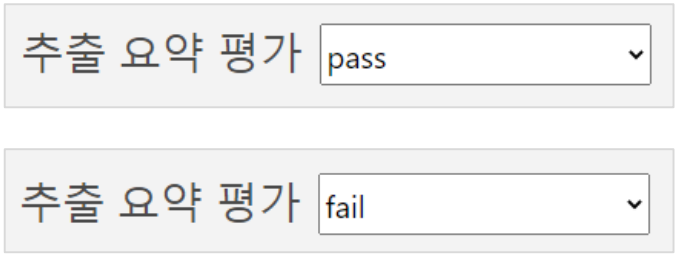
※ 검수 도구의 각 영역별 기능 소개입니다.

1. 사용자 작업 내역 목록
2. 사용자 메뉴(작업정보, 로그아웃 등)
3. 작업 가이드 라인
4. 문서 상태
5. 원문 상태 정보 표기
6. 기사 본문 및 추출요약
7. 생성요약
8. 추출 요약 평가
9. 생성 요약 평가
10. 요약 의견란
11. '건너뛰기 ' 버튼
12. '제출 ' 버튼
13. 본인에게 할당된 문서 좌측 이동 버튼
14. 본인에게 할당된 문서 우측 이동 버튼

03 검수 작업 환경

추출 요약 평가

추출 요약이 올바르면 pass, 올바르지 않으면 fail을 부여합니다.

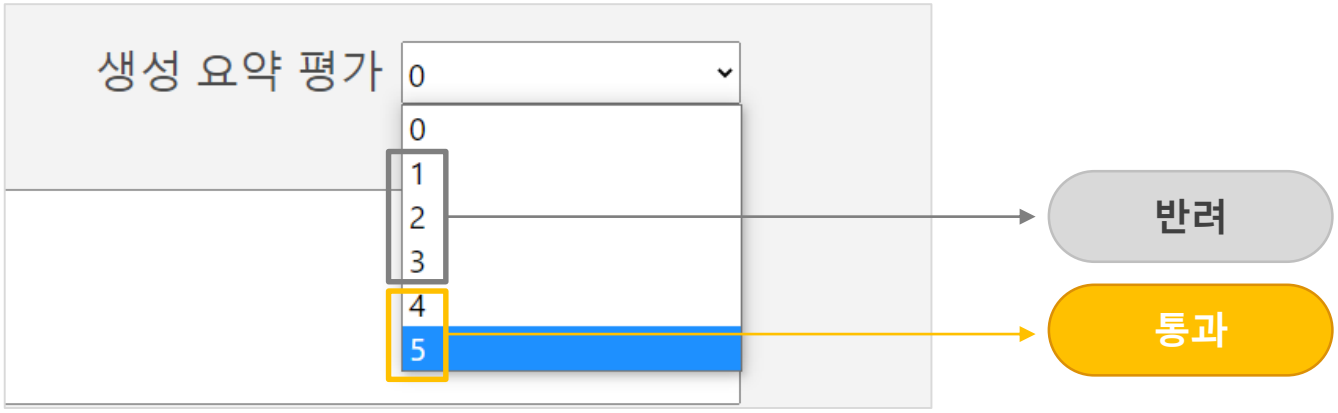


※ 추출요약 반려 사유 예시

- 완전한 문장을 추출하지 않은 경우 ('~다.'로 끝나지 않는 문장, 단어로 끝나는 구의 경우 추출요약 대상 제외)
- 문서의 키워드/핵심 문장을 포함하지 않은 경우
- 기사 원문과 상관없는 정보(신문사, 기자 이름, 사진 출처 등)가 추출 요약 문장에 붙어있는 경우 등

생성 요약 평가

생성 요약을 평가한 후 점수를 부여합니다. (1~3점 반려, 4~5점 통과)



04 데이터 학습 방법 및 시연

추출요약모델

뉴스 데이터 / BertSum 모델 기준으로 설명

명령어: sh scripts/train_bertsumext_news.sh

```
1  BERT_DATA_PATH='./datasets/kor_data/bertabs_data_news/news'
2  MODEL_PATH='./checkpoint/bertsum_original/ext/news/'
3  LOG_PATH='./logs/bertsum/ext/ext_news50k.log'
4  python src/train.py  \
5      -task ext \
6      -mode train \
7      -bert_data_path ${BERT_DATA_PATH} \
8      -ext_dropout 0.1 \
9      -model_path ${MODEL_PATH} \
10     -lr 2e-3 \
11     -visible_gpus 0,1 \
12     -report_every 50 \
13     -save_checkpoint_steps 1000 \
14     -batch_size 3000 \
15     -train_steps 50000 \
16     -accum_count 3 \
17     -use_interval true \
18     -warmup_steps 10000 \
19     -max_pos 512 \
20     -log_file ${LOG_PATH}
```

- 학습에 사용될 데이터 경로
- 학습된 모델이 저장될 경로
- 학습 과정 로그가 저장될 경로
- 변경 가능한 모델 학습 파라미터
- 사용 가능한 GPU

04 데이터 학습 방법 및 시연

생성요약모델

뉴스 데이터 / BertSum 모델 기준으로 설명

명령어: sh scripts/train_bertsumabs_news.sh

```
1  BERT_DATA_PATH='./datasets/kor_data/bertabs_data_news/news'
2  MODEL_PATH='./checkpoint/bertsum_original/abs/news/'
3  LOG_PATH='./logs/bertsum/abs/abs_news200k.log'
4  python src/train.py  \
5      -task abs \
6      -mode train \
7      -bert_data_path ${BERT_DATA_PATH} \
8      -dec_dropout 0.2  \
9      -model_path ${MODEL_PATH} \
10     -sep_optim true \
11     -lr_bert 0.002 \
12     -lr_dec 0.2 \
13     -save_checkpoint_steps 5000 \
14     -batch_size 140 \
15     -train_steps 200000 \
16     -report_every 50 \
17     -accum_count 10 \
18     -use_bert_emb true \
19     -use_interval true \
20     -warmup_steps_bert 20000 \
21     -warmup_steps_dec 10000 \
22     -max_pos 512 \
23     -visible_gpus 0,1 \
24     -log_file ${LOG_PATH}
```

학습에 사용될 데이터 경로

학습된 모델이 저장될 경로

학습 과정 로그가 저장될 경로

변경 가능한 모델 학습 파라미터

사용 가능한 GPU

05 학습된 모델 결과 확인 방법 및 시연

추출요약모델

뉴스 데이터 / BertSum 모델 기준으로 설명

명령어: sh scripts/test_bertsumext_news.sh

```
set -x

BERT_DATA_PATH='./datasets/kor_data/bertext_data_news/news'
OUTPUT_PATH='./outputs/bertsumext/news/news'
CHECKPOINT_PATH='./checkpoint/bertsum_original/ext/news/model_step_24000.pt'
LOGDIR='./logs/bertsum/ext/news.log'

python src/bertsum/train.py \
    -task ext \
    -mode test \
    -batch_size 1000 \
    -test_batch_size 1000 \
    -bert_data_path ${BERT_DATA_PATH} \
    -log_file ${LOGDIR} \
    -sep_optim true \
    -use_interval true \
    -visible_gpus 0 \
    -max_pos 512 \
    -max_length 200 \
    -alpha 0.95 \
    -min_length 50 \
    -test_from ${CHECKPOINT_PATH} \
    -result_path ${OUTPUT_PATH} \
    -report_rouge false \
    -block_trigram false

exec > >(ts "%m/%d/%Y %H:%M:%S"| tee -a ${LOGDIR}) 2>&1
python src/bertsum/cal_kor_rouge.py \
    --candidate_path './outputs/bertsumext/news/news_step24000.candidate' \
    --save_path './result/bertsumext/news/'
```

추론에 사용될 데이터 경로

요약 결과물이 저장될 경로

학습이 완료된 모델의 경로

학습 과정 로그가 저장될 경로

변경 가능한 모델 학습 파라미터

모델 성능(rouge score) 확인을 위해 필요한 경로들

05 학습된 모델 결과 확인 방법 및 시연

생성요약모델

뉴스 데이터 / BertSum 모델 기준으로 설명

명령어: sh scripts/test_bertsumabs_news.sh

```
set -x

BERT_DATA_PATH='./datasets/kor_data/bertabs_data_news/news'
OUTPUT_PATH='./outputs/bertsumabs/news/news'
CHECKPOINT_PATH='./checkpoint/bertsum_original/abs/news/model_step_130000.pt'
LOGDIR='./logs/bertsum/abs/news.log'

python src/bertsum/train.py \
    -task abs \
    -mode test \
    -batch_size 3000 \
    -test_batch_size 3000 \
    -bert_data_path ${BERT_DATA_PATH} \
    -log_file ${LOGDIR} \
    -sep_optim true \
    -use_interval true \
    -visible_gpus 0 \
    -max_pos 512 \
    -max_tgt_len 300 \
    -alpha 0.95 \
    -min_length 50 \
    -report_rouge false \
    -result_path ${OUTPUT_PATH} \
    -test_from ${CHECKPOINT_PATH}

exec > >(ts "%m/%d/%Y %H:%M:%S"| tee -a ${LOGDIR}) 2>&1
python src/bertsum/cal_kor_rouge.py \
    --candidate_path './outputs/bertsumabs/news/news.130000.candidate' \
    --save_path './result/bertsumabs/news/'
```

추론에 사용될 데이터 경로

요약 결과물이 저장될 경로

학습이 완료된 모델의 경로

학습 과정 로그가 저장될 경로

변경 가능한 모델 학습 파라미터

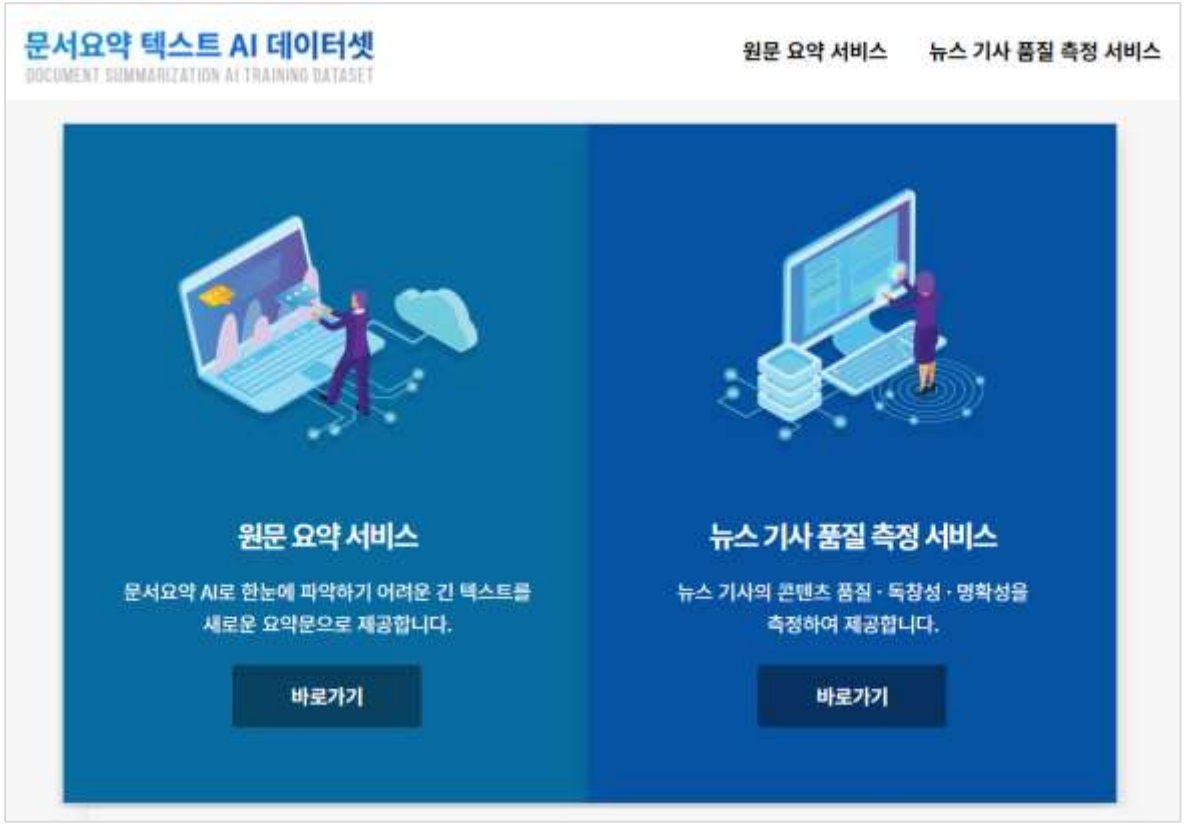
모델 성능(rouge score) 확인을 위해 필요한 경로들

06 서비스 소개 및 활용 방법

서비스 소개

저희 서비스는 본 과제에서 학습한 문서요약 모델을 웹페이지에서 누구나 직접 쉽게 사용할 수 있도록 개발되었습니다.
또한, 신문기사의 품질을 측정하는 도구를 추가로 제공합니다.

서비스 화면



06 서비스 소개 및 활용 방법

원문 요약 서비스

문서요약 AI로 긴 원문을 짧게 요약하는 서비스입니다.

원문 입력하기

1. 원문을 직접 입력하거나,
   원문 파일을 (.txt 또는 .docx) 업로드

2. 원문의 유형을
   신문, 기고, 잡지, 법률 중에 선택

3. 원문 요약 실행 버튼 클릭

06 서비스 소개 및 활용 방법

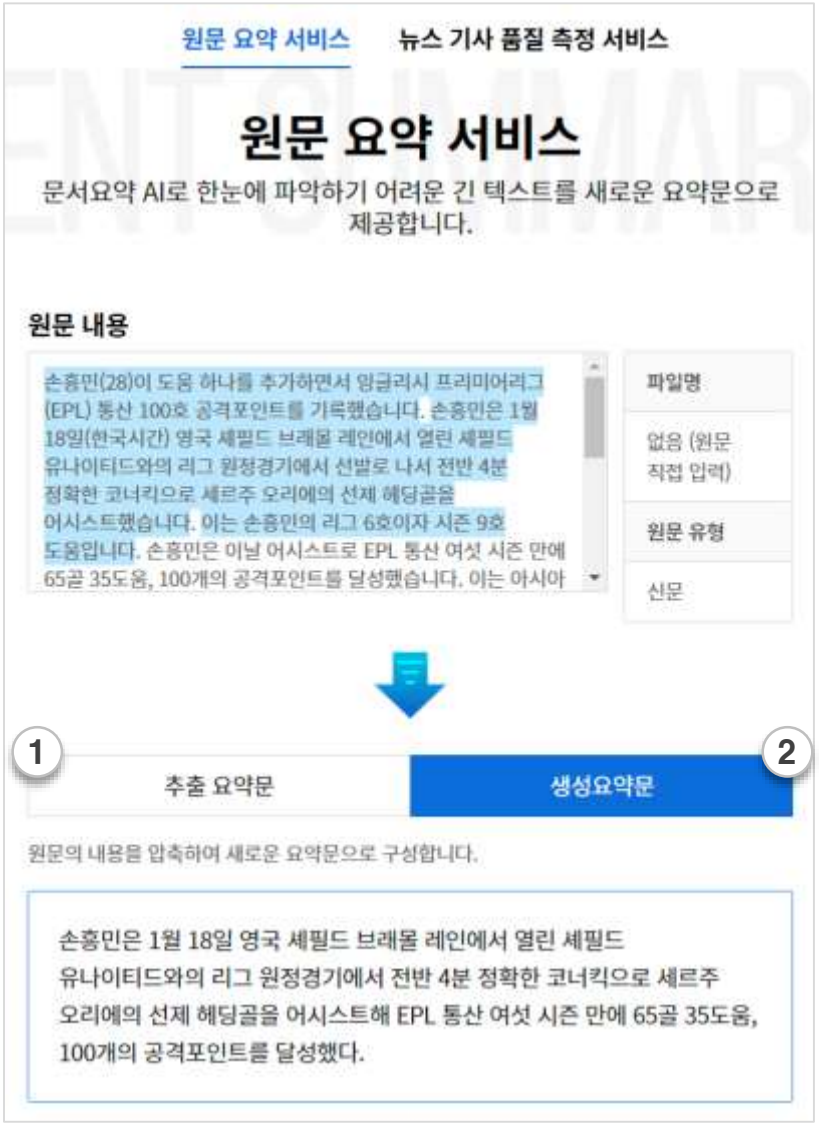
원문 요약 서비스

문서요약 AI로 긴 원문을 짧게 요약하는 서비스입니다.

결과 살펴보기

1. 추출요약문
   원문 중에서 가장 중요하다고 판단되는
   3개 문장을 추출.
   화면 중앙 원문 내용 창에는 해당 문장들을
   하늘색으로 표시함

1. 생성요약문
   원문의 내용을 압축하여
   아예 새로운 요약 문장을 생성해 제공함



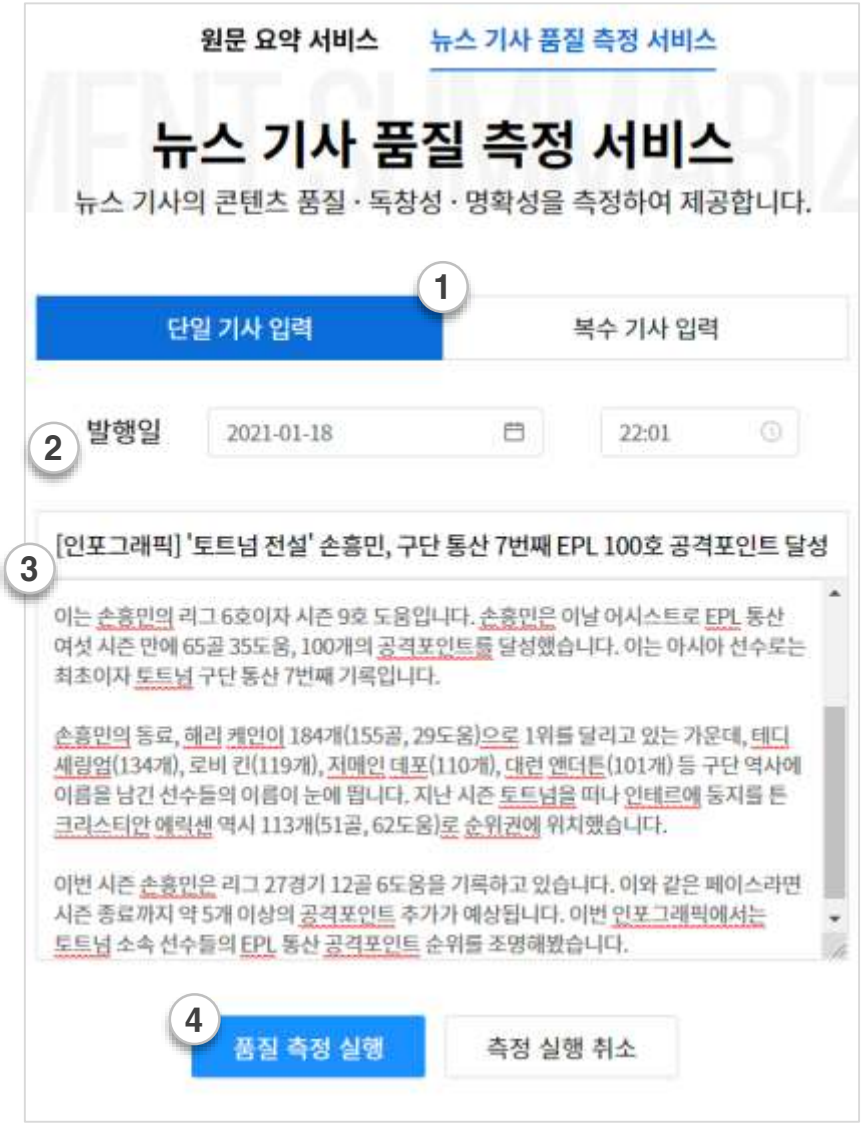
06 서비스 소개 및 활용 방법

뉴스 기사 품질 측정 서비스

입력하신 기사의 품질을 다양한 지표로
측정하여 제공하는 서비스입니다.

원문 입력하기

1. 하나의 원문을 입력하거나, 제공한 엑셀 양식 파일에 여러 원문의 내용을 입력
2. 뉴스 기사의 발행 날짜, 시간을 입력
3. 원문의 제목과 내용을 입력
4. 품질 측정 실행 버튼 클릭



06 서비스 소개 및 활용 방법

뉴스 기사 품질 측정 서비스

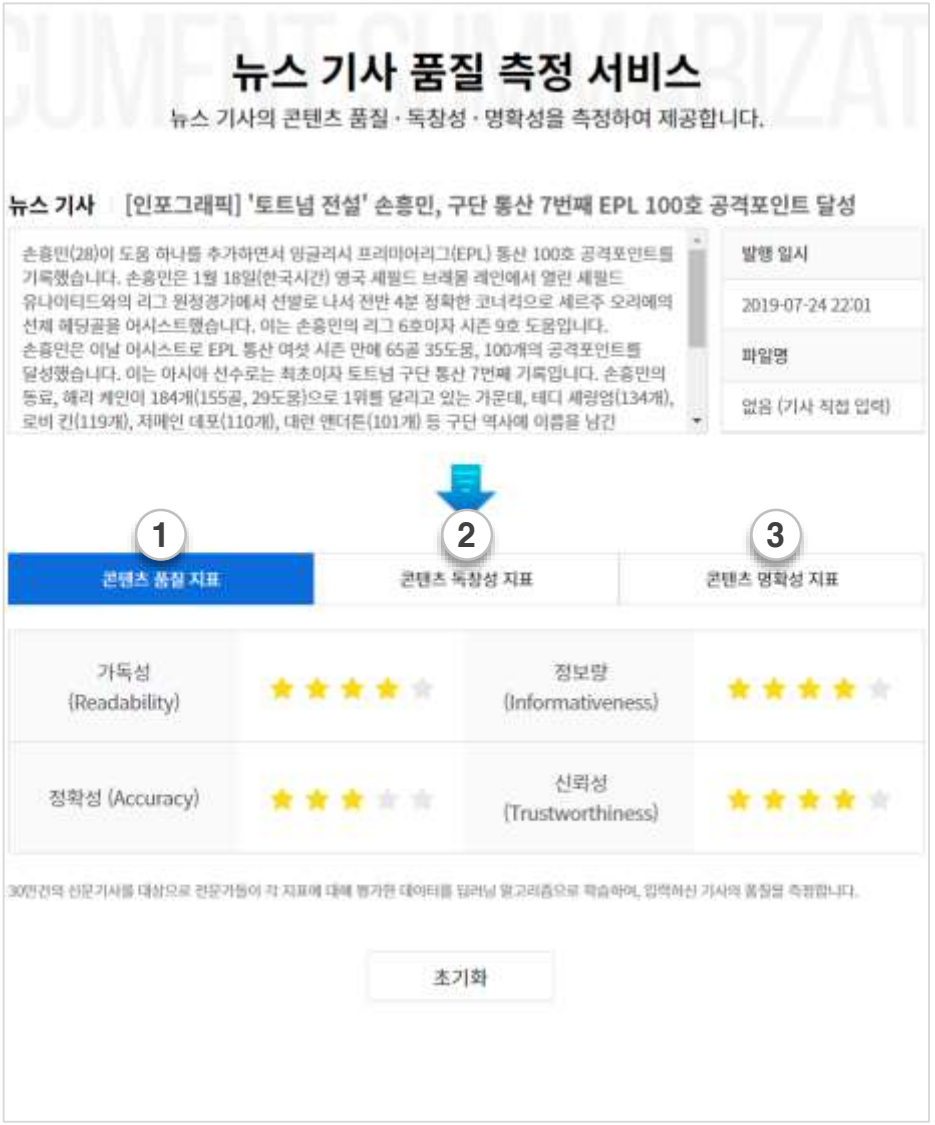
입력하신 기사의 품질을 다양한 지표로 측정하여 제공하는 서비스입니다.

결과 살펴보기

1. **콘텐츠품질지표**
   전문가들이 30만 건의 신문기사를 대상으로 각 지표에 대해 평가한 데이터를 딥러닝 알고리즘으로 학습하여, 해당 기사의 품질을 측정

2. **콘텐츠독창성지표**
   - 타 기사 대비 차별화 정도 : DB내의 기사들과 현재 기사와의 유사도를 기반으로 측정
   - 유사기사 중 발행시점 선행 정도 : 유사한 타 기사들 대비 얼마나 빨리 발행되었는지 수치화 (2020년 이전의 기사를 입력해야 의미있는 값이 도출됨)

3. **콘텐츠명확성지표**
   생성 요약문과, 요약의 신뢰도를 사용하여 기사의 명확성을 측정



대한민국 대전환 한국판뉴딜

감사합니다

비플라이소프트 ㈜
이현미

E-MAIL : rice127@bflysoft.com
PHONE : 070-7091-8562

과학기술정보통신부 NIA 한국지능정보사회진흥원

2021.03