제로베이스 24기

모여봐요 신문읽자

파이널 프로젝트 3조

고준환 김유현 김은비

목차

1. 프로젝트 기획 의도

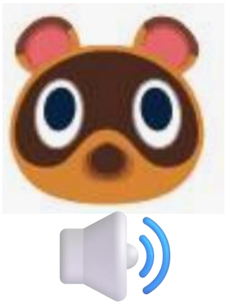
프로젝트 기획 의도

문제점 및 현황

1.인터넷이 발달한 세상에서 태어난 학생들 및 성인들의 **문해력이 낮다**
2.낮은 문해력은 학습과 업무에 좋지 않은 영향을 야기
3.한국 학생 정보 식별률 **25.6%** (회원국 평균 47.4%) **OECD(2023)**

해결 방안

1.유명 언론매체에서 문해력을 높이기 위해 **신문 읽기**를 강력 추천함
2.유아기때부터 꾸준한 신문 읽기 학습 시 문해력 증가 예상
3.신문 기사 요약 반복 학습을 통해 핵심 내용을 빠르게 파악 가능

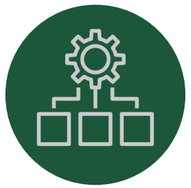


2. 데이터 수집 & EDA

데이터 수집

데이터 수집

AI허브 - 문서 요약 텍스트
json 형식의 파일



데이터 구성

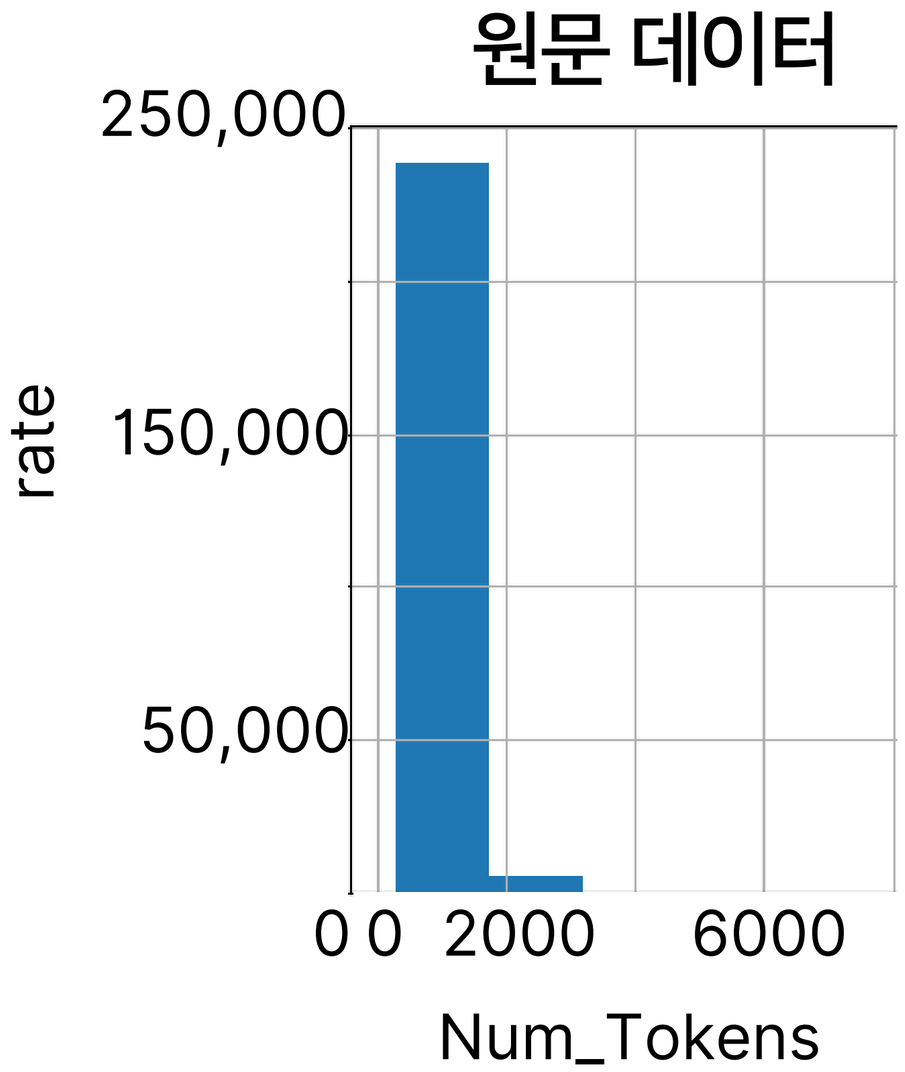
본문 / 요약문
(4개 문서 유형)

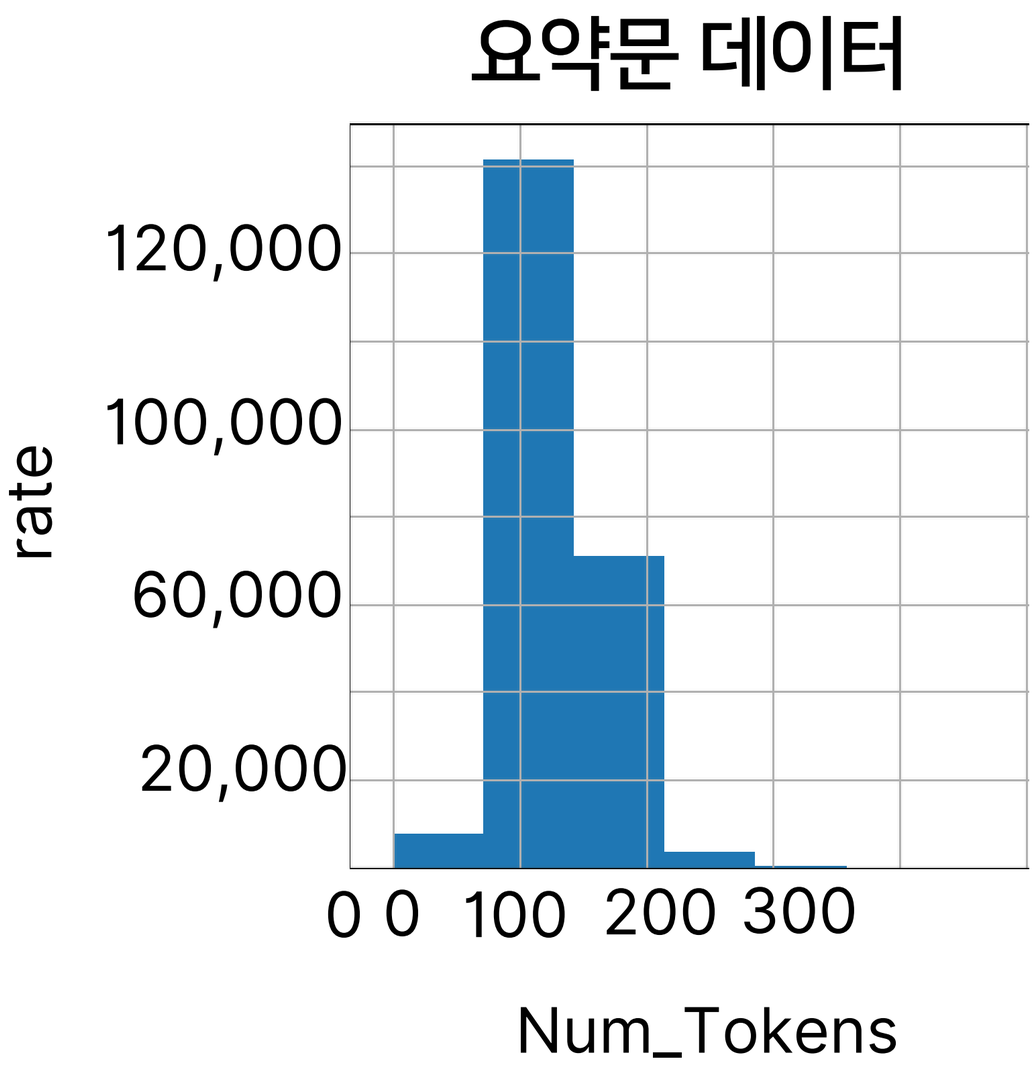
신문기사
기고문
잡지
법률

데이터 규모

json파일
800,000건(400MB)
-> 약 280,000건(266MB)

EDA


원문 데이터
250,000
150,000
50,000
rate
0 0
2000
6000
Num_Tokens


요약문 데이터
120,000
100,000
60,000
20,000
rate
0 0
100
200
300
Num_Tokens

3. 모델 학습 & 성능 평가

모델링 과정 요약

모델, 토크나이저 로드

gogamza/kobart-base-v2
등 다양한 모델 사용

텍스트 전처리

- 원문 / 요약문 토큰화
- 패딩으로 길이 조정
- 텐서 변환
- DataLoader 준비

모델 학습

- Batch size, learning rate,
- epoch, optimizer 등
- 하이퍼파라미터 튜닝

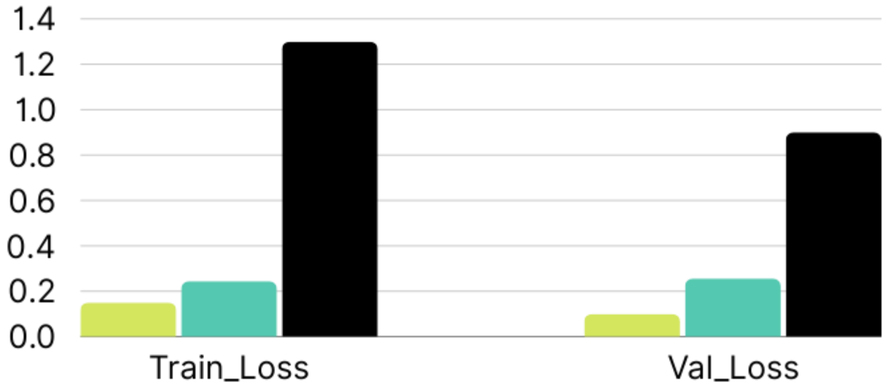
성능 평가

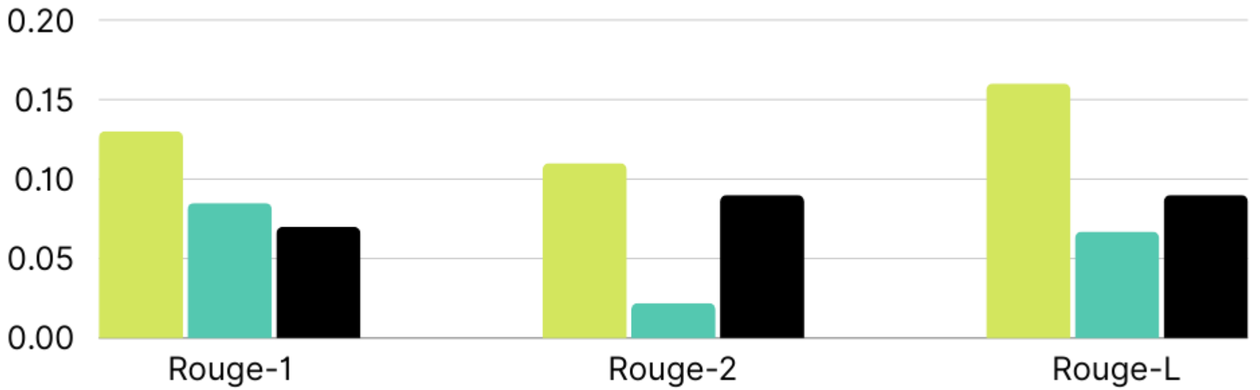
KoRouge 지표 활용
(한국어 요약 성능 평가)

모델 학습 결과

(10,000건씩 테스트)

Model	Train_Loss	Val_Loss	Rouge-1	Rouge-2	Rouge-L
Kobart	**0.15**	**0.1**	**0.13**	**0.11**	**0.16**
t5 -small	0.244	0.256	0.085	0.022	0.067
MiniLM	1.3	0.9	0.07	0.09	0.09





Kobart 학습 & 성능 평가

	batch	Epoch	learning_rate	Train_Loss	Val_loss	Rouge-1	Rouge-2	Rouge-L
20000개	8	10	e^-5	0.167	0.162	0.08	0.0002	0.08
30000개	**8**	**12**	**5e^-5**	**0.07**	**0.06**	**0.294**	**0.09**	**0.279**
40000개	16	3	3e^-5	0.27	0.27	0.2	0.112	0.183
80000개	16	5	3e^-5	0.18	0.2	0.221	0.12	0.198
80000개	16	10	3e^-5	0.14	0.178	0.226	0.122	0.2

optimizer : adam / loss : cross entropy / model_checkpoint : val_loss

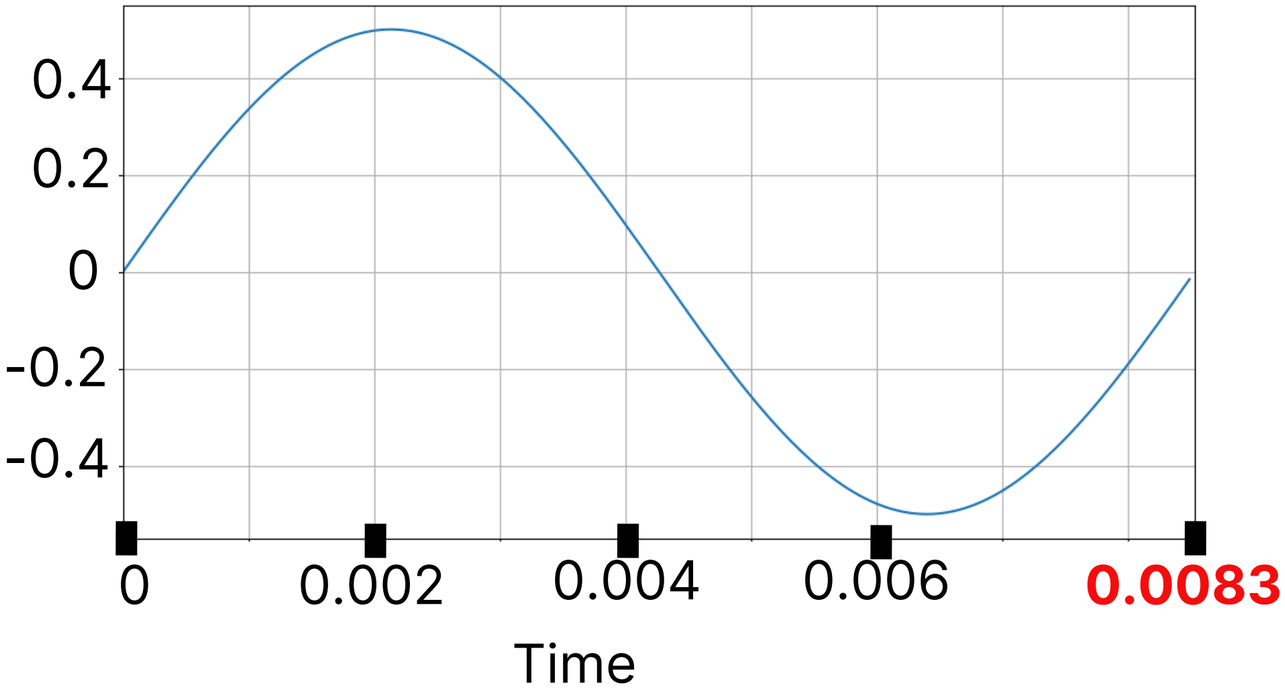
4. 음성처리 및 웹 업로드

음성처리

초성, 중성, 종성 (71개사용)

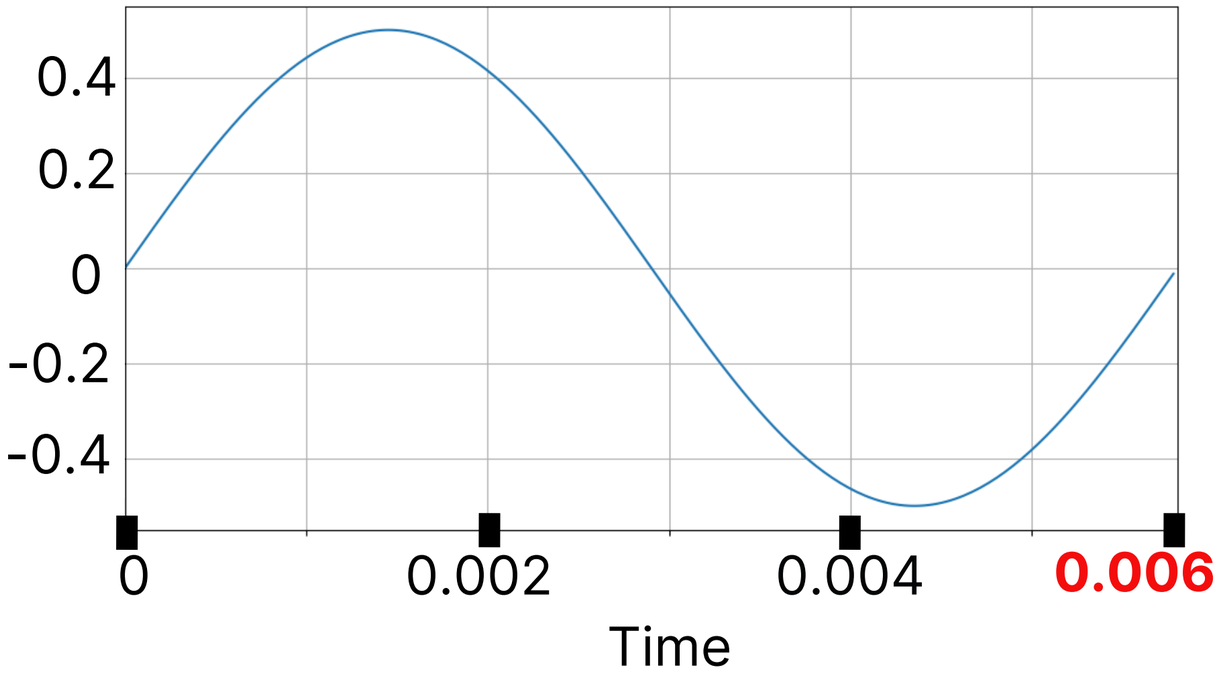
10만벨 남았어 구리

(보통 속도)

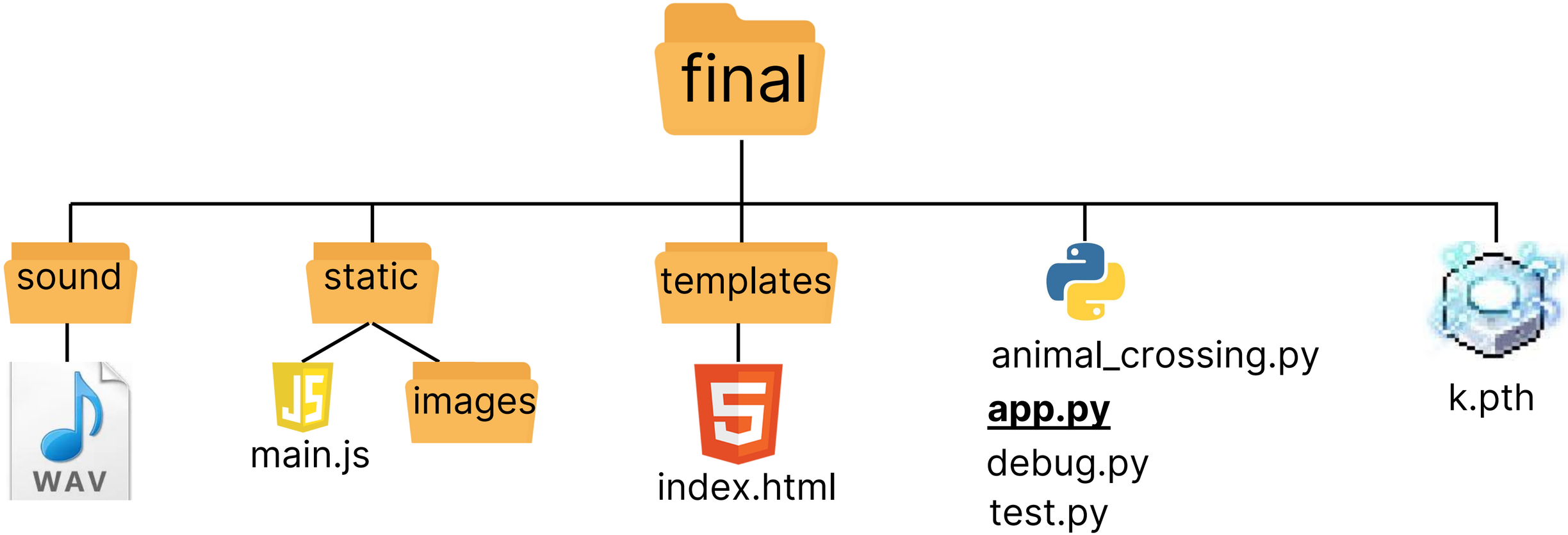


10만벨 남았어 구리

(동물언어화)



웹 업로드


final
sound
WAV
static
main.js
images
templates
index.html
animal_crossing.py
app.py
debug.py
test.py
k.pth

5. 모델 시연

모델 시연

1

어서와 구리 ʕ·ᴥ·ʔ

글을 입력해주세요

클릭 시 요약본이 출력 및 재생됩니다

2

어서와 구리 ʕ·ᴥ·ʔ

지구 반대편에는 불우한 환경에서 지내는 유아 및 청소년들이 많은데, 이들을 돕고자 다양한 범세계적 구호활동 단체들이 전국각지를 순회하며 이들을 위한 구호활동에 나서고 있다.
유니세프 제공 30일 세계청소년권익위원회가 지정한 '국제 우정

클릭 시 요약본이 출력 및 재생됩니다

조금만 기다리숑 약 40 초?

3

콩돌, 밤돌

유니세프 제공 30일 세계청소년권익위원회가 지정한 '국제 우정의 날(International Friendship Day)'을 맞아 적십자는 특별한 영상을 전 세계에 전격 공개했다. 이번 영상은 가정과 학교, 사회에서 다양한 폭력에 노출돼 있는 세계 청소년과 어린이들에 대한

0:13 / 0:31

6. 프로젝트 회고

프로젝트 회고

요약의 한계

- 신문기사를 학습시킨 모델의 특성상, 너무 긴 글이나 기사의 내용을 벗어난 글은 요약이 잘 수행되지 않음
- 다양한 카테고리의 글을 학습시키는 것이 추후 요약의 퀄리티를 높일 것으로 기대됨

시간 · 자원적 한계

- 학습 시간이 오래 걸리는 모델링 특성상 추가적인 하이퍼파라미터 최적화 작업이 충분히 이루어지지 못함
- 모델의 최적 성능을 발휘하지 못했을 가능성이 있음

모델 활용의 한계

- 텍스트를 읽어주는 음성을 게임 내 동물 캐릭터의 목소리로 변환하려는 목표는 달성되었으나,
  음성이 제대로 들리지 않고 인식하기 어려움

발표를 들어줘서 고마워 구리