LLM을 활용한 한국어 교육 솔루션

한류 컨텐츠를 활용한 한국어 교육 플랫폼

팀 명 : 뉴럴닥터스

김동완 오윤택 임수현 정슬기 조차선

목차

1. 개요

- 팀원 소개
- 일정 소개
- 스킬 셋

2. 서비스

- 서비스 기획 의도
- 서비스 주제
- 서비스 기반

3. 기술

- 모델 설명
- 홈페이지 제작 과정

4. 시연 및 정리

- 기대효과
- 개선점
- 시연

개요

- 팀원 소개
- 일정 소개
- 스킬 셋

팀원 소개

김동완

총괄, 발표, 모델 파인튜닝, 데이터 전처리

오윤택

EDA, 데이터 수집, 데이터 전처리

임수현

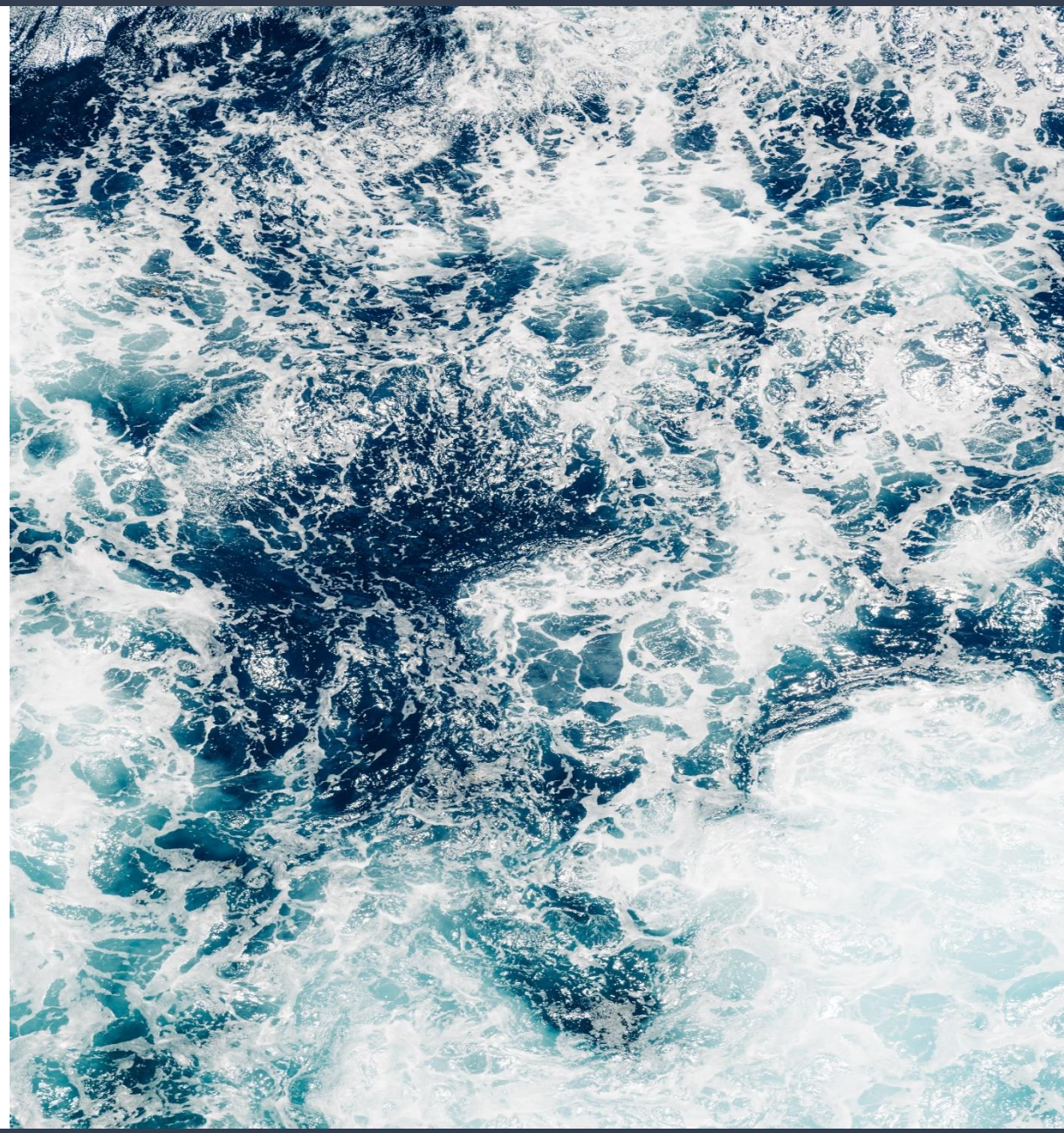
웹(플라스크), 모델 수집, 데이터 전처리

정슬기

프론트엔드, PPT 제작, 데이터 전처리

조차선

EDA, 데이터 수집, 데이터 전처리



프로젝트 기간 별 수행 절차

구분	기간	활동	비고
사전 기획	07/03(월) ~ 07/07(금)	• 프로젝트 기획 및 주제 선정 • 기획안 작성	+ 7/7(금) 1차 기획안 발표
모델 수집 및 말뭉치 수집	07/10(월) ~ 07/21(금)	• 허깅페이스에서 모델 찾기 • 필요한 말뭉치 수집 • 그라디오 웹 구현	
데이터 전처리	07/24(월) ~ 07/28(금)	• 말뭉치 데이터 전처리	
모델 파인튜닝	07/31(월) ~ 08/11(금)	• 모델 파인튜닝 및 중간 피드백	8/4(금) 중간 피드백
웹 구현	08/14(월) ~ 08/22(화)	• 플라스크 웹 구현	
총 개발 기간	**약 7주** (7/3(월)~8/23(수))		

스킬 셋

활용모델

Google BERT

Whisper

라이브러리 및 개발환경

Hugging Face

Keras

Flask
web development, one drop at a time

pandas

NumPy

TensorFlow

gradio

GitHub

웹

HTML
Markup Language
Content

CSS
Style sheet Language
Presentation

JS
Programming Language
Behavior

개발환경

colab

jupyter

python

서비스

- 서비스 기획 의도
- 서비스 주제
- 서비스 제공 기반

서비스 기획 의도 - 한류

"BTS·오징어게임 좋아요"...한류 콘텐츠 경쟁력 높아졌다

한류를 즐기는 외국인들, "한국어 배우자" 열풍

출처 - 한화문화타임즈

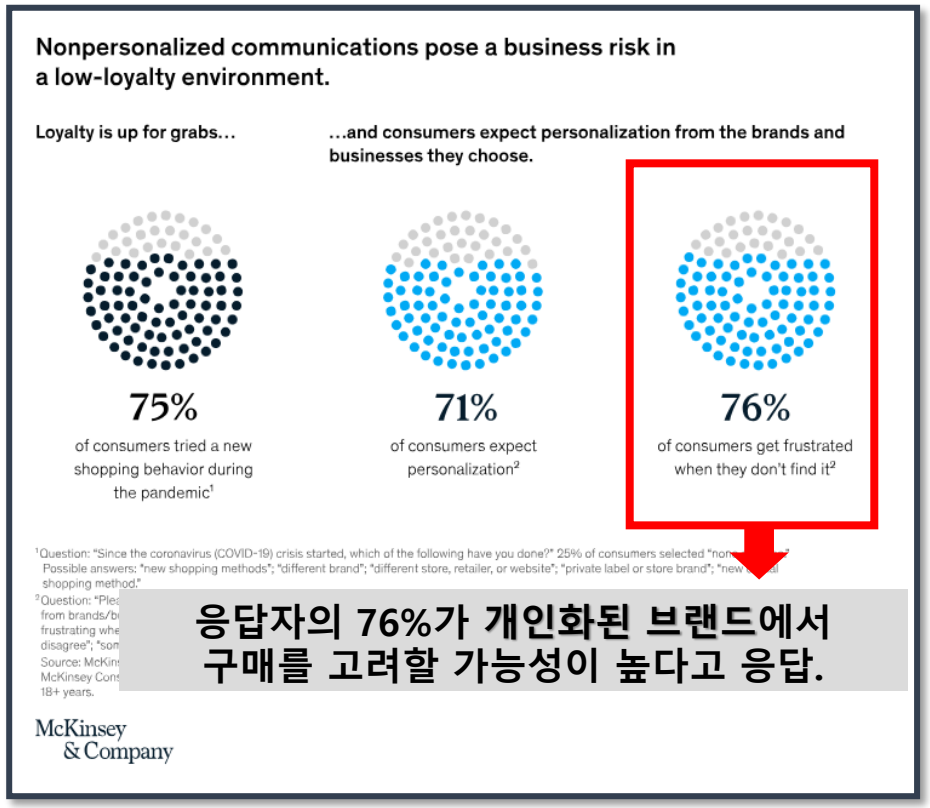
2021년 기준 문화 예술 저작권
7억 5천만 달러 흑자

역대 최대 기록

세종학당 한국어 보급 증가

2017년 52개국
2021년 82개국

서비스 기획 의도 – 개인 맞춤



응답자의 76%가 개인화된 브랜드에서 구매를 고려할 가능성이 높다고 응답.

개인 맞춤형 서비스의 니즈 증가.

맞춤형 서비스 제공을 위해 Ai 기술이 필수적.

출처 – 맥킨지 앤 컴퍼니 설문조사
https://www.mckinsey.com/featured-insights/mckinsey-explainers/what-is-personalization

서비스 주제

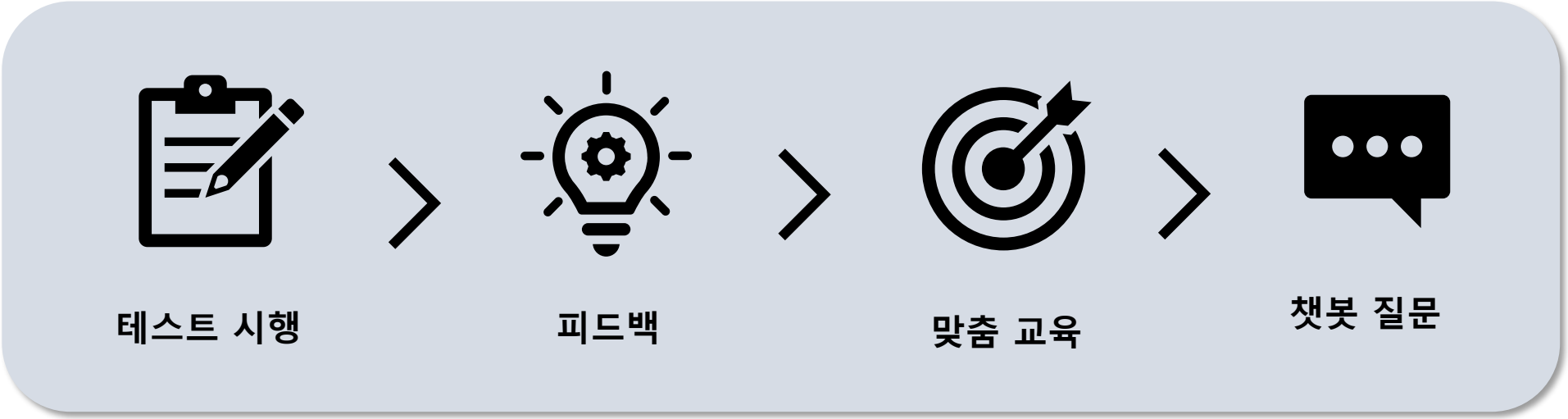
인공지능 딥러닝 기술을 활용하여

한국어를 배우고 싶어하는 외국인에게

맞춤형 한국어 교육을 제공하는 플랫폼 개발

서비스 시나리오

교육 커리큘럼 설명



서비스 기반 1

중급 수준의 학습자.
일상적이고 다양한 사회활동에 필요한 의사소통능력

서울대학교 언어교육원의 정규 교육과정을 기준

한국어에 대한 지식이 전혀 없거나 초급 수준의 학습자.
기본적인 의사소통능력 위주.

고급 수준의 학습자.
한국의 역사, 문화에 대해 폭넓은 지식 위주.

서비스 기반 2

본 연구의 목적은 웹드라마를 중심으로 영상 매체를 활용한 한국어 교육의 효과를 살펴보는 데 있다. 이를 위해 외국인 유학생 30명을 대상으로 24차시 동안 웹드라마를 활용한 수업을 실시하였다. 그리고 영상 매체를 활용한 한국어 교육의 효과를 검증한 11편의 연 구를 메타분석하고 결과를 비교하여 논의하였다. 본 연구의 주요 연구 결과는 다음과 같다. 첫째, 영상 매체 전체의 효과크기는 .733, 웹 드라마를 활용한 한국어 교육의 효과크기는 1.437로 나타났다. 이를 통해 웹드라마의 효과가 매우 크다는 사실을 알 수 있었다. 둘째, 숙 달도에 따른 효과크기를 분석한 결과 중・고급 학습자보다 초급 학습 자들에게 더 효과적이라고 나타났다(초급: 2.585, 중・고급: 1.188). 셋째, 매체 유형에 따른 효과크기를 분석한 결과 웹드라마의 효과크기가 가장 크고 드라마, 영화, 뉴스 순으로 효과크기가 큰 것으로 분석 되었다(웹드라마: 1.437, 드라마: .878, 영화: .881, 뉴스: .577). 본 연구는 웹드라마를 활용한 한국어 교육의 효과가 매우 크다는 사실을 검증한 데 의의가 있다. 또한 본 연구에서 제시된 메타분석 결과는 향후 수행될 효과성 검증 연구의 기초자료로 활용될 수 있을 것으로 기대된다.

드라마를 활용한 한국어 교육의 효과가 매우 크다는 사실이 검증됨.

요인'으로 나타났다. 박한별(2018)에서도 마찬가지로 싱가포르 한국어 학습자들의 학습 동기 요인으로 '여행 지향성'과 '한국 문화에 대한 관심'이 나타났으며 '학습 심리'와 '한국어에 대한 태도'가 탈동기 요인으로 나타났다. 또한 국외 중국인 대학생을 대상으로 한 우림(2018)에서는 학습자들이 '외국어에 대한 관심'이라는 동기 요인으로 한국어 학습을 시작하였음을 밝혀냈다.

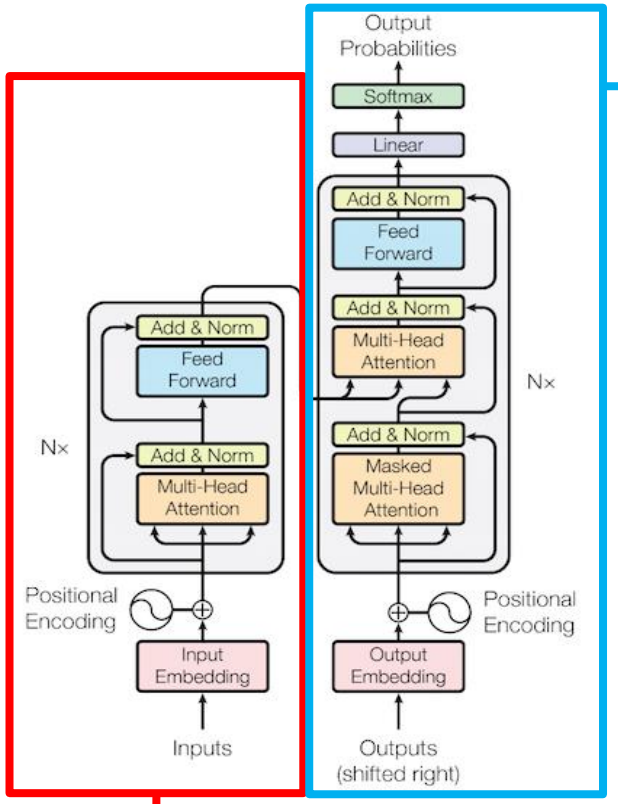
K-pop, 한국 드라마가 한국어 교육 동기에 매우 중요한 요소.

참고문헌
백재파. (2018). 영상매체를 활용한 한국어 교육의 효과 : 웹드라마를 중심으로. 한국어문화교육, 11(2), 61-83.
차 서신웨. (2019). 미얀마인 한국어 학습자의 학습 동기 분석 -한국어 비학위 과정 학습자를 대상으로-. 언어와 문화, 15(4), 235-262.

기술

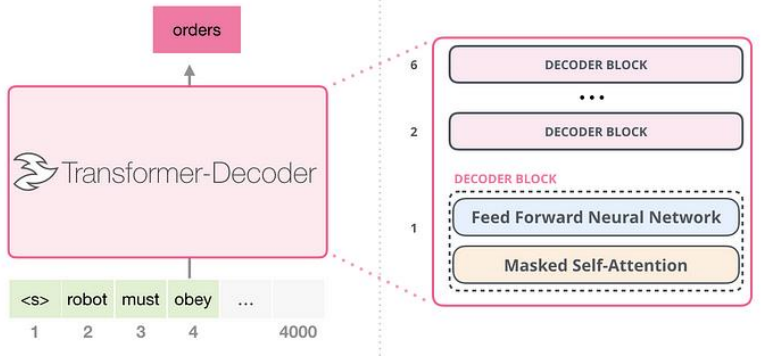
- 모델 설명
- 홈페이지 제작 과정

사용 언어 모델 - 기반

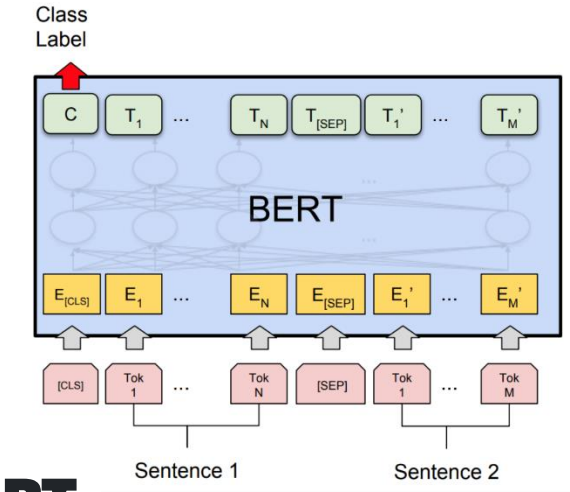


< Transformer >

GPT

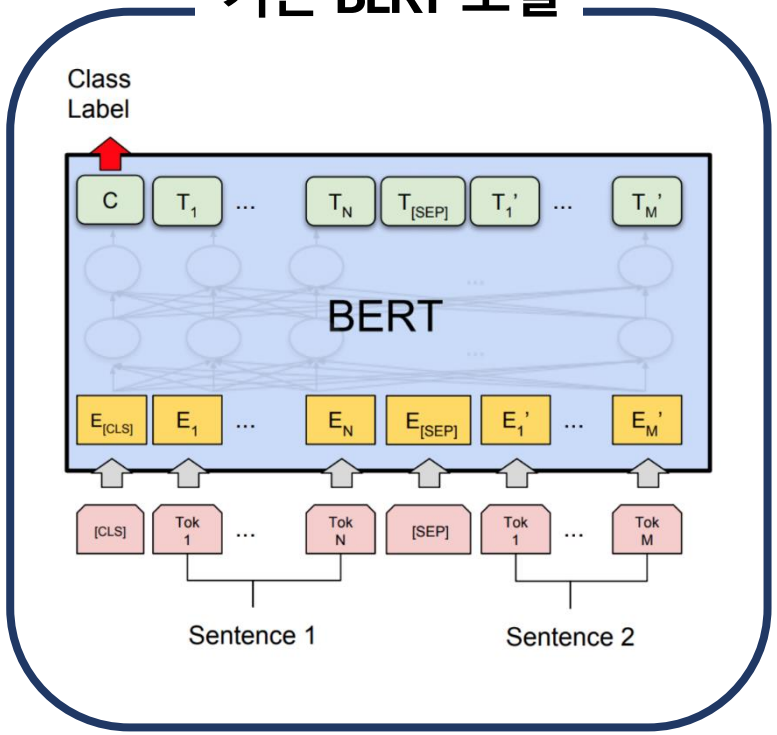


BERT



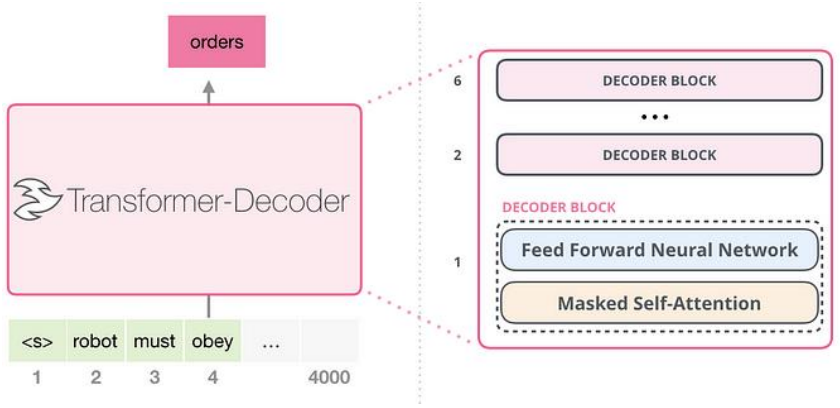
사용 언어 모델 – BERT

기존 BERT 모델



사용 언어 모델 – GPT

기존 GPT 모델



- 문장 요약 모델
- 기계 번역 모델
- 문장생성 모델(뉴스)
- CHAT BOT 모델

1. 분류 모델

```
            (LayerNorm): LayerNorm((768,), eps=1e-12, elementwise_affine=True)
            (dropout): Dropout(p=0.1, inplace=False)
          )
        )
        (intermediate): BigBirdIntermediate(
          (dense): Linear(in_features=768, out_features=3072, bias=True)
          (intermediate_act_fn): NewGELUActivation()
        )
        (output): BigBirdOutput(
          (dense): Linear(in_features=3072, out_features=768, bias=True)
          (LayerNorm): LayerNorm((768,), eps=1e-12, elementwise_affine=True)
          (dropout): Dropout(p=0.1, inplace=False)
        )
      )
    )
    (pooler): Linear(in_features=768, out_features=768, bias=True)
    (activation): Tanh()
  )
  (classifier): BigBirdClassificationHead(
    (dense): Linear(in_features=768, out_features=768, bias=True)
    (dropout): Dropout(p=0.1, inplace=False)
    (out_proj): Linear(in_features=768, out_features=7, bias=True)
  )
)
```

BERT 모델의 output값 중
전체 문장의 임베딩 값을 활용하여
클래스를 분류하는 모델

1. 분류 모델 - 개선 과정

```
id2label[2]

'사회'
```

문제점

정확성 부분

```
text = """
이달 들어 열흘 만에 주담대가 1조원 이상 늘어나는 등 가계 대출 증가세가 이어지는 가운데 50년 만기 등 초장기 주담대가 총부채원리금상환비율(DSR
김 위원장은 16일 '수출금융 종합지원 방안' 간담회 이후 기자들과 만나 "4월부터 주담대가 증가하고 있는데 (초장기 만기 주담대가) 어떤 연령대에서
금융당국은 50년 만기 주담대가 DSR 규제 우회 수단으로 활용되자 만 34세 미만으로 연령을 제한하는 방안 등을 검토하고 있는 것으로 알려졌다.
"""
```

개선

사용 말뭉치
Hugging face
klue_ynet 데이터

Train : 45,678
Validation : 9,107
총 54,785

title	label
다결정 소재 성능제어 기술 개발…반도체·배터리에 활용 기대	IT과학
전국 기능경기대회 순위 공개 금지해야	사회
학생 기다리는 초등학교 1학년 교실	사회
시부터 방랑기까지…서정주 전집 20권으로 완간	생활문화
미 트럼프이스라엘 네타냐후 통화…요르단계곡 합병 논의종합	세계
대전MBC 아나운서 채용 성차별 문제 대책위 발족	사회
獨 생존위해 노역 지원 게토 출신 900명에 보상	세계
홈런레이스 우승 이대호 상금은 종덕이와 동료에게종합	스포츠
주말 N 여행 강원권 그땐 그랬지…강릉서 감자전·닭강정 맛보며 추억여행	생활문화
기자협 창립 56주년 축사하는 민병욱 이사장	사회

파인튜닝

```
training_args = TrainingArguments(
    output_dir="topic_model",
    learning_rate=2e-5,
    per_device_train_batch_size=16,
    per_device_eval_batch_size=16,
    num_train_epochs=4,
    weight_decay=0.01,
    evaluation_strategy="epoch",
    save_strategy="epoch",
    load_best_model_at_end=True,
)

trainer = Trainer(
    model=model,
    args=training_args,
    train_dataset=tokenized_datasets["train"],
    eval_dataset=tokenized_datasets["validation"],
    tokenizer=tokenizer,
    # data_collator=data_collator,
    compute_metrics=compute_metrics,
)

trainer.train()
```

```
You're using a BertTokenizerFast tokenizer. Please note that with a fast tokeni
Attention type 'block_sparse' is not possible if sequence_length: 18 <= num glol
[11420/11420 21:01, Epoch 4/4]
```

Epoch	Training Loss	Validation Loss	Accuracy
1	0.395600	0.406281	0.864390
2	0.319600	0.393907	0.868782
3	0.243600	0.470033	0.861974
4	0.180000	0.487071	0.869661

1. 분류 모델 - 평가

```
ck_size =
경제
```

문제점

해결

```
text = """
이달 들어 열흘 만에 주담대가 1조원 이상 늘어나는 등 가계 대출 증가세가 이어지는 가운데 50년 만기 등 초장기 주담대가 총부채원리금상환비율(DSR
김 위원장은 16일 '수출금융 종합지원 방안' 간담회 이후 기자들과 만나 "4월부터 주담대가 증가하고 있는데 (초장기 만기 주담대가) 어떤 연령대에서
금융당국은 50년 만기 주담대가 DSR 규제 우회 수단으로 활용되자 만 34세 미만으로 연령을 제한하는 방안 등을 검토하고 있는 것으로 알려졌다.
"""
```

평가

사용 말뭉치
klue_ynet 데이터
9,107개

title (string)	label (class label)
"5억원 무이자 융자는 되고 7천만원 이사비는 안된다"	2 (사회)
"왜 수소충전소만 더 멀리 떨어져야 하나 한경연 규제개혁 건의"	2 (사회)
"항응고제 성분 코로나19에 효과…세포실험서 확인"	0 (IT과학)
"실거래가 가장 비싼 역세권은 신반포역…3.3㎡당 1억 육박"	1 (경제)
"기자회견 하는 성 소수자 단체"	2 (사회)
"모의선거 교육 불허 선관위·교육부 각성하라"	2 (사회)
"뮤지컬 영웅 합류한 안재욱 정성화와 다른 안중근 보여줄 것"	3 (생활문화)
"가짜뉴스 징벌적 손해배상제도 도입 변협 토론회"	2 (사회)
"MBN 노조 부동산 물적분할 중단하고 소유경영 분리해야"	2 (사회)
"포스코건설 11년 만에 더 샵 브랜드 로고 교체"	1 (경제)

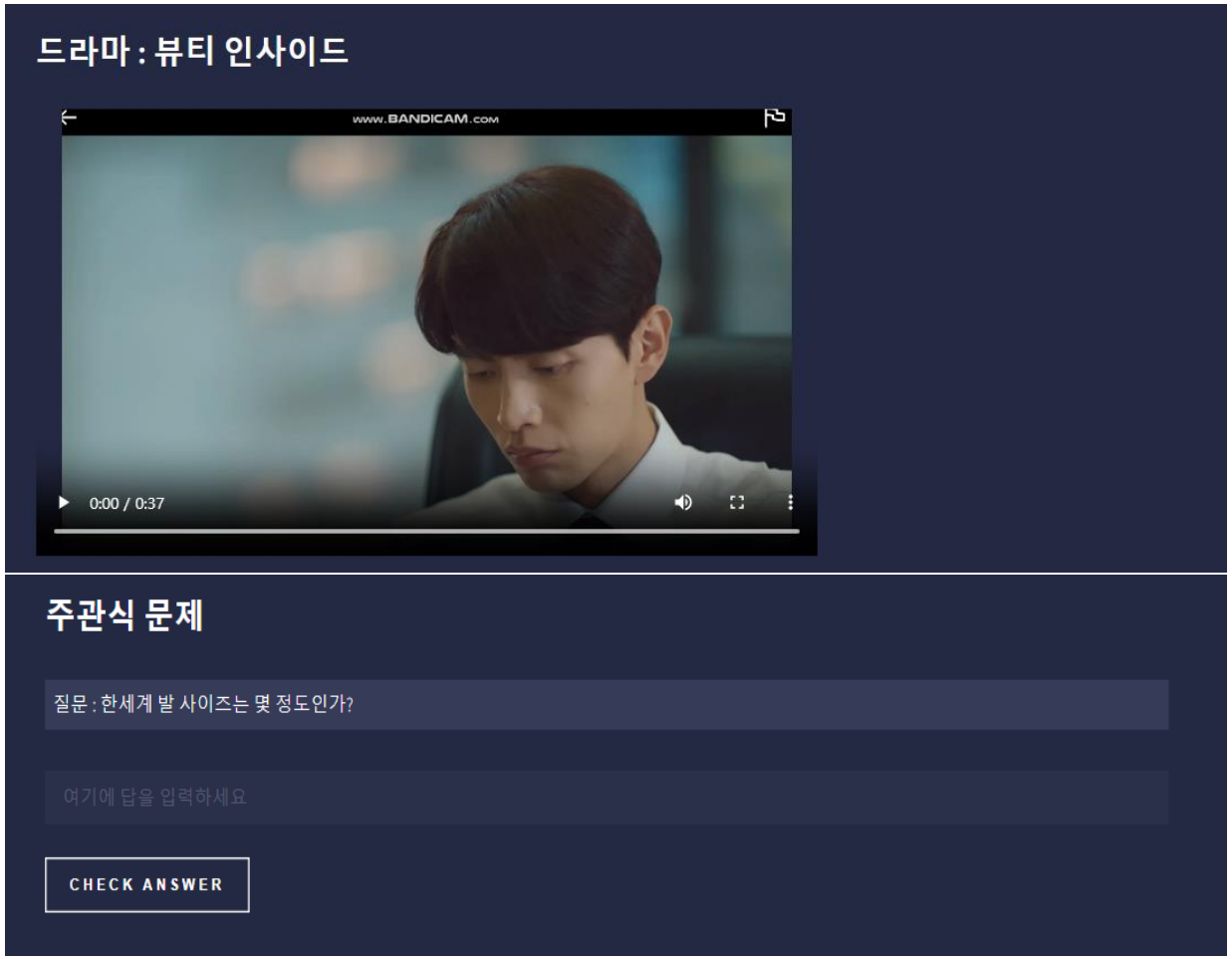
정확도 평가,
손실값

Validation Loss	Accuracy
0.431158	0.855276

2. 질의 응답 모델

```
                (query): Linear(in_features=1024, out_features=1024, bias=True)
                (key): Linear(in_features=1024, out_features=1024, bias=True)
                (value): Linear(in_features=1024, out_features=1024, bias=True)
                (dropout): Dropout(p=0.1, inplace=False)
              )
              (output): RobertaSelfOutput(
                (dense): Linear(in_features=1024, out_features=1024, bias=True)
                (LayerNorm): LayerNorm((1024,), eps=1e-05, elementwise_affine=True)
                (dropout): Dropout(p=0.1, inplace=False)
              )
            )
            (intermediate): RobertaIntermediate(
              (dense): Linear(in_features=1024, out_features=4096, bias=True)
              (intermediate_act_fn): GELUActivation()
            )
            (output): RobertaOutput(
              (dense): Linear(in_features=4096, out_features=1024, bias=True)
              (LayerNorm): LayerNorm((1024,), eps=1e-05, elementwise_affine=True)
              (dropout): Dropout(p=0.1, inplace=False)
            )
          )
        )
      )
    )
    (qa_outputs): Linear(in_features=1024, out_features=2, bias=True)
  )
```

**BERT 모델의 word별 output의 첫번째,
마지막 로직을 활용하여
답안을 예측하는 모델**



- 자동 생성된 문제의 답안지 제작에 사용.
- 사용자의 질문에 답변하는 모델로 사용.

2. 질의 응답 모델- 개선 과정

문제점

긴 문장에서는 결과가 잘 나오지 않음.

```
text1 = """서울특별시는 대한민국의 수도[18], 최대도시다. 평양시, 경주시, 개성시와 함께 오랜 역사를 가진 한
지방자치법의 특별법[19]으로 법률상 대한민국 제1의 도시로 규정되어 있다. 현재 한국에서 유일한 특별시고, 수
역사적으로도 백제, 조선, 대한제국의 소도이자 현재 대한민국의 수도로서 중요성이 높다. 기원전 18년 백제가 한
기원후 553년 신라가 서울 지역을 차지하고 신라의 한산주 한양군이 되었다. 이후 고려시대에는 약 250년간[22]
qestion = "고구려가 한성을 함락한 시기"

qa(text1,qestion)
```

```
''
```

개선

Hugging_face
korquad 데이터
Train : 93,557
Validation : 16,496
총 110,053

	text	question	answer
	2012년 11월 16일 네덜란드 헤렌빈에서 열린 2012-2013 국제빙상경기연맹...	김보름은누구를 제치고 종합우승을 달성하였는가?	{'answer_start': 1641, 'text': '마리스카 후이스만'}
	대장장이 클란이 자기 집에서 연회를 열고 콘코바르를 초대했다. 연회에 가기 전에 콘...	대장장이 클란이 연회에서 초대한 사람은 누구인가?	{'answer_start': 23, 'text': '콘코바르'}
	세 개의 오스트리아 부대는 마을의 반도 점거하지 못했다. 아스페른의 주요 거점 오세...	아스페른 에슬링 전투에서 프랑스군의 1차 돌격은 어느군대에 의해 격퇴당했나?	{'answer_start': 266, 'text': '오스트리아군'}
	1981년 3월 역시 제육관에서의 간접 선거를 통해 스스로 대한민국의 제12대...	국민들의 환심을 사기 위한 대표적인 유화책으로 컬러 텔레비전 보급 및 프로스 영화 ...	{'answer_start': 251, 'text': '35 정책'}
	추수감사절은 영국에서 매사추세츠의 플리머스 식민지로 이주한 필그림 파더스의 첫 수확...	1621년 가을 윔파노아그 족을 초대하여 같이 음식을 먹은 이들은?	{'answer_start': 219, 'text': '필그림 파더스'}
	..	..	..
77	각 올림픽 종목들은 IOC로부터 승인을 받은 국제경기연맹의 관리를 받는다. 35개의...	정식종목 선정의 방법은?	{'answer_start': 235, 'text': 'IOC위원들의 투표'}
78	현대의 초기 번역자들은 테트리그라마톤 대신에 "아도나이"로 읽는 것을 경시하였다. ...	현대의 초기 번역자들은 테트리그라마톤 대신 무엇으로 읽는것을 경시하였는가?	{'answer_start': 26, 'text': '아도나이'}
79	1985년 3월 제2대 총장에 권이혁 박사가 취임하고 520명의 신입생을 받아 제1...	한국교원대학교 1회 입학식 당시 신입생 수는?	{'answer_start': 30, 'text': '520명'}
80	15세에 그는 아동복 제작을 위한 스케치를 그린다. 모든 의상이 마르셀 조사(Mar...	장 폴 고티에가 아동복 제작을 위해 스케치를 그린 나이는?	{'answer_start': 0, 'text': '15세'}
81	프랑스 침공 제1단계 계획인 황색 작전(Fall Gelb)은 육군최고지휘관 발터 폰...	황색 작전은 1차 세계대전 때의 무엇을 답습한 것이었나?	{'answer_start': 129, 'text': '슐리펜 계획'}

파인튜닝

```
trainer(model,dic,optimizer,criterion,4)
```

```
에포크 수: 0 / 4
훈련 총 손실값: tensor(4.6273, device='cuda:0', grad_fn=<DivBackward1>)
벨류 총 손실값: 5.065569185447138
에포크 수: 1 / 4
훈련 총 손실값: tensor(4.9695, device='cuda:0', grad_fn=<DivBackward1>)
벨류 총 손실값: 4.7537631952253384
에포크 수: 2 / 4
훈련 총 손실값: tensor(5.0075, device='cuda:0', grad_fn=<DivBackward1>)
벨류 총 손실값: 4.592672390359074
에포크 수: 3 / 4
훈련 총 손실값: tensor(5.3103, device='cuda:0', grad_fn=<DivBackward1>)
벨류 총 손실값: 4.48261434337415565
```

2. 질의 응답 모델- 평가

문제점

해결

```
text1 = """서울특별시는 대한민국의 수도[18], 최대도시다. 평양시, 경주시, 개성시와 함께 오랜 역사를 가진 한반도
지방자치법의 특별법[19]으로 법률상 대한민국 제1의 도시로 규정되어 있다. 현재 한국에서 유일한 특별시고, 수장인
역사적으로도 백제, 조선, 대한제국의 소도이자 현재 대한민국의 수도로서 중요성이 높다. 기원전 18년 백제가 현 송
기원후 553년 신라가 서울 지역을 차지하고 신라의 한산주 한양군이 되었다. 이후 고려시대에는 약 250년간[22] 개경
qestion = "고구려가 한성을 함락한 시기 "

qa(text1,qestion)
```

```
'475'
```

평가

사용 말뭉치
Hugging_face
korquad 데이터
11,548

	text	start_index	end_index
0	임종석이 여의도 농민 폭력 시위를 주도한 혐의로 지명수배 된 날은?[SEP]1989...	42	54
1	1989년 6월 30일 평양축전에 대표로 파견 된 인물은?[SEP]1989년 2월 ...	162	165
2	임종석이 여의도 농민 폭력 시위를 주도한 혐의로 지명수배된 연도는?[SEP]1989...	42	47
3	임종석을 검거한 장소는 경희대 내 어디인가?[SEP]1989년 2월 15일 여의도 ...	394	404
4	임종석이 조사를 받은 뒤 인계된 곳은 어딘가?[SEP]1989년 2월 15일 여의도...	487	499
...	...	...	...
5769	중동 지역에서 갤럭시 S7 엣지가 폭발하는 사건이 발생한 년도는?[SEP]2016년...	41	46
5770	갤럭시 S7엣지를 만든 회사는?[SEP]2016년 9월 4일 중동 지역에서 갤럭시 ...	105	109
5771	2016년 9월 4일 갤럭시 S7엣지가 폭발한 사건이 발생한 지역은?[SEP]201...	55	60
5772	갤럭시 노트 7은 출시 며칠만에 기기 결함으로 터지기 시작하였나?[SEP]2016년...	327	330
5773	2016년에 갤럭시 S7 엣지가 폭발한 사건은 어느 지역에서 일어났는가?[SEP]2...	57	62

손실값

```
# 총 손실값 - 기존 모델
evaluate_model(model,optimizer,criterion,ex_dic,1)
```

```
1 / 1
총손실값 tensor(11.0990, device='cuda:0')
```

```
# 총 손실값 - 파인튜닝 후 모델
evaluate_model(model,optimizer,criterion,ex_dic,1)
```

```
1 / 1
총손실값 tensor(4.5101, device='cuda:0')
```

3. 요약 모델

```
            )
            (layer_norm): T5LayerNorm()
            (dropout): Dropout(p=0.1, inplace=False)
          )
          (1): T5LayerCrossAttention(
            (EncDecAttention): T5Attention(
              (q): Linear(in_features=768, out_features=768, bias=False)
              (k): Linear(in_features=768, out_features=768, bias=False)
              (v): Linear(in_features=768, out_features=768, bias=False)
              (o): Linear(in_features=768, out_features=768, bias=False)
            )
            (layer_norm): T5LayerNorm()
            (dropout): Dropout(p=0.1, inplace=False)
          )
          (2): T5LayerFF(
            (DenseReluDense): T5DenseGatedActDense(
              (wi_0): Linear(in_features=768, out_features=2048, bias=False)
              (wi_1): Linear(in_features=768, out_features=2048, bias=False)
              (wo): Linear(in_features=2048, out_features=768, bias=False)
              (dropout): Dropout(p=0.1, inplace=False)
              (act): NewGELUActivation()
            )
            (layer_norm): T5LayerNorm()
            (dropout): Dropout(p=0.1, inplace=False)
          )
        )
      )
    )
    (final_layer_norm): T5LayerNorm()
    (dropout): Dropout(p=0.1, inplace=False)
  )
  (lm_head): Linear(in_features=768, out_features=50358, bias=False)
)
```

사용모델 - GPT 모델
긴 문장을 input값으로 이용하여
짧은 문장을 생성하는 모델

뉴스로 학습하기

뉴스 기사 제목

"호신용품이 흉기로"...신림 성폭행범 '너클'에 피해자 위독

뉴스 기사 내용

[이데일리 박지혜 기자] 이른바 '신림동 공원 성폭행' 피의자가 범행에 '너클'을 사용한 것으로 알려지면

뉴스 기사 요약

서울 관악경찰서는 신림동 공원 여성을 때리고 성폭행한 혐의로 30대 남성 A씨를 현행범 체포했으며, A씨는 피해자와도 아는 사이가 아닌 것으로 파악되었다. 금속 재질의 너클은 항공기 내 반입 금지 물품이지만 국내에서는 온라인을 통해 제한 없이 구매할 수 있으며, 최근 묻지마 흉기 난동 사건이 연이어 발생하면서 너클 구매도 증가하고 있다.

소지나 사용을 금지하는 국가도 있다. 너클은 항공기 내 반입 금지 물품이기도 하다. 그러나 국내에선 호신용품으로 알려지면서 온라인을 통해 제한 없이 구매할 수 있다. 최근 '묻지마 흉기 난동' 사건이 잇따르면서 너클 구매도 증가했다. 너클 관련 범죄는 이번이 처음이 아니다. 올해 1월 경기도 수원에서 한 10대 운전자가 보행자에게 너클을 손에 낀 채 주먹을 휘두르고 달아났다가 붙잡히는 사건이 발생했다. 이 사건 피해자는 실명 위기에 놓였다. 당시 경찰은 너클을 쓴 폭행에 대해 특수상해, 흉기로 협박한 것에 특수협박 혐의를 적용해 가해자를 구속했다. 한편, 경찰은 A씨가 범행 도구를 미리 준비한 점 등으로 미뤄 사전에 범행을 계획한 것으로 보고 최근 올라온 신림동 살인 예고 글 등 관련성도 조사하고 있다. 경찰은 A씨를 상대로 범행 동기와 경위 등을 조사한 뒤 18일 구속영장을 신청할 방침이다.

- **뉴스, 논문 기사 요약**
- **드라마 대화내용 요약**

3. 요약 모델 - 개선 과정

문제점

과학논문자료 요약 능력 미흡

```
summarize(text)
```

'해양환경 정보화플랫폼의 자료와 위성 관측자료를 활용하여 해양오염 예측 도구를 개발하고자 함 딥러닝 기법의 효율성을 검증함 딥러닝 모델을 사용한 연구 사례를 소개함 딥러닝 기법의 추가적인 개발은 환경정책계획 수립에 기여할 수 있다.'

개선

사용 말뭉치
Ai hub에서 제공
기술과학
요약 데이터
Train:96,047
validation:12,145
총: 108,192

'context': ' LED 처리량에 따른 SD rat 간 조직에서 cytokine TNF-㎳(㎳alpha ㎳),IL-1 ㎳(㎳beta ㎳), IL-6의 함량을 측정하였다. Control rat 간조직의 TNF-㎳(㎳alpha ㎳)함량은 149 ㎳(㎳mathrm { ng } / ㎳mathrm { g } ㎳)으로 119 ㎳(㎳mathrm { ng } / ㎳mathrm { g } ㎳)인 Vehicle rat 간조직에 비해 유의성(p ㎳(㎳mathrm {< } ㎳)0.05)있게 증가하여 ㎳(㎳mathrm { CCl } _ { 4 } ㎳)의 독성이 유발 되었다. TNF-㎳(㎳alpha ㎳) 의 함량은 LED 전 처리 (200-400 BW)에서 Control rat에 비해 유의성 있게 감소되어(㎳((㎳mathrm { p }<0.05) ㎳), LED의 TNF-㎳(㎳alpha ㎳) 감소효과를 알 수 있었다. 이 감소효과는 Silymarin처리(120 ㎳(㎳mathrm { ng } / ㎳mathrm { g } ㎳)/)효과와 유사하였다. 그러나 LED 200, 300, 400 ㎳(㎳mathrm { mg } / ㎳mathrm { kg } ㎳)BW 처리간의 차이는 유의성이 없었다. 이와 같은 ㎳(㎳mathrm { mg } / ㎳mathrm { kg } ㎳) BW 처리에서 가장 효율적으로 ㎳(㎳mathrm { CCl } _ { 4 } ㎳)의 SD rat 간독성다. LED 200 ㎳(㎳mathrm { mg } / ㎳mathrm { kg } ㎳)BW 이상의 처리에서 이들 함량은 Vehicle rat수준은 아니지만 Silymarin 처리 rat의 수준으로 감소되었다. 이들 cytokine에 대한 결과는 LED ㎳(^ { 2 } ㎳)가 ㎳(㎳mathrm { CCl } _ { 4 } ㎳)유발 rat의 간 손상을 완화시킬 수 있고, LED 200 ㎳(㎳mathrm { mg } / ㎳mathrm { kg } ㎳)BW의 처리량이 가장 효율적임을 의미하였다. <h1>고찰</h1> ㎳(㎳mathrm { CCl } _ { 4 } ㎳)는 xenobiotics와 biomolecules를 대사하는 간에 독성을 유발하는 물질로서 간에 대한 독성을 연구하는데 널리 사용되고 있다. ㎳(㎳mathrm { CCl } _ { 4 } ㎳)는 간의 mixed function (cytochrome ㎳(㎳mathrm { P } _ { 450 } ㎳))oxidase system에 의해 free radical인 ㎳({ } ^ {㎳circ } ㎳mathrm { CCl } _ { 3 } ㎳)로 전환되고, 전환된 free radical은 간세포의 막 단백 thiol기와 공유결합하여 막의 지질과산화 반응을 촉진시키고 조직 내 지질과산화를 유도함으로써 지방산 조성을 변화시키고 궁극적으로는 세포독성을 유발시킨다. 본 연구에서 ㎳(㎳mathrm { CCl } _ { 4 } ㎳)로 간독성을 유발한 SD rat에서 LED는 간 독성을 완화하였는데, 이는 LED의 항산화력과 간 독성과 관련이 있는 cytokine 생성 억제에 기인하였다. ',
'context_length': 1542,
'summary': 'LED 처리량에 따라 SD rat 간 조직의 TNF-㎳(㎳alpha㎳), IL-1㎳(㎳beta㎳), IL-6 함량을 측정한 결과, Control rat 간조직의 TNF-㎳(㎳alpha㎳) 함량은 Vehicle rat 간조직에 비해 유의성 있게 증가하여 CCl4의 독성이 유발되었고, LED 전 처리 (200-400 BW)에서 Control rat에 비해 유의성 있게 감소하여 LED의 TNF-㎳(㎳alpha㎳) 감소효과를 알 수 있었으며, Silymarin 처리에 비해 유의성 있게 큰 감소효과를 나타냈다.',
'clue': [{'clue_text': 'LED 처리량에 따른 SD rat 간 조직에서 cytokine TNF-㎳(㎳alpha ㎳),IL-1 ㎳(㎳beta ㎳), IL-6의 함량을 측정하였다.',

파인튜닝

```
fine_tuning(model,dic,optimizer,4)
import numpy as np
%cd /gdrive/MyDrive
param1 = update_param[0].to("cpu").detach().numpy()
param2 = update_param[1].to("cpu").detach().numpy()
np.save("summarize_param_1.npy",param1)
np.save("summarize_param_2.npy",param2)
```

```
에포크 수: 1 / 4
훈련용 손실값:  1.1699388500091414
벨류값 손실값:  1.1005418923750294
에포크 수: 2 / 4
훈련용 손실값:  1.143553729733851
벨류값 손실값:  1.0906655586748952
에포크 수: 3 / 4
훈련용 손실값:  1.1290372989506128
벨류값 손실값:  1.0841190944329169
에포크 수: 4 / 4
훈련용 손실값:  1.1192988877936307
벨류값 손실값:  1.0798830372403931
/gdrive/MyDrive
```

3. 요약 모델 - 평가

문제점

해결

```
summarize(text)
```

'해양환경 정보화플랫폼의 자료, 위성 관측 자료, 기존의 수치 모형에서 계산된 방대한 물리적 자료를 딥러닝 기술에 적용하여 해양오염 예측 도구를 개발하고자 한다.'

평가

사용 말뭉치
Hugging_face
korean_science
데이터
10,000

	text	start_index	end_index
0	임종석이 여의도 농민 폭력 시위를 주도한 혐의로 지명수배 된 날은?[SEP]1989...	42	54
1	1989년 6월 30일 평양축전에 대표로 파견 된 인물은?[SEP]1989년 2월 ...	162	165
2	임종석이 여의도 농민 폭력 시위를 주도한 혐의로 지명수배된 연도는?[SEP]1989...	42	47
3	임종석을 검거한 장소는 경희대 내 어디인가?[SEP]1989년 2월 15일 여의도 ...	394	404
4	임종석이 조사를 받은 뒤 인계된 곳은 어딘가?[SEP]1989년 2월 15일 여의도...	487	499
...	...	...	...
5769	중동 지역에서 갤럭시 S7 엣지가 폭발하는 사건이 발생한 년도는?[SEP]2016년...	41	46
5770	갤럭시 S7엣지를 만든 회사는?[SEP]2016년 9월 4일 중동 지역에서 갤럭시 ...	105	109
5771	2016년 9월 4일 갤럭시 S7엣지가 폭발한 사건이 발생한 지역은?[SEP]201...	55	60
5772	갤럭시 노트 7은 출시 며칠만에 기기 결함으로 터지기 시작하였나?[SEP]2016년...	327	330
5773	2016년에 갤럭시 S7 엣지가 폭발한 사건은 어느 지역에서 일어났는가?[SEP]2...	57	62

손실값

```
evaluate(model,dataloader)
```

평가값: tensor(2.6605, device='cuda:0')

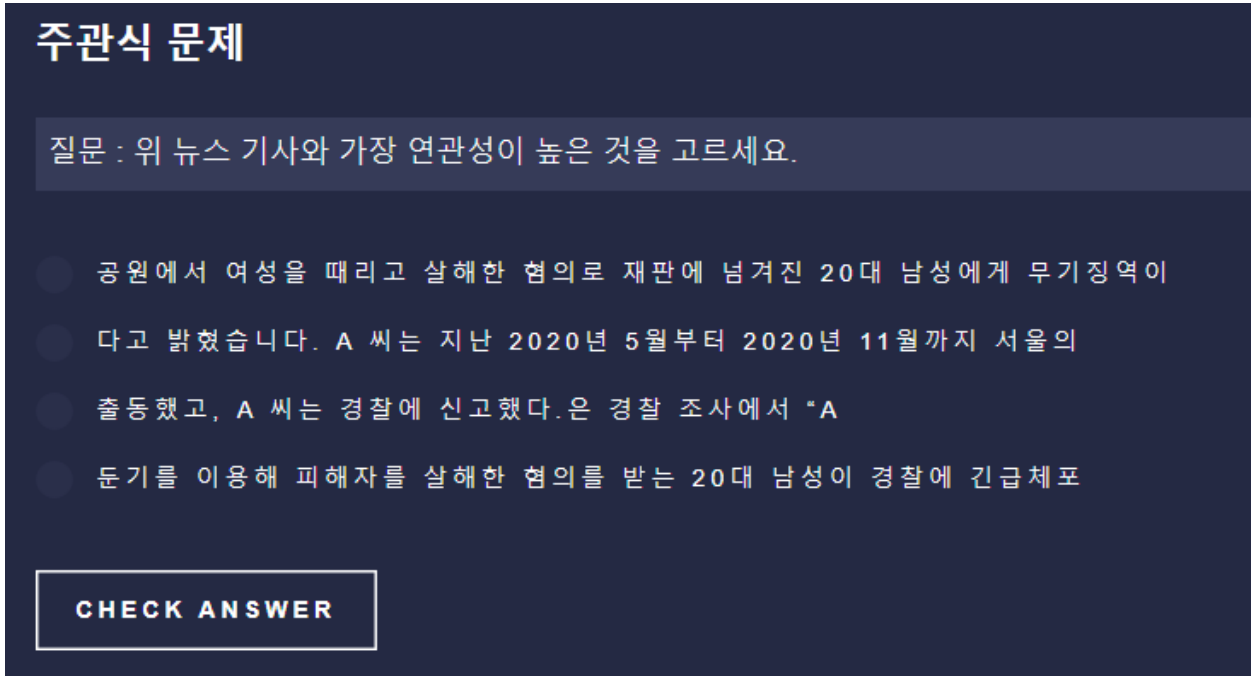
```
evaluate(model,dataloader)
```

평가값: tensor(2.2284, device='cuda:0')

4. 문장 생성 모델 - 사용

```
GPTNeoXForCausalLM(
  (gpt_neox): GPTNeoXModel(
    (embed_in): Embedding(30080, 2048)
    (emb_dropout): Dropout(p=0.0, inplace=False)
    (layers): ModuleList(
      (0-23): 24 x GPTNeoXLayer(
        (input_layernorm): LayerNorm((2048,), eps=1e-05, elementwise_affine=True)
        (post_attention_layernorm): LayerNorm((2048,), eps=1e-05, elementwise_affine=True)
        (post_attention_dropout): Dropout(p=0.0, inplace=False)
        (post_mlp_dropout): Dropout(p=0.0, inplace=False)
        (attention): GPTNeoXAttention(
          (rotary_emb): GPTNeoXRotaryEmbedding()
          (query_key_value): Linear(in_features=2048, out_features=6144, bias=True)
          (dense): Linear(in_features=2048, out_features=2048, bias=True)
          (attention_dropout): Dropout(p=0.0, inplace=False)
        )
        (mlp): GPTNeoXMLP(
          (dense_h_to_4h): Linear(in_features=2048, out_features=8192, bias=True)
          (dense_4h_to_h): Linear(in_features=8192, out_features=2048, bias=True)
          (act): GELUActivation()
        )
      )
    )
    (final_layer_norm): LayerNorm((2048,), eps=1e-05, elementwise_affine=True)
  )
  (embed_out): Linear(in_features=2048, out_features=30080, bias=False)
)
```

GPT 모델을 사용하여
작은 단어로 문장을 생성하는 모델



- 뉴스 문제 생성.
- 논문 문제 생성.

4. 문장 생성 모델 - 개선 과정

문제점

잘못된 문장 많이 배출함

```
text = "국민"
def gen(text,model,tokenizer):
  token = tokenizer(text,return_tensors='pt')
  model.eval()
  model = model.to("cpu")
  output = model.generate(input_ids = token["input_ids"])

  return tokenizer.decode(output[0])
```

```
gen(text,origin_model,origin_tokenizer)
```

```
The attention mask and the pad token id were not set. As a consequence, you may observe unexpected behavior. Please pass your inp
Setting `pad_token_id` to `eos_token_id`:2 for open-end generation.
'국민들은 다 알고 있다.<|endoftext|>'
```

개선

Ai hub제공
전문 용어
기계번역 데이터
Train : 120만
Val :14만
총 134만

```
print(len(train))
print(len(val))
```

```
1200002
149993
```

```
train[10]
```

```
'정몽헌은 대한민국의 기업인으로, 현대그룹 부회장을 역임했다.'
```

파인튜닝

```
에포크 수: 1 / 4
훈련 손실값:  2.5633180902642896
훈련용 손실값:  2.6935776536369325
에포크 수: 2 / 4
훈련 손실값:  2.3793940825823165
훈련용 손실값:  2.735653650671984
에포크 수: 3 / 4
훈련 손실값:  2.296974167868579
훈련용 손실값:  2.7770895330149283
에포크 수: 4 / 4
훈련 손실값:  2.2498750572232216
```

4. 문장 생성 모델 - 평가

문제점

해결

```
text = "국민"
def gen(text,model,tokenizer):
  token = tokenizer(text,return_tensors='pt')
  model.eval()
  model = model.to("cpu")
  output = model.generate(input_ids = token["input_ids"])

  return tokenizer.decode(output[0])
```

```
gen(text,finetune_model,finetune_tokenizer)
```

```
The attention mask and the pad token id were not set. As a consequence, you may observe unexpected behavior. Please pass your input's `attention_
Setting `pad_token_id` to `eos_token_id`:2 for open-end generation.
'국민의힘 윤석열 대선 후보가 지난 11일 서울 여의도 국회 소통관에서 '국민의힘'
```

개선

사용 말뭉치
AI hub
문장생성
평가 데이터
10,000

다리,힘들다,계단,디	다리를 다친 유민이에게 계단이 너무 힘들다.
유민이,다치다,계단	다리를 다친 유민이에게 계단이 너무 힘들다.
유민이,다치다,계단	다리를 다친 유민이에게 계단이 너무 힘들다.
유민이,다치다,계단	다리를 다친 유민이에게 계단이 너무 힘들다.
유민이,다치다,계단	다리를 다친 유민이에게 계단이 너무 힘들다.
종류,김치,다양하다	김치의 맛과 종류가 다양해졌다.
맛,다양하다,김치,종	김치의 맛과 종류가 다양해졌다.
맛,다양하다,김치,종	김치의 맛과 종류가 다양해졌다.
맛,다양하다,김치,종	김치의 맛과 종류가 다양해졌다.
맛,다양하다,김치,종	김치의 맛과 종류가 다양해졌다.
곱씹다,삼키다,숟가	아이는 밥 한 숟가락을 삼키지 않고 곱씹는다.
곱씹다,삼키다,숟가	아이는 밥 한 숟가락을 삼키지 않고 곱씹는다.
곱씹다,삼키다,숟가	아이는 밥 한 숟가락을 삼키지 않고 곱씹는다.
곱씹다,삼키다,숟가	아이는 밥 한 숟가락을 삼키지 않고 곱씹는다.
곱씹다,삼키다,숟가	아이는 밥 한 숟가락을 삼키지 않고 곱씹는다.
보이다,가능성,통하	그녀는 이번 드라마를 통해 가능성을 보여 주었다.
보이다,가능성,통하	그녀는 이번 드라마를 통해 가능성을 보여 주었다.

평가

```
evaluate(model,dataloader)
```

```
손실값:  tensor(13.3723, device='cuda:0')
```

Chat Bot

이미지 출처: KT enterprise

GPT-3.5 활용

맞춤형 Chat Bot

프롬프트 튜닝

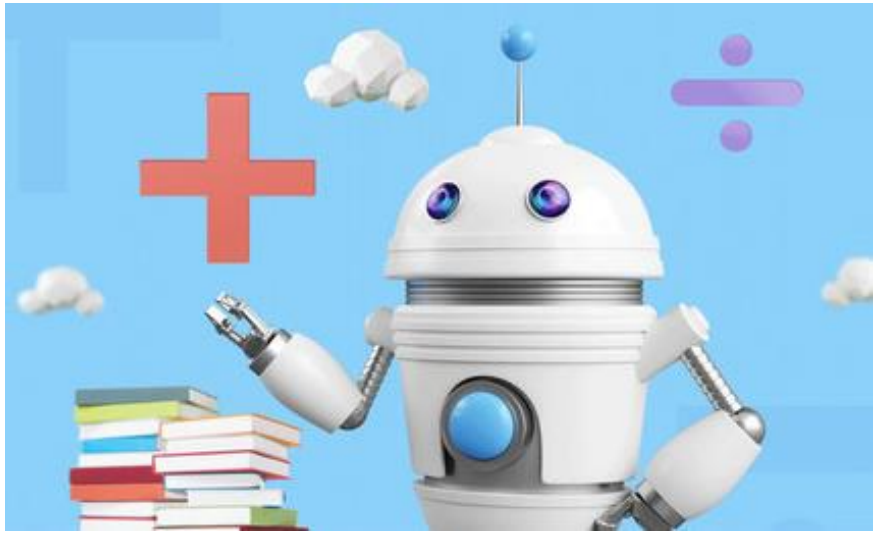
```
import openai
import os

from dotenv import load_dotenv, find_dotenv
_ = load_dotenv(find_dotenv()) # read local .env

openai.api_key  = os.getenv('OPENAI_API_KEY')
```

```
In [ ]:
```

```
def get_completion(prompt, model="gpt-3.5-turbo"
    messages = [{"role": "user", "content": promp
    response = openai.ChatCompletion.create(
        model=model,
        messages=messages,
        temperature=0, # this is the degree of re
    )
    return response.choices[0].message["content"
```

GPT API 활용하여 프롬프트 튜닝

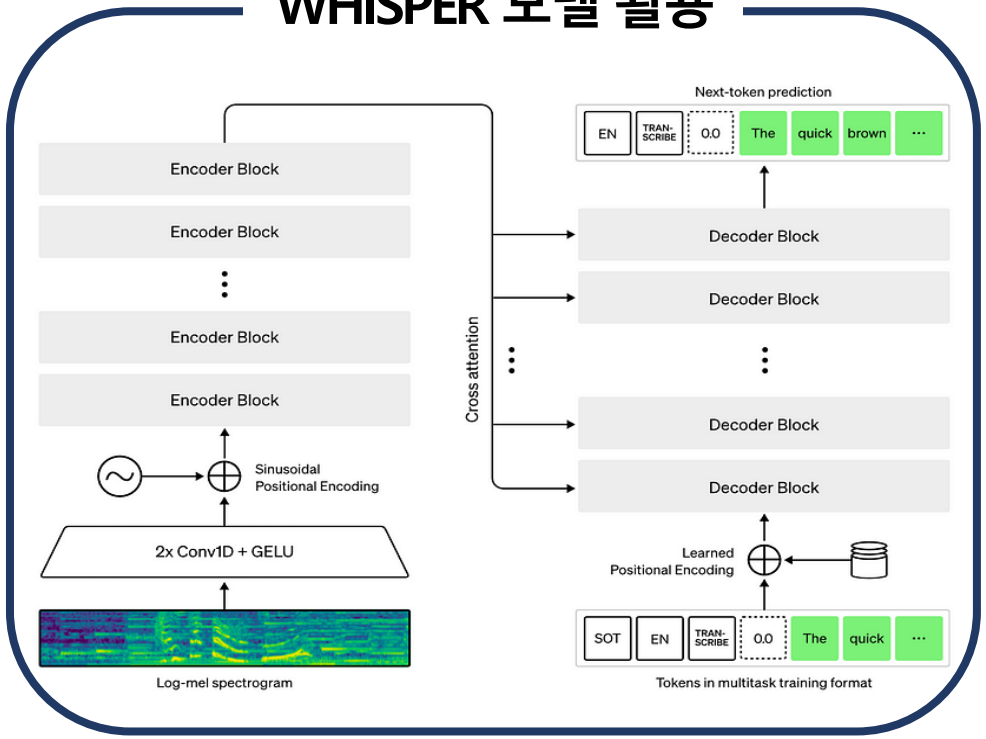
챗봇 레벨 구분

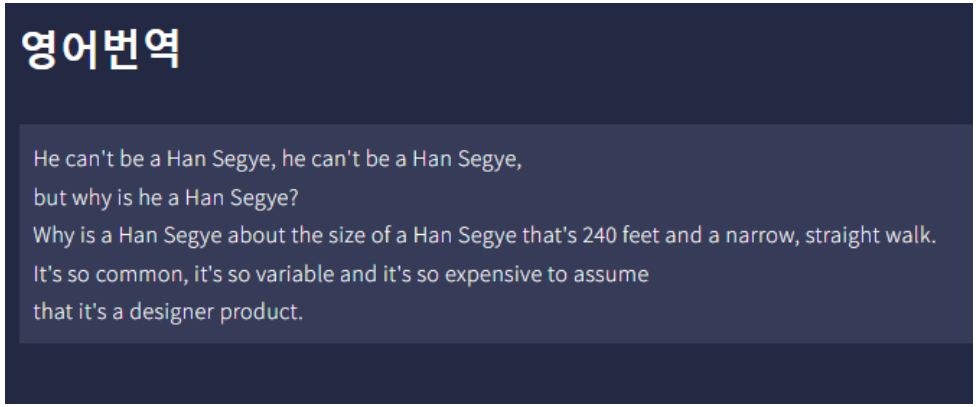


1. 하나부터 열까지 알려주는 선생님
2. 피드백 해주며 문제점을 찾아주는 선생님
3. 토론 하며 실력을 늘려주는 선생님

음성 인식 모델 소개

WHISPER 모델 활용





- 사용자 음성 인식 or 번역
- 드라마, KPOP음성 인식 or 번역

음성 인식 모델 활용

드라마 음성 인식

```
print(text1["text"])
print()
print(text2["text"])
print()
print(text3["text"])
print()
```

한석이 찾으려면 병원이 아니라 클럽 같은 데 뒤져야 되는 거 아니야? 호텔이나. 남자 많다며. 만나는 남자 알이

내가 카메라 앞에서 웃음파라 돈 버는 걸 다행인 줄이나 알아요. 경찰이었어 봐. 당장 연행했지. 두고 봐. 내 여

하는 짓이 거슬려서 더는 못 봐주겠네. 지금 그 말 나한테 하는 겁니까? 김 이사님. 이제 회사 나오지 마세요.

사용자 음성 인식

```
def transcribe_():
  path = record(sec=10)
  korean = whisper_model.transcribe(path)

  return korean
```

```
korean = transcribe_()
print("한국어: ",korean["text"])
```

```
/usr/local/lib/python3.10/dist-packages/whisper/transcribe.py:114: UserWarning: FP16 is not supported on CPU; using FP32 instead
  warnings.warn("FP16 is not supported on CPU; using FP32 instead")
한국어:  안녕하세요 코처입니다. 저희는 지금 피스퍼 모델을 이용해서 음성인식을 진행하고 있습니다.
```

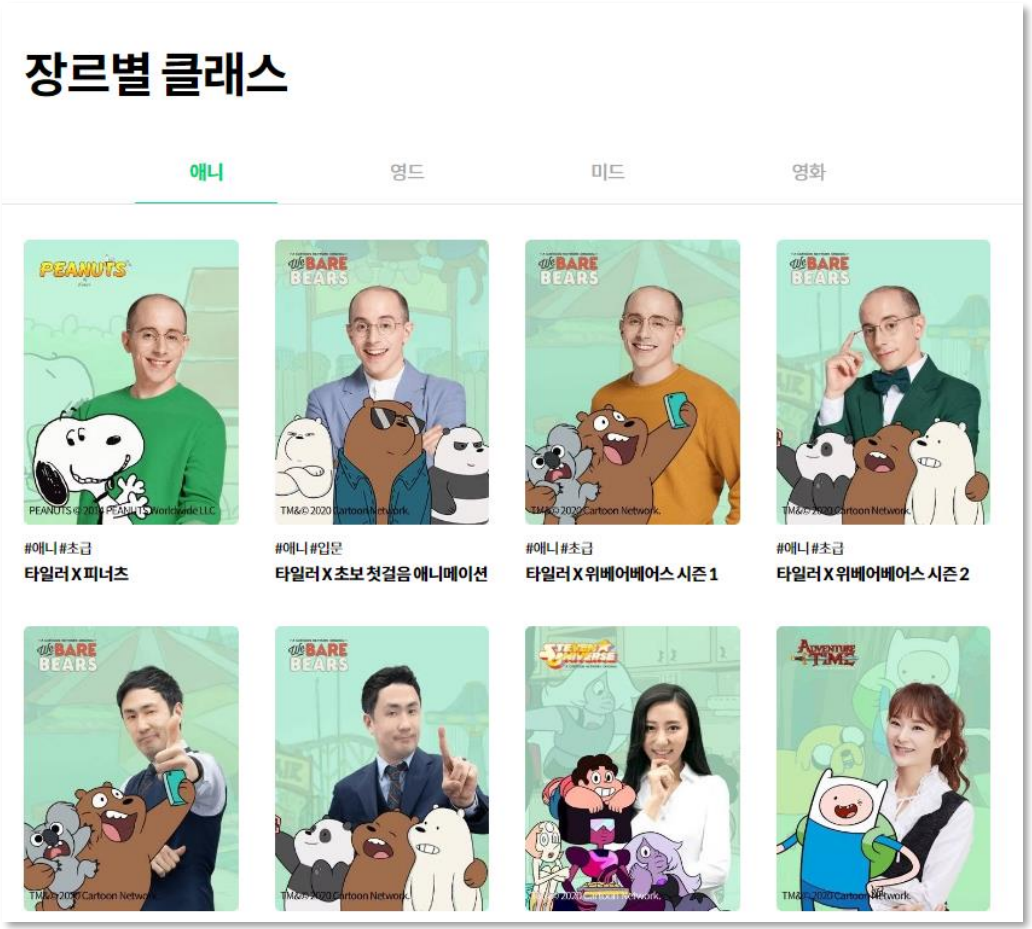
모델 활용 코드

```
# 메인코드
import random
import time
from PIL import Image
import matplotlib.pyplot as plt
from google.colab import files
import re
import requests
from bs4 import BeautifulSoup
import glob


class Koeic():
  def __init__(self, test_text, practice_text):
    print("저희 홈페이지에 접속해주셔서 감사합니다.")
    self.test_text = test_text
    self.practice_text = practice_text
    print("당신의 이름을 적으세요")
    name = input("")
    self.name = name
    print("반갑습니다.",name,'님')

  def start_page(self):
    user_point = 0
    while True:
        print("₩n원하는 작업을 선택하십시오.")
        print("1. 시험보기")
        print("2. 학습하기")
```

```
# 홈페이지 실행
Koeic([sample1,sample1],sample).start_page()
```

```
저희 홈페이지에 접속해주셔서 감사합니다.
당신의 이름을 적으세요
옥유리
반갑습니다. 옥유리 님

원하는 작업을 선택하십시오.
1. 시험보기
2. 학습하기
3. 질문하기
2

원하는 작업을 선택하세요.
1: 기존문제, 2:본인문제,3:발음대결
3
당신이 발음할 대사입니다.
앞 집 팥죽은 붉은 팥 풋팥죽이고, 뒷집 콩죽은 햇콩단콩 콩죽,
우리집 깨죽은 검은깨 깨죽인데 사람들은 햇콩 단콩 콩죽 깨죽 죽먹기를 싫어하더라.

10초 동안 발음하십시오.
끝
분석중입니다.

당신의 점수 입니다.
0.48768592
현재 랭킹은?

1 위:  김동완, 점수: 0.9871688
2 위:  정슬기, 점수: 0.50805974
3 위:  옥유리, 점수: 0.48768592
```

홈페이지 제작

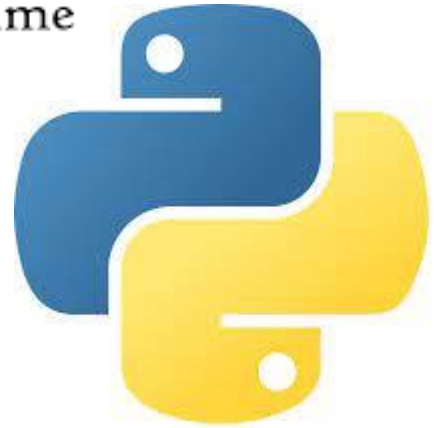


다른 학습 플랫폼처럼
사용자가 이용할 수 있도록 홈페이지 제작

홈페이지 제작 과정 1

HTML, CSS , JS 이용해서
홈페이지 생성

홈페이지 제작 과정 2

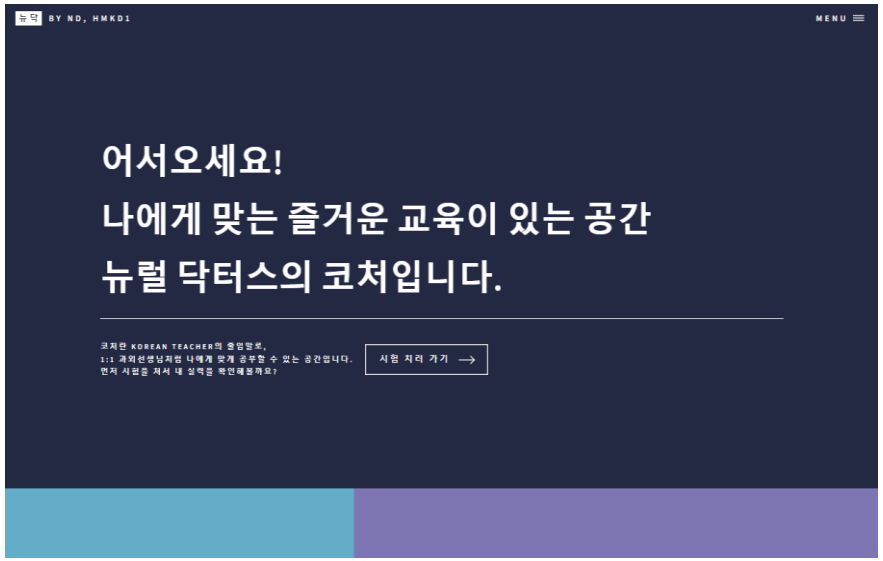


플라스크를 활용하여
파이썬과 HTML 연결

홈페이지 예시

```
[ ] # 홈페이지 실행
    Koeic([sample1,sample1],sample1).start_page()
```

```
저희 홈페이지에 접속해주셔서 감사합니다.

원하는 작업을 선택하십시오.
1. 시험보기
2. 학습하기
3. 질문하기
2

원하는 작업을 선택하세요.
1: 기존문제, 2:본인문제
1
풀고 싶은 문제를 고르세요. 1.K-POP, 2.K-DRAMA,3.뉴스,4.논문
3
```



시연 및 정리

- 기대효과
- 개선점
- 시연

기대효과

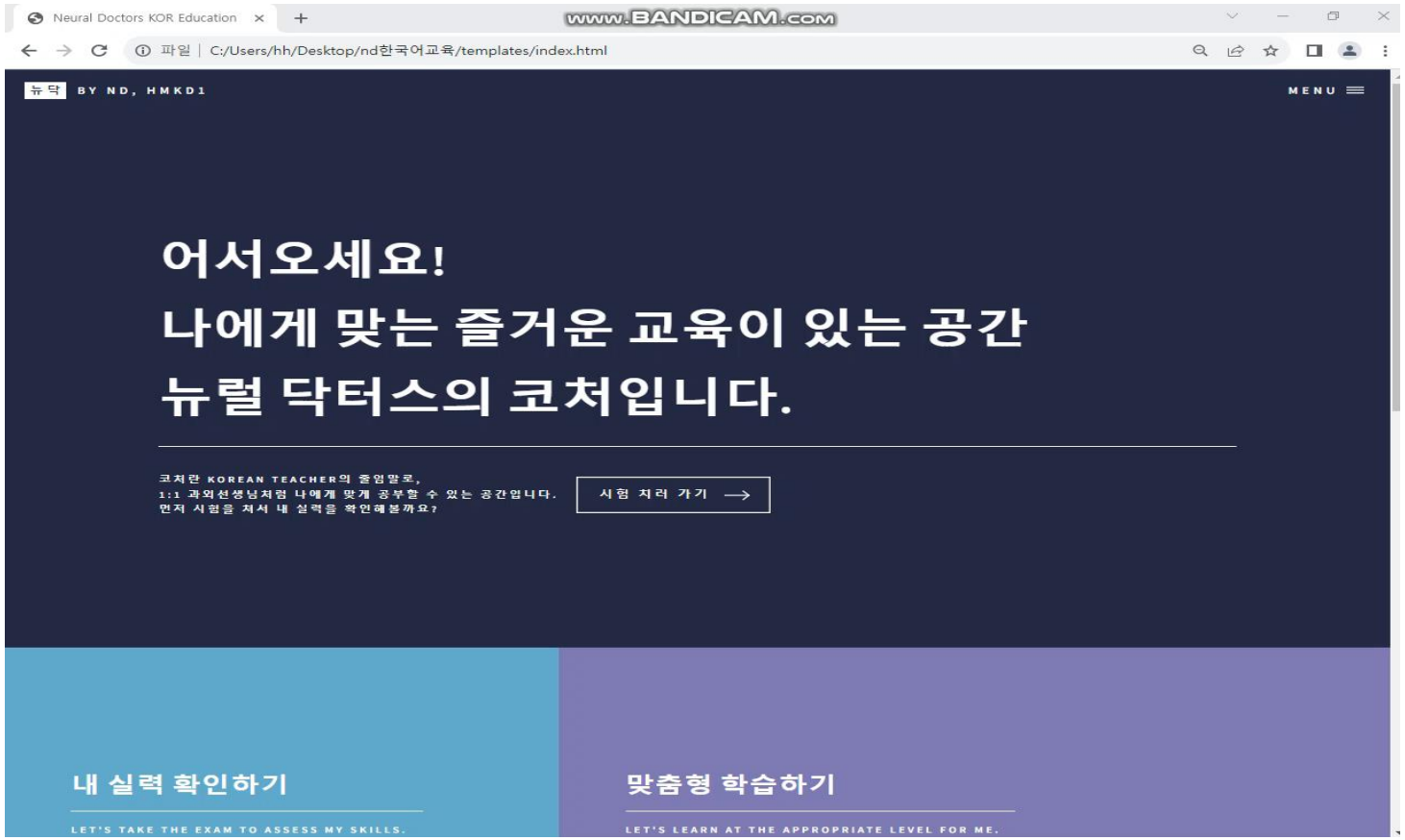
효과 1	외국인에게 개인 맞춤형 한국어 교육을 제공할 수 있다.
효과 2	노래 및 드라마를 활용한 쉬운 접근성으로 학습 능률을 올린다.
효과 3	AI기술을 활용한 서비스 제공 시장에 접근한다.

개선점

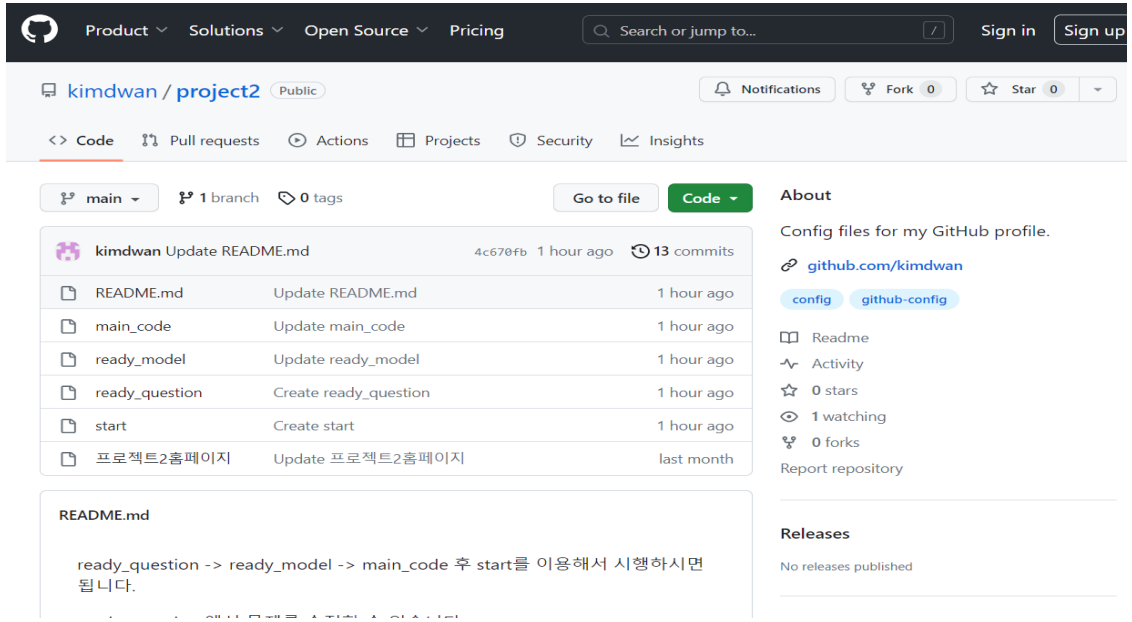
1. 말뭉치 데이터 부족

2. 플라스크 활용

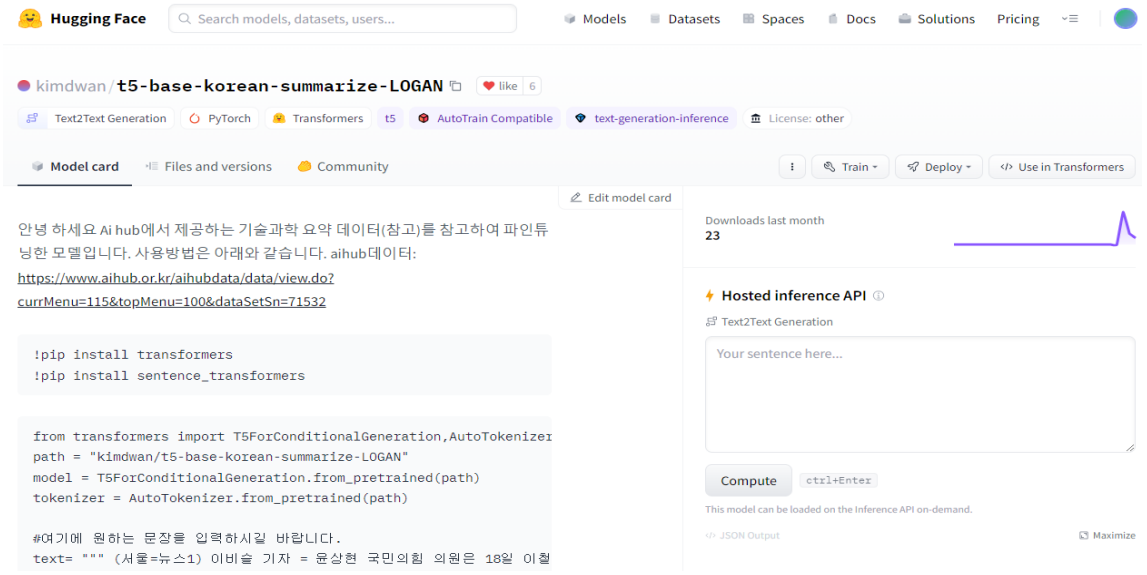
시연영상


Neural Doctors KOR Education
www.BANDICAM.com
파일 | C:/Users/hh/Desktop/nd한국어교육/templates/index.html
뉴닥 BY ND, HMKD1
MENU
어서오세요!
나에게 맞는 즐거운 교육이 있는 공간
뉴럴 닥터스의 코처입니다.
코처란 KOREAN TEACHER의 줄임말로,
1:1 과외선생님처럼 나에게 맞게 공부할 수 있는 공간입니다.
먼저 시험을 쳐서 내 실력을 확인해볼까요?
시험 치러 가기
내 실력 확인하기
LET'S TAKE THE EXAM TO ASSESS MY SKILLS.
맞춤형 학습하기
LET'S LEARN AT THE APPROPRIATE LEVEL FOR ME.

주소



코드 주소: https://github.com/kimdwan/project2



문장 요약 모델 주소: https://huggingface.co/kimdwan/t5-base-korean-summarize-LOGAN
질의 응답 모델 주소: https://huggingface.co/kimdwan/klue-roberta-finetuned-korquad-LOGAN
문장 생성 모델 주소: https://huggingface.co/kimdwan/polyglot-ko-1.3b-Logan

Q&A

Github

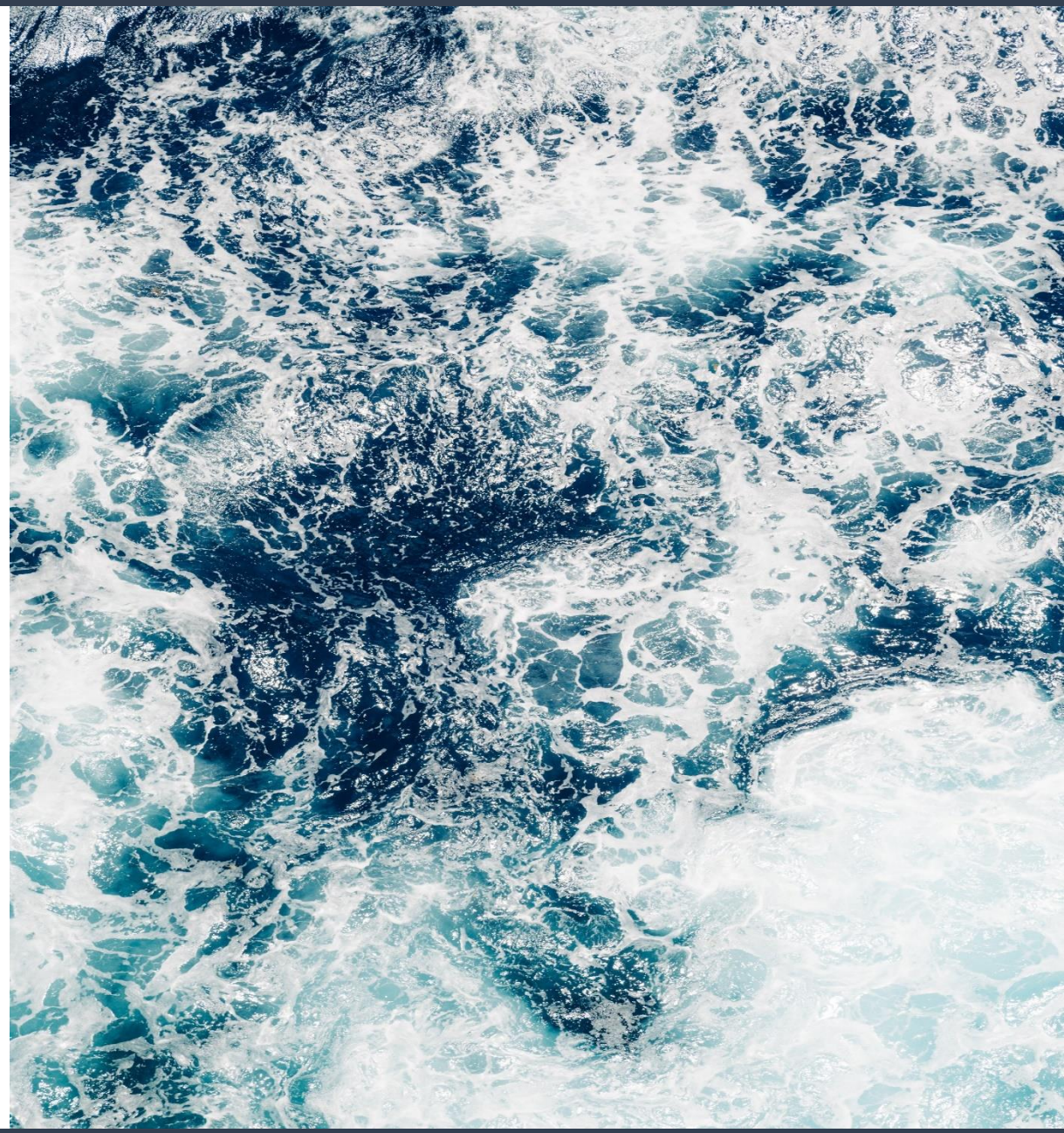
김동완 https://github.com/kimdwan

오윤택 https://github.com/ohyunteak

임수현 https://github.com/fortis001

정슬기 https://github.com/wjdtmfrl

조차선 ---



감사합니다 :)

팀 명 : 뉴럴닥터스

김동완 오윤택 임수현 정슬기 조차선