

"신청서 작성하는 방법 좀 알려주세요. "



공공민원·보건의료

BERT기반 챗봇 프로젝트



다양한 산업에서 적용가능한 챗봇
자연어처리(NLP) 분야 화두인 BERT 활용

contents

목차



chapter 01

프로젝트 개요

-주제 선정 이유
-프로젝트 목표 및 주요기능 소개

chapter 02

기술스택

- 자연어처리(NLP)
- BERT

chapter 03

프로젝트 과정

-데이터수집>모델링>챗봇구현>웹페이지배포

chapter 04

UX/UI

-사용자 친화적 인터페이스 환경

chapter 05

발전과제

-사업 확장성

chapter 06

후기

프로젝트를 마치며..

chapter 01

프로젝트 개요



chapter 01

프로젝트 개요



✓ 프로젝트 기간

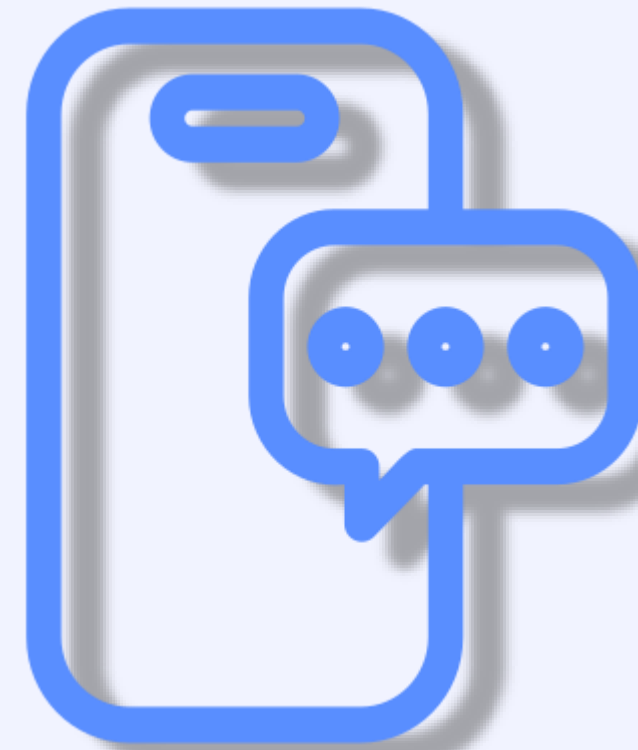
2024.09.23 ~ 2024.10.21

✓ 참여 인원

김민정(조장), 김명일, 김한나, 김현민, 신다연,
이지한

✓ 주요기능

BERT기반 챗봇서비스 ㄴ 모욕어 필터링



chapter 01

담당자 소개



김한나

PM

프로젝트 기획과 관리



김민정

DEV

챗봇구현
API설계 및 관리



이지한

DEV

웹사이트 구축

chapter 01

담당자 소개



김명일
ENGINEER

챗봇 모델 개발



신다연
ENGINEER

챗봇 모델 개발

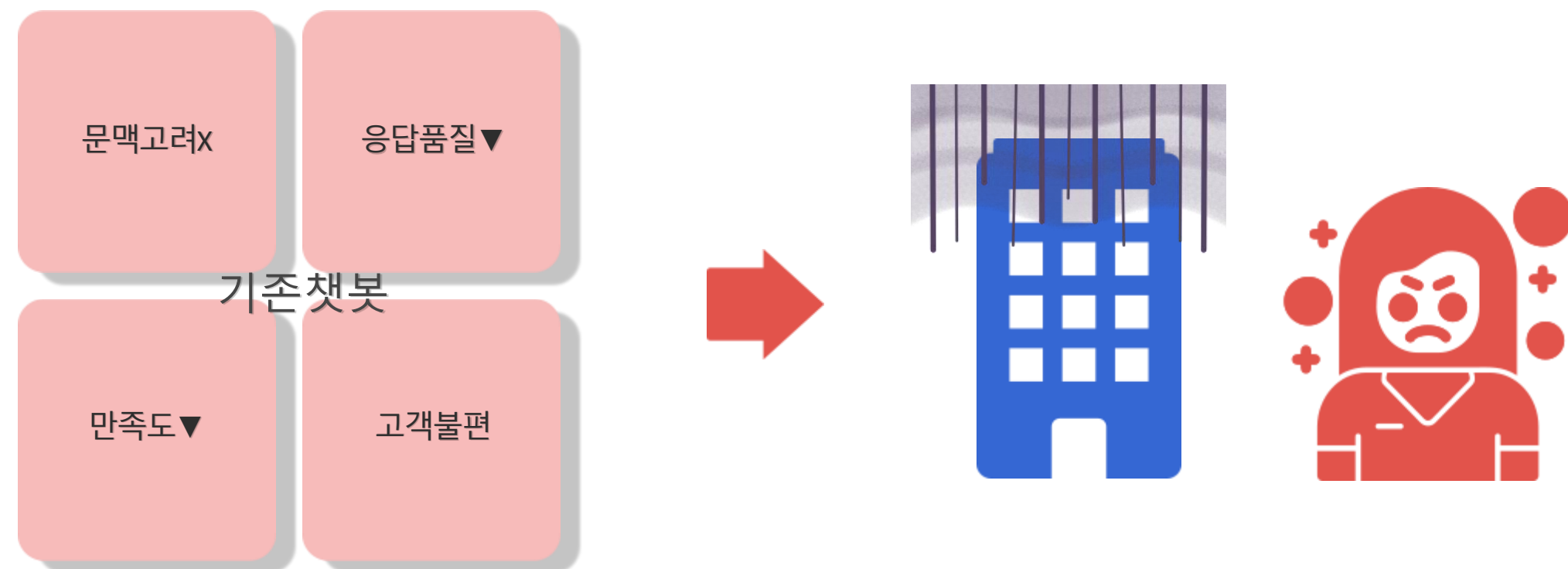


BATCAT
ENGINEER

클라우드 서버
구축 및 배포

chapter 01

주제 선정 이유



> 문맥을 파악하고 자연스러운 응답을 반환하는 BERT기반 챗봇 구현

chapter 02

기술스택



chapter 02

기술스택



DB



MySQL

모델링



BERT
양방향 자연어처리 모델

웹프레임



django

django
확장성 / API개발 용이

chapter 03

프로젝트 과정



chapter 03

프로젝트 일정



구분	1주차	2주차	3주차	4주차	5주차
주제선정	프로젝트 기획 / 주제선정				
DB설계		데이터베이스 설계			
데이터수집		데이터수집 · 전처리			
모델링			모델개발 및 학습		
챗봇구현		챗봇 설계 및 개발 / 테스트 검증			
웹배포					사이트 배포
					프로젝트발표

chapter 03

01) 기획



화면 ID	ch-01	화면명	메인 화면	작성자	김현민
PAGE	3	페이지 정보		작성일	2024-09-26
		<p>화면 설명</p> <p>1. 채팅창을 모바일 그림으로 표현하기는 했으나, 필요이나 다른 작은 창으로도 대체 가능하다. 고객이 상담사와 상담하다가 비속어를 썼을 경우 또는 챗봇 창이다.</p> <ul style="list-style-type: none">- 처음 뜨는 화면의 선택 메뉴는 ai hub 페이지에서 받은 메뉴를 기준으로 적기는 하였으나 구현 가능한 구문으로 대체 가능하다.- 5개의 메뉴 중 하나를 선택하면 하위 메뉴가 뜨고 그에 따른 로직이 수행된다. <p>2. 왼쪽 모바일은 고객 화면이고 오른쪽 모바일은 상담사 화면이다. 비속어의 가림 처리를 위하여 분리해서 보여준다.</p>			
		Page 3			

사용자분석

적용할 산업 및 사용자 분석

기능정의

✔ 질문에 적절한 답변 반환 - BERT

✔ 모욕어 필터링

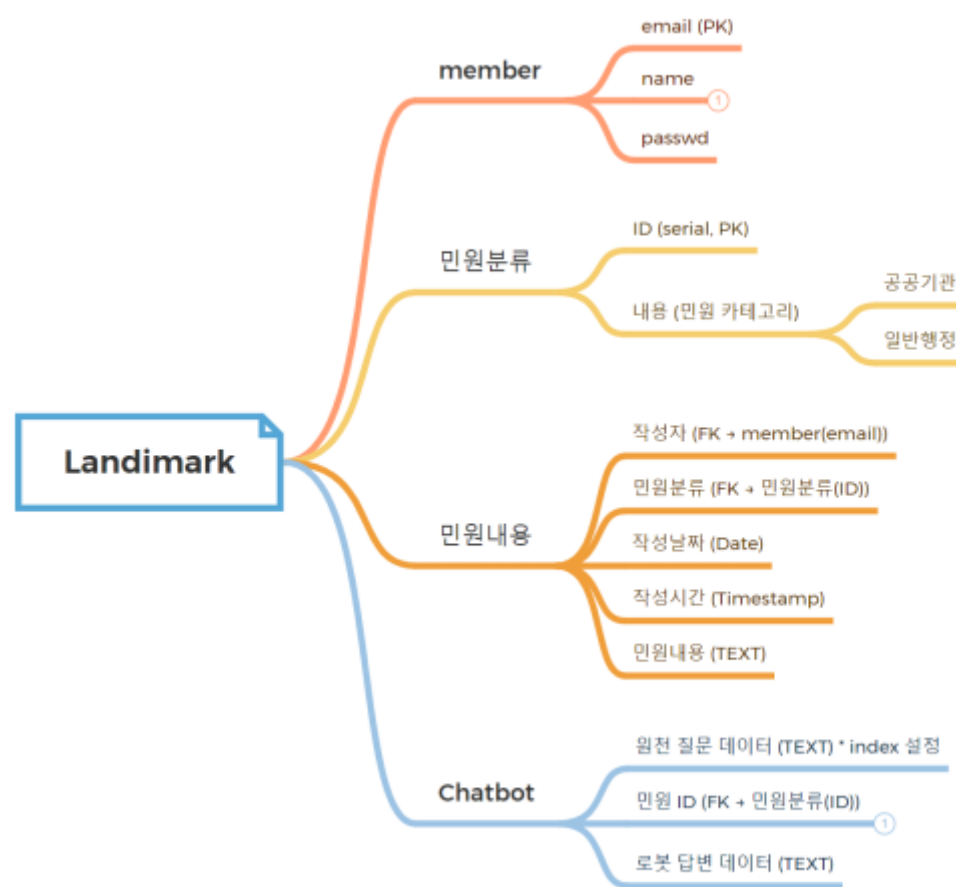
기술 스택 결정

자연어 처리(NLP), 딥러닝, 챗봇 플랫폼 서버 및 데이터베이스

> 다양한 산업으로 확장이 용이한 챗봇 프로젝트 기획

chapter 03

02) 데이터베이스 설계 - 주요기능



chapter 03

03) 데이터 수집/전처리



공공기관

- ① 입장 및 민원서비스 이용안내
- ② 민원별 번호표 발급 및
창구번호 안내
- ③ 준비서류 확인
- ④ 민원신청서 작성안내
- ⑤ 민원 신청서 용어 질문
- ⑥ 관내 복지/지원금 정보 질문

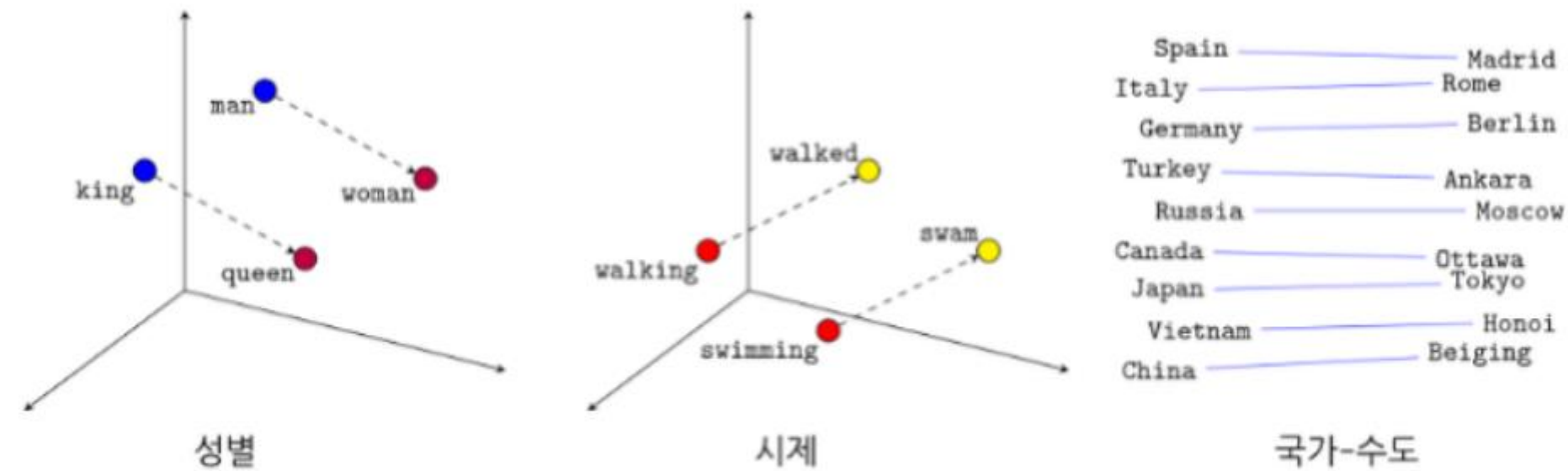
건강/의료

- ① 진료접수 안내
- ② 증상 상담
- ③ 대기실 및 진료실 위치안내
- ④ 약국 위치 문의
- ⑤ 수납 방법 안내
- ⑥ 증빙서류 발급
- ⑦ 다음 진료일 예약

> 전체 30만건 中 공공민원 3.8만건, 건강/의료 2.7만건 사용

chapter 03

임베딩이란?



이미지 출처 : <https://datascienceschool.net/view-notebook/6927b0906f884a67b0da9310d3a581ee/> 데이터사이언스스쿨

- > 단어를 공간상의 벡터로 표현
- > 비슷한 의미를 가진 단어들이 서로 가까운 곳에 배치

04) 모델링 - 2차 전처리 / 모욕어 문장 필터링



모욕어모음 : kaggle

[korean-hate-chat-data]

657908	657907	평화주의자님은	0.0
657909	657908	너는 개새끼 고정이야 설화야	1.0
657910	657909	지훈	0.0
657911	657910	후후후후후후후후	0.0
657912	657911	밑에 만들생각은 없름	0.0
657913	657912	경관 is fu cking sexy	0.0
657914	657913	그건 아니다	0.0
657915	657914	가정교사?	0.0
657916	657915	채팅좀 쳐봐주게니	0.0
657917	657916	시원=?누구냐고요	0.0
657918	657917	10674	0.0
657919	657918	..)	0.0
657920	657919	발신자 번호로하고	0.0

토큰화

konlpy패키지의 okt 형태소 분석기
이용불용어처리

BERT학습 > 임베딩값 계산

BertModel.from_pretrained('beomi/KcBERT-
base')

모욕어 문장 필터링

언더샘플링 / 오버샘플링
LogisticRegression 모욕어 문장(1) / 정상문장(0)

chapter 03

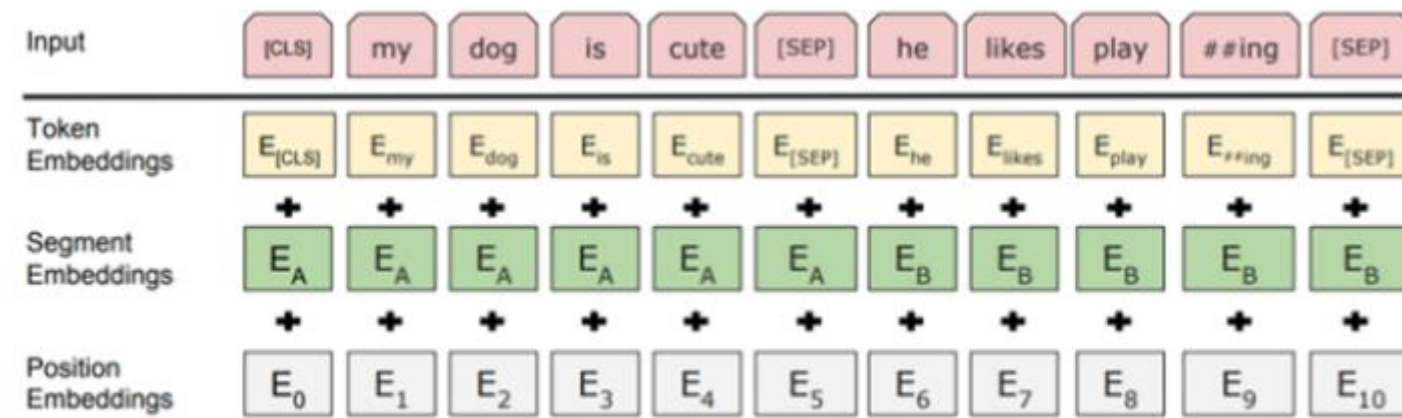
04) 모델링 - 2차 전처리 / 모욕어 문장 필터링



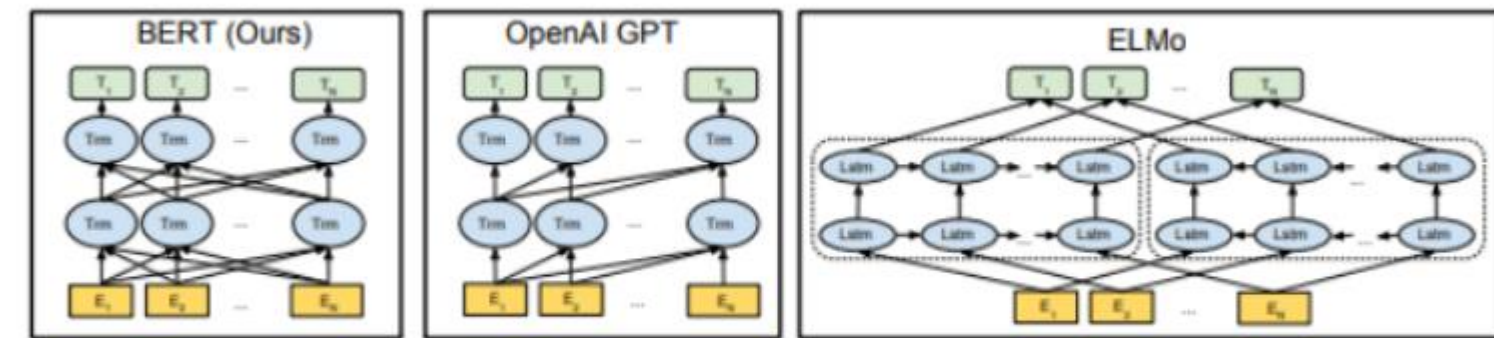
구분	모욕어 필터링
F1 스코어	0.9
ROC-AUC	0.96
테스트 데이터 정확도	0.9

chapter 03

BERT란?



Devlin, Jacob, et al의 figure 2



Devlin, Jacob, et al의 figure3

> 자연어 처리 부분에서 가장 좋은 성능 > 문맥을 양방향으로 해석

chapter 03

BERT란?



" 병원에서 처전방을 받아왔요. 약국은 있나에 어디요? "

chapter 03

BERT란?



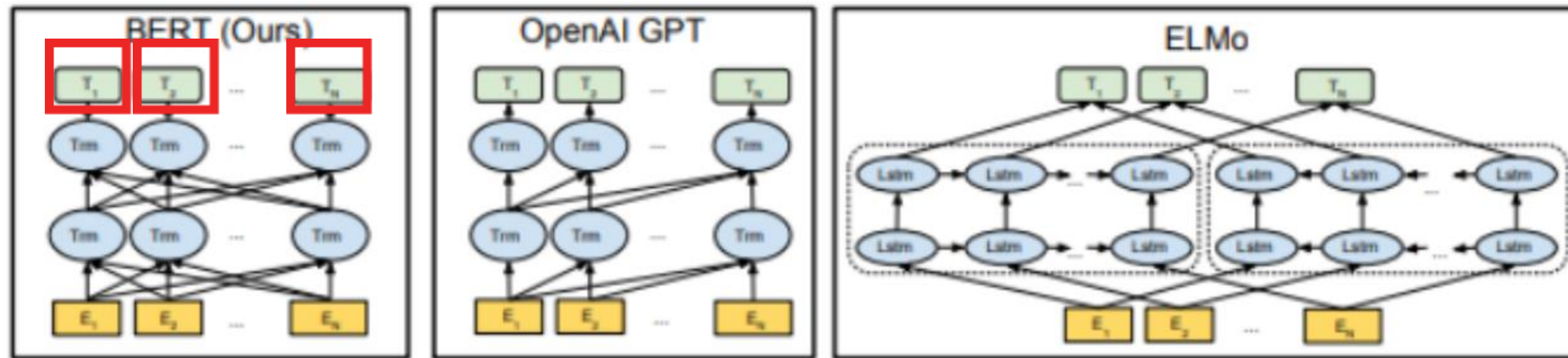
chapter 03

BERT란?



chapter 03

BERT란?



 Devlin, Jacob, et al의 figure3

04) 모델링 - 모델비교



구분	S-BERT	KcBERT-Autotrained	KcBERT-Auto
BERT 임베딩 생성 방식	- paraphrase-multilingual-MiniLM-L12-v2 (Mean Pooling)	- beomi/KcBERT-base (Mean pooling)	- beomi/KcBERT-base (Mean pooling)
답변 유추방식	- 질문 임베딩과 사전 계산된 임베딩 간 코사인 유사도 비교	-1차 라벨 분류 후, 코사인유사도 비교	-질문 임베딩과 사전계산된 임베딩 간 코사인유사도 비교
모델 활용 방식		-질문과 답변의 관계에 높은 가중치 부여 -질문을 주제에 따라 재라벨링	

>가장 높은 정확성을 보인 KcBERT-Autotrained 모델 사용

chapter 03

04) 모델링 - 모델비교



```
# 챗봇 인터페이스
while True:
    user_input = input("사용자 입력 (종료하려면 'exit' 입력): ")
    if user_input.lower() == 'exit':
        break

    response, mrr, precision_k, map_score = get_best_response(user_input, dfp, dfp_embeddings,
                                                             print(f"대답: {response}")

    # 평가지표가 None이 아닐 경우에만 출력
    if mrr is not None and precision_k is not None and map_score is not None:
        print(f"MRR: {mrr:.4f}, Precision@3: {precision_k:.4f}, MAP: {map_score:.4f}")

✓ 4m 24.8s Python
```

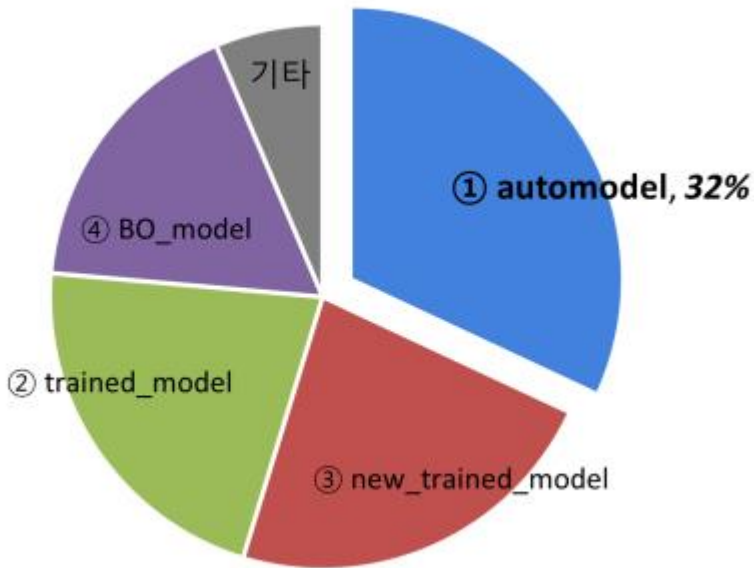
대답: 어서 오십시오. 원하는 서비스를 말씀해 주시기 바랍니다.
MRR: 0.0011, Precision@3: 0.0000, MAP: 0.0011
대답: 번호표 발급 후 저를 따라오세요. 세무 1과 창구로 안내해 드리겠습니다.
MRR: 0.5000, Precision@3: 1.0000, MAP: 0.5000
대답: 발급 창구로 안내해 드리겠습니다.
MRR: 1.0000, Precision@3: 1.0000, MAP: 1.0000
대답: 교육비는 무통장 입금, 신용카드도 가능하며 납부 가능 여부 알아보겠습니다.
MRR: 1.0000, Precision@3: 1.0000, MAP: 1.0000
대답: 범죄의 재발을 방지하고 사회의 원활한 복귀를 위해 특별 교육을 실시하고 기초 생활 가정에 지원금을 제공
MRR: 0.5000, Precision@3: 1.0000, MAP: 0.5000
대답: 세대주를 포함하여 세대 일부가 이주신고 하는 경우에만 작성해 주시면 됩니다.
MRR: 1.0000, Precision@3: 1.0000, MAP: 1.0000
대답: 신청서와 개인정보 동의서, 직업 동의서, 농어업이행 계약서와 농업경영체 등록 확인서가 필요합니다.
MRR: 0.0001, Precision@3: 0.0000, MAP: 0.0001
대답: 확인서 보여주시면 안내 도와드리겠습니다.
MRR: 1.0000, Precision@3: 1.0000, MAP: 1.0000
대답: 그 질문에 대해 더 자세히 설명해 주실 수 있나요?
대답: 어서 오십시오. 원하는 서비스를 말씀해 주시기 바랍니다.
MRR: 0.0009, Precision@3: 0.0000, MAP: 0.0009

MRR (Mean Reciprocal Rank)	사용자 입력에 대해 정답 문장이 얼마나 높은 순위로 반환되었는지 평가
Precision@k	상위 k개의 검색 결과 중에서 정답 문장이 포함된 비율 평가
MAP (Mean Average Precision)	여러 쿼리에서 검색된 상위 문장들이 얼마나 정확하게 정렬되었는지 평가

04) 모델링 - 모델비교



input_question	답변	한나	답변	명일	답변	민정
화상연고도 처방이 되나요?	저희 병원에서 사용되는 의약품은 아	1	화상은 보이는 것	4	화상은 보이는 것보다 심	4
오늘 두시 삼십분에 예약했어요. 언제 들어가나요?	접수처는 이곳 한 곳뿐입니다. 대기번	4	어서 오십시오. 성	1	어서 오십시오. 성함과 인	1
예약한 시간보다 조금 늦었어요. 다시 접수해야 하나요?	앞선 예약이 없어서 가능합니다. 예약	2	앞선 예약이 없어	2	대기하시기 전에 접수증	1
피가 계속 나는데, 저 죽는 걸까요?	긴급상황이라면 응급실로 이동하셔서	1	긴급상황이라면 :	1	긴급상황이라면 응급실로	1
접수 재등록해도 되나요?	안녕하세요 생년월일 말씀해 주시면 !	4	안녕하세요 생년	4	안녕하세요 생년월일 말	4
접수 시간이 지났는데 혹시 대기순번이 넘어갔나요?	확인을 도와드리겠습니다. 합의된 상	2	확인을 도와드리	2	확인을 도와드리겠습니다	2
접수가 왜 이렇게 오래 걸려요?	고객님 기다리게 해서 죄송합니다. 빠	3	고객님 기다리게	3	앞서 이용해 주시는 손님	1
다음에 들어가는 순번인데, 잠깐 화장실 다녀와도 되나요?	네 알겠습니다. 다녀오시면 앞에 보이	1	네 알겠습니다. 드	1	다녀오신 후 대기석에 있	2
접수하려고요.	가능합니다. 현재 진료 보시는 분 존	4	네 처음 오셨으면	3	네 처음 오셨으면 종이에	3
진료 접수 어떻게 하죠?	창구에서 성함 확인 후 치료 예약 도	4	창구에서 성함 확	4	바로 도와드리겠습니다.	3
오래된 상처 흉터도 없어질까요?	미용 목적의 수술의 경우 진단서 발급	1	미용 목적의 수술	1	잠시만 기다려 주십시오.	3
여기 한의원도 같이 하나요? 침을 맞고 싶어요.	1층 입구에서 왼쪽으로 가시면 진료 ?	3	1층 입구에서 왼	3	1층 입구에서 왼쪽으로 :	3
목 디스크 때문에 잠을 잘 수 없어요. 어떤 치료를 받을 수 있나요?	움직이지 못하는 만큼 근육의 힘이 약	4	움직이지 못하는	4	진료는 진료실에서 전문	1
탈모 초기 인 것 같아요. 머리카락이 자꾸 빠지는데, 이것도 진료를 볼 수 있나요?	검사 접수 도와드리겠습니다. 지금 드	2	검사 접수 도와드	2	검사 접수 도와드리겠습니다	2
저는 아니고, 제 아이가 열이 자주 오르다 말다 해요. 소아과도 진료 하나요?	평범한 감기일 수도 있으니까 너무 걱	2	평범한 감기일 수	1	평범한 감기일 수도 있	2



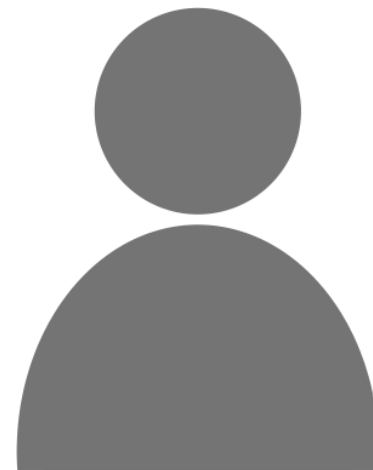
> 가장 높은 정확성을 보인 KcBERT-Auto모델 사용

chapter 03

04) 모델링 - 챗봇



"다음 예약은 이번주 금요일로 잡아주세요."



input question 임베딩값 계산

tokenizer로 토큰화 > 임베딩값 계산

같은 카테고리 임베딩 추출

`AutoModelForSequenceClassification.from_pretrained("./fine_tuned_model")`

임베딩값 계산

`max : find_most_similar_answer_cosine()`

관련 답변 반환

`best_answer` 반환

chapter 03

04) 모델링 - 질문:응답 반환



사용자 질문

"화장실이 어딘가요?"



로봇답변

" 제 가 안 내 해
드리겠습니다. 저를 따라
이 동 해 주 시 기
바랍니다."

chapter 03

05) 챗봇구현



```
# 사전 계산된 질문 임베딩 및 csv 파일 로드
question_embeddings = torch.load(question_embeddings_path)
df = pd.read_csv(csv_file_path)
answers = df['answer'].tolist()

class ChatbotConsumer(AsyncWebsocketConsumer):
    async def connect(self):
        await self.accept() # 웹소켓 연결 수락

    async def disconnect(self, close_code):
        pass # 연결 해제 처리

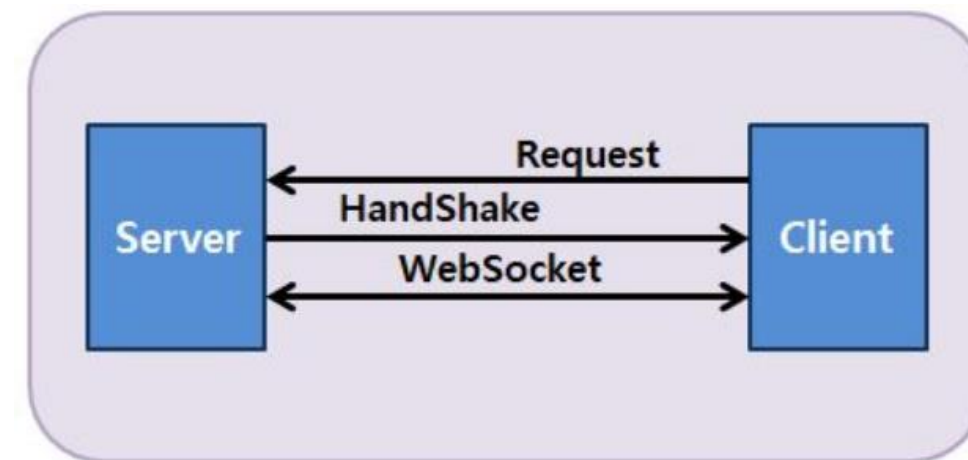
    async def receive(self, text_data):
        data = json.loads(text_data)
        input_question = data['question'] # 사용자의 질문

        # 모욕어 감지 비동기 처리
        label = await insult_detection(input_question)

        if label == 1:
            # 모욕어 감지된 경우 경고 메시지 전송 (is_offensive 추가)
            await self.send(text_data=json.dumps({
                'answer': '모욕어가 감지되었습니다. 다시 입력해 주세요.',
                'is_offensive': True # 이 값을 추가하여 프론트엔드에서 빨간색 표시하도록
            }))
            return

        # 유사 답변 찾기 비동기 처리
        best_answer, cosine_similarity = await asyncio.get_event_loop().run_in_executor(
            None, find_most_similar_answer, input_question, question_embeddings, answers
        )

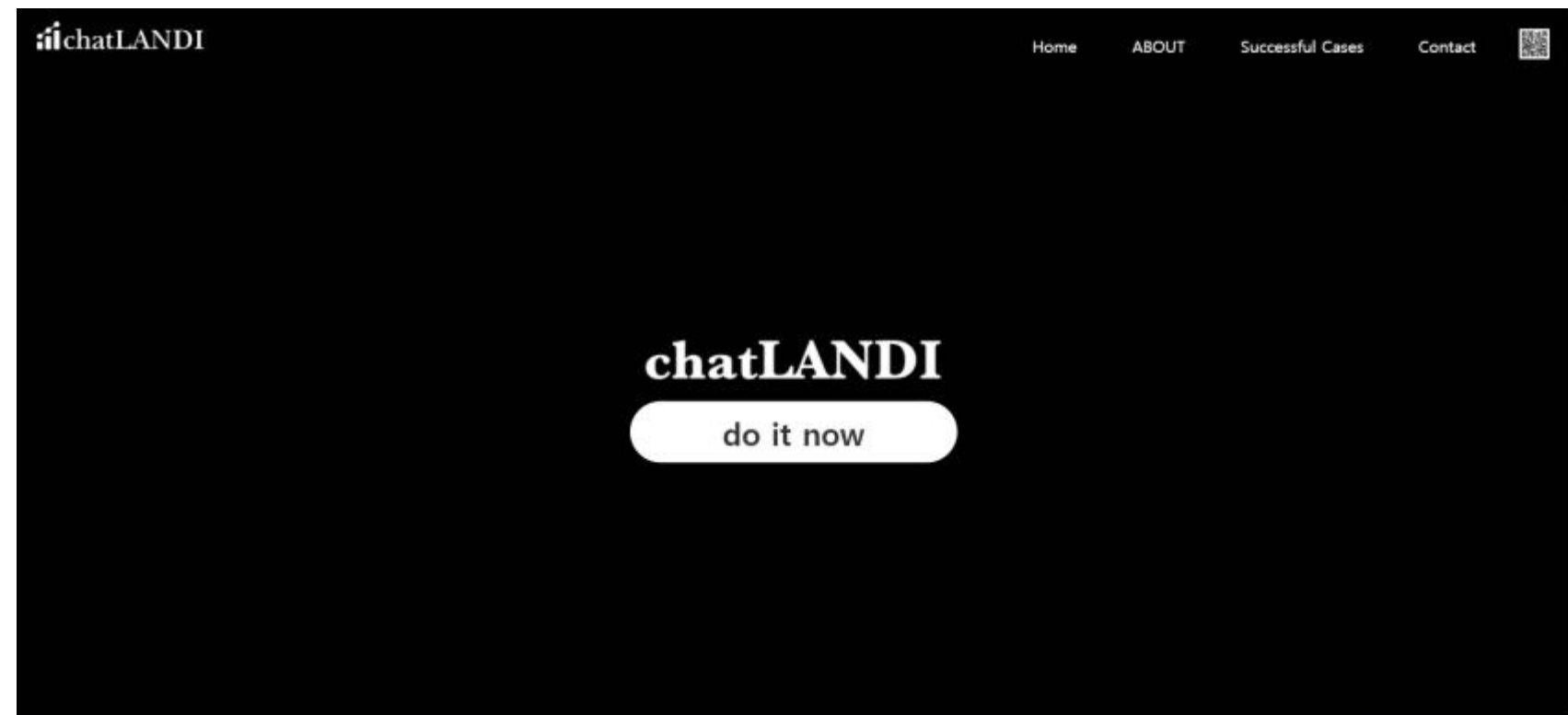
        # 답변 전송 (모욕어가 아닌 경우 is_offensive는 False로 설정)
        await self.send(text_data=json.dumps({
            'answer': best_answer,
            'similarity': cosine_similarity,
            'is_offensive': False # 모욕어가 아닐 때는 False로 설정
        })))
```



> 비동기 실시간 화면처리 구현

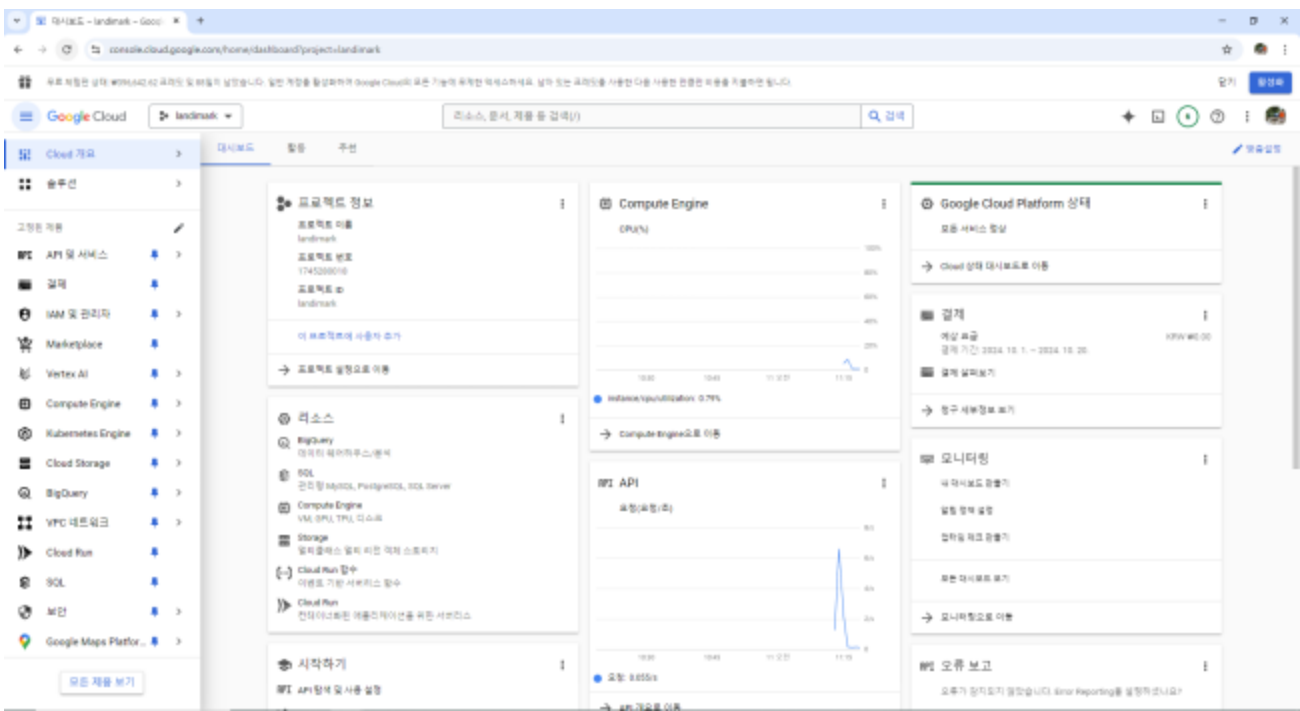
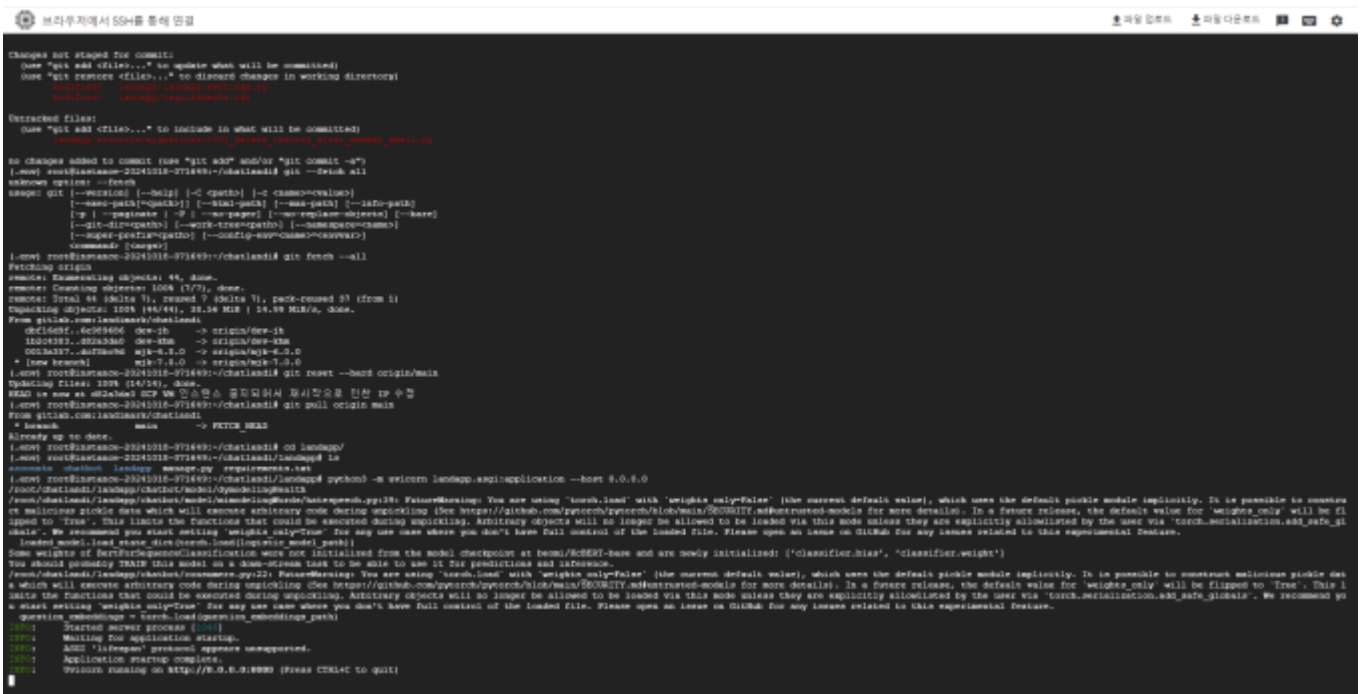
chapter 03

06) 웹사이트 구축



> 확장성과 편리함, 시각적으로 매력적인 UX/UI

07) 웹 배포



> 구글클라우드플랫폼(GCP) 이용하여 배포

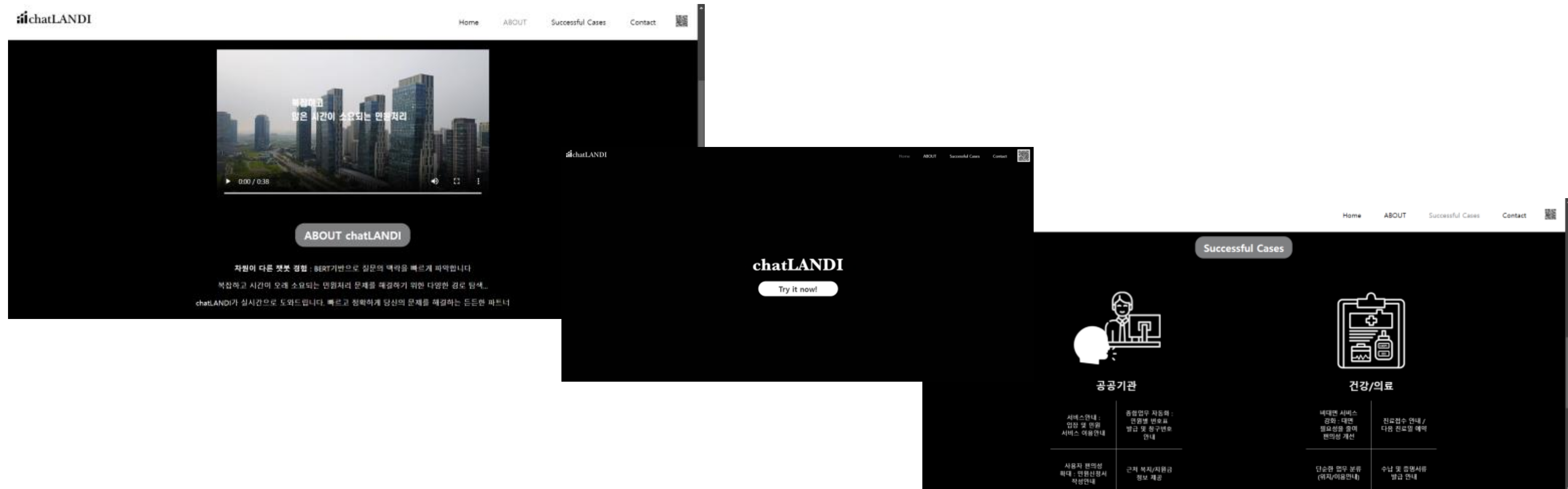
chapter 04

UX/UI



chapter 03

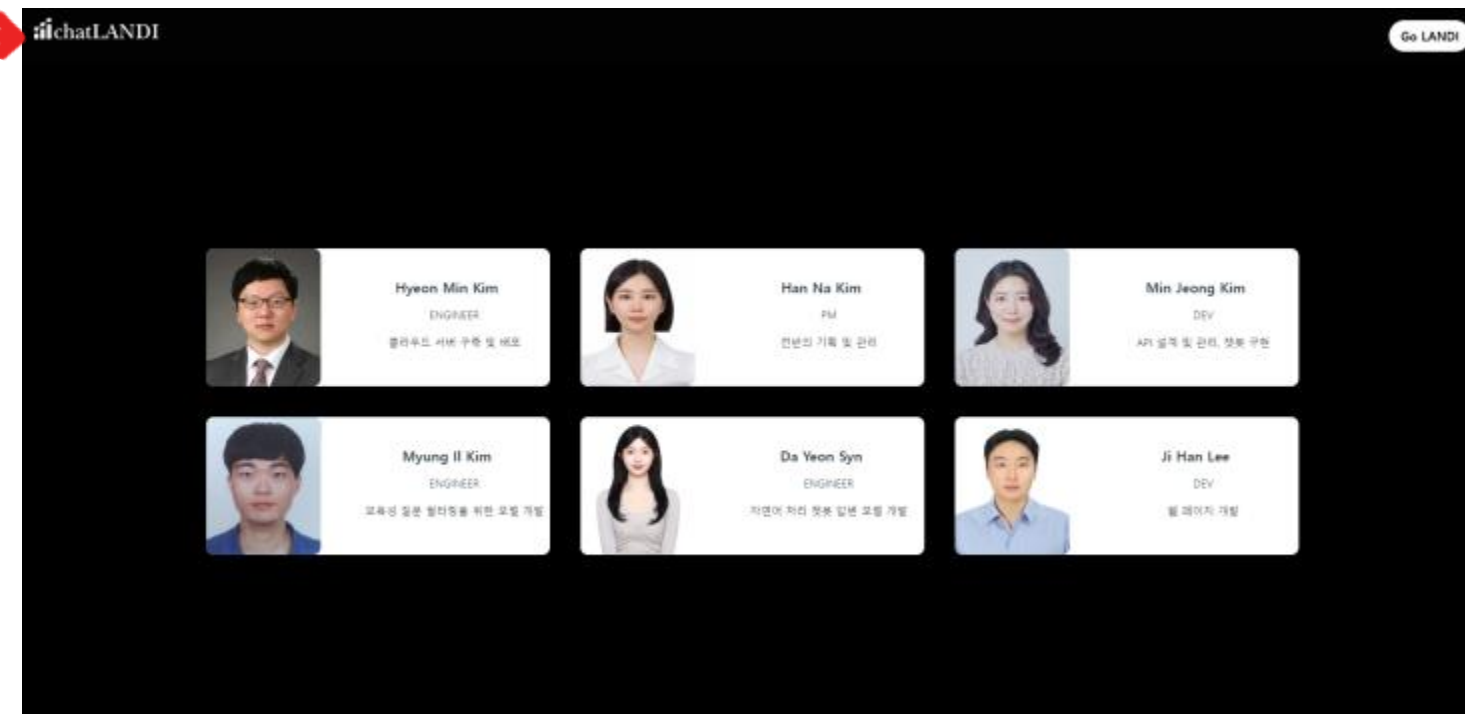
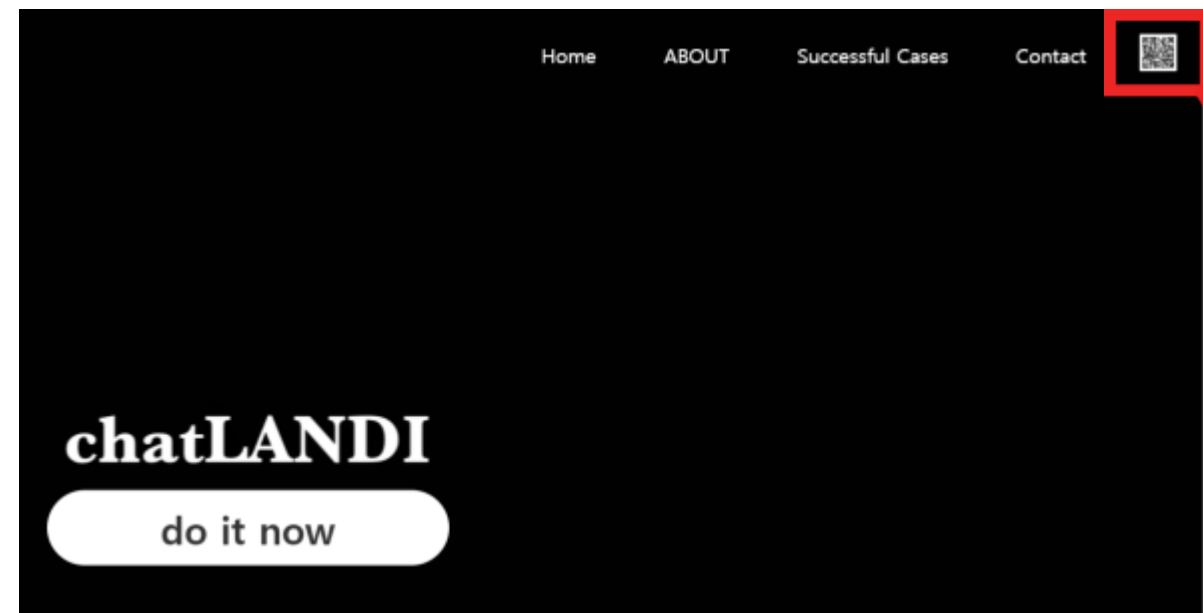
01) 메인페이지



> 사이트의 핵심기능과 목적을 자연스럽게 노출

chapter 03

01) 메인페이지 - QR개발자소개



> QR로 개발자소개페이지 접근 : 모바일 중심의 사용자 경험 강화

챗봇 시연



① "도대체 얼마나 기다려야돼요 짜증나네."

② "약국은 여기서 멀어요?"

③ "예약 시간 알 수 있는 방법있음?"

④ "접수비 영수증 어디서 받아요?"

⑤ "현금 밖에 없는데 현금 결제 되나요?"

chapter 03

01) 챗봇페이지

**채팅
사이드바(Conversation
Sidebar)**

- 대화 히스토리 저장
- 이전 질문 내용 클릭 시
답변 히스토리 확인 가능

chatLANDI

- 병원 서류가 필요한데 어디서 발급 받아요?
서류 발급은 안내 데스크에서 확인 부탁드립니다.
- 병신아
- 따로 예약할 수 있는 방법이 있나요?
- 약국은 여기서 멀어요?

병원 서류가 필요한데 어디서 발급 받아요?

서류 발급은 안내 데스크에서 확인 부탁드립니다.

병신아

모욕어가 감지되었습니다. 다시 입력해 주세요.

따로 예약할 수 있는 방법이 있나요?

병원 홈페이지 진료 예약 메뉴에 들어가시면 진료 예약이 가능합니다.

약국은 여기서 멀어요?

건물 밖으로 나가시면 바로 보이는 버스 정류장 옆에 약국이 있습니다.

궁금한 사항을 입력해 주세요.

**메인 콘텐츠(Main
Content)**

- 메시지 입력창
- 이전메세지 내용

**메시지
입력창(Message Input
Field)**

- 메시지 입력필드

chapter 04

사용자 친화적 UX/UI



반응형 웹사이트

데스크탑, 태블릿, 모바일 등
다양한 기기에서 편리하게 접근

QR개발자소개

복잡한 절차 없이 바로 개발자 소개
페이지로 이동

UX/UI

시각적 일관성 / 접근성

통일감 있는 디자인 요소
클릭하거나 터치할 때, 즉각적인 시각적 피드백

데모 영상/로고
디자인

직관적이고 시각적으로 매력적인 UI를 구축
플랫폼의 목적과 방향 제시

chapter 05

발전과제



chapter 05

발전과제



AS - IS [현재 상황]

모욕어필터링 - 단순 챗봇

공공민원/건강의료 주제



TO - BE [이상적 지향점]

설명가능한AI(XAI)

다양한 산업으로 확장

> 다양한 산업으로 확장가능한 플랫폼 구축

chapter 06

후기



chapter 06

후기



PM
Product Manager



이름 김한나 | 관리자

팀원들의 아이디어를 조율하고 전체적인 흐름을 잡는 과정에서 큰 보람을 느꼈습니다. 새로운 기술을 시도하면서 유연한 사고와 팀워크의 중요성을 다시 한번 깨달았고, 팀원들의 의견을 반영해 함께 목표를 이루어가는 경험이 매우 값졌습니다.

DEV
Backend-Developer



이름 김민정 | 개발자

이번 프로젝트에서 다양한 새로운 시도를 할 수 있어 좋았습니다. 특히 평소 관심 있던 데이터베이스와 백엔드를 파이썬으로 경험해 볼 수 있어 재미있게 참여했습니다. 팀원들의 열정적인 참여 자세에서 많은 것을 배웠습니다.

DEV
Frontend Developer



이름 이지한 | 개발자

프론트엔드를 집중적으로 맡았는데 배운 거에 비해 더 많은 기능과 작업들을 해야해서 어려움이 많았지만 그 기능들을 구현함으로써 프론트 엔트에 대해 더 많은 배움과 흥미를 느끼고 한층 더 성장한 것 같습니다

chapter 06

후기



ENGINEER

Machine Learning Engineer



이름 김명일 | 엔지니어

모델 설정과 최적화 과정에서 원하는 결과를 얻지 못해 반복 수정이 쉽지 않았지만, 그만큼 달성했을 때의 성취감이 컸습니다. 팀원들의 헌신적인 모습이 큰 동력이 되어 끝까지 집중할 수 있었습니다.

ENGINEER

Machine Learning Engineer



이름 신다연 | 엔지니어

처음 모델을 설계할 때부터 다양한 기술과 알고리즘을 적용하며 많은 시행착오를 겪었습니다. 이 과정에서 데이터의 중요성을 다시 한번 깨달았고, 어떻게 하면 더 정확하고 효율적인 모델을 만들 수 있을지 깊이 고민하게 되었습니다.

ENGINEER

System Engineer



이름 김현민 | 엔지니어

팀 내에서 활용 가능한 리소스를 최대한 활용하여 프로젝트 목표를 달성할 수 있는 방향으로 진행했습니다. 하둡 (Hadoop) 과 카프카 (Kafka) 는 팀원들이 만든 소스와 통합하기가 적합하지 않을 것 같아 적용하지 않았고, 클라우드 서버 구축 및 배포 작업에 집중하였습니다.

감사합니다.