개발의민족(1조)

수집 데이터 기반 BERT 질의응답

김규희(조장), 김수연, 김지혜

목차

추진목표

□ 정답 데이터 정확도 향상을 위해 국가 데이터를 AI(Artificial Intelligence, 인공지능) 언어 처리 모델에 적용하기 위한 적정한 데이터 세트 수집 및 분석

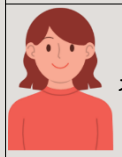
□ 수집한 국가 데이터 기반으로 BERT(Bidirectional Encoder Representations from Transformers, 버트)를 통한 정답 데이터 출력

기대효과

□ 데이터를 정제하여 정보의 부정확성을 해소

□ 쉬운 접근성으로 원하는 답을 찾을 수 있도록 생산성을 높임

팀원 구성 및 역할

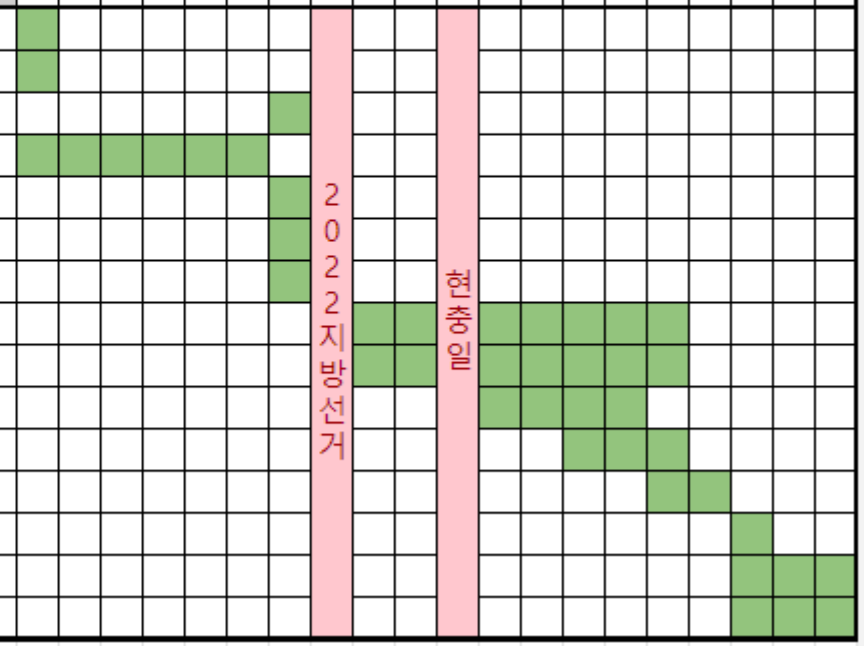
	WBS 관리 및 작성	프로젝트B 데이터 정제 (전처리)	질의응답 데이터 작성	데이터 무결성 대조 검사	DB설계	DB 데이터 삽입	UI화면 설계	Butter Block 작동 테스트	Butter Block API 연동	Django 프론트엔드 구현	Django 백엔드 구현	통합 테스트	개발산출물 작성
 김규희	V	V	V	V	V			V	V		V	V	
 김수연	V	V	V	V		V	V	V				V	V
 김지혜	V	V	V	V		V	V	V		V		V	

- WBS(Work Breakdown Structure)
- DB(Database)
- UI(User Interface)
- API(Application Programming Interface)

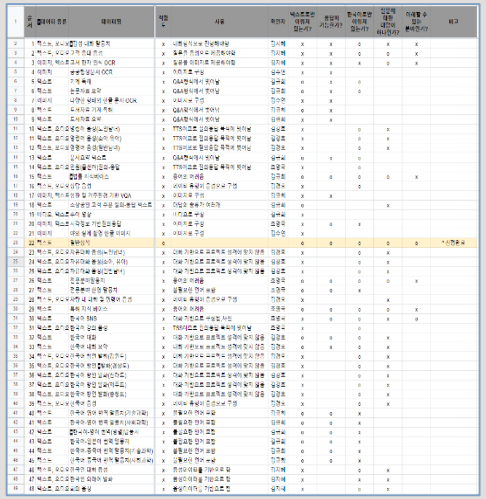
Work Breakdown Structure

1조 프로젝트A Work Breakdown Structure

단계	업무	시작일	예상종료일	종료일	기간	완료도
분석	프로젝트 배경 및 시나리오 이해	2022.05.23	2022.05.25	2022.05.23	1	100%
	개발 환경 구축	2022.05.23	2022.05.25	2022.05.23	1	100%
	요구 사항 정의	2022.05.31	2022.05.27	2022.05.31	1	100%
	Butter Block 연결 테스트	2022.05.23	2022.05.27	2022.05.30	6	100%
설계	시스템 구조 설계	2022.05.31	2022.06.01	2022.05.31	1	100%
	DB 설계	2022.05.31	2022.06.01	2022.05.31	1	100%
	UI 화면 설계	2022.05.31	2022.06.03	2022.05.31	1	100%
구현	프론트 개발	2022.06.02	2022.06.08	2022.06.13	7	100%
	백엔드 개발	2022.06.02	2022.06.10	2022.06.13	7	100%
	데이터 마운팅	2022.06.07	2022.06.10	2022.06.10	4	100%
	API 연계	2022.06.09	2022.06.10	2022.06.14	3	100%
테스트	통합 테스트	2022.06.13	2022.06.13	2022.06.14	2	100%
발표 및 수정보완	프로젝트 발표	2022.06.15	2022.06.15	2022.06.15	1	100%
	수정 보완	2022.06.15	2022.06.17	2022.06.17	3	100%
	과제완료보고서작성	2022.06.15	2022.06.17	2022.06.17	3	100%



프로젝트 도메인 선정

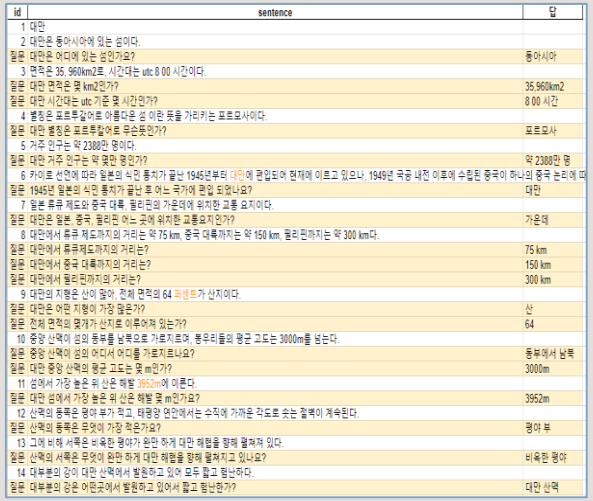


데이터 획득 & 데이터 정제1

데이터 정제2

데이터명	특징	비고
베트남	역사, 베트남어 표기 다수	중 48개의 유효한 문장
중화인민공화국	sub category -> 공산당 , 41번 42번 문장은 이어짐 , 50번째 문장 단위 표현 확인 필요	중 49개의 유효한 문장
타이완	섬의 특징 多, 불필요 문장 다수 식별	중 50개의 문장 중 (34~40번 문장 1문장에 각 섬이름)
인도	역사	중 49개의 유효한 문장
말레이시아	역사, 통계치 다수	중 49개의 유효한 문장
이란	역사(기원전부터)	중 49개의 유효한 문장
인도네시아	나라 특징, 역사, 47번줄 오타 (돌다 -> 돌 다)	중 49개의 유효한 문장
터키	역사, 제1차 세계 대전	중 49개의 유효한 문장
이스라엘	sub catagory -> 이스라엘의 기후 및 역사	중 50개의 유효한 문장
필리핀	역사(미국과의 관계)	중 50개의 유효한 문장
한국	대한민국의 sub category로 들어가는게 맞을수있음	데이터 양이 충분하지 않음
싱가포르	질의 응답 단순	중 48개의 유효한 문장
대한민국	sub category -> 서울, 한국, 조선민주주의인민공화국, 49번 50번째 문장 데이터의 사실성 확인 필요	중 49개의 유효한 문장
러시아	역사, 분쟁 지역	중 49개의 유효한 문장
조선민주주의인민공화국	sub category -> 대한민국	중 50개의 유효한 문장
타이	sub category -> 전쟁과 공산당 관련 질문 가능성 높음	중 49개의 유효한 문장

선정, 규모, 분석

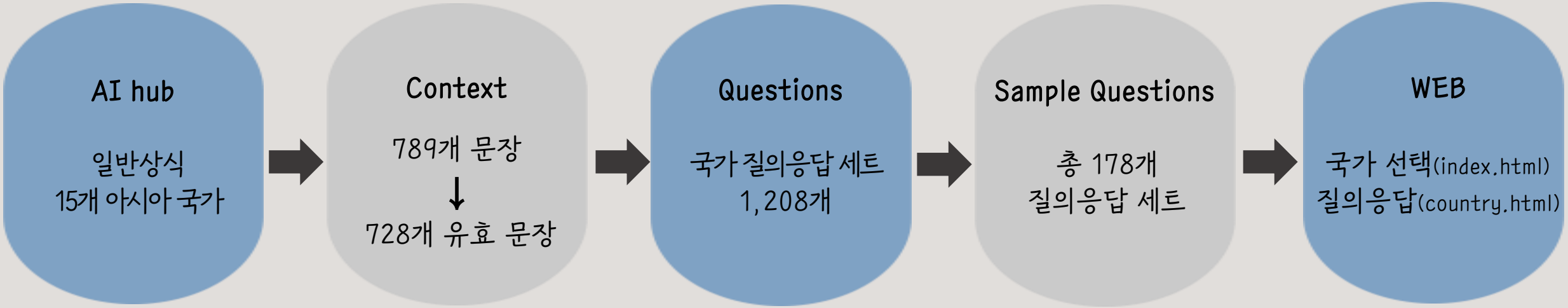


데이터 가공

시각화(word cloud)

데이터 획득	데이터 정제1	데이터 정제2	선정	규모	분석	데이터 가공
✓ AI-hub 학습용 데이터 총 48개 획득	✓ 적합성 판별 (영상, 음성, OCR 제외) ✓ 질의 응답 구조로 구현 불가능한 데이터 제외 ✓ 데이터 적합성 대조 검사 실시	✓ '일반 상식' 데이터 선정 ✓ 데이터내 128개의 파일을 카테고리화	✓ 카테고리 중 국가 도메인 선정, 15개의 아시아 국가 대상 ✓ 총 789개의 문장	✓ 745개의 유효 문장	✓ 데이터 최신화 확인 ✓ JSON 구조 분석 ✓ 분석한 어노테이션 구조로 모듈 구현 : json_to_txt ✓ 분석한 json 파일 문장 추출	✓ 불필요한 문자 이외의 외국어를 제거 ✓ 통계치 포함 모든 수치는 22년 기준 최신화 ✓ 국가별 80-100개 질의응답 세트 구성

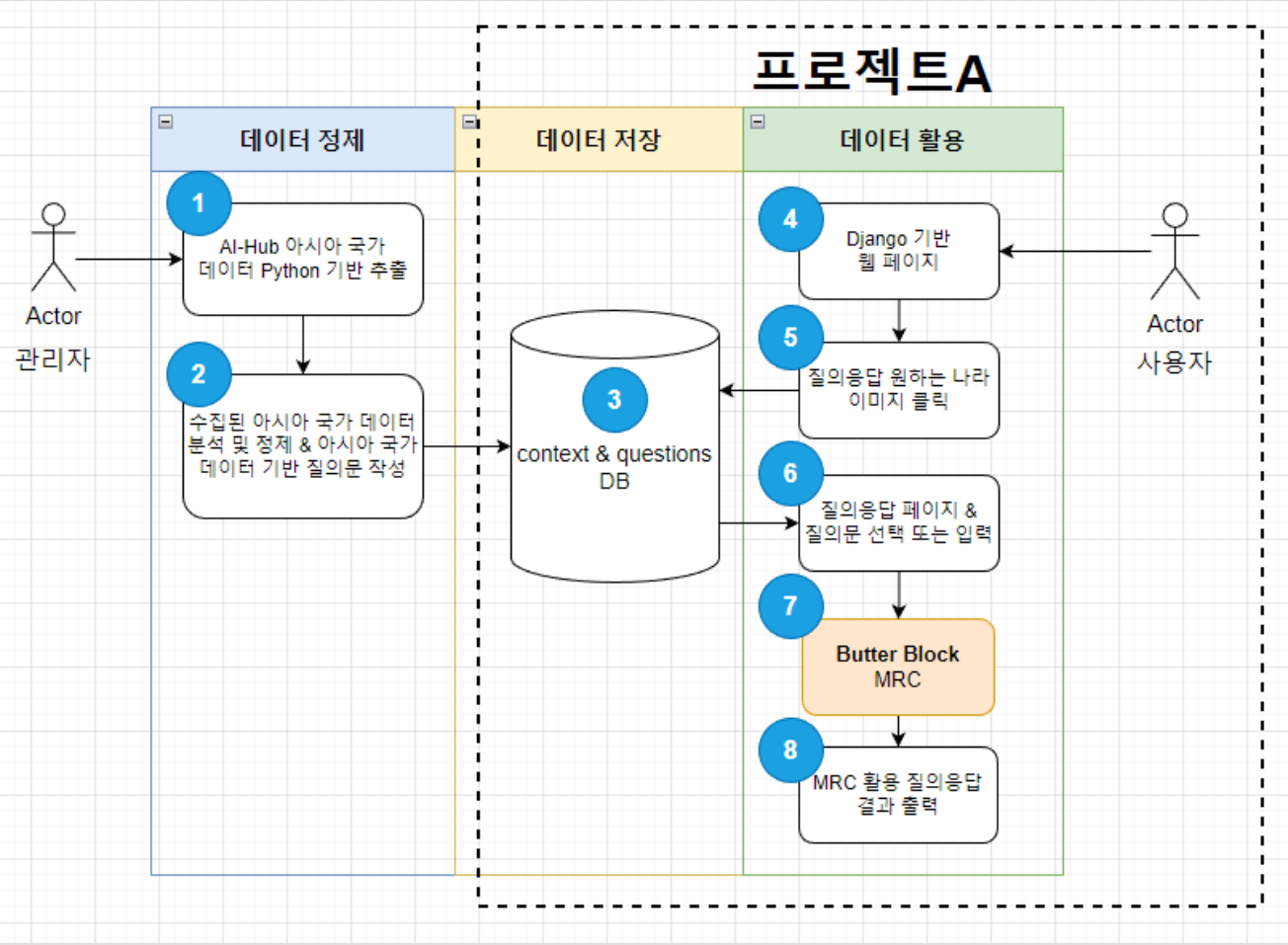
프로젝트 수행 범위


AI hub
일반상식
15개 아시아 국가
Context
789개 문장
↓
728개 유효 문장
Questions
국가 질의응답 세트
1,208개
Sample Questions
총 178개
질의응답 세트
WEB
국가 선택(index.html)
질의응답(country.html)

프로젝트 설계 요구사항 정의

- 국가 정보에 대한 질의응답 시스템 구축
- 프로젝트B의 수집된 데이터를 질의응답 데이터로 활용
- 사용자가 직관적으로 국가 본문 내 빈도수 높은 명사를 파악하기 위한 시각화 활용 – 워드 클라우드(word cloud)
- 사용자가 직접 작성하지 않고 이용할 수 있는 질문 리스트(questions) 필요
- 사용자가 질의응답 할 국가를 선택한 후 DB에서 본문(context)과 질문(questions) 출력 또는 직접 작성
- 웹 UI에서 사용자가 질문 입력 시 Butter Block의 MRC(Machine Reading Comprehension, 기계독해) 답변 수신
- Python 기반 웹 인터페이스에 대한 기본 설계 및 구현 – Platform : Django, DB : MySQL

프로젝트 설계 시스템 구조 정의



데이터 정제(Python)

1. AI-Hub 아시아 국가 15개국 데이터 추출
2. 각 국가의 문장을 파인 튜닝 후 질의응답 데이터 세트 작성

데이터 저장(DB)

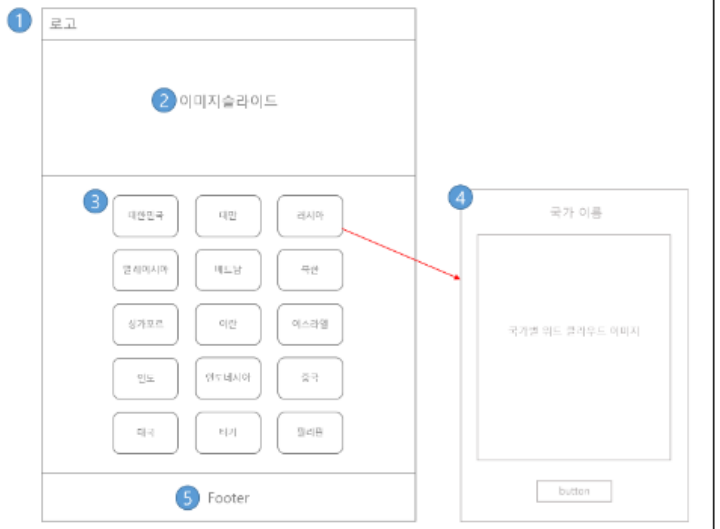
3. context & questions 작성된 MySQL DB

데이터 활용(Django)

4. 사용자가 url 입력하면 웹페이지 출력
5. 웹페이지에서 질의응답을 원하는 국가 이미지 클릭
6. 질의응답 페이지로 이동하고 그 페이지에서 질문을 선택 또는 입력
7. Butter Block의MRC를 활용해 해당 질문에 대한 정답 도출
8. 웹페이지에 입력된 질문과 해당 질문에 대한 정답 출력

프로젝트 설계 화면 정의

INDEX			
화면ID	QA_UI_010	화면명	국가 선택
관련 유스케이스 ID	QA_UCD_010	파일명	index.html
화면 URL	http://127.0.0.1:8000/qa		
화면유형	선택	메뉴 경로	국가 선택
화면개요	15개의 아시아 국가 중에서 질의 원하는 국가를 선택하는 화면		



1. 로고를 클릭 시 index.html로 이동
2. 국가별 대표 이미지 슬라이드 출력
3. 질의응답할 국가 클릭 시 4번 모달창 출력
4. 국가별 워드 클라우드 이미지 출력
   및 button 클릭 시 country.html로 이동
5. Butter Block 제작사 투블럭 AI와 프로젝트 담당 팀 저작권 표시

COUNTRY			
화면ID	QA_UI_020	화면명	질의응답
관련 유스케이스 ID	QA_UCD_020	파일명	country.html
화면 URL	http://127.0.0.1:8000/qa/country/?country=%EB%8C%80%ED%95%9C%EB%AF%BC%EA%B5%AD		
화면유형	질의응답	메뉴 경로	국가 선택 > 질의응답
화면개요	사용자가 본문과 질문을 입력하면 MRC를 사용하여 정답을 알아내는 화면		

1. 로고를 클릭 시 index.html로 이동
2. 국가별 대표 이미지 클릭 시 해당 국가 context 출력
3. 국가 context 출력 또는 사용자가 직접 context 입력
4. 미리 작성된 질문 목록 선택 또는 사용자가 직접 질문 입력 후 답변 찾기 클릭
5. 질문에 해당하는 답변 출력
6. Butter Block 제작사 투블럭 AI와 프로젝트 담당 팀 저작권 표시

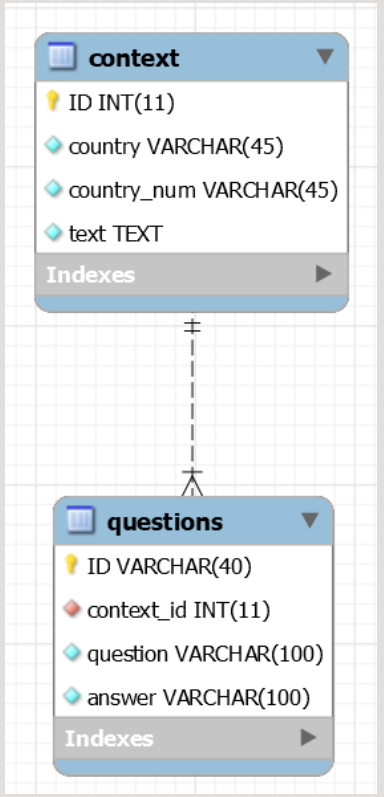
프로젝트 설계 데이터베이스 정의

테이블 정의서

주제명	수집한 국가 데이터 기반으로 BERT를 통한 질의응답					작성일	2022.05.31	작성자	1조 개발의민족	
테이블 한글명	본문					테이블 영문명	context			
테이블 설명	아시아 국가별 정보로 구성된 테이블									
컬럼한글명	컬럼영문명	데이터 타입	길이	NotNull	DEFAULT	AUTO_INCREMENT	PRIMARY KEY	FORIGN KEY	UNIQUE	CHECK
나라_고유번호	ID	INT		Y			Y			
나라_문장번호	country_num	VARCHAR	45	Y						
나라명	country	VARCHAR	45	Y						
내용	text	TEXT		Y						

테이블 정의서

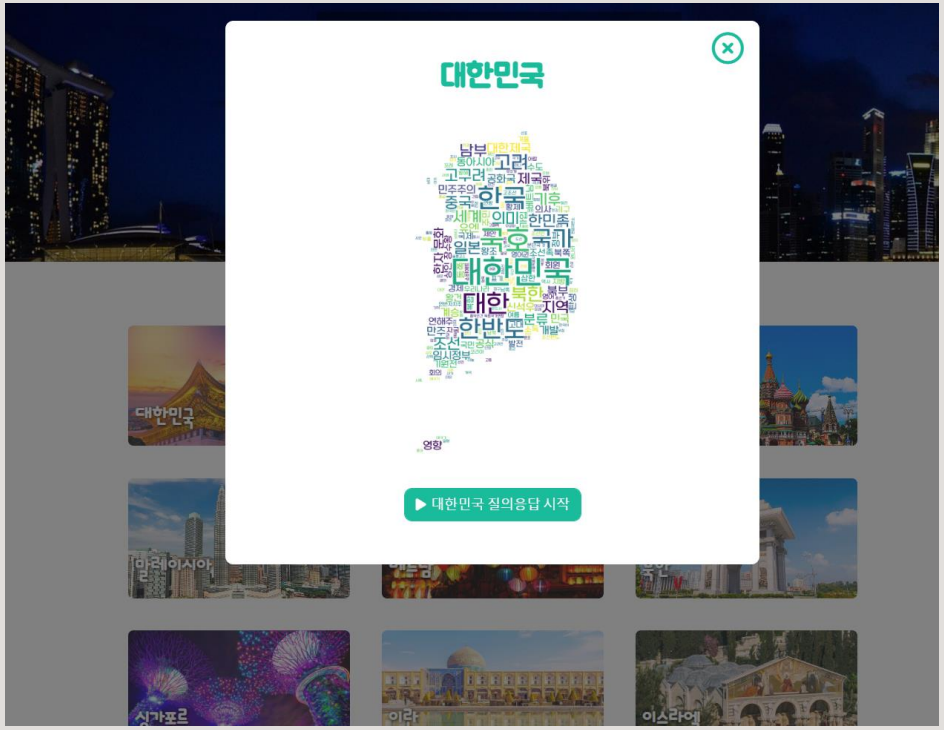
주제명	수집한 국가 데이터 기반으로 BERT를 통한 질의응답					작성일	2022.05.31	작성자	1조 개발의민족	
테이블 한글명	질문					테이블 영문명	questions			
테이블 설명	아시아 국가별 질의응답으로 구성된 테이블									
컬럼한글명	컬럼영문명	데이터 타입	길이	NotNull	DEFAULT	AUTO_INCREMENT	PRIMARY KEY	FORIGN KEY	UNIQUE	CHECK
질문번호	ID	VARCHAR	40	Y			Y			
나라_고유번호	context_id	INT		Y				Y		
질문	question	VARCHAR	100	Y						
답변	answer	VARCHAR	100	Y						



프로젝트 구현

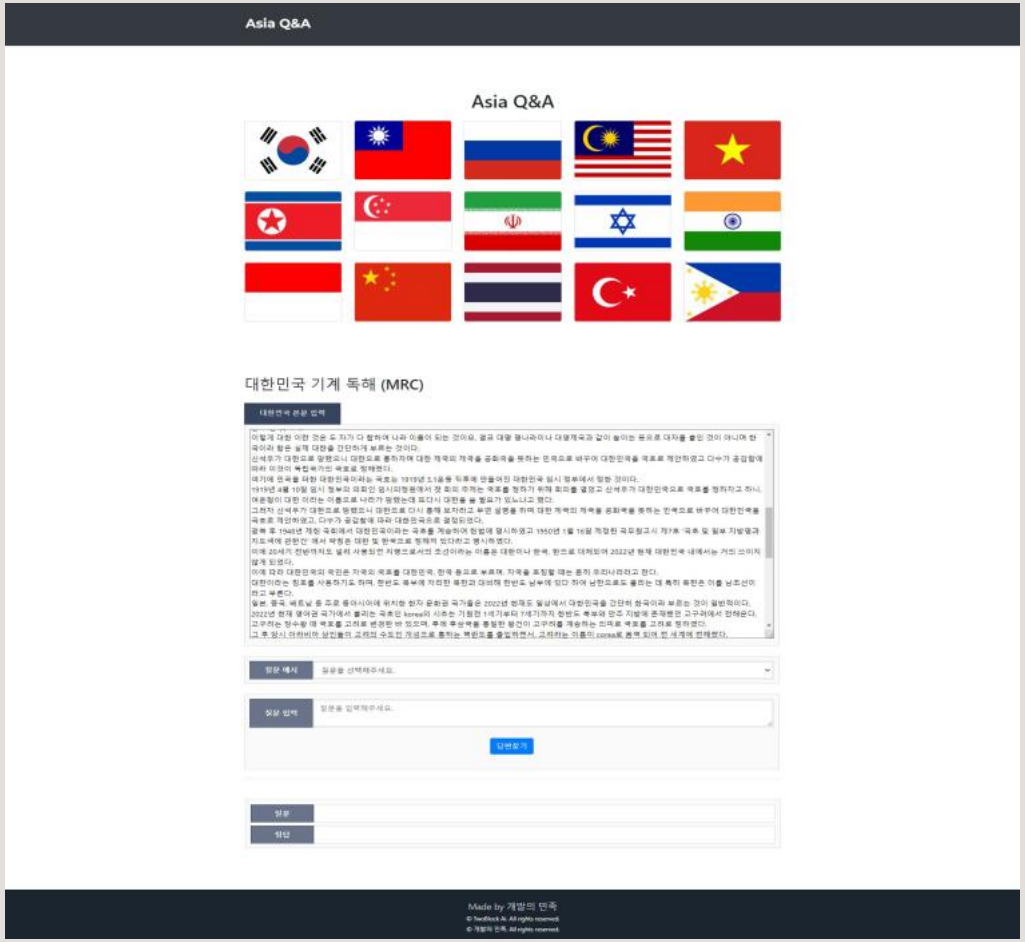
index

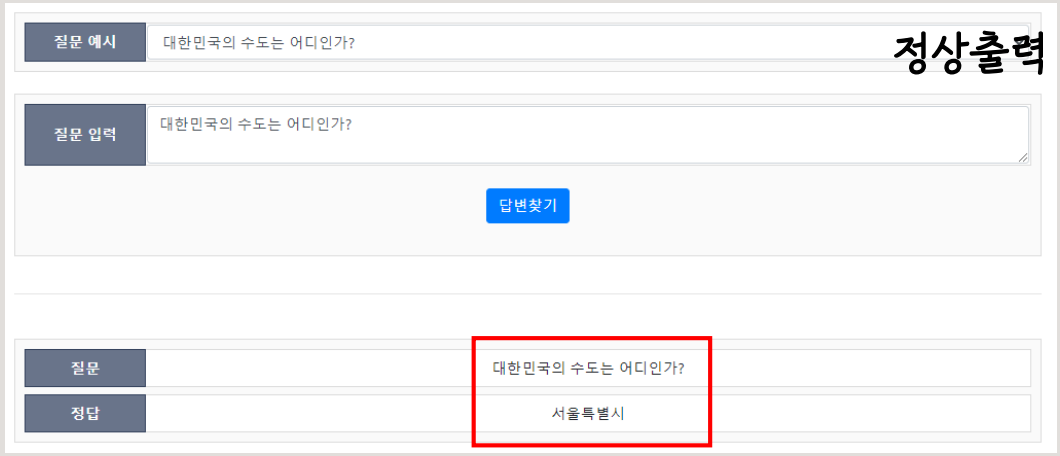
index - modal

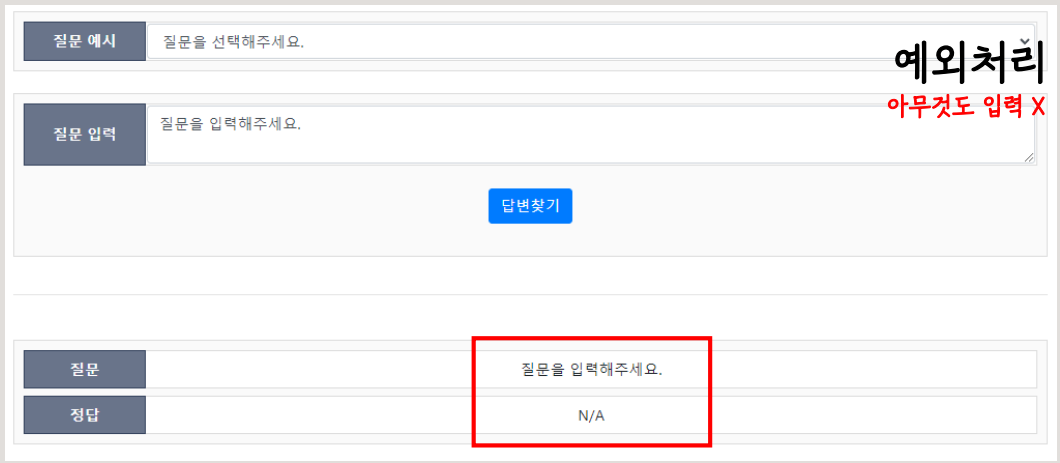


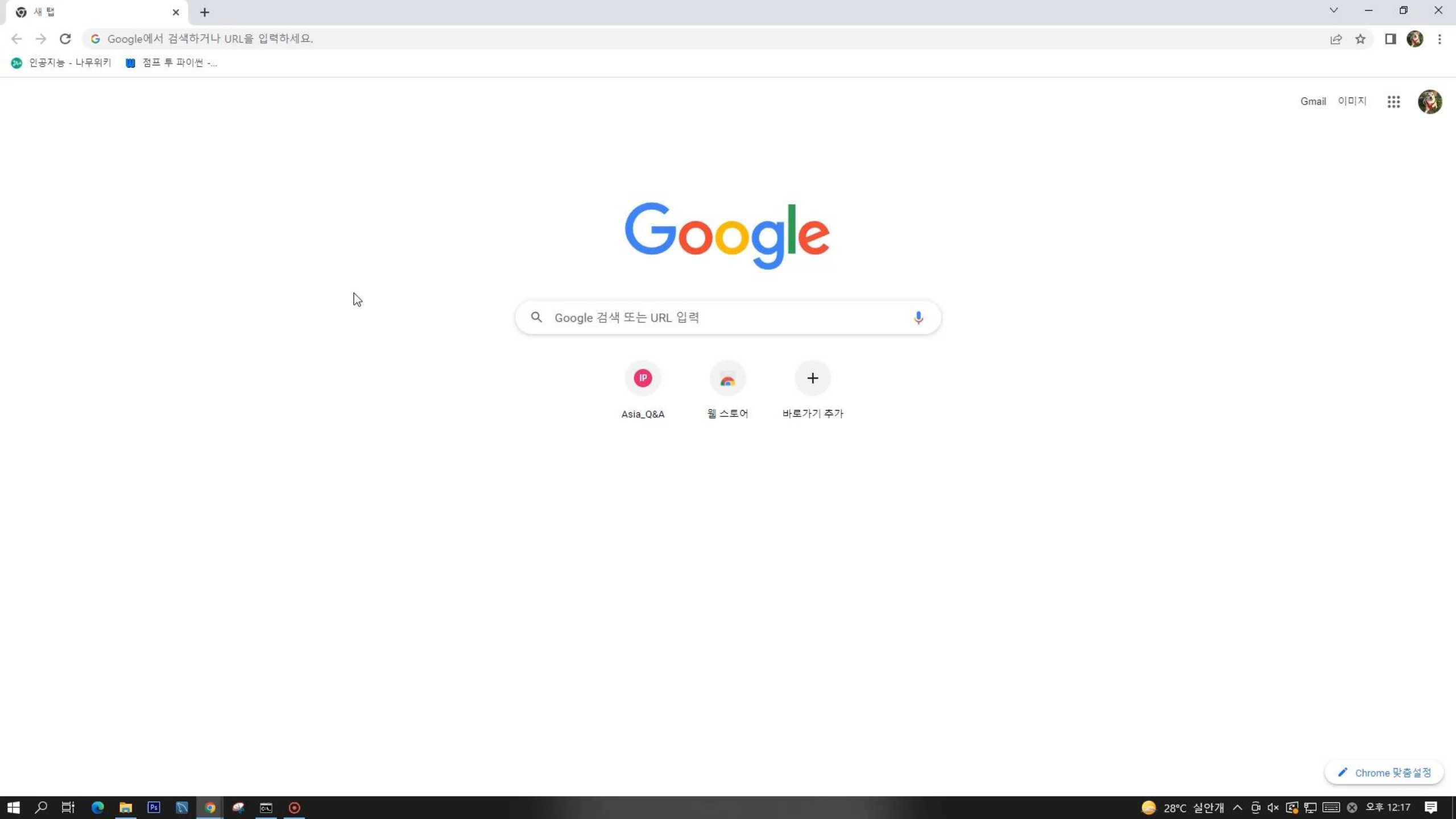
프로젝트 구현

country








새 탭
Google에서 검색하거나 URL을 입력하세요.
인공지능 - 나무위키
점프 투 파이썬 -...
Gmail
이미지
Google
Google 검색 또는 URL 입력
Asia_Q&A
웹 스토어
바로가기 추가
Chrome 맞춤설정
28°C 실안개
오후 12:17

프로젝트 후기

김규희(조장)

데이터를 얻는 과정은 쉬웠지만 수집된 데이터를 자세히 들여다보니 업데이트가 안 된 데이터, 사실과 다른 데이터, 맞춤법과 띄어쓰기가 되어 있지 않은 데이터, 한국말로 적힌 외국어 데이터들이 많았습니다. 프로젝트에 적합하지 않은 데이터들을 butter block이 이해할 수 있게 정제하는 과정이 까다롭고 오랜 시간이 걸린다는 것을 느꼈습니다.
그리고 처음 다뤄본 Django 프레임워크에 대한 전반적인 구조를 이해할 수 있어 여러모로 유익한 경험이었습니다.

김수연

수집한 데이터의 맞춤법, 문법, 시제, 사실확인, 외래어 표기가 다소 정확하지 않아 자료 확인 및 수정에 많은 시간이 소요되었습니다. 원하는 데이터를 얻기 위해 어떻게 정제해야 하는지에 대해 여러 시행착오를 겪으며 데이터 정제의 중요성을 느낄 수 있었습니다. 또 Django를 이용한 웹 페이지 구현을 이해할 수 있었고 많은 공부가 필요하다는 것을 다시 한 번 깨달았습니다.

김지혜

데이터를 선정하는 과정은 어렵지 않았으나 선정한 데이터의 사실 유무 확인, 외래어, 한문, 특수문자 처리, 맞춤법, 시제 정리 등 정제할 사항이 많아 데이터 전처리 과정에서 많은 시간을 보냈습니다. 이 과정을 통해 데이터의 정확성과 정제가 중요하다는 것을 느꼈습니다. 그리고 Django를 이용하여 데이터베이스와 연동하고 웹에서 서비스를 이용할 수 있다는 점을 배울 수 있어 유익한 경험이었습니다.

감사합니다