개발의민족(1조)

수집 데이터 기반 BERT 질의응답

김규희(조장), 김수연, 김지혜

목차

○1 추진 목표(기대효과)

02 팀원 구성 및 역할

03 WBS

○4 프로젝트 도메인 선정

05 프로젝트 수행 범위

06 프로젝트 설계

07 프로젝트 구현

08 프로젝트 후기

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

팀원구성및역할

	WBS 관리 및 작성	프로젝트B 데이터 정제 (전처리)	질의응답 데이터 작성	데이터 무결성 대조검사	DB설계	DB 데이터 삽입	UI화면 설계	Butter Block 작동테스트	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	23		W	6 2	7 30	2w 30 31 1 2 3			3 6	5 7	3w	 10	4w 13 14 15 16 17		17	
	프로젝트 배경 및 시나리오 이해	2022.05.23	2022.05.25	2022.05.23	1	100%																П
 분석	개발 환경 구축	2022.05.23	2022.05.25	2022.05.23	1	100%		П											一	十	\top	П
世令	요구 사항 정의	2022.05.31	2022.05.27	2022.05.31	1	100%													\neg	\top	\top	П
	Butter Block 연결 테스트	2022.05.23	2022.05.27	2022.05.30	6	100%																
	시스템 구조 설계	2022.05.31	2022.06.01	2022.05.31	1	100%							2									
설계	DB 설계	2022.05.31	2022.06.01	2022.05.31	1	100%							0								Т	П
	UI 화면 설계	2022.05.31	2022.06.03	2022.05.31	1	100%							2		ė	4					\Box	
	프론트 개발	2022.06.02	2022.06.08	2022.06.13	7	100%							지		ð	5				Т	Т	
구현	백엔드 개발	2022.06.02	2022.06.10	2022.06.13	7	100%							방		2	빌						
T 2	데이터 마운팅	2022.06.07	2022.06.10	2022.06.10	4	100%							선									
	API 연계	2022.06.09	2022.06.10	2022.06.14	3	100%							거									
테스트	통합 테스트	2022.06.13	2022.06.13	2022.06.14	2	100%																
발표 및	프로젝트 발표	2022.06.15	2022.06.15	2022.06.15	1	100%																
물표 및 수정보완	수정 보완	2022.06.15	2022.06.17	2022.06.17	3	100%																
1876	과제완료보고서작성	2022.06.15	2022.06.17	2022.06.17	3	100%																

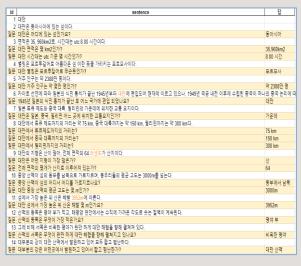
개발의민족 : 김규희(조장), 김수연, 김지혜

프로젝트 도메인 선정

1	는 전 10대에대 8년	4049	42	서용	학연자	역스트로만 여워져 있는게?	봉답회 기능관기?	한국의로반 이렇지 없는가?	원문제 대한 대한데 하나인가?	역생활수 일본 본의연시7	112
	1 적소등, 모디	보통3성 대회 환경적	×	대회성식으로 관심해야함	당시하	×	x	0	×	×	
3	2 텍스트, 오디	이 그런 중대 중성	×	질문은 문영으로 제공하다함	경지하	×	×	0		×	
4	3 00070, 900	로 교사 한자 영식 OCR	×	질문을 이미지로 제공하여할	경지해	×		×		×	
5	4 0[0[3]	공공합성문서 00명	×	이미지로 구성	848	×					
6	5 945	기계 독대	×	Q84정식에서 댓어날	결근실		×	0			
	6 먹스트	는문자로 보작	×	QEA행식에서 뜻이남	경구회	0	×	0			
ĕ	7 0000	나양한 형태의 선물 문제 OCR	×	이미지로 구성	장수원	×	×				
9	용 텍스트	도세자를 가져 죽히	×	GRASHINA TON	884		×				
10	9 목소로	도시자료 요약	×	GRA합의에서 첫이당	日音和	×					
11	11 945, 20	의명성이 용성(노인당대)		TTS이므로 실적용답 목적이 됐어요	982	×		0			
17		요명한이 음성(소이 원이)	×	TTS이므로 정의용달 목적에 벗어날	202	×			×		
15	12 목소트, 모디	보험하여 음성(발판당대)	×	TTS이므로 필의용답 목적에 벗어난	285	×		0	×		
11	13 No.5	2499 965		GRA전심에서 되어남	9-2						
15	14 905 25	요한 음악 이 전에 가장 이 의 의 의 의 의 의 의 의 의 의 의	X	TTS이므로 정의용달 목적이 됐어날	488	×		0			
15	15 ♥△=	한민준 지식비디스	×	용어의 여권용	275					×	
17	16 50.00		×	전에의 유럽이 급했으로 구성	286	×					
13		ADV INTERPRETATION	×	이미지로 무섭	884	×					
19	18 945	소상공의 교석 주문 일의 동안 박스트	-	아님의 숙우가 여러가	948						
29	19 11/12, 90	240 99		unas 28	994	×					
21		보시작정보 기반장의용답	Ŷ	이미지로구성	299	×		¥			
22	21 001	야외 설계 중합 선물 이미지	į,	이미지를 구설	248	×					
23	22 MAE	\$954						0		0	*ARR
24		유자용대회 중성(노인남의)	Ý.	대학 기반으로 프로젝트 성격이 맛지 않는	SAR	×	-		×		000
25		요가유대의 중앙(소아, 유야)	Ŷ	이와 기반으로 프로젝트 성격이 당지 않는	985	*					
35		27879 88(259H)	Ŷ	이와 가진으로 프로젝트 성격이 잊지 않음	982	×			- î		
27	26 Q.A.E.	2. 単単の記事で	-0	용어의 이권을	295	0				×	
23	27 No.8	경문분의 원양 달음의	Ŷ	설립보선 전에 보완	135			*		- 1	
20		오전한 내 대한 및 한편이 있었	÷	이이의 흥명이 건성으로 구성	Das	×			×		
30	29 90.5	목하 지식 바이스	Ŷ	892 988	394					×	
24	30 405	THO SNS	÷	대학 기반으로 무성될 사진	284		- 1				
12		일부대 SRS 일부급이 공원 음성	Ŷ	TSS이므로 중의용을 목하여 맛이날	282	×					
33	お替合品	한국이 대화	÷	대화 기반으로 프로젝트 성격이 맞지 않음	202	- 1			×		
26	33 TOR	선명이 대회 보다	÷	대화 기반으로 프로젝트 성격이 맞지 않음	228				×		
35		인적에 해외 보기 유원국의 원원 발제(참원도)	× ×	대학 기환으로 프로젝트 성격에 맞지 않는 대학 기환으로 프로젝트 성격에 맞지 않는	580 580	Y Y		0	×		
16		2340 82588(38E)	*	대한 기원으로 프로젝트 성격이 당지 않는	985	×					
"		요한국의 왕인 날씨(신라도)		대학 기원으로 프로젝트 성식이 맞지 않는	982			0			
35			X	대학 기반으로 프로젝트 성격이 맞지 않음 대한 기반으로 프로젝트 성격이 당지 않음	992	×		0			
30		요한국이 왕인 영화(제주도)				×		0	1		
30 41		의원국의 당인 발화(충정도)	×	대화 기원으로 프로젝트 성격이 맞지 않음	202	×			×		
	38 적소등, 모디		×	네이티 유용이 음성으로 구성		×			×		
41	4) 택스트	전국와·영어 함역 활동자(기송과활)	×	생물으라 영어 구장	334	0		×			
42	41 MAE	한국어-명이 번째 말문지(사회과학)	X	종필요한 전에 프랑	200	0		X			
	Q 목소트	출한국어-당이 선택(생활/설생시	X	출신으한 단위 호텔	914	0					
44	43 QAE	한국어-일본에 원칙 당용지	×	불립요한 단어 표정	924	0		Y			
45	44 핵소트	선국어-중국에 전역 현용자(기술과학)		불원보산 연기 포함	명규회	0		X			
45	45 144.5	선물의 중국에 관직 활용자(사회관약)		물통보면 전에 포함	813			×			
47		오전국인 대학 중성		중성의의되를 기반으로 잘	집지하	×		0	×		
43		요한국인 의해이 발화		음성이 이다를 기반으로 합	용지해	×		×	×		
42	48 9445, 90	2 2 2 3 3 2		음성이 이렇을 가난으로 전	470	*					

	rist	885	이상은 운성(본건)	
	아르헨티나	49		전체리가 조금 필요하나 전반적으로 데이터 일은 양보
아메리카	역시코	50		Méjco(소리한학)/metal('당')와 xiall('배문') 참석X
	보라질	49	6번째문장 - 가디X	Republica Federativa do Brasil, Brasilla(東亜年間の)
n-erowan	게냐	- 25		테이터 규모가 학통
아프리카	이집트	49	9번-10번 경제대용 15번~22번 이집트	나용은 출분한거 같으나, 해석 불가능한 이집로 명칭. 언어가 있음
오세아니아	오스트레일리아	50		데이하의 규모 중분, 질의용답 보통
	별기에	48	43~46년 언어	48문장 중 별기에 관한 내용이 부족합 필의용답 부작합
	영국	49	11번째 인간비 최신	영어의 규모 發車
	형가리	49		테이터 규모는 양호하나 전쟁에 많이 지우쳐있음
	이탈리아	50	20년 2011년	과거 자료이기 때문에 업데이르가 필요해보이지만 데이터 내용이 골고루 있어 필의용답 가능
	원준도	50	10년 2014년 39년 해석? 47-50년 주소개	Nan(원란도이), 필의용답 가능
	스웨덴	50	26~32 스웨덴 종리	데이터 규모 당호, 절의용답 보통
	네엘란드	50	45번 역도날드	데이터 규모 양호, 정의용답 보통
유럽	면마크	49		데이터 규모 당보, 필의용답 보통
	두마니아	49		데이터 규모 양호, 철의용답 양호
	스위스	48	44번 2011년 41번 대한근제	데이터 규모 중분, 절의용답 가능
	아일랜드	50		데이터 규모 중분, 절의용답 당호
	포르투갈			초기 역사에 대한 내용만 있음.
	품장도	50	기후(기온에대한정	역사는 간략하게 정리되어있음. 철의 응답 보통
	독일	49		데이터 규모는 항로한편이며, 절의용답 구조로 확합
	프랑스	50	5번째 문장	중세까지의 역사만 있음(반문 원본어 프랑스어(Painte)
	스페인	50		스페인어,그리스어표기법이 있음
	베트남	48		레트남어 종종 등장, 데이터 규모 중분, 절의등답 적절
	중화인민공화국	49		데이터 규모는 용분하나 질의 용답 구조가 가능할지 의문
	타이완	49	34-48번 불필요	성 관련 데이터 제외하고 나마지는 교두 철의 용답 구조에 적절
	인도	49		B연구수, 개인소육 관련 회신화된 자료 필요. 필의 중합 구조로는 데이터 가치는 중하 정도
	달레이시아	50		데이하의 내용은 다양하나 질문을 주출하기에는 어렵다
	이란	50		데이터 규모 충분하지만 문장들이 필의용답에 어려워보면다.
	인도네시아	49	47번 맞중법 요류	기후, 환경과 관련하여 절의 등답 구조로 제작이 쉬워보임. 이 외에 정치 신명은 버려질 테이터
	타키	48	13-16분 언어관련	데이타 규모는 중분한거같지만, 문장들이 필의용답하기 데메하다.
	이스라엘	50		데이터 규모 종분, 질의 응답 보통, 현재의 데이터와 일지하는지 과약 필요
	타이	50	46년째 PPP 49년째 최근?	한문, 약에처리(PPP), 역사 데이터 위주이며 대부분 철의용답 가능
0FA)0F	월리편	50	7번 설 경수 8번 인구 순위	정보의 정확성이 떨어지는 데이터들이 있음. 정보 없데이트가 필요해보임 (섬 것수 정확하지 3
	조선민주주의연민공화국	. 51		데이터 중 북한 정부 수립 전 남북한 공통의 역사 내용(고구려, 신리~ 대한제국(15~32문장 등)
	한국	48	10~48년 언어관련 19-23년 일본이 24-28년 중국이	요국에서 한국을 뜻하는 안에 혹은 언어의 표기 요국에가 주된 내용이라 필의 용답하기 어려움

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





데이터 획득 & 데이터 정제1

데이터 정제2

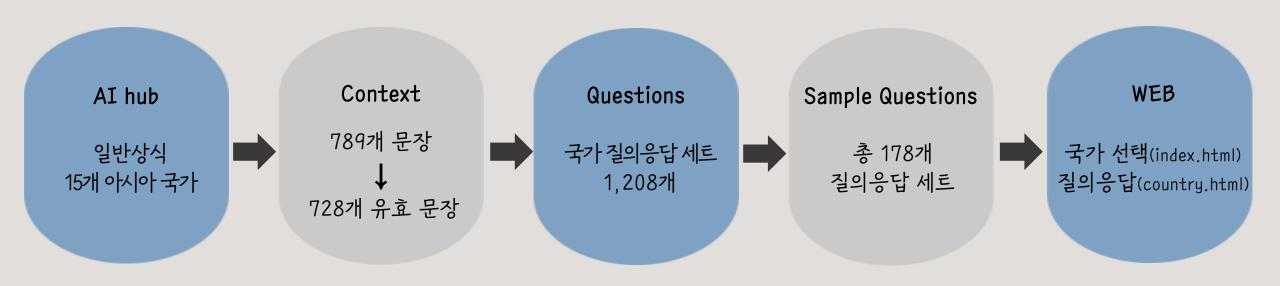
선정, 규모, 분석

데이터 가공

시각화(word cloud)

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

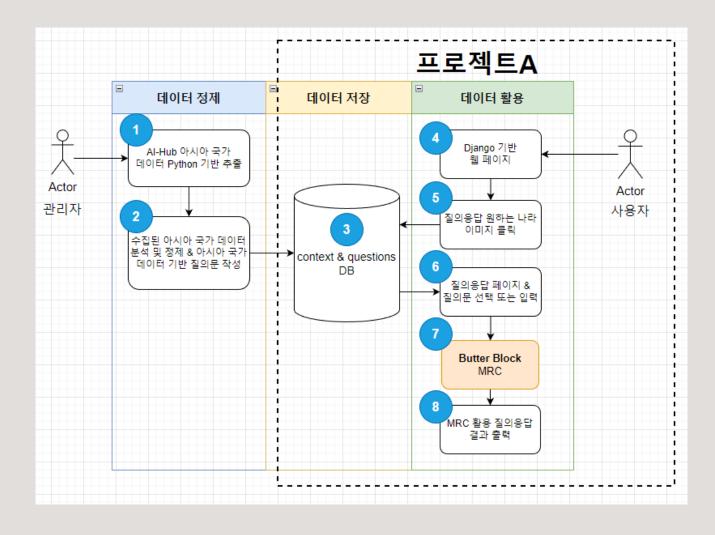
프로젝트 수행 범위



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)

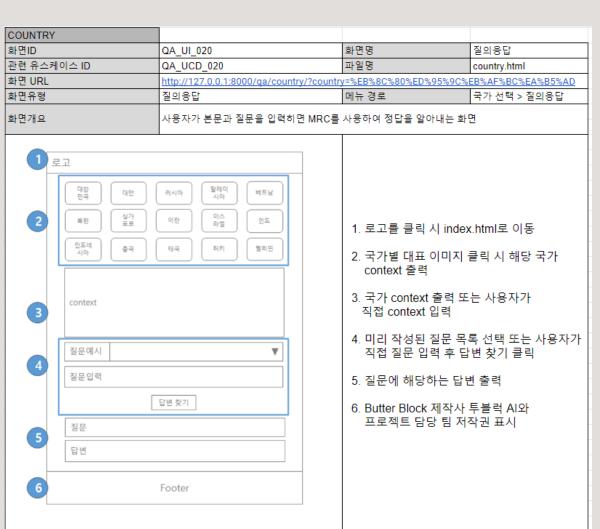
- 4. 사용자가 url 입력하면 웹페이지 출력
- 5. 웹페이지에서 질의응답을 원하는 국가 이미지 클릭
- 6. 질의응답 페이지로 이동하고 그 페이지에서 질문을 선택 또는 입력
- 7. Butter Block의MRC를 활용해 해당 질문에 대한 정답 도출
- 8. 웹페이지에 입력된 질문과 해당 질문에 대한 정답 출력

프로젝트 설계 화면정의





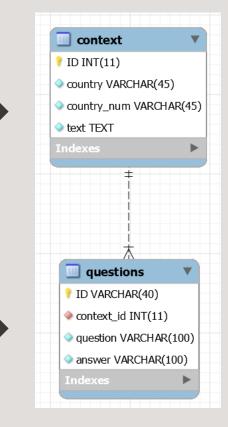
- 1. 로고를 클릭 시 index.html로 이동
- 2. 국가별 대표 이미지 슬라이드 출력
- 3. 질의응답할 국가 클릭 시 4번 모달창 출력
- 4. 국가별 워드 클라우드 이미지 출력 및 button 클릭 시 country.html로 이동
- 5. Butter Block 제작사 투블럭 AI와 프로젝트 담당 팀 저작권 표시



프로젝트 설계 데이터베이스 정의

	테이블 정의서													
주저	ll 명	수집한 국가 데이	터 기반으로 BEF	RT를 통한 질의응	답	작성일	2022.05.31	작성자	1조 개발의	민족				
테이블	한글명	본문				테이블 영문명	context							
테이블	불설명	아시아 국가별 정보로 구성된 테이블												
컬럼한글명	컬럼영문명	테이터 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													
주저	ᅨ명	수집한 국가 데이	터 기반으로 BEF	RT를 통한 질의응	답	작성일	2022.05.31	작성자	1조 개발의민족					
테이블 한글명 질문 테이블 영문명 questions														
테이블 설명 아시아 국가별 질의응답으로 구성된 테이블														
컬럼한글명	컬럼영문명	테이터 타입	길이	NotNull	DEFAULT	AUTO_INCREMENT	PRIMARY KEY	FORIGN KEY	UNIQUE	CHECK				
질문번호	ID	VARCHAR	40	Y			Y							
나라_고유번호	context_id	INT		Y				Υ						
질문	question	VARCHAR	100	Y										
답변	answer	VARCHAR	100	Y										

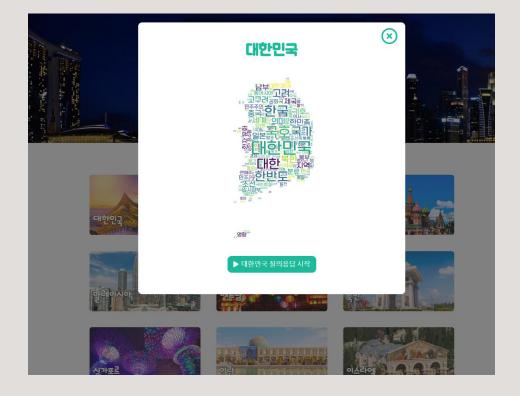


프로젝트 구현

index

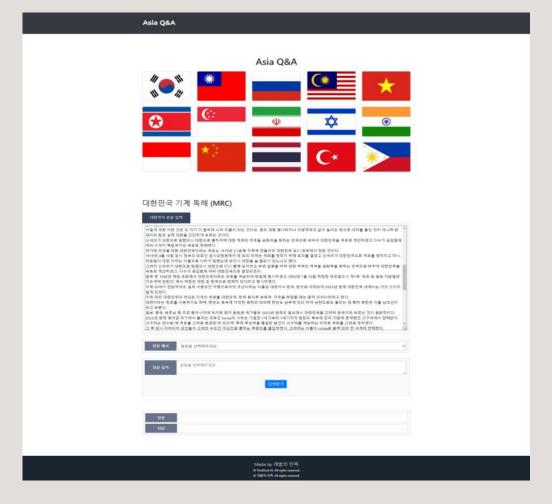


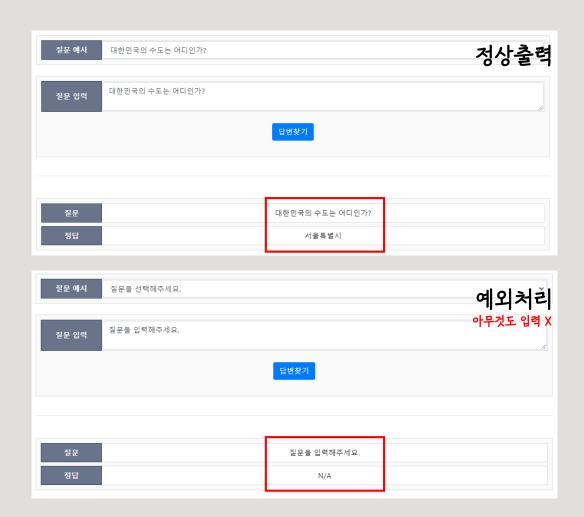
index - modal



프로젝트 구현

country









S

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 이용할수 있다는 점을 배울 수 있어 유익한 경험이었습니다.



감사합니다