개발의민족(1조)

수집 데이터 기반 BERT 질의응답

김규희(조장), 김수연, 김지혜

추진목표

□ 정답 데이터 정확도 향상을 위해 국가 데이터를 AI언어 처리 모델에 적용하기 위한 적정한 데이터 세트 수집 및 분석

□ 수집한 국가 데이터 기반으로 BERT를 통한 정답 데이터 출력

기대효과

□ 데이터를 정제하여 정보의 부정확성을 해소

□ 쉬운 접근성으로 원하는 답을 찾을 수 있도록 생산성을 높임

프로젝트 수행 과정

프로젝트B AI hub 일반상식 15개 아시아 국가
789개 문장→728개 유효 문장

프로젝트A

샘플	질문	Q13	1948년 5.10 총선거를 통해 구성된 국회는?	제헌국회
샘플	질문	Q14	대한민국정부가 출범한 날짜는?	1948년 8월 15일
	질문	Q14	대한민국의 민주주의 지수는 몇번째인가요?	2번째
	질문	Q15	대한민국의 민주주의 지수는?	8.16점
	질문	Q16	대한민국이 합법 정부로 유엔으로부터 승인 받은 시기는?	1948년 12월
샘플	질문	Q17	대한민국은 언제 유엔에 가입했나요?	1991년
	질문	Q18	대한민국이 세계적인 강국으로 발전한 시기는?	1990년대
	질문	Q19	대한민국의 1인당 국민 총소득은?	31,489달러
	질문	Q20	대한민국의 인간 개발 지수는 세계에서 몇 위인가?	22위
샘플	질문	Q21	IMF에서는 대한민국을 무엇으로 분류하고있나요?	선진국
	질문	Q22	대한민국의 명목 국내 총생산은?	2조 4,478억 달러
	질문	Q23	대한민국의 명목 국내 총생산의 순위는?	10위
샘플	질문	Q24	대한민국이 회원국으로 활동하고있는 기구는?	g20, 경제 협력 개발기구 oecd, 개발
	질문	Q26	삼한은 무엇인가요?	마한, 진한, 변한
샘플	질문	Q28	고종이 대한 제국을 선포한 연도는?	1897년
	질문	Q29	최남선의 책 이름은?	조선상식문답
	질문	Q30	한국은 무엇을 간단하게 부르는 말인가요?	대한
	질문	Q31	신석우가 제안한 국호는?	대한민국
	질문	Q32	공화국을 뜻하는 것은?	민국
	질문	Q33	대한민국이라는 국호는 언제 정해졌나요?	1919년 3.1 운동 직후
	질문	Q34	대한민국이라는 국호를 정한 곳은?	대한민국 임시 정부
샘플	질문	Q35	대한민국으로 국호를 정하자고 한 사람은?	신석우

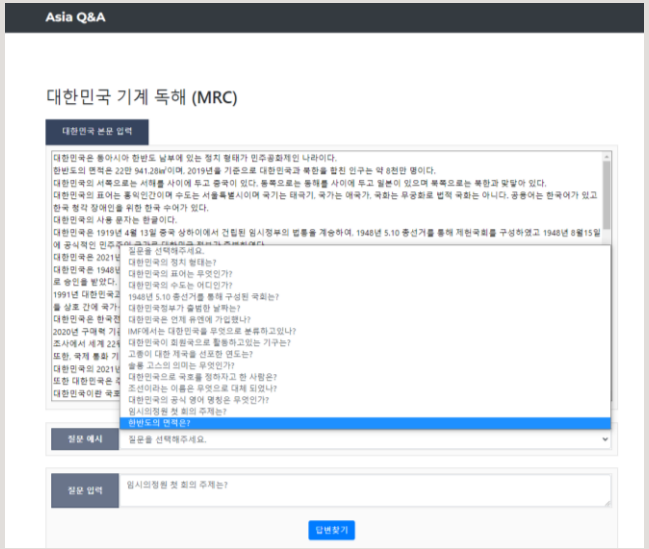
Questions

국가 질의응답 세트 1,208개

ID	context_id	question	answer
한국01	1	대한민국의 정치 형태는?	민주공화제
한국02	1	대한민국의 표어는 무엇인가?	홍익 인간
한국03	1	대한민국의 수도는 어디인가?	서울특별시
한국04	1	1948년 5.10 총선거를 통해 구성된 국회는?	제헌국회
한국05	1	대한민국정부가 출범한 날짜는?	1948-08-15
한국06	1	대한민국은 언제 유엔에 가입했나?	1991년
한국07	1	IMF에서는 대한민국을 무엇으로 분류하고있나?	선진국
한국08	1	대한민국이 회원국으로 활동하고있는 기구는?	g20, 경제 협력 개발기구 oecd, ...
한국09	1	고종이 대한 제국을 선포한 연도는?	1897년
한국10	1	을룽 고스의 의미는 무엇인가?	무지개 뜨는 나라
한국11	1	대한민국으로 국호를 정하자고 한 사람은?	신석우
한국12	1	조선이라는 이름은 무엇으로 대체 되었나?	대한이나 한국, 한
한국13	1	대한민국의 공식 영어 명칭은 무엇인가?	republic-of-korea
한국14	1	임시의정원 첫 회의 주제는?	국호
한국15	1	한반도의 면적은?	22만 941.28㎢
대만01	2	대만은 어디에 있는 섬인가?	동아시아
대만02	2	대만 거주 인구는 몇 만 명인가?	2388만 명
대만03	2	가장 긴 하천의 이름은 무엇인가?	줘 수이강
대만04	2	대만의 기후는 어떤 기후로 대변 되나?	온난 습윤 기후
대만05	2	섬 전체의 연평균 기온은 어떻게 되나?	23°C
대만06	2	대만의 기후에 영향을 주는 것은 무엇인가?	높고 험한 타이완 산맥과 계절풍
대만07	2	대만 섬 전체의 연간 강수량은 몇 mm인가?	2540mm
대만08	2	대만에 분포하는 식물은 몇 종인가?	3800종
대만09	2	일본이 실효 지배하고 있는 섬 이름은?	댜오위 섬
대만10	2	고대부터 이어져 온 대만은 원래 어떤 원주민들...	오스트로네시아어족
러시아01	3	러시아의 위치는 어디인가?	동유럽과 북아시아
러시아02	3	러시아의 공용어는?	러시아어
러시아03	3	러시아의 수도는?	모스크바

Sample Questions

총 178개 질의응답 세트



Django

Index : main 국가 선택 /Country : 질의응답

팀원 역할 및 후기

김규희(조장)

역할

JSON 파일에서 문장 추출 코드 작성, 도메인 선정을 위한 AI-HUB 데이터 수집 및 분석, 질의응답 데이터 작성, 선별된 데이터 정제(전처리), Django 백엔드 구현, Butter Block 연동, WBS관리 및 작성, DB설계, Butter Block 연결 테스트

후기

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

역할

도메인 선정을 위한 AI-HUB 데이터 수집 및 분석, 질의응답 데이터 작성, 선별된 데이터 정제(전처리), 데이터 무결성 대조 검사, 정제된 데이터 TXT파일 저장, WBS관리 및 작성, 문서 작성, DB 데이터 삽입, UI화면 설계, Butter Block 연결 테스트

후기

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

역할

도메인 선정을 위한 AI-HUB 데이터 수집 및 분석, 질의응답 데이터 작성, 선별된 데이터 정제(전처리), 데이터 무결성 대조 검사, Django 프론트엔드 구현, WBS관리 및 작성, DB데이터 삽입, UI화면 설계, Butter Block 연결 테스트

후기

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