

자연과학 Q&A 챗봇 개발 문서

[3조 : 장하영 김태호 박서현 윤주완]

1. 프로젝트 개요


- **프로젝트명:** 자연과학 Q&A 챗봇 개발
- **목표:**
 - 사용자가 자연과학 및 기술에 대한 질문을 하면 AI가 적절한 답변을 제공하는 챗봇 개발
 - 한국어 질문도 지원하여 더욱 폭넓은 사용자 경험 제공
- **대상 사용자:**
 - 과학을 공부하는 학생 (초·중·고·대학·대학원생)
 - 연구자 및 실험실 종사자
 - 관련 업계 전문가 (환경공학, 에너지산업, 바이오테크 등)
 - 과학에 관심 있는 일반 사용자
- **데이터 출처**

microbiology qna

Data Card Code (0) Discussion (0) Suggestions (0)

microbiology_qna.csv (15.84 MB)

Detail Compact Column

#	context	question	answer
 0 12.4k	4006 unique values	11941 unique values	12008 unique values
0	Human activities not only directly produce pollutants, as seen in mining or agriculture, but also in...	How does burning fossil fuels contribute to pollution?	Burning fossil fuels increases atmospheric carbon dioxide levels, leading to global warming, which i...
1	Human activities not only directly produce pollutants, as seen in mining or agriculture, but also in...	What are the consequences of inadequate waste management and disposal?	Inadequate waste management and disposal can result in the presence of pathogenic microorganisms in ...

출처: <https://www.kaggle.com/datasets/moonstone34/microbiology-qna>

2. 문제 정의 (4W)

1) What (무엇이 문제인가?)

- 자연과학 및 기술에 대한 방대한 정보 속에서 **정확한 답변을 찾는 것이 어렵고 시간이 많이 소요됨**
- 신속하고 신뢰할 수 있는 답변을 제공하는 AI 챗봇이 필요함

2) Why (왜 이 문제를 해결해야 하는가?)

- 과학 정보는 최신성을 요구하며, 이를 빠르게 탐색하는 것이 중요함
- 기존 검색 엔진은 시간이 오래 걸리거나 신뢰성이 부족할 수 있음
- AI 챗봇을 통해 즉각적인 응답을 제공함으로써 학습 및 연구의 효율성을 높일 수 있음

3) Who (누구를 위한 해결책인가?)

- 학생 (과학을 배우는 초·중·고·대학생 및 대학원생)
- 연구자 및 실험실 종사자
- 관련 업계 전문가 (환경공학, 에너지산업, 바이오테크 등)
- 과학에 관심 있는 일반 사용자

4) Where (이 문제가 발생하는 환경은?)

- 학생들이 과학 학습 중 궁금한 내용을 빠르게 찾지 못하는 상황
 - 연구자들이 논문이나 실험을 진행할 때 **신뢰할 수 있는 정보**를 빠르게 확인해야 하는 경우
 - 기업 및 산업 종사자들이 기술적 문제를 해결하기 위해 **과학적 정보**를 필요로 하는 경우
 - 일반 사용자들이 과학적 궁금증을 해결하기 어려운 경우
-

3. 기획 의도

- 자연과학 및 기술 질문에 대해 **신속하고 정확한 답변**을 제공하여 학습 및 연구를 지원
 - **AI 기술을 활용하여 질의응답 시스템 자동화**, 사용자 편의성 향상
 - 사용자의 질문 패턴을 분석하여 **데이터베이스를 지속적으로 보완 및 개선**
-

4. 기대 효과

- 과학 학습 및 연구의 효율성 증가
 - 정확하고 신뢰할 수 있는 정보 제공을 통한 사용자 만족도 향상
 - AI 챗봇을 활용한 지식 공유의 새로운 방식 도입
 - 자연과학 및 기술의 다양한 주제에 대한 접근성 향상
-

5. 모델 개발 및 저장

1. 모델 선택:

- **Sentence-BERT (구체적으로 all-MiniLM-L6-v2)**를 사용하여 질문 임베딩 생성

주요 코드 흐름:

- **CSV 파일에서 데이터를 로드하여 질문과 답변 리스트 생성**

- Sentence-BERT를 이용하여 질문 임베딩 계산
- 계산된 임베딩과 원본 질문, 답변 데이터를 qna_model.pkl 파일로 저장

2. 모델 목적:

- 사용자가 입력한 질문과 가장 유사한 질문을 임베딩 유사도 기반으로 검색하고, 해당 질문에 매핑된 답변을 제공

2. 챗봇 구현 방향

- 사용 기술:
- 프론트엔드: Streamlit을 활용하여 웹 UI/UX 구현
- 백엔드 및 모델: Python, Sentence-BERT (all-MiniLM-L6-v2)
- 번역 기능:
- Google Translate API를 활용하여 한글 입력을 영어로 변환하고, 영어 답변을 한글로 재번역하여 사용자에게 제공
- 구체적으로, deep_translator 라이브러리의 GoogleTranslator 클래스를 활용

4. 운영 과정:

- 사용자가 질문 입력
- 입력된 한글 질문을 영어로 번역
- 영어 질문 기반으로 임베딩 계산 및 FAQ 질문과의 코사인 유사도 평가
- 가장 유사한 FAQ 질문을 검색 후 해당 답변 선택
- 선택된 영어 답변을 한글로 재번역하여 사용자에게 출력

6. 성능 평가

챗봇의 성능은 MRR(Mean Reciprocal Rank) 및 Top-K Accuracy를 기준으로 평가함.

1) 성능 지표 정의

- **MRR (Mean Reciprocal Rank, 평균 역순위)**
 - 사용자가 원하는 정답이 챗봇이 반환한 답변 리스트에서 몇 번째에 위치했는지를 평가
 - 각 테스트 케이스에서 정답의 순위의 역수를 구하고 평균을 내어 계산
 - MRR 값이 1에 가까울수록 챗봇이 높은 순위에서 정답을 제공
- **Top-K Accuracy (정확도@K)**
 - 챗봇이 반환한 상위 K개의 답변 중 하나가 사용자가 기대한 답변과 일치하는 비율

- Top-1, Top-3, Top-5 성능 평가

2) 챗봇 성능 평가 결과

영어 원본 데이터 기준 성능

✅ TF-IDF (영어 데이터 기준)

평가 지표	성능
MRR	0.9900
Top-1 Accuracy	0.9800
Top-3 Accuracy	1.0000
Top-5 Accuracy	1.0000

✅ SBERT (영어 데이터 기준)

평가 지표	성능
MRR	1.0000
Top-1 Accuracy	1.0000
Top-3 Accuracy	1.0000
Top-5 Accuracy	1.0000

한글 질문 (번역된 데이터) 기준 성능

모델	MRR	Top-1 Accuracy	Top-3 Accuracy	Top-5 Accuracy
TF-IDF 기반 챗봇	0.7603	0.6800	0.8400	0.8700
Sentence-BERT 기반 챗봇	0.9273	0.9100	0.9400	0.9600

7. 결론 및 향후 개선 방향

1) 결론

- TF-IDF 모델은 단순한 키워드 매칭 방식으로 한글 번역 데이터에서 성능이 저하됨
- Sentence-BERT 모델은 의미 기반 비교가 가능하여 한글 질문에서도 높은 정확도를 유지
- MRR 및 Top-K Accuracy 결과에서 Sentence-BERT 모델이 더 우수한 성능을 보임

2) 향후 개선 방향

- ✅ 데이터셋 확장: 더 많은 과학 관련 질문 데이터 추가
- ✅ 하이브리드 모델 개발: TF-IDF와 BERT의 가중치 조정을 통해 최적화된 성능 도출
- ✅ 모델 경량화: Sentence-BERT 모델을 압축하여 실시간 응답 속도 개선
- ✅ Streamlit UI 개선: 사용자 경험 향상을 위해 인터페이스 최적화

최종 정리

- ✅ Streamlit을 이용하여 챗봇 UI 구현 완료
- ✅ 한글 질문 번역 기능 추가 및 성능 비교 완료
- ✅ Sentence-BERT 모델이 가장 우수한 성능을 보임
- ✅ 향후 모델 최적화 및 데이터셋 확장 진행 예정