기업 요구사항 기반의 문제해결 PJT

챗봇 백과사전

6조 Wiki-Taka

https://github.com/tnals545/Final_PJT

목차

01 프로젝트 기획

02 프로젝트 수행 절차

03 기대효과 및 향후 과제

04 개발후기 및 느낀점

01

프로젝트 기획

1.1 배경

1.2 개요 및 목적

1.3 흐름도

1.4 구성원 및 역할

1.1
기획배경

"원하는 정보만 얻을 수 없을까?"

위키백과
우리 모두의 백과사전

위키백과 검색

대문
최근 바뀜
요즘 화제
임의의 문서로
기부

사용자 모임
사랑방
사용자 모임
관리 요청

편집 안내
도움말
정책과 지침
질문방

도구
여기를 가리키는 문서
가리키는 글의 최근 바뀜
파일 올리기
특수 문서 목록
고유 링크
문서 정보
이 문서 인용하기
위키데이터 항목
언어 간 링크 편집

인쇄/내보내기
책 만들기
PDF로 다운로드
인쇄용 판

다른 프로젝트
위키미디어 공용

언어
이 위키백과에서 언어 링크는 문서 제목 건너편의 문서 최상단에 있습니다. 최상단으로 이동합니다.

문서 토론 읽기 편집 역사 보기

4월 30일부터 5월 6일까지 2022년 상반기 에디터톤 공모전의 주제 투표가 진행됩니다.
대문에 좋은 글을 노출시킬지에 대한 토론이 진행 중입니다.

은닉 마르코프 모형

27개 언어

위키백과, 우리 모두의 백과사전.

은닉 마르코프 모형(영어: hidden Markov model, HMM)은 통계적 마르코프 모형의 하나로, 시스템이 은닉된 상태와 관찰가능한 결과의 두 가지 요소로 이루어졌다고 보는 모델이다. 관찰 가능한 결과를 야기하는 직접적인 원인은 관측될 수 없는 은닉 상태들이고, 오직 그 상태들이 마르코프 과정을 통해 도출된 결과들만이 관찰될 수 있기 때문에 '은닉'이라는 단어가 붙게 되었다. 은닉 마르코프 모형은 동적 베이지안 네트워크로 간단히 나타낼 수 있으며, 은닉 마르코프 모형의 해를 찾기 위해 전향-후향 알고리즘을 제안한 스트라토노빅[1]의 최적 비선형 필터링 문제와 밀접한 관련이 있다. 한편 은닉 마르코프 모형에 사용된 수학적 개념들은 바움(L. E. Baum)과 그의 동료들에 의해 정립되었다.[2][3][4][5][6]

마르코프 연쇄와 같은 단순한 마르코프 모형에서는 상태를 관찰자가 직접적으로 볼 수 있으며, 그러므로 상태가 전이될 확률은 단순히 모수(parameter)들로 표현될 수 있다. 반면 은닉 마르코프 모형에서는 상태를 직접적으로 볼 수 없고 상태들로부터 야기된 결과들만을 관찰할 수 있다. 각각의 상태는 특정 확률 분포에 따라 여러가지 결과를 도출해 낼 수 있으므로, 은닉 마르코프 모형로부터 생성된 결과들의 나열은 기저에 은닉된 상태들에 대한 정보들을 제공하고 있다고 생각할 수 있다. 여기서 단어 은닉(Hidden)이 모델의 모수를 가리키는 것이 아니라 모델이 거쳐가는 연속된 상태를 지칭하는 것에 주의해야한다. 은닉 마르코프 모형에서 모수들이 정확히 알려졌음에도 불구하고 여전히 '은닉' 마르코프 모형로 불리는 이유는 결과를 야기하는 상태들이 근본적으로 은닉되어있어 관찰할 수 없기 때문이다.

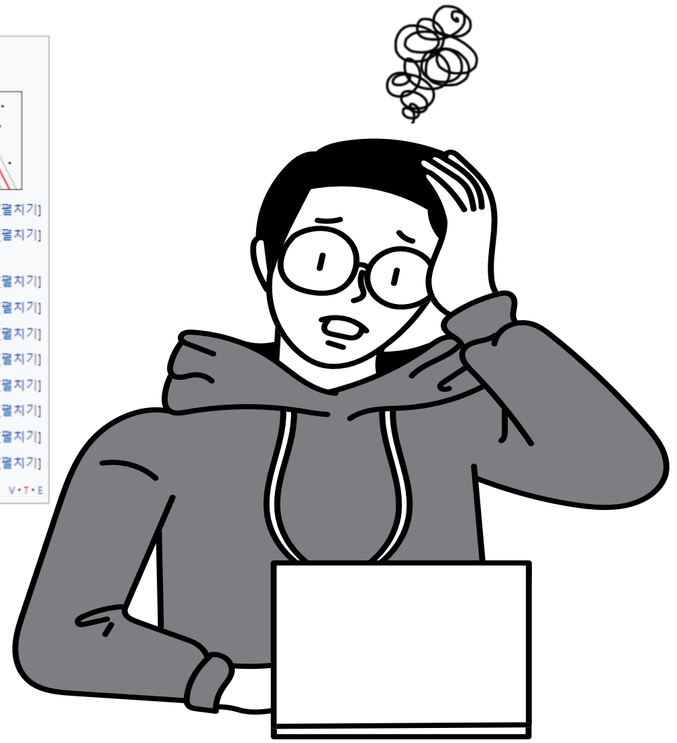
은닉 마르코프 모형은 시간의 흐름에 따라 변화하는 시스템의 패턴을 인식하는 작업에 유용하다. 예를 들어 음성 인식, 필기 인식(en:Handwriting recognition), 동작 인식(en:Gesture Recognition),[7] 품사 태깅(en:Part-of-speech tagging), 악보에서 연주되는 부분을 찾는 작업,[8] 부분 방전(en:Partial discharge),[9] 생물정보학 분야에서 이용된다.

은닉 마르코프 모형은 은닉 변수가 독립되지 않고 마르코프 과정을 통해 변화하면서, 각 과정에서 혼합 요소를 선택하는 혼합 모델의 일반화로 볼 수 있다. 최근 이러한 은닉 마르코프 모형은 더 복잡한 자료 구조들과[10][11] 안정적이지 않은 데이터들을 모델링할 수 있도록[12][13] 이중 마르코프 모형, 삼중 마르코프 모형 등으로 일반화되고 있다.

목차 [숨기기]
1 역사
1.1 마르코프 연쇄
1.2 은닉 마르코프 모형로의 발전

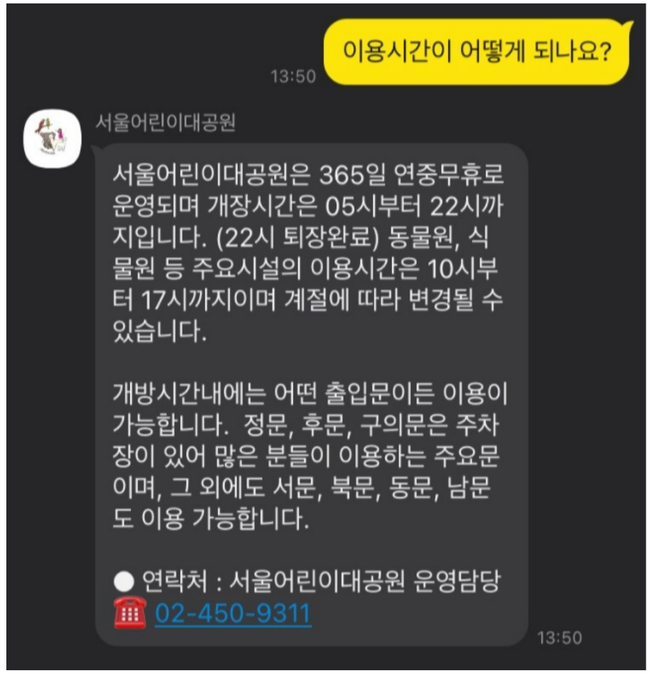
기계 학습과 데이터 마이닝

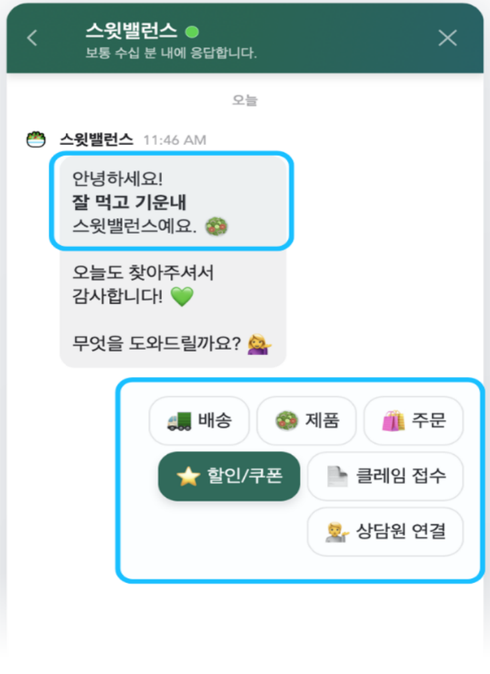
문제 [펼치기]
지도 학습 (통계적 분류 • 회귀 분석) [펼치기]
클러스터 분석 [펼치기]
차원 축소 [펼치기]
그래프 모형 [펼치기]
이상 탐지 [펼치기]
인공 신경망 [펼치기]
강화 학습 [펼치기]
이론 [펼치기]
관련 문서 [펼치기]
v • t • e

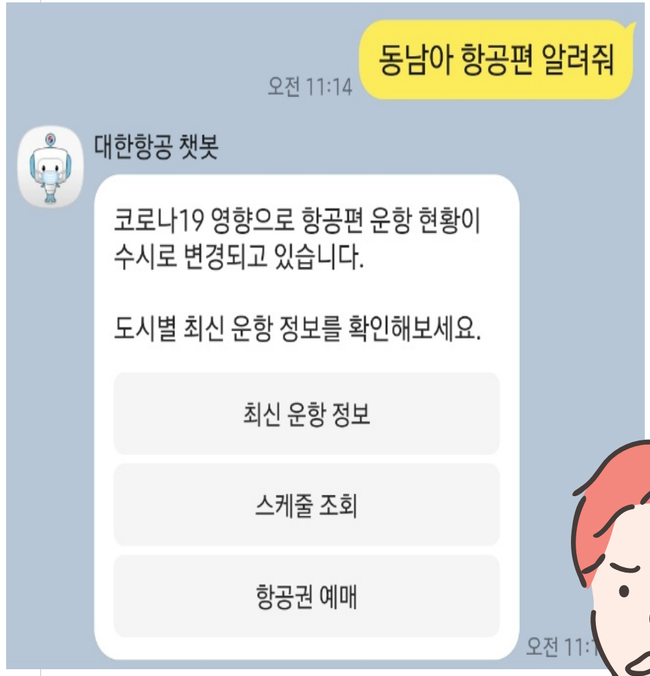


1.1
기획배경

“다양한 분야에 대한 정보를 얻을 수 없을까?”



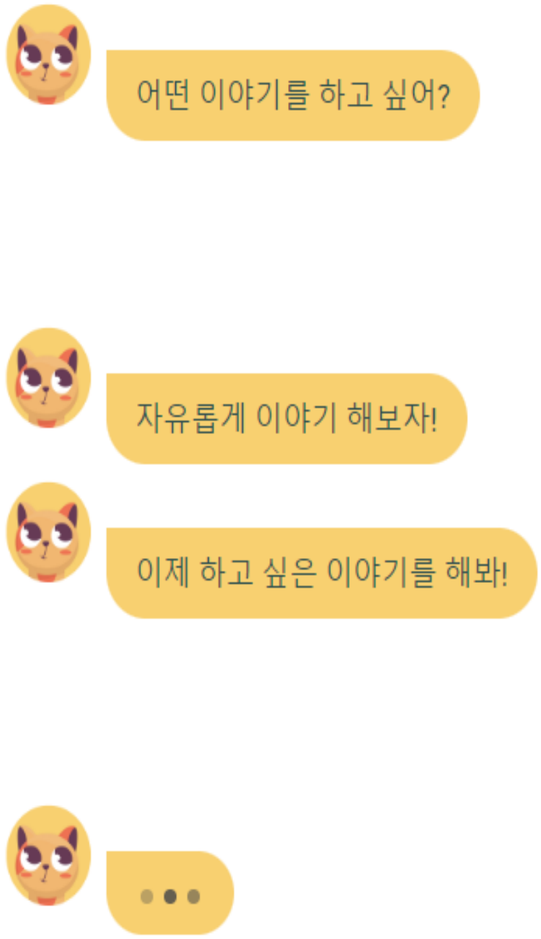


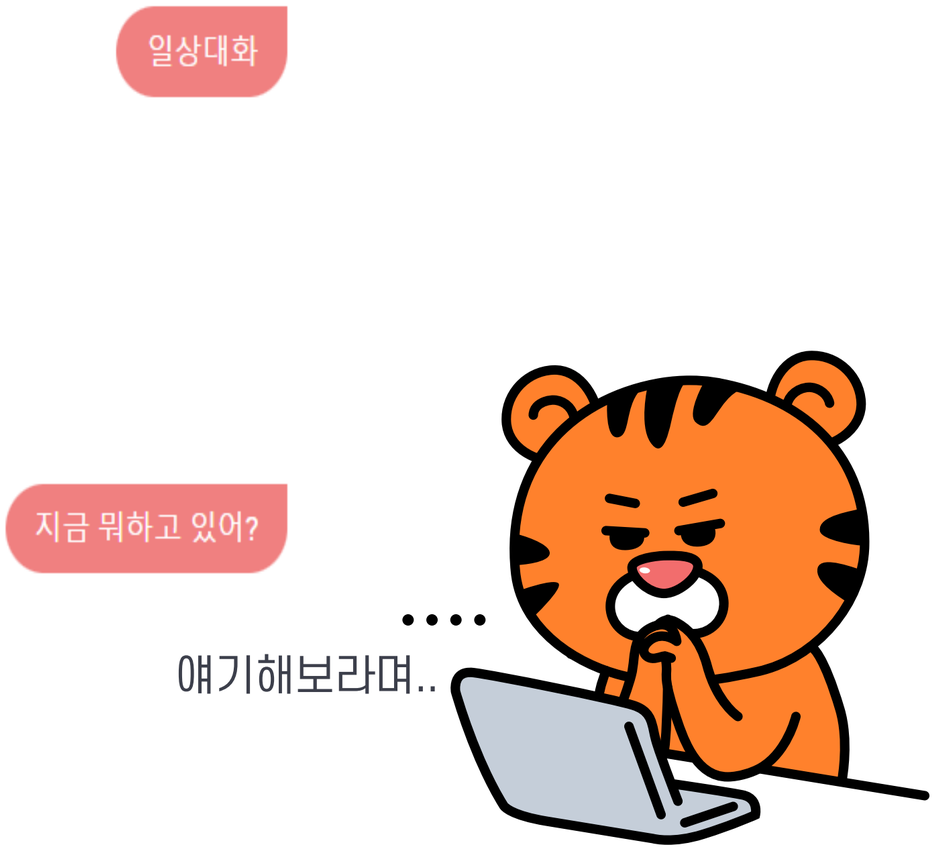


인물, 사회, 경제, 역사 등
다양한 분야의 답변을 얻고 싶은데...

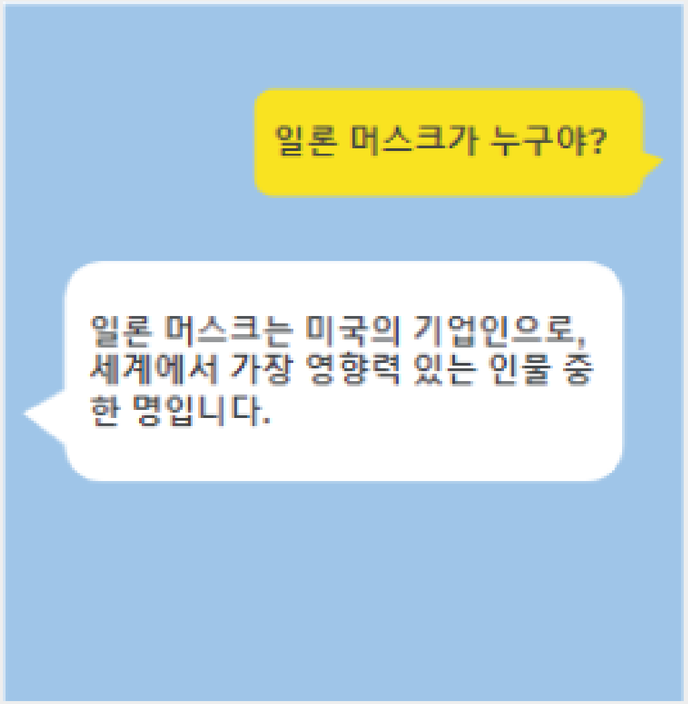
1.1
기획배경

"내 질문을 이해하는 챗봇은 없을까?"

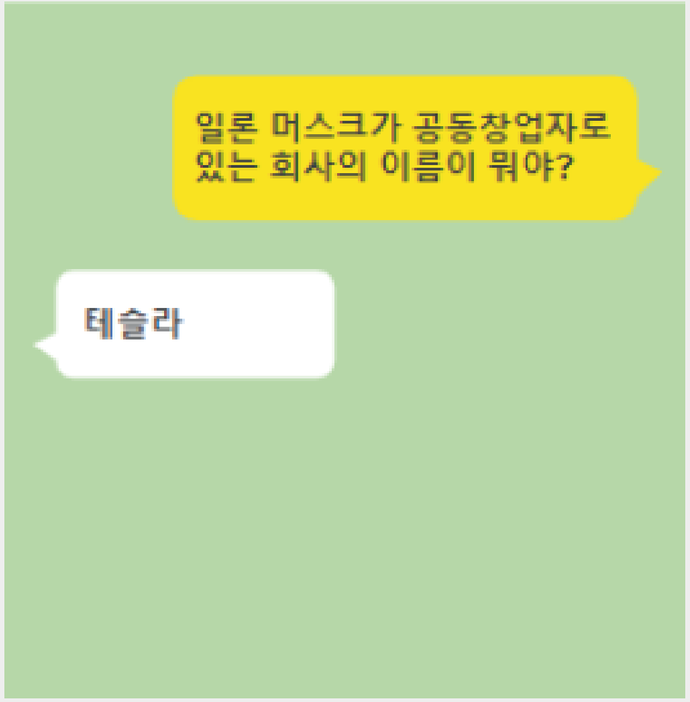




1.2
개요



단순한 검색어에 대해
일반적이고 개괄적인 정보를
얻고 싶을 때 사용



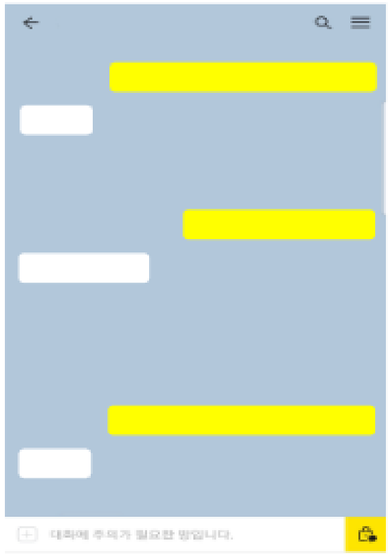
구체적인 질문에 대한 답을
얻고 싶을 때 사용

1.2
목적

- ☑ NLP를 활용해 짧게 원하는 정보만을 제공하여 사용자가 정보를 탐색하는 시간과 노력을 최소화함.
- ☑ 위키백과의 DB를 사용하여 특정 분야에 한정되지 않는 다양한 정보 제공 가능
- ☑ 음성인식을 통해 사용자가 채팅을 입력할 수 없는 환경에서도 이용할 수 있도록 편의성 제공

1.3
흐름도

Chatbot Service



FastAPI

HTML

JavaScript

RestAPI

질문

답변

Colab Server

DB

Q&A Data

Tools(Model, API service)

T5
KeyBERT
Naver Chatbot

의도분석 (유사도)

HUGGING FACE

Google BERT

NAVER Cloud

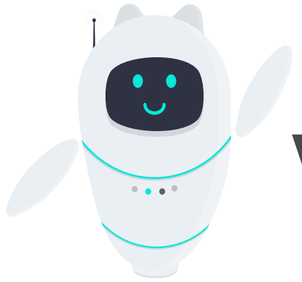
키워드 추출

개체명 추출

시맨틱 분석

구문 분석

1.4
팀원 구성 및 역할



wiki-taka

권태헌	박소연	양동욱	유승희	이시내	조혜원	최수민
- SBERT 모델 - 서버 구축 및 운영 - 웹페이지 제작	- 전처리 - T5 모델 -Naver Chatbot	- 전처리 -토픽 모델링 -모델 성능평가	- 전처리 - T5 모델 - 웹페이지 제작	- 전처리 -토픽 모델링 - 모델 성능평가	- 전처리 - KeyBERT 모델 - SBERT 모델	-데이터 수집 - LSTM 모델 - KeyBERT 모델

02

프로젝트 수행 절차

2.1 데이터 수집
2.2 데이터 전처리
2.3 모델링
2.4 성능평가
2.5 웹서비스 구현

2.1
데이터 수집

	KorQuAD 1.0	KorQuAD 2.0	위키백과 우리 모두의 백과사전
문서 형태	- 문단, 제목 - 질문, 답변	- 본문, 제목 - 질문, 답변	- 본문, 제목
데이터 개수	-10,645개의 문단 - 66,181개의 질의응답쌍	- 47,957개의 문서 - 102,960개의 질의응답쌍	- 510,442개의 문서
파일 타입	json	json [문서 내용: HTML 형태]	txt
적용 모델	LSTM Naver Chatbot	T5	KeyBERT

2.2
데이터 전처리

선수	성별	나라	종목	로고	올림픽 참가 기간	1 금	2 은	3 동	합계
마이클 펠프스	남성	미국	수영		2000년~2016년	26	3	2	31
라리사 라티니나	여성	소련	체조		1956년~1964년	9	5	4	18

각주 [편집]

1. ↑ [1], 기사 확인
2. ↑ Judith Swaddling (2000). 《The Ancient Olympic Games》 [고대 올림픽 경기] (영어). 텍사스 대학교 출판부(University of Texas Press).
3. ↑ Young (2004), p. 12
4. ↑ Pausanias, "Elis 1", VII, p. 7, 9, 10; Pindar, "Olympian 10", pp. 24–77

context	answer_text
1839년 바그너는 괴테의 파우스트을 처음 읽고 그 내용에 마음이 끌려 이를 소재로...	교향곡
1839년 바그너는 괴테의 파우스트을 처음 읽고 그 내용에 마음이 끌려 이를 소재로...	1악장
1839년 바그너는 괴테의 파우스트을 처음 읽고 그 내용에 마음이 끌려 이를 소재로...	베토벤의 교향곡 9번

01 표 제거
(모델을 활용한 정보탐색 어려움)

02 특수문자, 링크 등 제거

03 정보 탐색이 쉽도록
데이터프레임 형태로 정리

정제된
데이터셋

2.3.1 모델링

탈락 모델

LSTM

문제점 답변에 특정 문구가 반복적으로 나타남

현재까지 발견된 행성의 대부분은 어떤 행성인가?
2010년 7월 1일 kbs 1일 kbs 통해보던 화상이 수 있다

김유정은 드라마 메이퀸에서 어느 지역의 사투리 연기를 선보였을까?
2010년 7월 1일 kbs 1일 kbs 통해보던 화상이 수 있다

2010년 아시안 게임 대한민국 VS 파키스탄 전 승리한 투수는 누구인가?
2010년 7월 1일 kbs 1일 kbs 통해보던 화상이 그대로 안 될 위해

카카오미니의 모델명은 무엇인가요?
2010년 7월 1일 kbs 1일 kbs 14일 않으면 안 될 때 컴퓨터 기능

SBERT

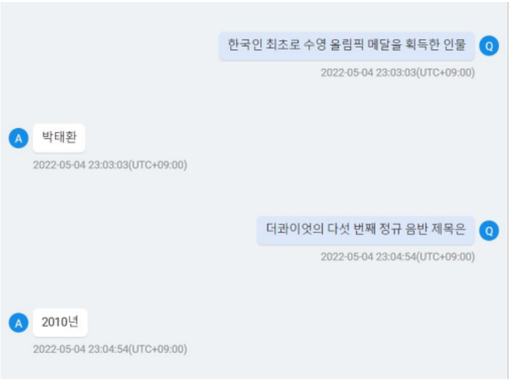
문제점 질문 전체를 임베딩하기 때문에 질문이 조금만 달라져도 전혀 다른 답변을 출력함

```
return_answer2('짱구는 못말려의 일본어 제목은?')
'제1대 미다이 도코로의 이름은 무엇인가?'

return_answer2('김유정이 메이퀸에서 사용한 사투리는?')
'박지윤'
```

Naver Chatbot

문제점 학습시키는 질문의 데이터가 많을수록 정확도가 떨어짐



2.3.1 모델링

최종 모델

T5

선정 이유: 구체적인 질문에 대한 답을 본문 내용(context)에서 찾아 줄 수 있음

```
val_df.iloc[16]
```

```
question                장기하와 얼굴들의 앨범 매진으로 첫날 긴급하게 추가제작한 앨범수량은 얼마인가?
context          《장기하와 얼굴들》은 발표 다음일에 주요 온라인 판매처인 YES24, 알라딘, 인터...
answer_text                                                  1만장
```

```
sample_question = val_df.iloc[16]
generate_answer(sample_question)
```

```
'1만장'
```

KeyBERT

선정 이유: SBERT 기반에 추가적으로 keyword를 추출하여 사용자 질문의 의도를 파악할 수 있음

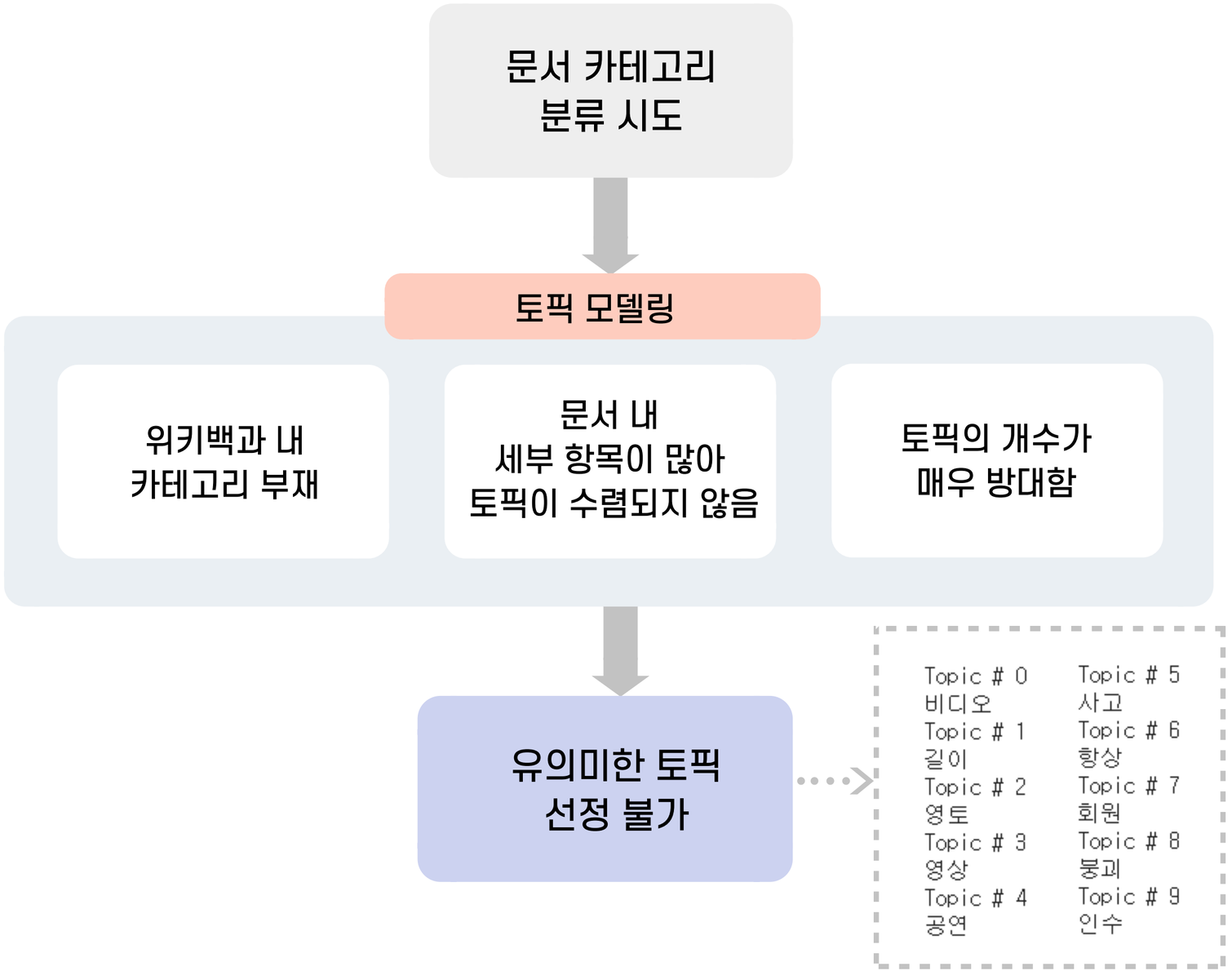
```
question = '지미 카터가 누구야?'
```

```
print(title)
print(subs_li)
```

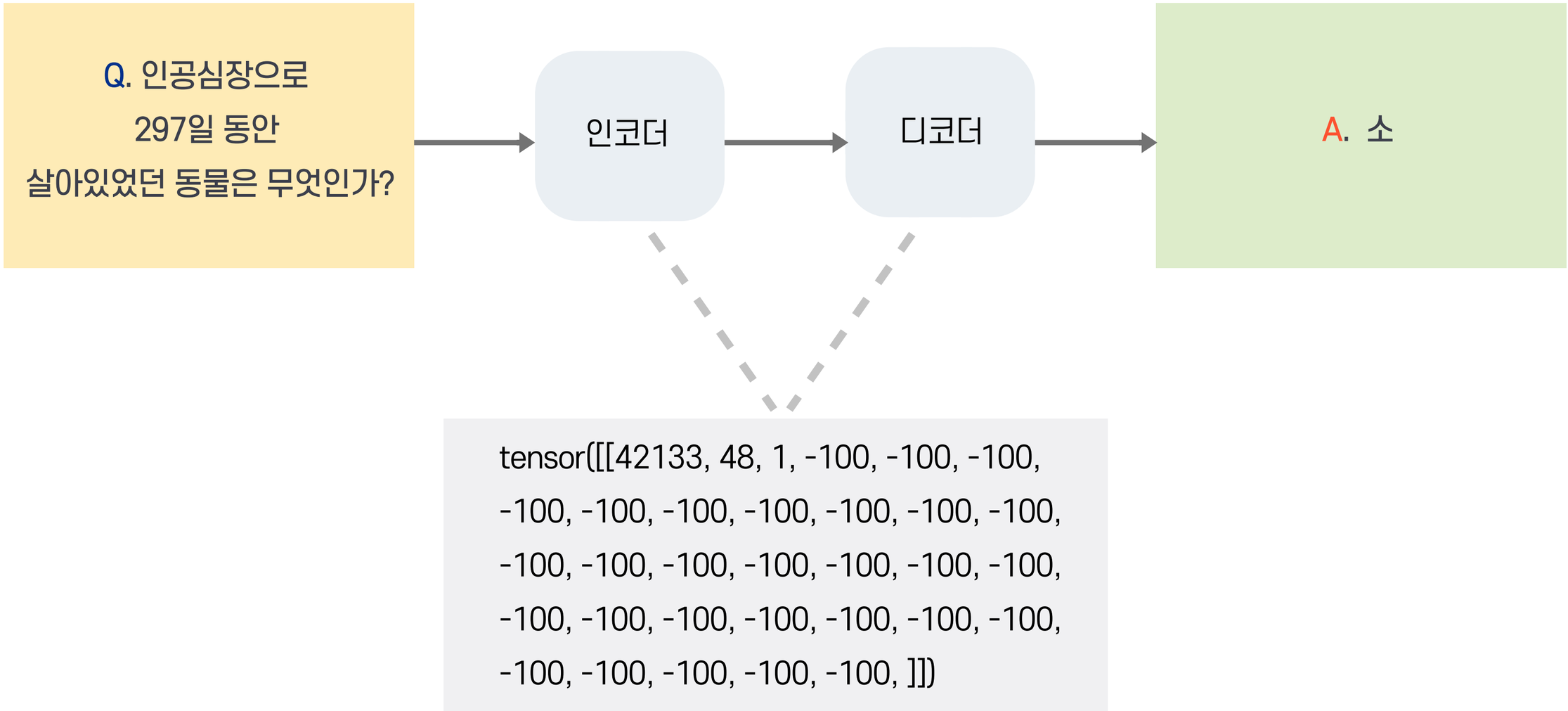
```
지미 카터
['기본', '어린 시절', '정계 입문', '대통령 재임', '외교 정책', '퇴임 이후', '평가']
```

```
'지미 카터는 조지아주 섬터 카운티 플레인스 마을에서 태어났다. 조지아 공과대학교를 졸업하였다. 그 후 해군에 들어가 전함 · 원자력 · 잠수함의 승무원으로 일하였다. 1953년 미국 해군 대위로 예편하였고 이후 땅콩 · 면화 등을 가꿔 많은 돈을 벌었다. 그의 별명이 "땅콩 농부" (Peanut Farmer)로 알려졌다.'
```

2.3.1 모델링


문서 카테고리
분류 시도
토픽 모델링
위키백과 내
카테고리 부재
문서 내
세부 항목이 많아
토픽이 수렴되지 않음
토픽의 개수가
매우 방대함
유의미한 토픽
선정 불가
Topic # 0
비디오
Topic # 1
길이
Topic # 2
영토
Topic # 3
영상
Topic # 4
공연
Topic # 5
사고
Topic # 6
항상
Topic # 7
회원
Topic # 8
붕괴
Topic # 9
인수

2.3.2
모델링 : T5


Q. 인공심장으로
297일 동안
살아있었던 동물은 무엇인가?
인코더
디코더
A. 소
tensor([[42133, 48, 1, -100, -100, -100,
-100, -100, -100, -100, -100, -100, -100,
-100, -100, -100, -100, -100, -100, -100,
-100, -100, -100, -100, -100, -100, -100,
-100, -100, -100, -100, -100,]])

2.3.2
모델링 : T5

Q. 00중학교의 2019년 졸업생 수는?

01 문서 제목 in 입력 질문

02 문서 질문 입력 질문 > 맨하튼 유사도

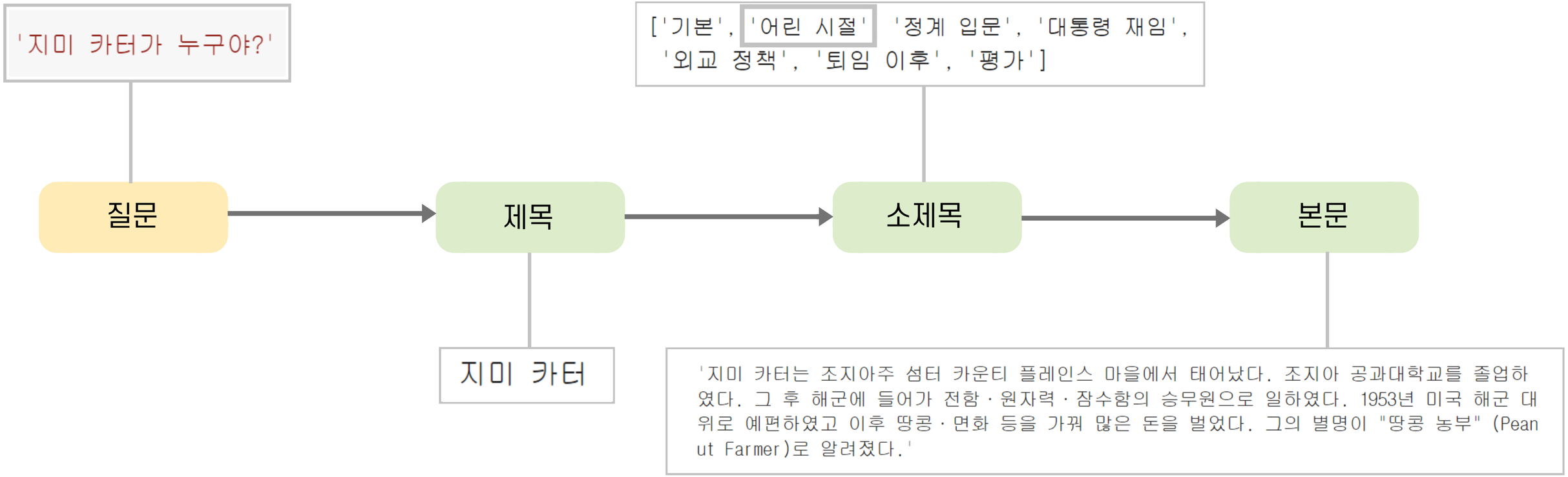
03 문서 본문 입력 질문 > 코사인 유사도

04 유사도 높은 본문 입력 질문 > 인코딩

제목	질문	본문
00중학교	2019년 졸업생 수는?	00중학교에서는 000명이 2019년에 졸업했다

A. 100명

2.3.2
모델링 : KeyBERT



2.3.2
모델링 : KeyBERT

질문 : 지미 카터가 누구야? → 제목 → 소제목 → 본문

01 질문 토큰화

```
[('지미', 'Noun'), ('카터', 'Noun'), ('가', 'Josa'),
 ('누구', 'Noun'), ('야', 'Josa'), ('?', 'Punctuation')]
```

02 키워드 추출

```
tokenized_nouns : 지미 카터
```

→ `['지미', '카터']`

03 후보 문서 선정

```
same # keys가 모두 포함된 문서 제목

[[0, '지미 카터'], [349165, 'USS 지미 카터']]
```

	title	desc
0	지미 카터	제임스 얼 카터 주니어(1924년 10월 1일 ~)는 민주당 출신 미국 39대 대...
1	USS 지미 카터	USS 지미 카터 (SSN-23)는 미국 해군의 시울프급 잠수함의 마지막 3번함이다.

04 코사인 유사도 계산

```
지미 카터 [[1.]]
USS 지미 카터 [[0.34059158 0.7622613  0.9666749  1.
  0.6993183 ]]
```

→ 지미 카터

2.4
성능평가

T5

question	answer_text	pred_a	result
송계월의 사망 당시 나이는?	23	39세	0
최정희의 여성작가 그룹 결성을 반대한 사람은 누구야?	송계월	송계월	1
곽규석이 연극배우로 첫 데뷔한 해는 언제일까요?	1948년	1948년	1
자일대우BC의 엔진 구동 방식은 어떻게 되는가?	후부 엔진-후륜 구동	후부 엔진-후륜 구동(RR) 방식	1

KeyBERT

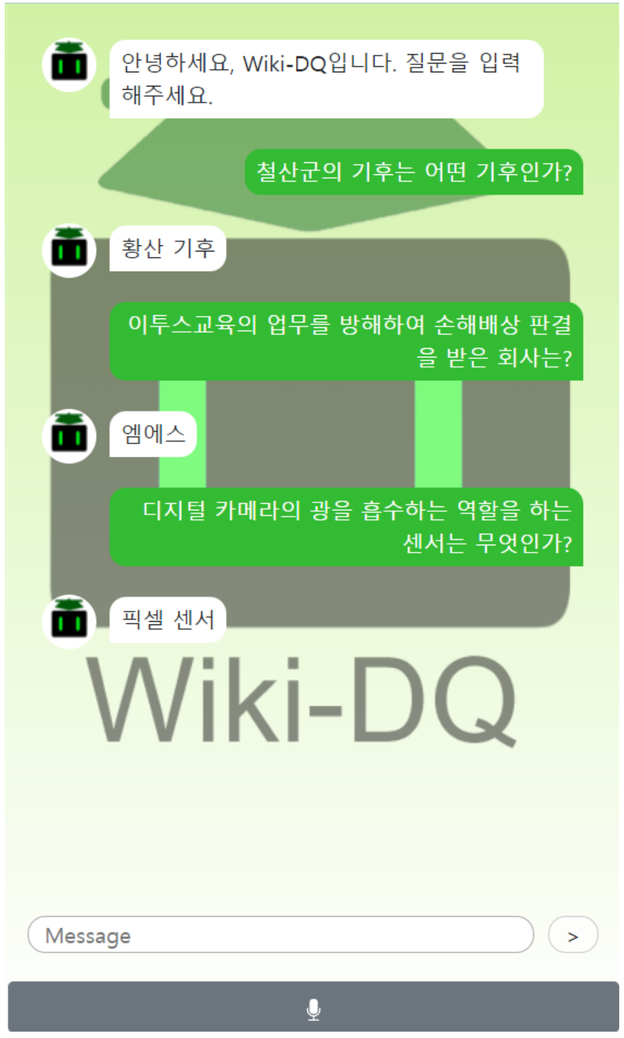
question	title	output	result
2016년 KBO 올스타전에 대해서 알아?	2016년 KBO 올스타전	밀리언 달러 콰르텟	0
JFN	JFN	JFN	1
안양 관양동 존속 살해 사건에 관해 알고 싶어	안양 관양동 존속 살해 사건	안양 관양동 존속 살해 사건	1
댓 터치 오브 밍크에 관련해 알고 있어?	댓 터치 오브 밍크	댓 터치 오브 밍크	1

2.4
성능평가

모델명	질문형태	성능평가 환경	속도	정확도
T5	**구체적인 질문** ex) 자일대우BC의 엔진 구동 방식은 어떻게 될까?	Colab Pro+	9,860(s) / 4798 ≒ 2.055(s)	(1486 / 4798) * 100 ≒ 30.97%
KeyBERT	**간단한 질문** ex) 지미 카터에 대해 알려줘	Colab Pro+	1,794(s) / 4798 ≒ 0.374(s)	(3363 / 1435) * 100 ≒ 70%

2.5
웹서비스 구현


안녕하세요, Wiki-Taka입니다. 질문을 입력 해주세요.
인공심장에 대해 알려줘
찾으신 인공 심장에 대한 정보입니다.
인공심장(人工心臟, artificial heart 또는 mechanical heart)은 심장의 펌프기능을 대신해주는 인공장기이다.
지미 카터에 대해 알려줘
찾으신 지미 카터에 대한 정보입니다.
Wiki-Taka
기본
어린 시절
정계 입문
대통령 재임
외교 정책
퇴임 이후
평가
Message
>


안녕하세요, Wiki-DQ입니다. 질문을 입력 해주세요.
철산군의 기후는 어떤 기후인가?
황산 기후
이투스교육의 업무를 방해하여 손해배상 판결을 받은 회사는?
엠에스
디지털 카메라의 광을 흡수하는 역할을 하는 센서는 무엇인가?
픽셀 센서
Wiki-DQ
Message
>

03

기대효과 및 향후과제

3.1 기대효과

3.2 보완할 점

3.3 후속 프로젝트

3.1
기대효과

위키백과 DB	- 위키백과는 **매우 다양한 분야의 정보**가 있어 활용도가 높음 ex) 초등학생 숙제도우미 초등학생에게 모르는 단어, 궁금한 점에 대한 답변 제공
NLP 모델	- 단순 검색으로는 찾기 어려운 **세부적인 질문에 대한 답변** 제공이 가능 ex) 드라마 '즐거운 우리집'에서 결혼 후 생계를 책임져온 사람은?
	- 문서 전체를 훑어볼 필요가 없어 정보 탐색 **시간과 노력 감소**
음성 인식	- 음성인식 기능이 있어 운전, 요리 등 **다른 활동중에도 사용가능**

3.2
보완할 점

고유명사를 인식하지 못해 알맞은 키워드를 만들어내지 못함

ex) 이성은 -> 이성(명사) 은(조사)

- 향후에 발전된 한국어 토크나이저를 이용하여 개선

표 데이터를 활용하지 못함

- 표에 관한 질문을 구분하는 기준을 더 모색
- 표 질문을 구분하게 되면 이미지나 데이터프레임으로 제공할 수 있을 것으로 보임

속도 및 정확도의 문제

- 추가적인 전처리 과정을 통해 데이터를 모델 학습에 더 적합한 형태로 만듦
- 토픽 모델링 외에 문서의 카테고리를 분류할 수 있는 방안을 모색
- 해당 문제를 해결할 수 있는 적절한 알고리즘 구현을 시도

3.3
후속 프로젝트

초등학생 검색용 챗봇

- 수업 중에 혹은 숙제를 할 때 모르는 내용을
질문해 해당 정보를 얻을 수 있다.
- 초등학생의 눈높이에 맞게 한자어 등을
쉬운 단어로 바꾸고, 부드러운 말투 탑재

노인 대상 음성인식 챗봇

- 작은 글씨를 읽기 어려운 노인을 위해
질문에 대한 답변을 잘 들리는 음성으로 제공할 수 있다.
- 노인음성 데이터를 활용하여 STT 인식율 개선

드라마 백과사전

- 검색으로 알기 어려운 특정 드라마의 디테일한
내용을 답변할 수 있다.
- DB의 범위를 드라마에 한정해 정확도 개선

선생님 챗봇

- 일상 회화 기능과 정보검색 기능을 합친
더 인간적인 챗봇의 구현

04

개발후기 및 느낀점

4
개발 후기 및 느낀점

박소연

자연어에 대해 잘 모르는 상태에서 시작했지만 LSTM,BERT와 같은 여러 모델을 사용해보고 자연어 처리에도 다양한 방법이 있다는 걸 공부하게 되었습니다. 열심히 공부하는 조원들과 함께했기에 부족하더라도 챗봇이라는 결과를 낼 수 있었습니다.

양동욱

이번 프로젝트를 하면서 전처리가 또 중요하다는 것을 알았습니다. 기획을 하고 어떤 모델링이 있고 정하는데 있어 다같이 많이 공부하면서 조금은 자연어처리를 알게되어서 매우 유익한 시간이였습니다. 모델링을 했을때 시간이 오래걸리고 결과가 만족스럽게 잘 나오지 않는 날이 많았습니다. 하지만 팀장님, 부팀장님, 팀원들 덕분에 프로젝트를 마무리 할 수 있었던것 같습니다. 감사합니다

이시내

이전 프로젝트에서 시도하지 못했던 언어모델들을 공부하고 사용해 보면서 많이 배울 수 있었고, 그 과정에서 좋은 답변을 주는 챗봇을 만드는 것이 매우 어려운 작업이라는 것을 몸소 느낄 수 있었던 시간이었습니다. 또한 혼자 공부한 것도 도움이 되었지만, 중간중간 생기는 문제들을 조원분들과 이야기하면서 하나씩 해결해 나갔던 과정이 저에게 더 큰 공부가 되었습니다.

최수민

이번에 처음으로 자연어 처리를 접하게 되어서 챗봇을 구현해봤는데 생각보다 많이 어려웠습니다. 어려운 주제를 가지고 했지만 조원 분들이 맡은 일을 잘 해주셔서 다행히 결과를 낼 수 있었습니다. 비록 만족스러운 결과를 내지는 못했지만 자연어 처리에 대해 공부할 수 있었던 좋은 시간이었습니다.

권태헌

여러 작은 프로젝트를 하면서도 그렇고 이번 최종 프로젝트에서도 느끼지만 항상 모든 프로젝트에는 해냈다는 성취감과 더 잘할 수 있었을텐데 하는 아쉬움이 뒤따르는 것 같다. 지금까지 배운 지식들을 총동원하고, 부족한 부분은 더욱 찾아가고 배우면서 점차 완성되어가는 결과물을 바라보는 것은 언제나 뿌듯하다. 텅 빈 맨땅에서의 막막한 기분으로 시작한 처음이 항상 떠오른다. 이번 프로젝트 역시 아쉬움과 뿌듯함이 많고, 이를 통해 또 한 걸음 성장한 기분이 든다. 마지막으로 한 달여간 함께 고생해준 팀원들께 감사한다는 말을 남기고 싶다.

유승희

파이널프로젝트를 하면서 넘지못할 것 같은 산을 여러번 만났던 기억이 난다. 그것들을 해결하려는 과정에서 많이 배웠고 팀원들에게 많은 도움을 받았다. 관심이 있던 자연어처리를 프로젝트 주제로 삼아 최신 언어모델을 사용해봤다는 것이 큰 의미가 있었다. 좀더 공부해서 이번 프로젝트에서 미흡했던 점들을 해결해보고 싶다.

조혜원

서비스 개발만 고민했던 지난 번과 달리 구현까지도 고려하면서, 실제 현업 프로젝트에서는 고려할 점이 훨씬 많겠다는 생각이 들었습니다. 또한 전처리, 알고리즘, 자연어 모델링 등 다양한 내용을 깊게 고민하고 공부하는 시간이었습니다. 처음에는 지식이 부족해 헤매기도 했지만 팀원분들과 함께 공부하며 문제를 해결할 수 있어 좋았고, 부팀장으로서 더 적극적으로 참여하려 노력한만큼 많은 것을 얻어가는 시간이었습니다

감사합니다.
Q&A