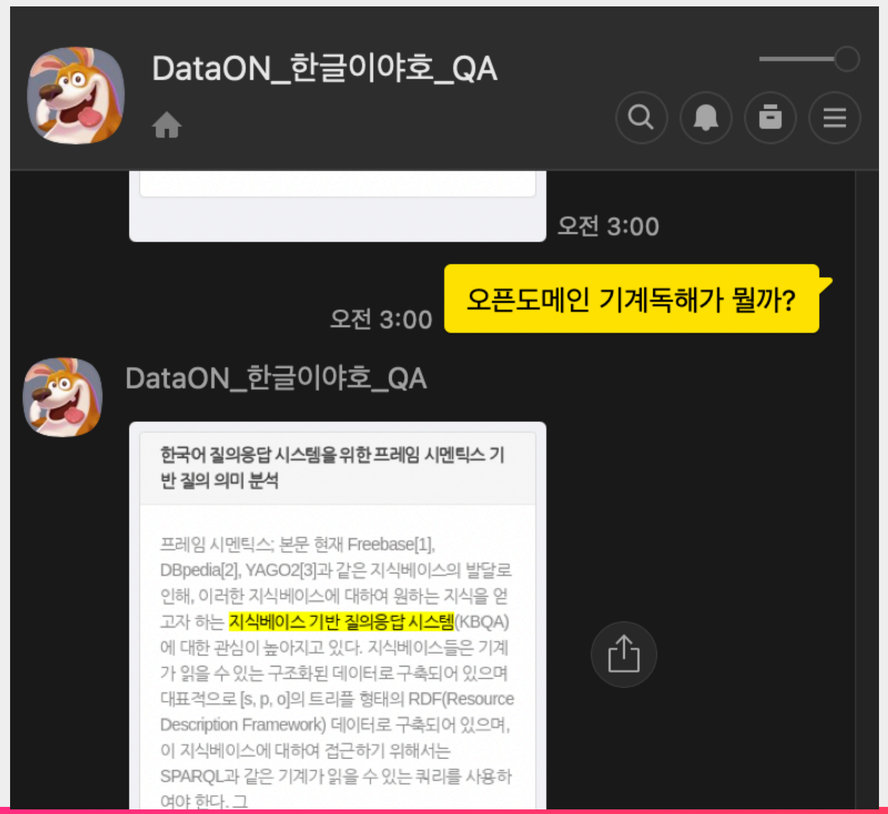
21’ 논문 연구데이터 분석 경진대회

한국어논문 오픈도메인 기계독해 수행을 위한 REST API기반 서버 구축



팀명 : 한글이 야호
발표 : 이경임
팀원 : 이지훈, 이경임, 유원준

목차

01
Overview

오픈도메인 기계독해란?

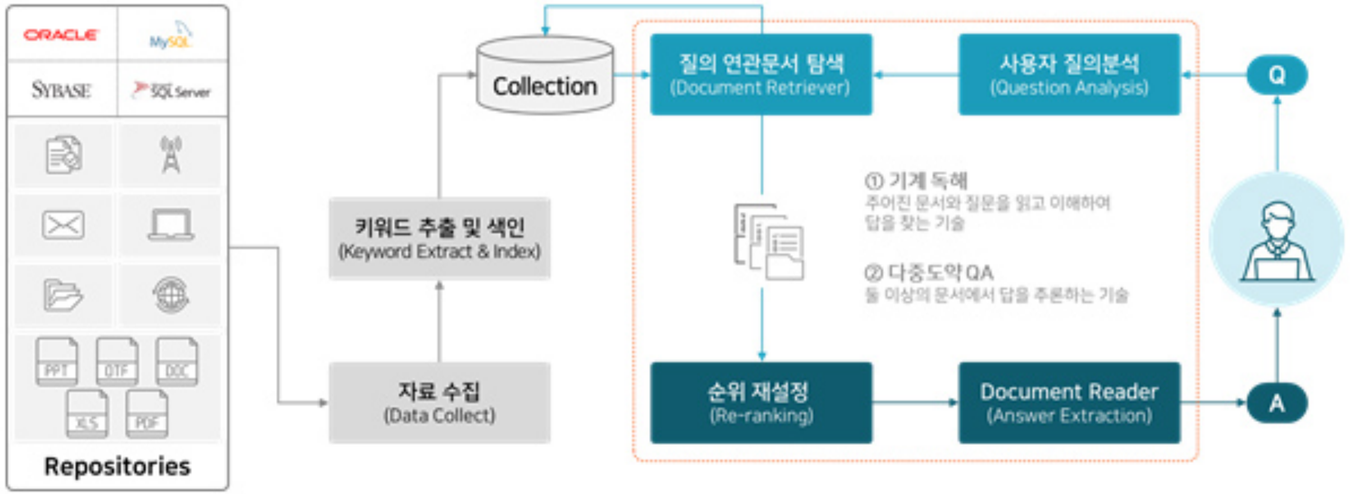
MRC(Machine Reading Comprehension)

- 자연어 질의와 주어진 대상문서에 대해 기계가 이해하고 자동으로 질의에 대한 정답을 찾아 대답하는 기술
- 영문 위키피디아 기반 제작된 SQuAD, 한글 위키피디아 기반 KorQuAD 데이터셋이 대표적
- KorQuAD 2.0의 경우 표, 리스트가 포함된 HTML문서상에서도 답 찾을 수 있도록 제작됨

ODQA(Open-Domain Question Answering)

- 지문이 따로 주어지지 않은 질의에 대해 지식베이스 기반 답변 찾아내는 Task
- 국내에서도 다양한 주제(위키데이터, 뉴스, 코로나 데이터 등)에 대한 오픈도메인 기계독해 질의응답 시스템 연구된 바 있음

오픈도메인 기계독해란?



[그림 1] 기계독해 QA 구성

ODQA 구성

- 질의 답변위한 정보 따로 주어지지 않으므로 사전 구축된 Knowledge resource 구축 필요
  - 일반적으로 뉴스, 지식백과, Wiki등 데이터 활용해 특정 주제에 대한 질의응답 시스템 구현
  - 본 시스템에서는 대회 데이터로 주어진 국내 논문 질의응답 QA 데이터셋 활용
- 기본적으로 (1) 질의 관련 문서검색기, (2) 정보에 대한 문서독해기로 이루어진다.

아이디어 개요

인공지능 논문검색 봇

- 73만개 논문 질의응답 데이터 기반, 국내 한국어 논문 정보 효과적으로 검색, 분석하기 위한 시스템 구축
- 한국어 논문 특화 모델 및 전처리 방식 적용

확장 가능한 기계독해 서버

- 논문 정보 활용성 증대 위해 다양한 형태의 클라이언트 서비스와 결합 가능한 시스템 제작.
- 질의 응답, 논문 검색 등 기계독해의 다양한 과정을 REST API형태로 호출하여 서비스에 결합 가능.
- 성능, 확장성 고려하여 Node기반 Restful API 서버와 인공지능 예측 수행 파이썬 Flask 서버로 별도 구성

다양한 형태의 응답 반환

- 질의 정답 텍스트, 이미지, 관련 논문정보 등 서비스 활용가능한 다양한 형태의 응답 반환
- 카카오 챗봇과 연동하여 검색된 정보에 대한 추가 정보 확인 가능한 시나리오 제작. 사용 확장성 증대

02
시스템 개요

개발 상세

Development stack

- Server : Node.js + Express(REST API서버), Flask(AI Inference서버)
- Search Engine : Elastic Search, Nori Analysis(Korean morphology)
- Model : KLUE RoBERTa, KoBigbird QA

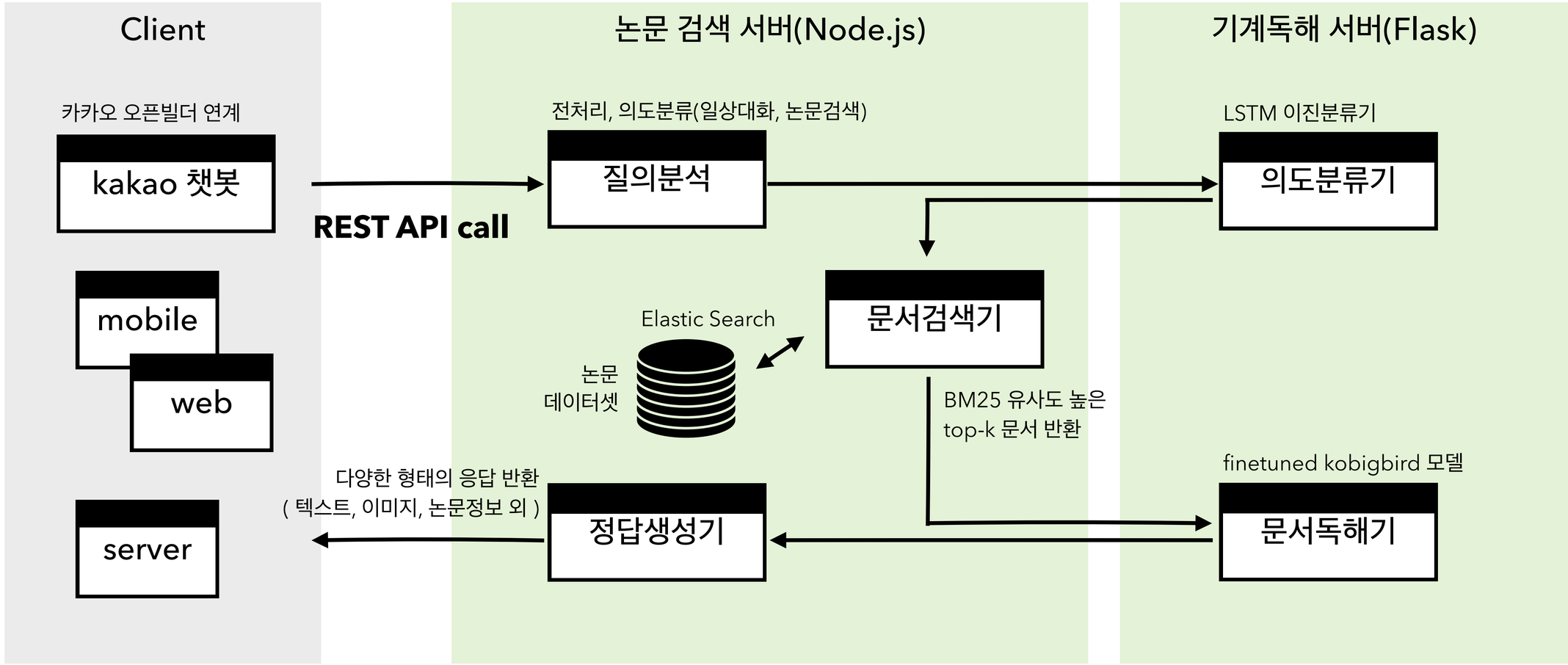
VM Resource

- Google Colab Pro Plus
- Cloud Server(GPU 2개, GPU 메모리 64GB)

Additional Dataset

- KISTI 국내 논문전문 텍스트 데이터셋
- KISTI 국내 논문 QA 데이터셋
- AiHub 감성대화 말뭉치 데이터셋

시스템 구조


Client
논문 검색 서버(Node.js)
기계독해 서버(Flask)
카카오 오픈빌더 연계
kakao 챗봇
mobile
web
server
REST API call
전처리, 의도분류(일상대화, 논문검색)
질의분석
LSTM 이진분류기
의도분류기
문서검색기
Elastic Search
논문
데이터셋
BM25 유사도 높은
top-k 문서 반환
finetuned kobigbird 모델
문서독해기
다양한 형태의 응답 반환
(텍스트, 이미지, 논문정보 외)
정답생성기

03
Retriever

의도분류기

자연어 질의 의도분류

- 사용자 질의에 대한 의도 분류 : 일상대화, 논문검색
- 감성대화 데이터셋과 논문 QA데이터셋의 질문 데이터셋 7만여개 활용해 학습
- 카카오 챗봇의 경우 응답 속도 향상을 위한 카카오 챗봇 기본 일상대화 분류기 적용

모델 구성

- Bert Tokenizer 사용
- LSTM 기반 이진 분류기로 제작

Search Engine

Elastic Search + BM25

- 국내 논문 전문 데이터셋 적재 후 Elastic Search 이용해 질의 관련 문서 검색
- 질의에 대해 문서 집합에서 BM25 유사도가 높은 상위 K개 문서 검색

Elastic Search + BM25 + Nori tokenizer

- 전처리 과정에 따라 문서 검색기 성능 달라짐
- 한국어 형태소분석기 Nori tokenizer 추가하여 사용자 질의에 대해 형태소 분석하여 명사 추출후 이를 이용해 인덱싱.

Multi-match index

- 질문만으로 조회시 검색성능 낮음
- 질문, 답변, 논문 원문에 대해 검색

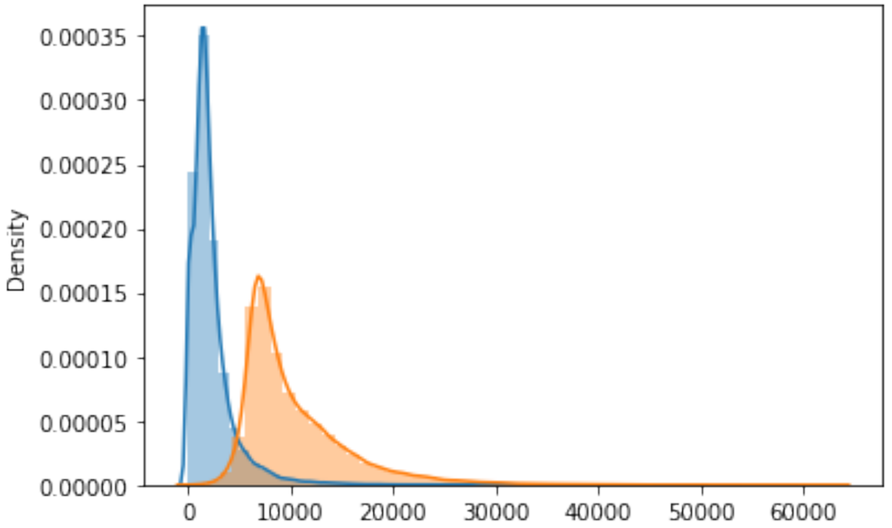
대상 Context / 정확도	top-1	top-2	top-3	top-5
question	44.17%	51.9%	55.5%	60.12%
context + question + answer	87.88%	91.00%	92.22%	93.48%

04
Reader

논문 데이터 전처리

전체 Context 기반 학습/예측 문제점

- 논문 길이 최대 6만자. 대부분 1만자 내외로 분포.
- 논문에서 답변 위치는 대부분 1만자 이전에 존재
- 대부분의 한국어 자연어 모델 Max seq len은 1024
  - 하나의 논문에 대해 답변 학습, 예측시 (논문길이 / max_seq_len) 만큼 반복 소요.



논문 컨텍스트 길이 조정

- 논문 원문에서 도표, 특수문자, 기호 제거 전처리 후 답변 구간 포함 1500자 추출.
- 답변 재조정 => 데이터셋의 답변문장과 실제 컨텍스트의 답변 문장 다른 케이스 다수 존재

```
curans = preprocess(curctx[int(ans_starts[idx]):int(ans_starts[idx])+len(answers[idx])])
curctx = preprocess(curctx)
ans_start = curctx.find(curans) - offset_st
ans_end = ans_start+len(curans)
```

KLUE RoBERTa

Model / Scoring	F1	MRC ROUGE-W	MRC-EM
XML-RoBERTa-base	81.55	53.93	27.48
KoBERT-base	84.58	58.54	48.28
koELECTRA-base	84.59	66.05	59.82
KLUE-BERT-base	85.73	68.51	62.32
KLUE-RoBERTa-base	85.07	73.98	68.67

- 21년 5월 발표된 RoBERTa 기반 한국어 자연어처리 모델
- KLUE 자체 벤치마크 점수 기준 기계독해 포함 대부분의 자연어 처리 태스크에서 우수한 성능 보임

Model	Embedding	Hidden sz	# Layers	# Heads
KLUE-RoBERTa-base	768	768	12	12

KoBigBird

Model / Dataset	TyDi QA	Korquad 2.1	Fake News	Modu Sentiment
KLUE-RoBERTa-base	76.80 / 78.58	55.44 / 73.02	95.20	42.61
KoBigBird-BERT-Base	79.13 / 81.30	67.77 / 82.03	98.85	45.43

- 21년 11월 11일 올라온 한국어 Bigbird 자연어처리 모델
- BigBird: Transformers for Longer Sequences에서 소개된 **sparse-attention** 기반의 모델로, 일반적인 BERT보다 **더 긴 sequence**를 다룰 수 있다.
- BERT의 8배인 최대 4096개의 token까지 다룰수 있다.
- 1024 이상의 Long Sequence에서 KLUE-RoBERTa보다 우수한 성능 보임
- 학습 느리고 모델 무거움. 최대 sequence 4096으로 학습하기 위해서는 하드웨어 고성능 필요

성능 향상

인공지능 연산 서버 분리, 연산속도 향상

- 웹 기반 TF.js 라이브러리, npm question-answering 라이브러리 등을 활용해 자바스크립트 생태계에서도 인공지능 연산 가능
- koBigBird 모델 특성상 연산 무거워 웹 라이브러리만으로 연산 어려움.
- 다양한 모델 로드해 Inference 연산 수행할 수 있는 별도 파이썬 서버 제작. 모델 연산시 내부적으로 파이썬 서버 호출하여 연산 시행

학습 설정(KoBigbird)

learning_rate : 3e-05
weight_decay : 2e-05
max_seq_len : 2048
doc_stride : 128

05
Answerer

답변생성기

Sentence Transformers

- 이미지, 텍스트에 대한 임베딩 생성하는 파이썬 프레임워크. 다국어 지원
- 생성된 임베딩을 활용해 문장간 코사인 유사도 등을 계산할 수 있다.
- 반환된 답변의 원문 논문 텍스트와 질문의 코사인 유사도를 계산해 실제 질문과 관련있는 논문인지 여부 확인가능
- Elastic Search 통해 반환된 논문 리스트에 대해 Sentence Transformer로 질문 <> 원문 간 유사도 비교하여 관련없는 문장은 제거 후 모델 예측에 넣을 수 있다.
- Long Sequence인 논문 원문에 대해 적용 어려움(연산시간)

Model output logit 비교

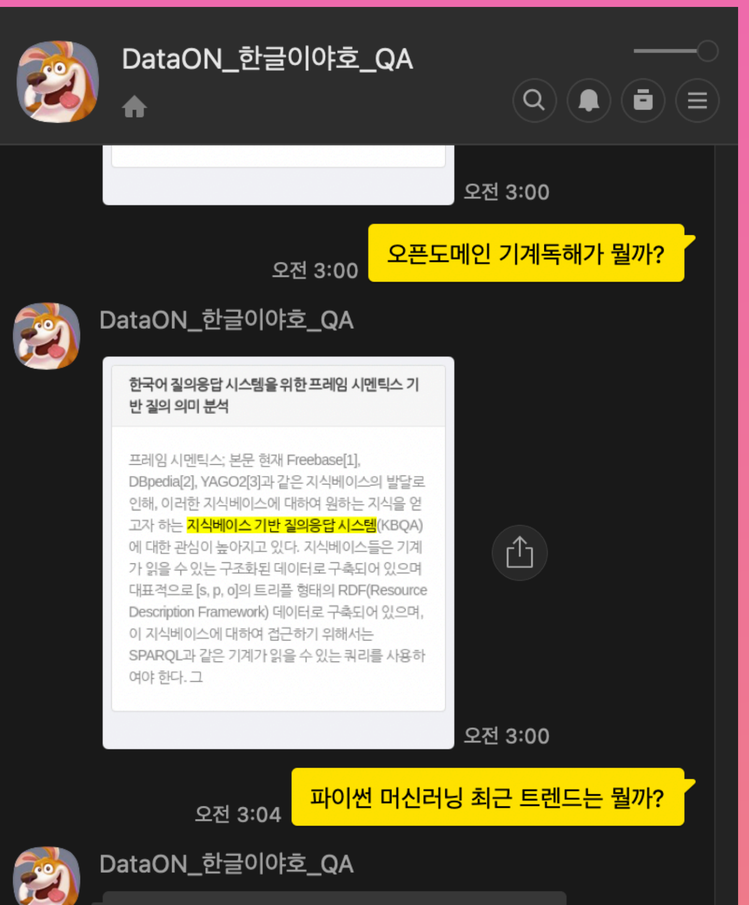
- max start logit과 max end logit을 softmax적용해 answer score로 저장. 추후 값 반환시 score 가장 높은 정답 반환

To-do Ideas

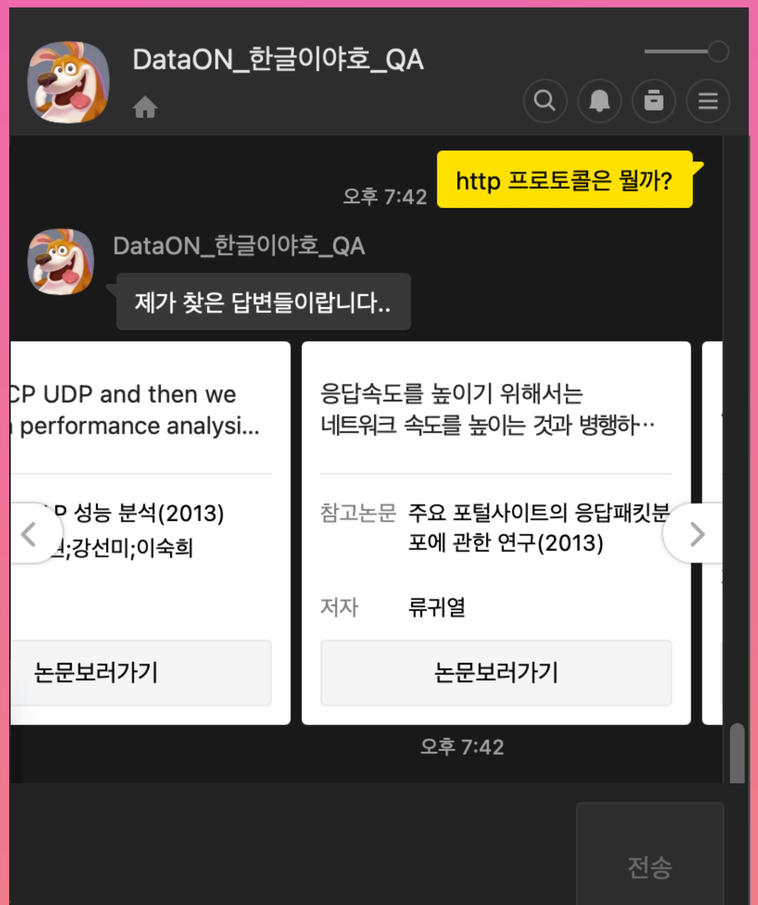
- 논문 정보와 결합해 검색, 답변 반환시 가중치 추가 : 피인용횟수, 논문 발표일, 저자 논문수 총합등

답변 출력 종류

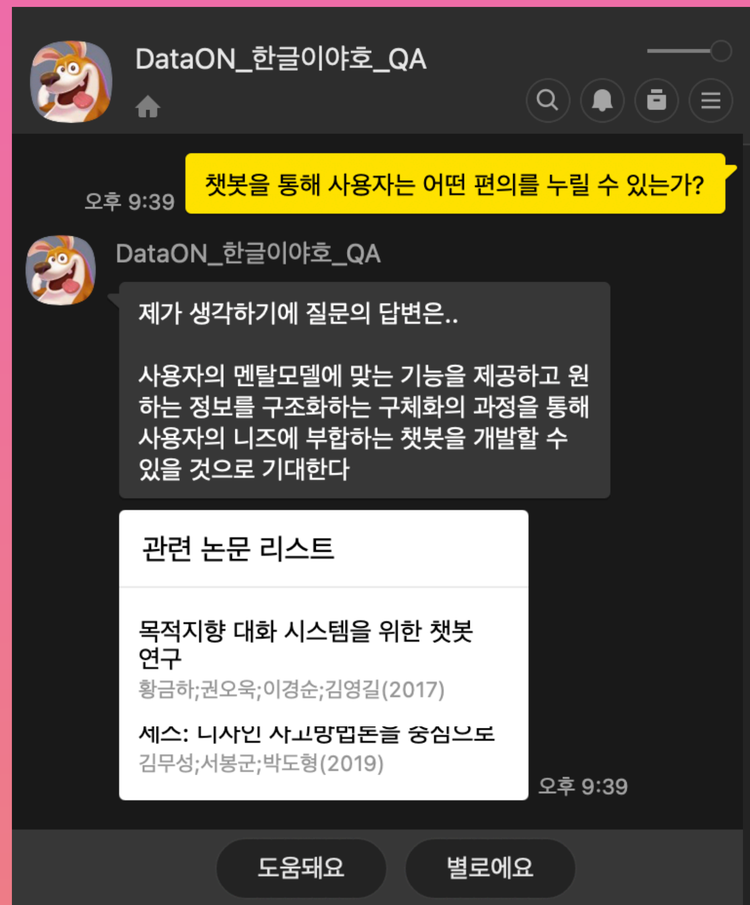
이미지형
유사도가 높은 답변 원문과 함께 이미지형태로 제공



논문정보형
Top 3답변과 각 답변 발췌된 논문 제공



기본형
가장 유사도가 높은 답변과 관련 논문리스트 제공



Demo

QnA

감사합니다