



말랑뉴스 로고

뉴스 요약 프로젝트

17기 홍여빈, 조성윤
18기 장원준

>> 목차

1 주제 설명

2 데이터셋

3 모델

4 시연

1

주제 설명

CHAOS!

“뉴스가 너무 많아!”

DIFFICULT!

“어려운 말은 어려워!”

I DON'T KNOW!

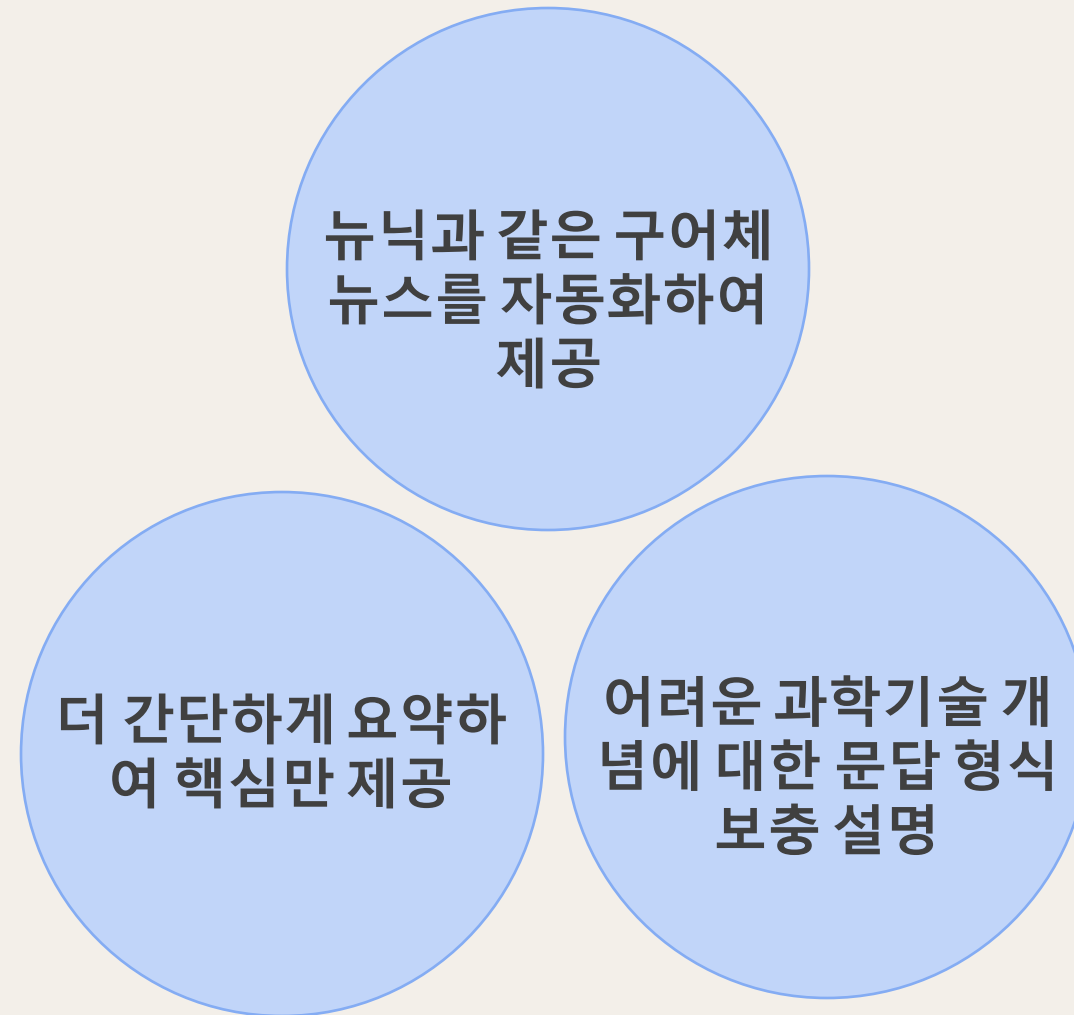
“앞뒤 맥락을 모르겠어!”

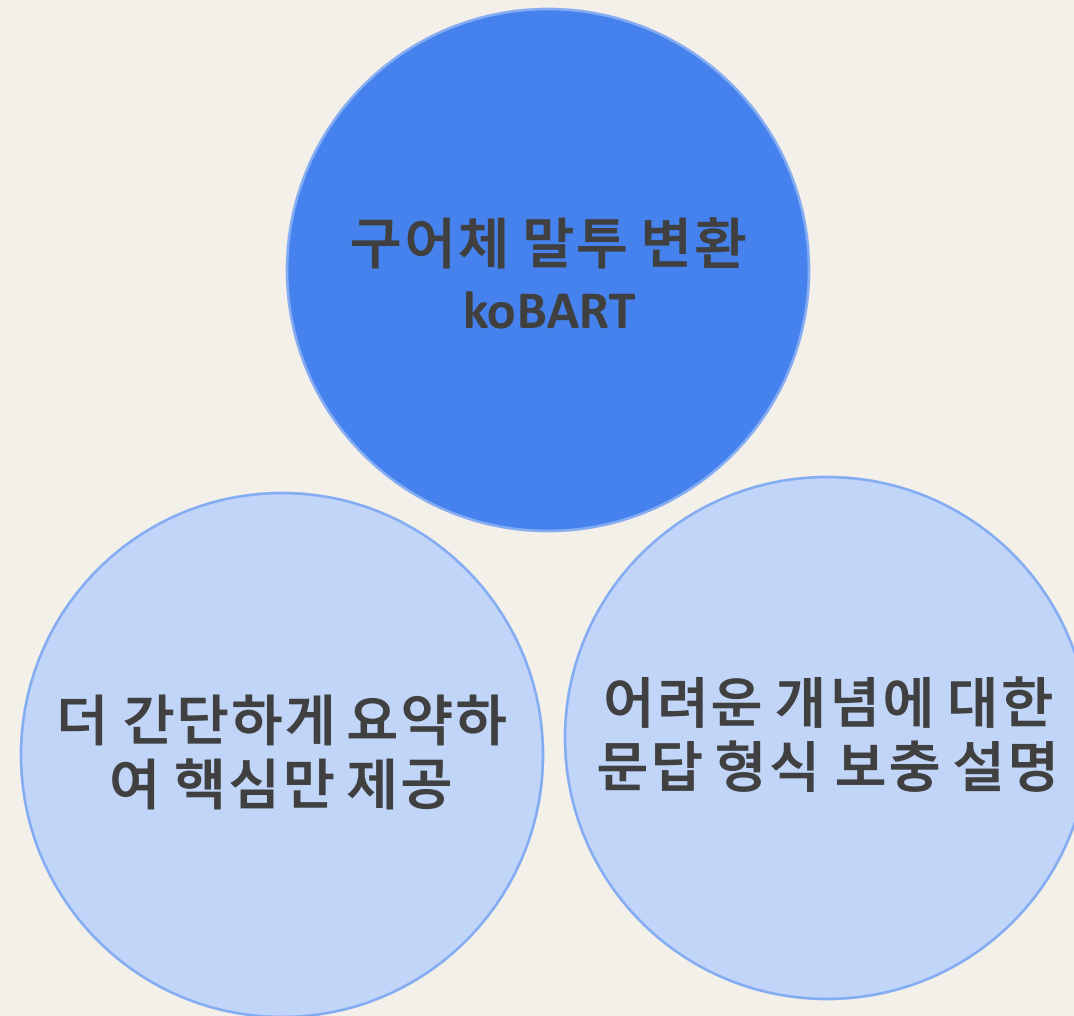
NEWNEEK

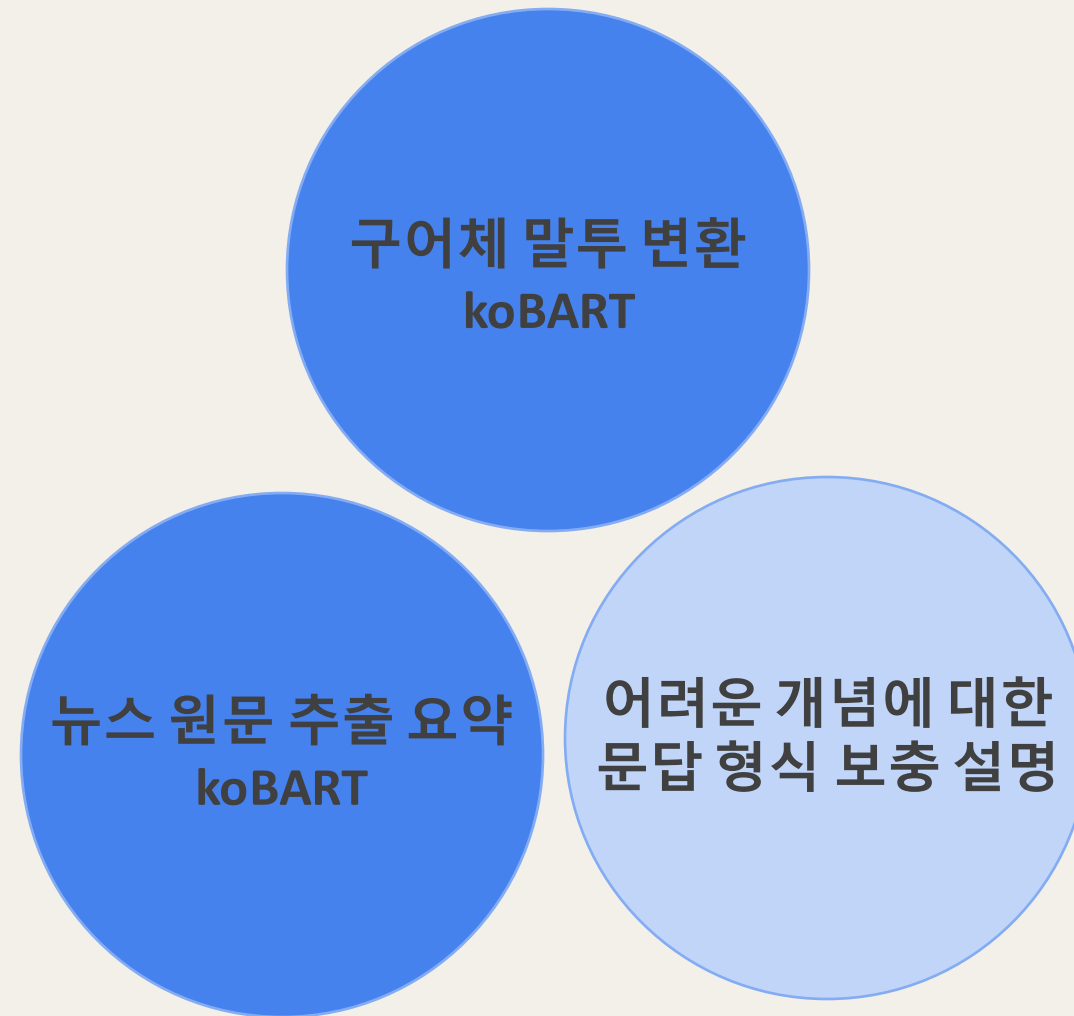


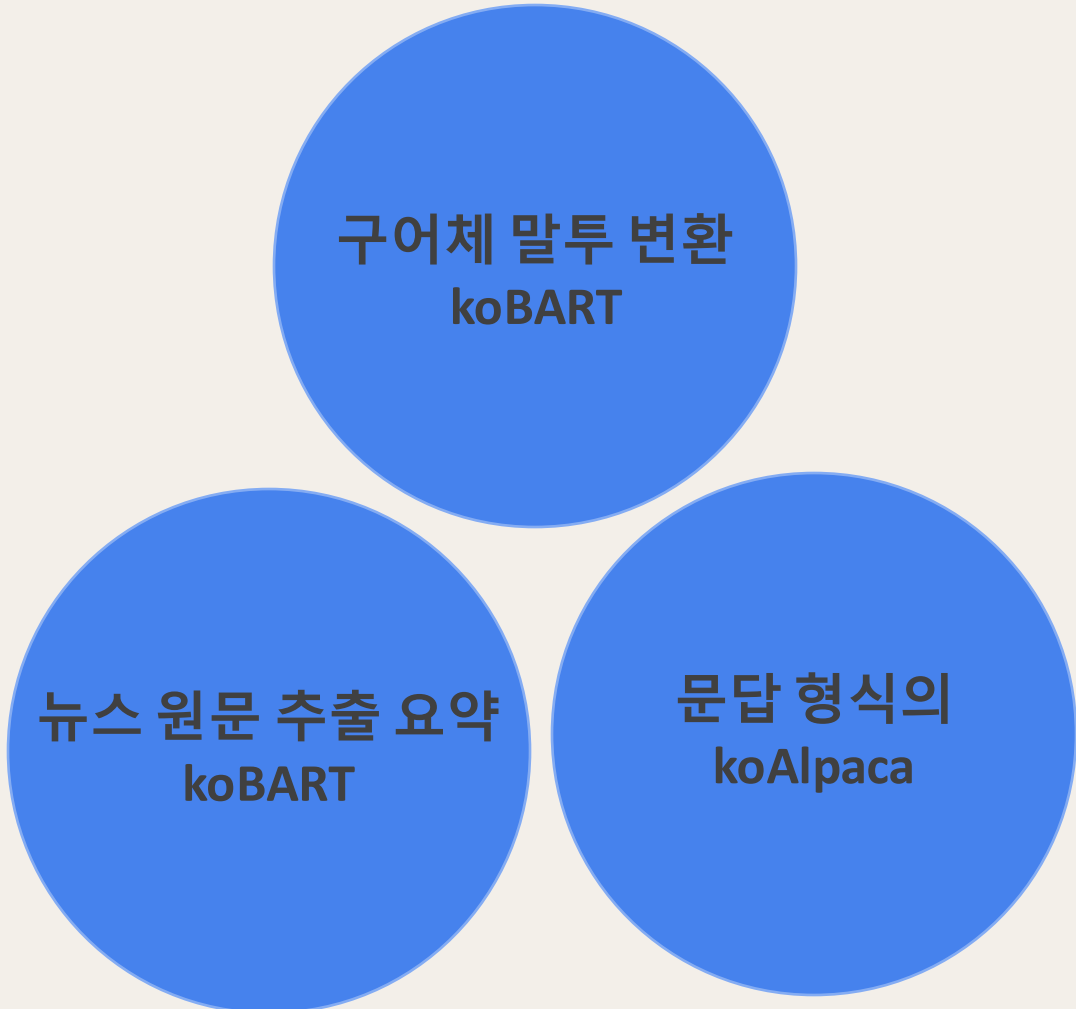
뉴닉은 꼭 알아야 할 이슈만 골라, 쉽고 재밌는 대화
처럼 풀어내는 뉴스레터 구독 서비스.

세 가지 핵심 기능









구어체 말투 변환
koBART


뉴스 원문 추출 요약
koBART

문답 형식의
koAlpaca

2

—

데이터셋

시사저널  구독



PICK ①

시사저널 헤드라인

박원순 前시장 모습, 손...유족 "반드시 색출"
🕒 1시간전

챗GPT 어디까지 써봤니? [김정희의 아하! 마케팅]

입력 2023.08.29. 오전 11:09 기사원문

 2  댓글

개인 영역뿐 아니라 비즈니스...
아임닭·베스킨라빈스 등 챗...

시작 장면: 도시의 아침 풍경
친구들과 얘기하며 즐기는 모...

"어떤 에너지를 원하십니까?"

장면 1:(탄산음료를 따르는 것
...

장면 2:(신선한 채소들의 근거리
돋보임) "...또는 지속적인 활...

장면 3:(트렌디한 주스 바에서
다운 그라데이션을 이루며, 무...

루를 시작하세요."

장면 4:(젊은이들이 채소 주스를 마시며 다양한 활동을 하는 모습. 스케이트보드를 타거나, 도서관에서 공부하거나, 피트니스센터에서 운동하는 장면 등) "진짜 에너지는 자연에서 옵니다."

마지막 장면 (도시의 아침 풍경. 젊은이들이 카페에서 주스를 마시며 대화를 나누는 장면 등)

 **본문 요약봇** ①

✕


자동 추출 기술로 요약된 내용입니다. 요약 기술의 특성상 본문의 주요 내용이 제외될 수 있어, 전체 맥락을 이해하기 위해서는 기사 본문 전체보기를 권장합니다.

챗GPT 어디까지 써봤니? [김정희의 아하! 마케팅]

닭가슴살 브랜드 아임닭이 챗GPT에 "아임닭 광고 콘티 짜줘, 재밌게.... 그리고 이 광고의 마케팅 타겟은 자취생, 아이 엄마, 직장인이야"라고 질문하고, 챗GPT가 답변한 대본대로 광고를 제작하는 모습을 담은 광고 영상으로 재미를 선사했다.

SPC 베스킨라빈스는 지난 4월 이달의 맛 '복숭아로 피치 올려' 출시를 위해 챗GPT에 '마이멜로디'와 '쿠로미'가 주인공인 동화 초안을 요청한 후, 이를 각색해 만든 광고를 공개해 화제를 모았다.

비용을 절감하면서도 영향력을 높일 수 있는 방법을 찾는 기업과 마케터에게 챗GPT는 방대한 데이터 분석으로 시장을 인간보다 더 빠르게 잘 이해해 인사이트를 제공하고, 더 적합한 콘텐츠를 만들어내는, 생산성과 수익성을 높일 수 있는 효율적인 마케팅 솔루션으로 인식되고 있다.

 네이버 뉴스추천

네이버뉴스 IT/과학 헤드라인 뉴스
원문 + 본문 요약봇(추출 요약)

"question": "Euler 방정식 및 Navier-Stokes' 방정식을 이용한 비정상 유동 문제를 적용하여 개선된 Interpolating Wavelet 기법 성능비교에 대한 결과는?"

"answer": "기존 수치 기법이 가지고 있던 공간 및 시간 정확도가 유지되면서, Shock-Vortex Interaction 문제의 경우 약 2배, Viscous Shock Tube 문제의 경우 약 1.7배 계산 효율이 향상됨을 확인하였다."

국내 논문 QA 데이터셋

기계가 과학기술 문헌을 읽고 이해하는 능력을 평가하기 위한 질의응답 데이터셋

3

모델

추출 요약

추출 요약은 원문에서 중요한 핵심 문장 또는 단어구를 몇 개 뽑아서 이들로 구성된 요약문을 만드는 방법. 3개의 문장을 추출해 요약

예시) 기사 제목 : 코스피, 강달러 지속에 1%대 하락...2430선 마감

>> 23일 코스피는 전 거래일 대비 27.16포인트(1.10%) 내린 2435.34에 거래를 마쳤다.

이날 13.19포인트 하락 출발한 지수는 장 초반 2440선 위쪽에서 움직였지만 오후 2시를 기점으로 낙폭을 확대하면서 1% 넘게 떨어졌다.

원·달러 환율이 강세 흐름을 이어가면서 시장 전반의 투자심리가 위축됐다.

생성 요약

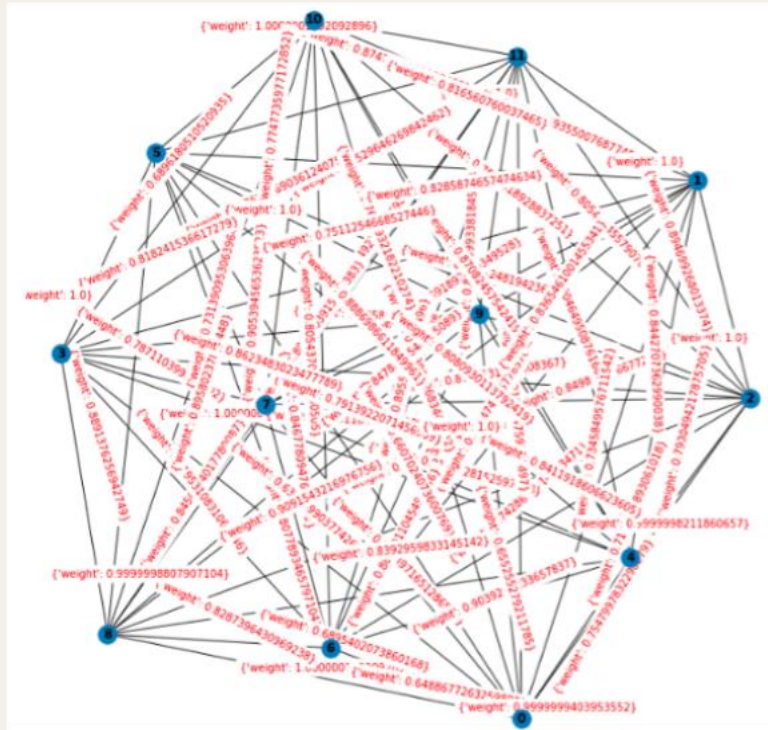
추상적 요약(생성 요약)은 원문에 없던 문장이라도 핵심 문맥을 반영한 새로운 문장을 생성해서 원문을 요약하는 방법.

예시) 기사 제목 : 코스피, 강달러 지속에 1%대 하락...2430선 마감

>> 코스피가 전 거래일보다 27.16포인트(1.10%) 내린 2435.34에 거래를 마쳤으며 원·달러 환율이 강세 흐름을 이어가면서 시장 전반의 투자심리가 위축되어 원·달러 환율은 전 거래일보다 5.7원 오른 1345.5원에 거래를 마쳤다.

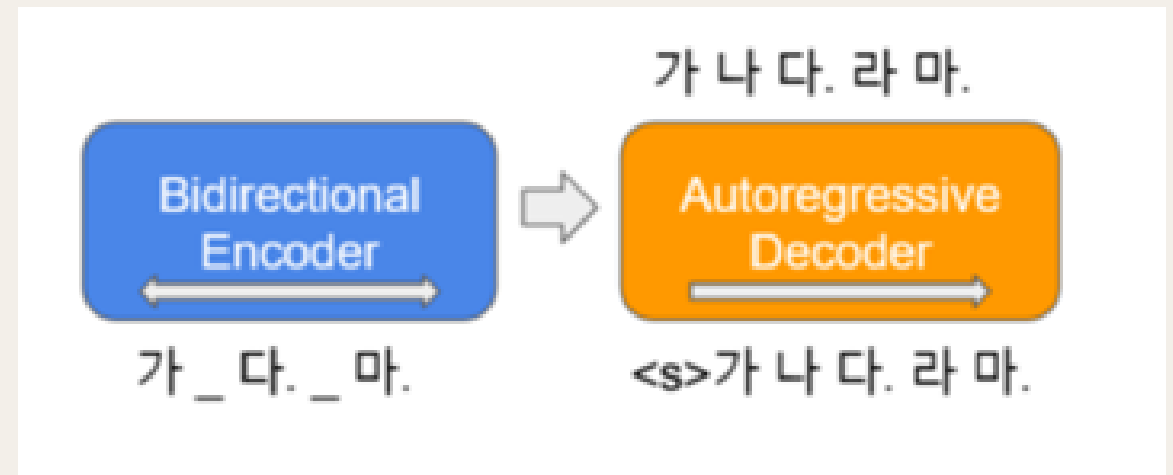
뉴스 원문의 의미를 해치지 않기 위해 **추출 요약** 진행

TextRank 기반 추출 요약



$$S(V_i) = (1 - d) + d * \sum_{j \in In(V_i)} \frac{1}{|Out(V_j)|} S(V_j)$$

KOBART 기반 추출 요약



TextRank 기반 추출요약

- 그래프 기반의 랭킹 모델로 샘플의 한 문장마다 순위를 매겨서 요약문을 추출하는 방법
- 문장과 문장 사이의 가중치를 구하고 텍스트랭크 점수를 계산하여 내림차순으로 정렬, 상위 n개의 문장으로 요약
- TF-IDF로 문장의 중요도를 계산, 중요도가 높은 n개의 문장으로 요약
- 정답 요약문을 보고 학습하는 것이 아니라는 한계

KOBART 기반 추출요약

BERT와 GPT가 합쳐진 형태의 모델로, 입력 텍스트 일부에 노이즈를 추가하여 이를 다시 원문으로 복구하는 AutoEncoder 형태로 학습

KoBART는 BART의 text infilling 노이즈 함수를 이용하여 한국어 텍스트에 대해서 학습한 encoder-decoder 언어 모델

KOBART 기반 추출요약

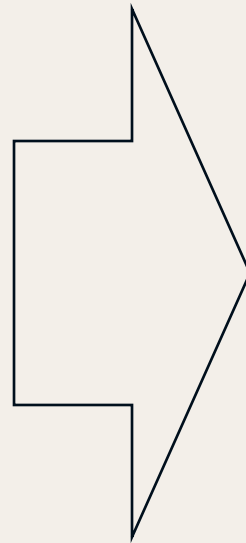
{'passage': '더불어민주당이 뉴딜 펀드 띄우기를 시작했다.Wn 이해찬 대표를 비롯한 민주당 지도부는 5일 서울 여의도 한국거래소에서 현장간담회를 열고 민간 참여를 당부했다.Wn 이 대표는 인사말에서 "(한국형 뉴딜 사업에) 2025년까지 약 160조원이 투자될 거라고 보는데 상당 부분은 정부 재정이지만, 민간 부분에서도 10% 가까이 투자를 하는 구상을 하고 있다"고 말했다.Wn 문재인 정부의 임기 말 중점 사업들에 뉴딜 펀드로 민간 자금 16조원 이상을 끌어오겠다는 것이다.Wn 민주당은 이날 간담회에 기획재정부, 금융위원회 등 정부 부처와 금융투자협회, 은행연합회, 경영자총연합회, 신용보증기금 등 금융권 주요 협회 관계자를 두루 불렀다.Wn 최영권 우리자산운용 대표, 김재익 KDB인프라자산운용 대표 등 펀드 운용사 대표들도 나와 뉴딜 펀드의 구체적 설계 방향을 제시하고 토론에 참여했다.Wn "개인뿐 아니라 연기금, 기관투자자 모두에게 좋은 '1석 3조' 펀드"(홍성국 민주당 의원)라는 메시지를 강조하려는 의도다.Wn 이호형 은행연합회 전무이사는 간담회 직후 통화에서 "부동산 대책 이후에 이런 (그린·디지털 뉴딜 등) 신산업 쪽으로 물꼬를 틀겠다는 상징적 의미의 행사였다"고 했다.Wn ……

'summary': '비공개 토론에서 뉴딜 펀드에 대한 몇 가지 의문이 제기됐는데 전례 없는 형태의 관제 인프라 펀드라 시중 유동성이 큰 지금 같은 상황이 유지될지가 논의의 핵심이었다.'



'pred': ['더불어민주당 지도부는 5일 서울 여의도 한국거래소에서 열린 한석, 등과 함께 함께 같이 함께 하는 구상을 하고 있다....., 1 1 석 석 으로 만들어졌다 다시 제로를가 시작했다 시작하면서 하며 또한 이러한 다양한 형태의 해가 해가 대통령 선거에 보조주가 들을 적인 요소가 요소 것들을 들이 것들도 일어나면 일어나고 일어날 일어났다 일어날 가능성이 판단이 하기 ']]

키워드 추출
KeyBERT

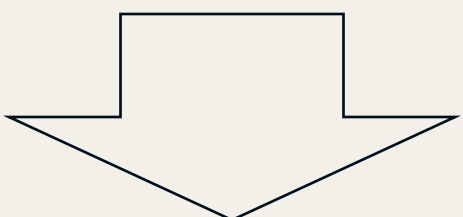


키워드 설명
koAlpaca



Koalpaca instruction tuning

"question": "..."
"answer": "..."
국내 논문 QA 데이터셋



아래의 template을 이용하여 instruction tuning

```
""" 질문: {x[ 'instruction' ] }\n\n""" 답변:\n{x[ 'output' ] }<|endoftext|>"""
```

- Instruction Tuning은 instruction으로 표현된 NLP 작업을 수행하는 방법을 모델에 가르쳐 자연어를 처리하고 이해하는 능력을 향상시킴.
- GPT-3.5(text-davinci-003), ChatGPT, Claude 및 Bing Chat과 같은 instruction-following model들이 많아짐.
- 질의응답 데이터를 활용하여 instruction tuning시키면 모델의 추론 성능 뿐만 아니라 사용자 입장에서 질의응답 양식으로 이해를 도울 수 있을 것으로 기대함.

Koalpaca instruction tuning using LoRA

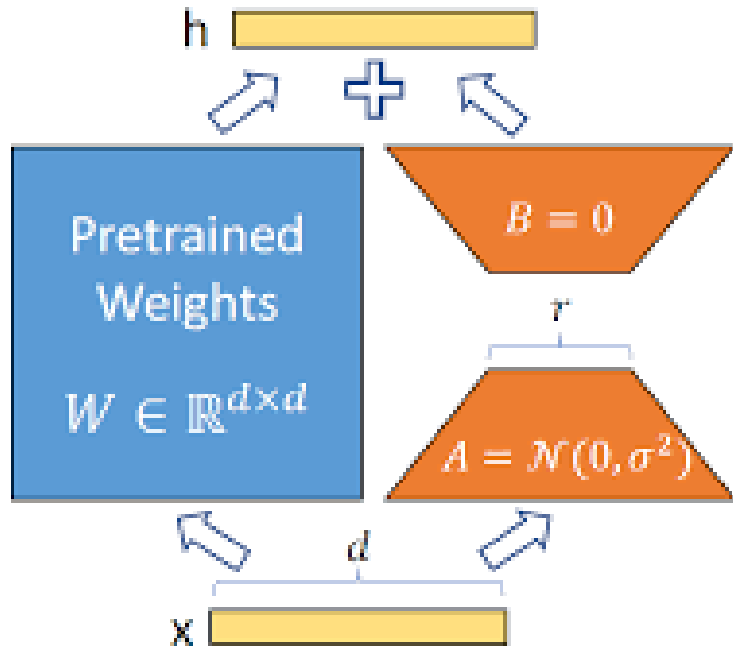


Figure 1: Our reparametrization. We only train A and B .

- Low-rank decomposition을 통해 두 개의 더 작은 행렬을 사용하여 가중치 업데이트.
- 전체 변경 횟수를 낮게 유지하면서 새로운 데이터에 적응하도록 훈련. 원래 가중치 행렬은 고정된 상태로 유지.
- 최종 결과를 생성하려면 원래 가중치와 조정된 가중치가 모두 결합
- LoRA는 훈련 가능한 매개변수의 수를 줄여 제한된 GPU 환경에서 koAlpaca를 fine-tuning할 수 있었음.
- LoRA를 사용하여 미세 조정된 모델의 성능은 완전히 미세 조정된 모델의 성능과 비슷함.

korean SmileStyle Dataset



SmileStyle:
Parallel Style-variant Corpus for
Korean Multi-turn Chat Text Dataset

Smilegate AI에서 공개하는 한국어 문체 스타일 변환 "SmileStyle" 데이터셋입니다.

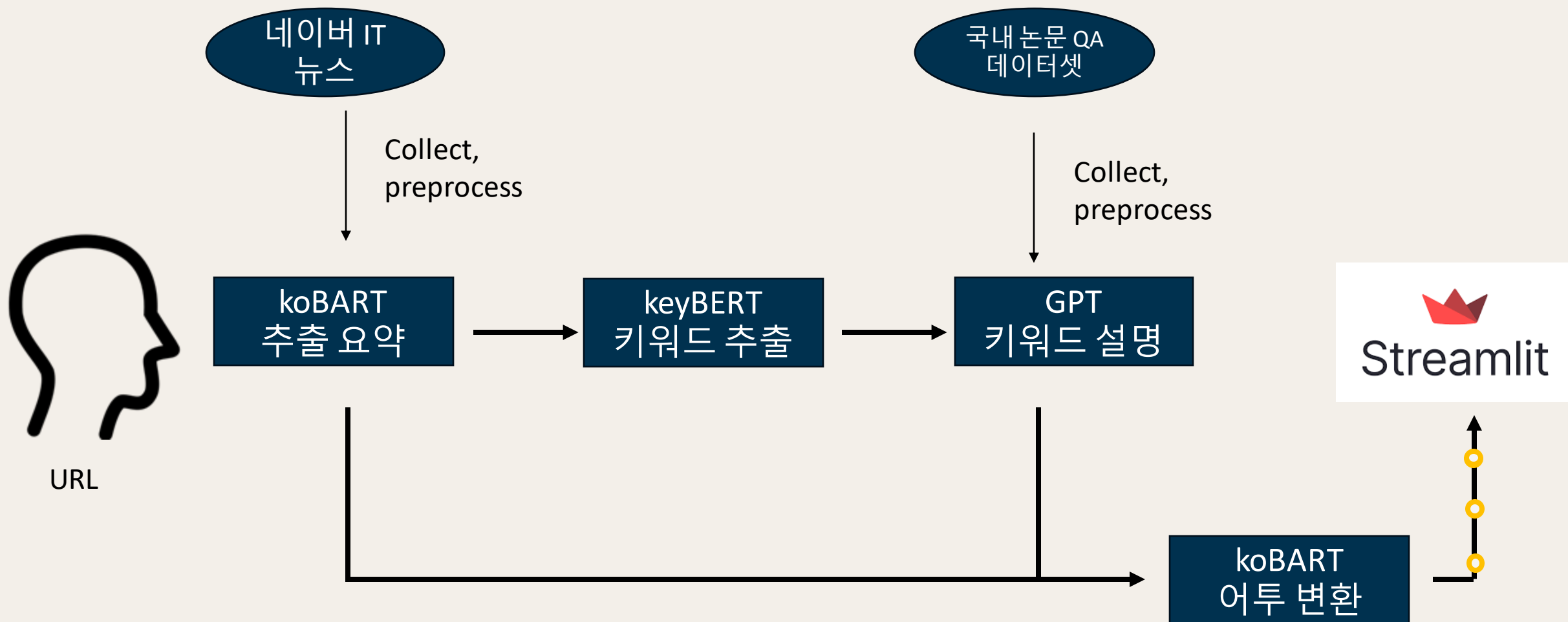
- 한국어 문체 스타일 변환 SmileStyle 데이터셋으로 학습한 pretrained koBART 사용
- 사용자 customizing이 가능한 말투 변환 기능 제공

- [formal]
존댓말 스타일입니다. (e.g. 나이가 어떻게 되시는데요?)
- [informal]
반말 스타일입니다. (e.g. 몇 살이야?)
- [android]
로봇의 답변 스타일입니다. (e.g. 휴먼. 나이. 무엇.)
- [azae]
연장자 스타일입니다. (e.g. 거 나이가 어떻게 되나?)
- [chat]
챗봇의 채팅 스타일입니다. (e.g. 몇 살임?)
- [choding]
어린이 말투 스타일입니다. (e.g. 몇 살임?)
- [emoticon]
반말 스타일에 이모티콘을 부착하였습니다. (e.g. 몇 살이야???? (´ω`)?)
- [enfp]
외향적인 스타일입니다. (e.g. 올해 몇살이양~?!?)

4

시연

말랑 뉴스



시연 예시

말투 선택

고급 기능 🌸

어떤 말투로 설명해드릴까요?

원문



원문

구어체

나루토

enfp

추가 질문 정도

Keyword를 뽑는 threshold

Keyword는 0~1의 중요도를 가짐
Threshold = 1 - 설정한 추가 질문 정도
0에 가까울수록 질문하지 않음
1에 가까울수록 질문을 많이 함
(0.4-0.6 구간 추천)

물론 많이 물어볼수록 대기 시간 증가..

추가 질문은 얼마나 할까요?

0.50

0.00

1.00

시연 예시

말랑뉴스 🧠

뉴스를 풀어서 설명해드려요.

URL:

<https://n.news.naver.com/mnews/article/020/0003517617?sid=105>

Submit

“틱톡, 인스타 제치고 페북 추격...2년뒤 美 최대 SNS 될 것”

원문: 입력2023.08.29. 수정2023.08.29. 지민구 기자
말랑뉴스에 의해 요약된 뉴스입니다.

29은 정보기술 업계에 따르면 미국 시장조사기관 인사이더 인텔리전스는 23일 틱톡이 2025년에는 미국 18세 이상 성인이 하루에 틱톡을 사용하는시간이 총 51억9000만 분으로 페이스북을 제칠 것으로 예상했으며, 틱톡의 미국 내 이용 시간은 2020년 58억200만 분으로 가장 높았지만 점차 감소하는 추세로 틱톡이 2025년 최대 소셜네트워크서비스로 성장할 것이라는 전망이 나왔다.

페이스북에 대해 설명해줘.

🧠 페이스북은 전 세계에서 가장 인기 있는 소셜네트워크서비스 중 하나로 알려져 있습니다. 사람들은 페이스북을 사용하여 친구들과 가족들과 소통하고, 사진과 동영상을 공유하며, 뉴스와 정보를 얻기도 합니다. 이 뉴스 기사에서의 페이스북은 기존에 소셜네트워크서비스 시장을 선도하고 있었지만, 틱톡이라는 중국 플랫폼의 성장으로 인해 그 지위가 위협받고 있다는 의미입니다.

문서에서 추출된 키워드: 페이스북, 소셜, 성인, 네트워크, 시간, 미국, 정보, 최대, 기술, 이용
추출된 키워드에 대해 설명이 필요하다면 추가 질문 정도를 높여보세요!

입력한 URL

제목

날짜 및 기자

요약된 본문

KoBART

키워드 질문


GPT3.5

키워드에 대한 답변

후보 키워드

KeyBERT

시연 예시

말랑뉴스 

뉴스를 풀어서 설명해드려요.

원문

URL:

<https://n.news.naver.com/mnews/article/020/0003517617?sid=105>


Submit

“틱톡, 인스타 제치고 페이스북 추격...2년뒤 美 최대 SNS 될 것”


원문: 입력2023.08.29. 수정2023.08.29. 지민구 기자
말랑뉴스에 의해 요약된 뉴스입니다.

29은 정보기술 업계에 따르면 미국 시장조사기관 인사이더 인텔리전스는 23일 틱톡이 2025년에는 미국 18세 이상 성인이 하루에 틱톡을 사용하는시간이 총 51억9000만 분으로 페이스북을 제칠 것으로 예상했으며, 틱톡의 미국 내 이용 시간은 2020년 58억200만 분으로 가장 높았지만 점차 감소하는 추세로 틱톡이 2025년 최대 소셜네트워크서비스로 성장할 것이라는 전망이 나왔다.

페이스북에 대해 설명해줘.

 페이스북은 전 세계에서 가장 인기 있는 소셜네트워크서비스 중 하나로 알려져 있습니다. 사람들은 페이스북을 사용하여 친구들과 가족들과 소통하고, 사진과 동영상을 공유하며, 뉴스와 정보를 얻기도 합니다. 이 뉴스 기사에서의 페이스북은 기존에 소셜네트워크서비스 시장을 선도하고 있었지만, 틱톡이라는 중국 플랫폼의 성장으로 인해 그 지위가 위협받고 있다는 의미입니다.

문서에서 추출된 키워드: 페이스북, 소셜, 성인, 네트워크, 시간, 미국, 정보, 최대, 기술, 이용
추출된 키워드에 대해 설명이 필요하다면 추가 질문 정도를 높여보세요!

말랑뉴스 

뉴스를 풀어서 설명해드려요.

구어체

URL:

<https://n.news.naver.com/mnews/article/020/0003517617?sid=105>


Submit

“틱톡, 인스타 제치고 페이스북 추격...2년뒤 美 최대 SNS 될 것”


원문: 입력2023.08.29. 수정2023.08.29. 지민구 기자
말랑뉴스에 의해 요약된 뉴스입니다.

2020년까지 정보기술업계에서 가장 영향력이 큰 시장조사기관인 미국 시장조사기관 인텔리전스는 틱톡이 2025년까지 미국 18세 이상 성인이 하루에 틱톡을 사용하는시간이 총 51억9000만 분으로 페이스북을 제칠 것으로 예측했어. 틱톡의 미국 내 이용 시간은 2020년 58억200만 분으로 가장 높았지만, 이후 감소하는 추세로 틱톡이 2025년 최대 소셜네트워크 서비스로 성장할 것이라는 예측이 나왔어.

소셜...?

 소셜의의 의미는, 사람들이 관계를 형성하고 유지하는 과정과 그 결과물을 의미해. 사람들이 서로 간의 상호작용과 소통을 통해 형성되는 네트워크야. 이 뉴스에서 소셜네트워크서비스라는 용어를 사용해서, 인터넷을 통해 사람들이 서로 연결되고 소통하는 플랫폼을 의미해.

페이스북은(는) 뭐야?

 페이스북의 가장 큰 의미로, 세계 최고의 소셜네트워크 서비스 중 하나야. 2004년에 설립된 소셜미디어 플랫폼이야. 이 뉴스 기사에서는 페이스북이 개인 프로필을 만들고 친구들과 사진, 동영상 등을 공유하며 소통할 수 있는 서비스로 유명해.

문서에서 추출된 키워드: 소셜, 페이스북, 정보, 성인, 인텔리전스, 네트워크, 이용, 기술, 사용, 시간
추출된 키워드에 대해 설명이 필요하다면 추가 질문 정도를 높여보세요!

시연 영상

https://youtu.be/0Mwv_NaXx0E

A modern kitchen interior featuring a wooden island with a white countertop. Two white bar stools with chrome bases are positioned in front of the island. To the left, a dark grey sofa is partially visible. The background shows a kitchen sink and a wooden cutting board on the counter. The overall lighting is warm and ambient.

Q&A



감사합니다