제3회 한국가스공사 (KOGAS)

빅데이터·인공지능 스타트업 경진대회

가스·에너지분야 문서요약 모델개발

문서 요약 생성 플랫폼 SUMMER

비정형 팀플

- 팀장 : 권순기
- 팀원 : 유진이, 한보영

한국가스공사

1. 배경 및 개요

배경

가스・에너지 분야의 활용 배경
- 국내.외 에너지 분야의 성장 -

1. 국내 기술의 발전

- 국내 최초 100% LNG 냉열 활용 콜드체인 클러스터 조성

2. 해외 협력

- 독일 지멘스에너지와 해외 그린수소 개발 협력
- 글로벌 녹색성장연구소와 MOU 체결 해외 그린수소 개발, 주도적 추진 발판 마련

출처 : KOGAS webzine vol.34

자동 문서요약 모델 활용 배경
- 자연 언어 처리의 자동 문서 요약에 대한 중요성 대두 -

1. 자동 문서 요약 국내외 AI NLP 서비스 확산

- 챗봇의 발전으로 인해 문서로 되어 있는 내용을 읽고 요약하여 정보를 제공하기 위한 문서 자동 요약(Automatic Text Summarization)에 대한 수요가 증가

출처 : BERT 임베딩과 선택적 OOV 복사 방법을 사용한 문서 요약

2. 자동 문서 요약의 효율성 검증

- 자동 문서 요약은 사람이나 시스템이 다양한 텍스트 데이터를 효과적으로 처리할 수 있게 도와 줌

출처 : 신문기사와 소셜 미디어를 활용한 한국어 문서요약 데이터 구축

개요

- 목표 : 가스・에너지분야의 기사, 법률문서 등을 요약해 시장분석에 활용할 수 있는 모델 개발
- 일반 및 가스・에너지분야 국어 문서 원문으로부터 생성 요약문을 도출해 내는 인공지능 개발 → 모델 개발을 통한 사업성 확장

2. 사용 데이터

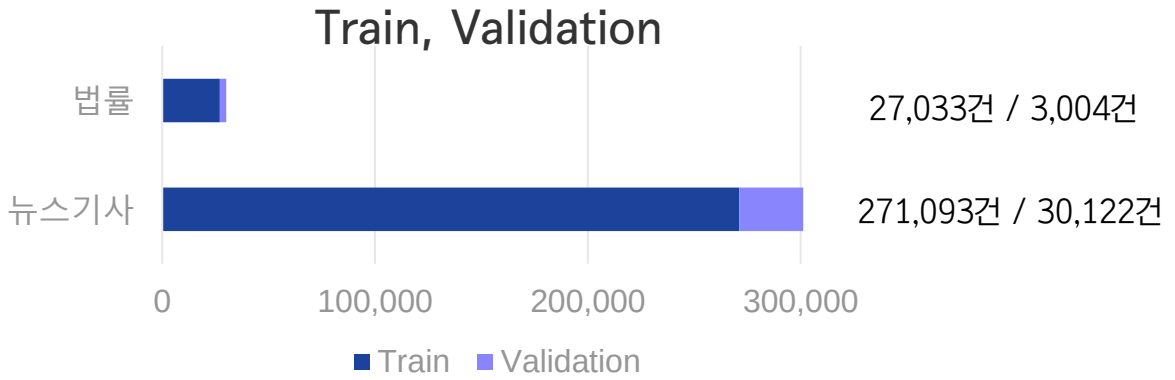
Ai hub 데이터셋

- 원문데이터 40만 건 (신문기사 30만 건, 기고문 3만 건, 잡지기사 1만 건, 논문 초록 3만 건, 법원 판결문 3만 건) 활용
  → 추출요약 40만 건, 생성요약 40만 건, 총 80만 건의 요약문 도출

활용 데이터 수집 및 출처

데이터	데이터 설명
뉴스기사 (신문기사)	요약 AI 알고리즘의 핵심 데이터로서 10개 언론사로부터 30만 건의 원문데이터를 확보
법률	법률은 공공데이터 포털 (open.law.go.kr)을 통하여 판례 원문의 오픈 API를 제공 받아 원문데이터를 확보

활용 데이터 구성

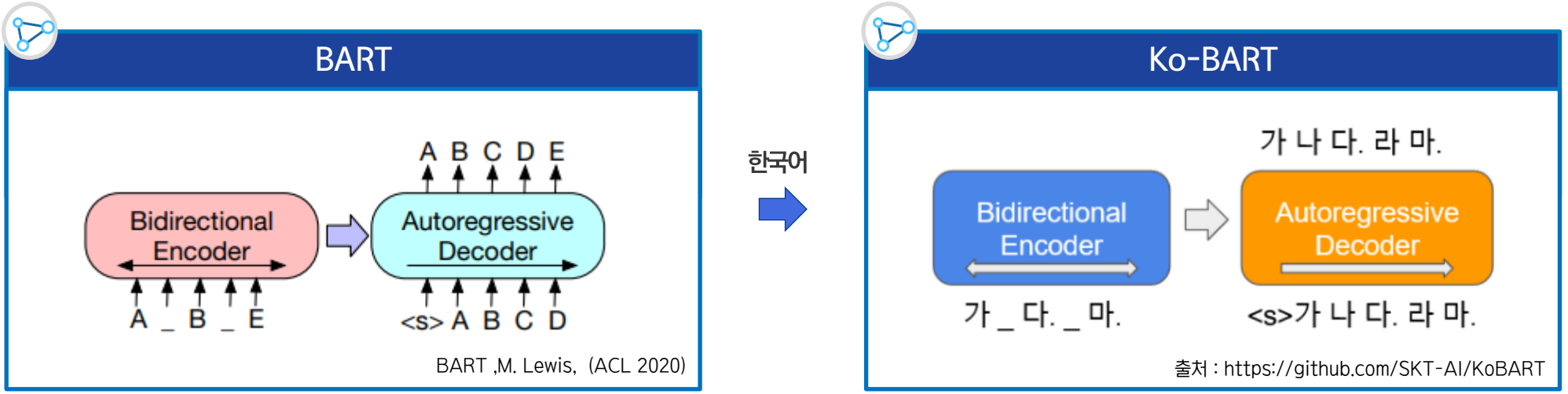


데이터	데이터 설명
뉴스기사 (신문기사)	수집 데이터로부터 종합면 30%, 정치 20%, 경제 20%, 사회 20%, 문화 및 스포츠 기타 10%의 비율로 구성
법률	민사, 형사 등 다양한 사건 판례로 구성

출처 : aihub 데이터셋 https://aihub.or.kr/aidata/8054

3. 사용 모델기술

BART 및 Ko-BART



- Ko-BART는 논문에서 사용된 Text Infilling 노이즈 함수를 사용해 40GB 이상의 한국어 텍스트에 대해서 학습한 한국어 BART 모델
- 한국어 위키 백과, 뉴스, 책, 모두의 말뭉치 v1.0(대화, 뉴스, ...), 청와대 국민청원 등의 다양한 데이터가 모델 학습에 사용

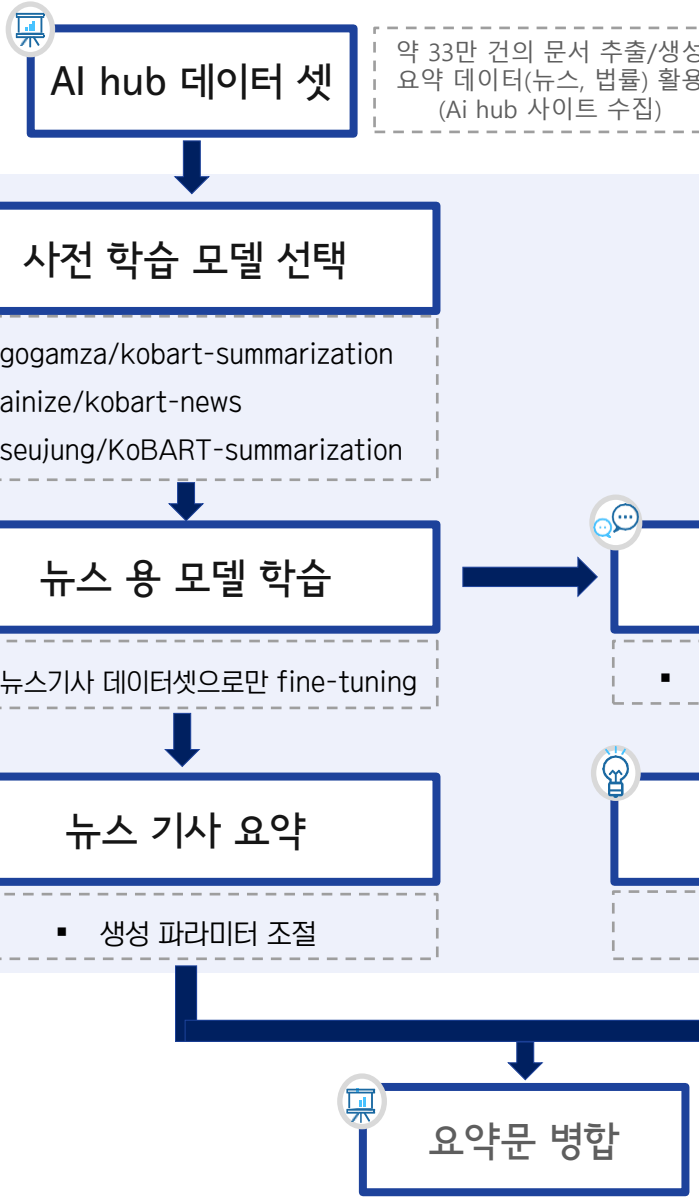
Ko-BART tokenizer

```
>>> from kobart import get_kobart_tokenizer
>>> kobart_tokenizer = get_kobart_tokenizer()
>>> kobart_tokenizer.tokenize("안녕하세요. 한국어 BART 입니다.🤣:)l^o")
['_안녕하', '세요.', '_한국어', '_B', 'A', 'R', 'T', '_입', '니다.', '🤣', ':)', 'l^o']
```

- Tokenizer는 character BPE 사용

3. 사용 모델기술

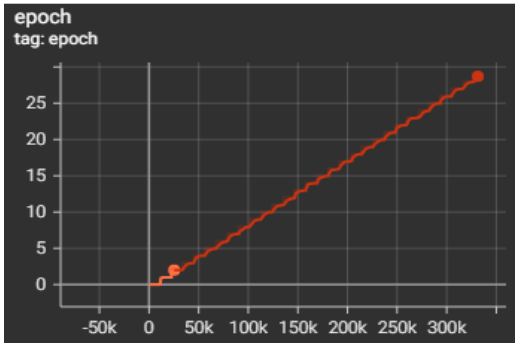
모델 프레임 워크

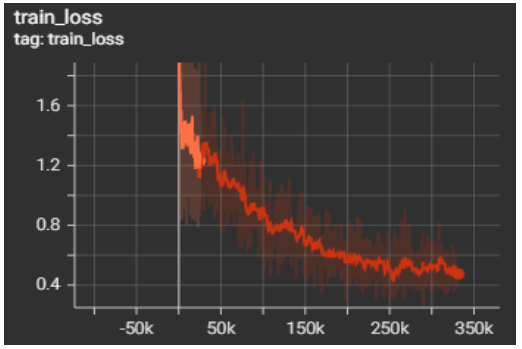


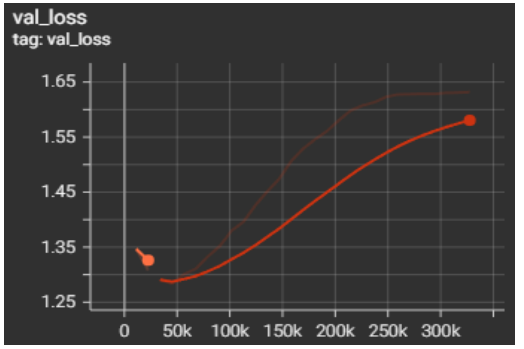
사전 학습 모델 출처
: seujung https://github.com/seujung/KoBART-summarization
: gogamza https://huggingface.co/gogamza/kobart-summarization
: ainize https://huggingface.co/ainize/kobart-news

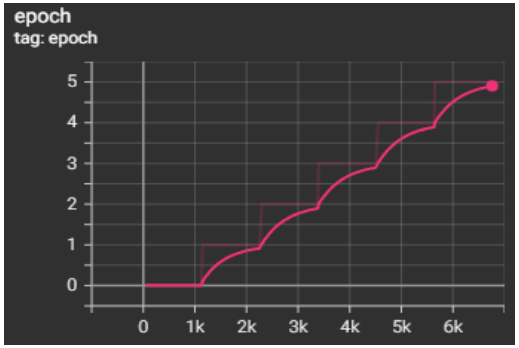
4. 모델링 결과

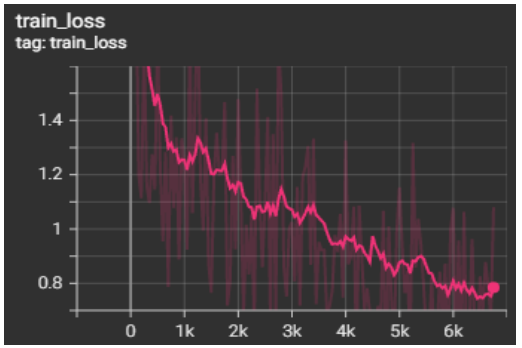
학습 로그

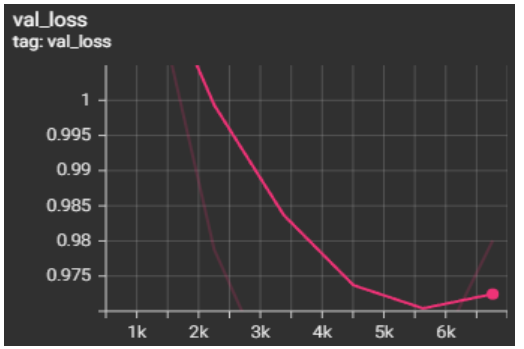












뉴스 기사 요약 모델

- fine tuning epoch 별 요약문의 뉘앙스가 달라짐
- 전체 30 epoch fine tuning 중 15 epoch 모델 선택
- repeat_ngram_size와 repetition_penalty등의 파라미터 조절을 통해 반복되는 단어 제한

법률 문서 요약 모델

- 이미 학습 된 상태이므로 warm_up_ratio를 크게 설정

4. 모델링 결과

모델 성능 평가 표

Model	Epoch	ROUGE-1	ROUGE-2	ROUGE-L
News	15	0.36942322	0.18042625	0.27775251
News + Repeat penalty	15	0.37514423	0.17915764	0.28152715
News + Repeat penalty	21	0.37318989	0.17572182	0.27976938
News + Repeat penalty	28	0.37281542	0.17567922	0.27922505
News + Repeat penalty + Law	15 / 6	0.37541149	0.17935329	0.28181497

- Dacon 가스.에너지 분야 문서요약 개발 경진대회 정량 평가 결과

5. 특장점(차별성/우수성)

기존 모델의 특징

- 기존 BART 기반의 한국어 생성요약 모델 :

다양한 분야의 비즈니스에 공통으로 적용이 가능하도록 한글 위키피디아, 신문 기사, 책 줄거리 등 일반적인 분야의 문어체 위주 한글 텍스트로 학습됨

본 모델의 차별성 / 우수성

- 도메인 별 상황에 맞는 모델을 생성하여 요약문의 정확도 향상
- 법률 문서 형식의 가스 문건 요약을 위해 법률 데이터셋으로 fine-tuning을 통한 정확도 향상

대회에서 제공된 가스 문건 예시

1. 개정이유" "친환경 기술개발에 대한 인센티브 부여 제도인 에코이노베이션 크레딧 인정량 상한을 현행 14.0g/km에서 17.9g/km로 확대하기 위함(에너지소비효율은 현행 3.5km/L에서 4.5km/L)" "2. 주요 개정내용" "가. 온실가스 배출저감 기술 실적인정량 상향 (안 제9조 개정) - 자동차 연비 및 온실가스 배출량 측정시 실내시험(동력계사용)에서 반영되지 못하는 온실가스 저감기술에 대한 실적인정량(제작사 평균값)을 상향 조정" "※ 각각의 기술 항목에 대한 개별 인정량은 변화 없음

6. 발전(사업화)계획

모델 발전 계획

모델의 한계점

- **모델의 Input Sequence를 512로 제한**
  - fine-tuning 시 컴퓨팅 자원의 한계로 최대 input sequence를 512로 지정 -> 원문이 길 경우 뒷부분 내용이 학습에 반영되지 못함

- **생성 요약 모델의 불확실성으로 인한 사업화 실현 어려움**
  - 기존 리뷰 요약, 뉴스 요약 등의 서비스의 경우 추출요약으로만 서비스 제공
  - 정보의 제공 측면에서 생성요약의 신뢰성 확보 필요

모델 발전 계획

- **모델의 Input Sequence를 확장을 통한 모델 개선**

- **추출요약 활용을 통한 모델 개선 방안**
  - 추출요약 모델을 통해 Top K개의 evidence text 추출

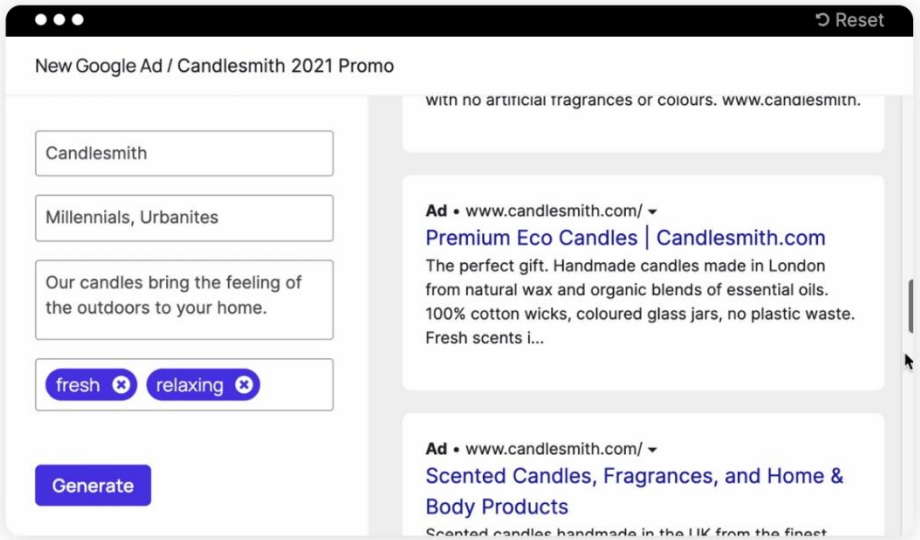
  1. 추출된 evidence text로 생성요약 진행 or 가중치를 부여하여 생성요약 진행
  2. 추출된 evidence text로 생성 요약문과의 문장 유사도 분석(Semantic Textual Similarity)을 통한 정확도 향상

6. 발전(사업화)계획

사업화 발전 계획

선행 사업 검토

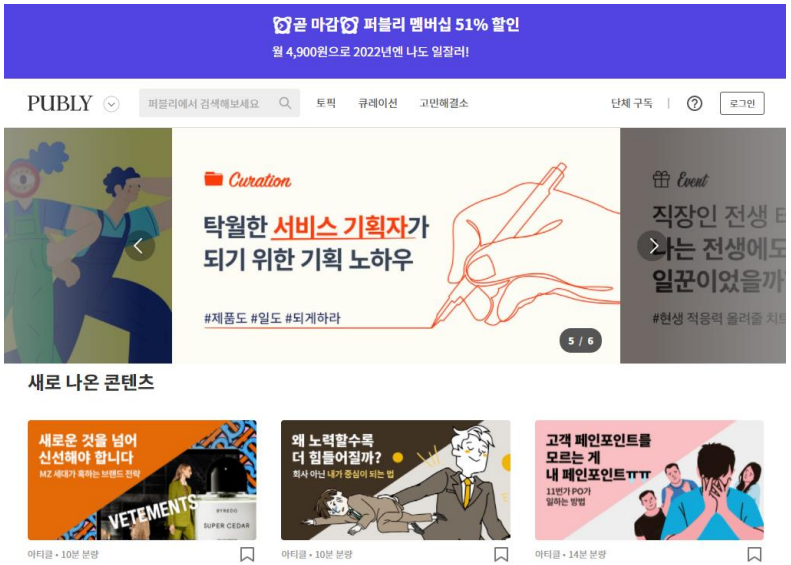
copysmith



출처 : https://copysmith.ai/

- 미국 스타트업 회사
- 기업 이름, 타겟 고객, 제품에 대한 간단한 설명을 입력
  → 자동으로 광고 문구를 출력해주는 플랫폼
- GPT3모델을 사용하여 자연어를 생성

PUBLY



출처 : https://publy.co/

- B2B 형태로 이미지 사업화를 통해 선별적 콘텐츠를 제공하는 플랫폼
- 구독 서비스에 해당
- 큐레이션 서비스를 통해 유용한 컨텐츠 추천

6. 발전(사업화)계획

사업화 발전 계획 | 제안 사업

한국어 요약/생성에 강점을 보이는 모델 개발 → 오픈소스 API 제공 회사

➢ 사업 목표

1.질 좋은 자료에 대한 수요 충족

- 요즘 대중들 뿐 아니라 기업에서도 선별적인, 질 좋은 자료에 대한 수요가 높음
- 선별적인 자료, 차별적인 자료를 얻기 위해 구독 서비스 이용
- 따라서, 선별적이고, 차별적인 서비스와 함께 글에 대한 요약 서비스 제공

2. 시간 대비 높은 가치 제공(확장 가능성)

- B2B 형태로 이미지 사업화를 하고 있는 '퍼블리' 와 같은 선별적 콘텐츠를 제공하는 업체가 있음
- 위와 같은 업체와 MOU 협약을 통해 이미 가지고 있는 선별적 콘텐츠를 요약해서 고객들에게 제공을 해주고, 기업에게 수익을 분배
- 고객들에게는 좋은 퀄리티의 글을 요약해서 제공 가능

➢ 주요 고객

- 문서요약이 필요한 개인, 기업
- 질 좋은 자료를 얻고자 하는 개인, 기업

6. 발전(사업화)계획

사업화 발전 계획 | 제안 사업

➢ 수익창출 방식

고객 유형	기본 고객	프리미엄 고객	고객 예시
기업	월 정액 구독 서비스 Enterprise 전용 API 유료 제공	월 정액 구독 서비스 기업 별 맞춤 데이터로 fine-tune 된 API 제공	대기업, 중소기업, 법률 사무소 등 문서가 많은 회사
개인	▪ 개인용 API 무료 배포 ▪ 동시 최대 3개 문서 요약 가능 ▪ 광고 1회 시청 후 다시 사용 가능 ▪ 월 30개 요약 제한	▪ 월 정액 구독 서비스 ▪ 광고 없이 사용 가능 ▪ 동시 요약 개수 제한 없음 ▪ 월 사용량 제한 없음	▪ 학생, 연구원 등 ▪ 요약하고자 하는 서류가 있는 개인

➢ 자금 조달 방식

자금 조달	조달 계획	비고
자가 출자	30,000,000 원	-
창업 공모전	5,000,000 원	공모전 입상
정부 지원자금	5,000,000 원	청년 스타트업 지원금
총계	40,000,000 원	-

6. 발전(사업화)계획

사업화 발전 계획 | 제안 사업

➢ 소요예산 계획 (6개월 창업준비 + 1개월 시연)

구분	세목 - 산출근거	비고	소요예산
창업 활동비	특정업무경비	7개월 동안의 창업활동비	3,500,000 원
광고선전비	일반용역비	웹디자인(홈페이지 + BI 포함)	5,000,000 원
		홍보영상 제작	1,500,000 원
		온라인 홍보물 제작	1,500,000 원
인건비	상용임금	2개월 동안의 디자이너	1,800,000 원 / 월(최저시급, 주 40시간 근무) *2개월
기계장치비	자산취득비	디자이너 노트북	2,500,000 원
	API 서버비용	AWS API Gateway 사용	4,000,000 원 / 월 * 4개월
총 계			33,600,000원

7. 결론(정리)

한국어 추출/생성 요약에 강점을 보이는 Ko-BART 모델 사용

뉴스 데이터로 파인튜닝한 경우와 법률 문서로 파인튜닝한 경우 모델 성능 향상을 보임

모델 발전 계획

1. 추출 요약문을 evidence text로 하여 생성 요약을 통한 성능 개선
2. 가중치를 부여한 생성 요약 진행을 통한 성능 개선
3. 모델의 input sequence 확장을 통한 성능 개선

Ko-BART 모델을 활용한 문서 요약 생성 플랫폼 SUMMER

질 좋은 자료에 대한 수요 증가
시간 대비 높은 가치 제공 가능

주요 고객	
개인	기업고객
▪ 기본 고객 : 무료 제공 (광고 시청 시 재 사용 가능) ▪ 프리미엄 고객 : 광고 제거 등 혜택 제공	▪ 기본 고객 : Enterprise 전용 API 제공 ▪ 프리미엄 고객 : Enterprise 맞춤형 모델 제공

8. 팀원 역할 및 참여도

팀장 권순기

서울과학기술대학교 산업공학과 학사 재학
서울과학기술대학교 ICT인공지능 부전공

- 데이터 분석 캡스톤 경진대회 대상 수상
- 딥러닝 기반 거북목 디텍션 서비스 개발
- 맑음 팩토리 데이터 분석 프로젝트 참여
- 데이터 기반 사용자 분석 연구실 재직

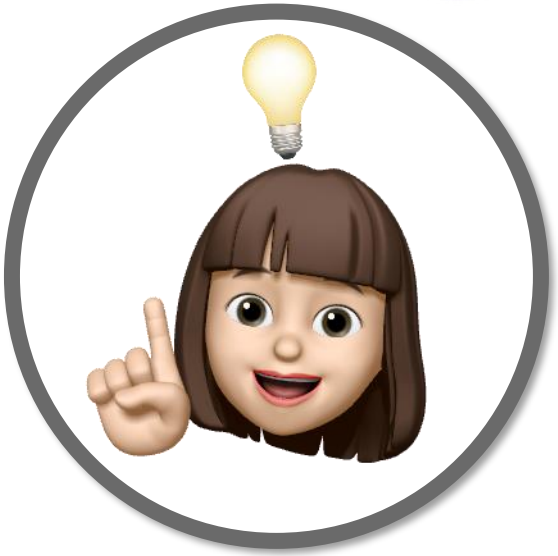
모델 구현 및 PPT 제작

유진이

서울과학기술대학교 산업공학과 학사 졸업
서울과학기술대학교 데이터사이언스학과 석사 재학

- 데이터 분석 캡스톤 경진대회 금상 수상
- 캠퍼스 특허 유니버시아드 장려상 수상
- 데이터 분석 연합 동아리 BITAMIN 수료
- 맑음 팩토리 데이터 분석 프로젝트 참여
- 데이터 기반 사용자 분석 연구실 재직

모델 구현 및 PPT 제작



한보영

서울과학기술대학교 IT Management 학사 졸업
서울과학기술대학교 데이터사이언스학과 석사 재학

- 데이터 분석 캡스톤 경진대회 장려상 수상
- 트위터 데이터 분석 진행
- 데이터 분석 프로젝트 다수 진행
- 정보보안 연구실 재직

모델 구현 및 PPT 제작

제3회 한국가스공사 (KOGAS)

빅데이터·인공지능 스타트업 경진대회

가스·에너지분야 문서요약 모델개발

Q&A

제3회 한국가스공사 (KOGAS)
빅데이터·인공지능
스타트업 경진대회

감사합니다