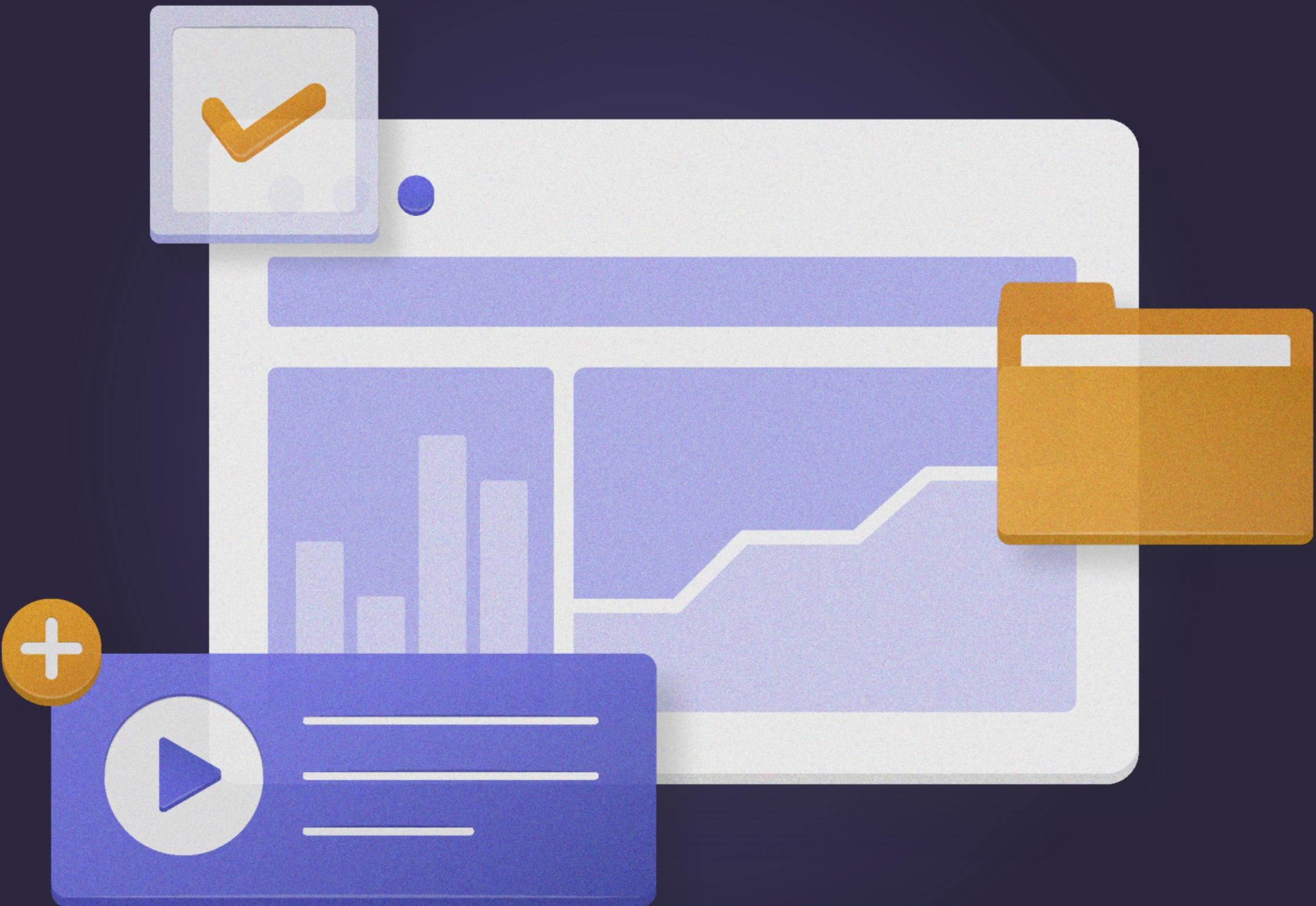


새로운 영상 데이터 요약 서비스, retrievAI

프로젝트 G조



주제 선정 동기

영상 데이터의 중요성 증가

Problem

경험 전달의 효율성이 중요도가 높아지며
데이터 해상도가 높은 영상 데이터의 중요성 증가

Solution

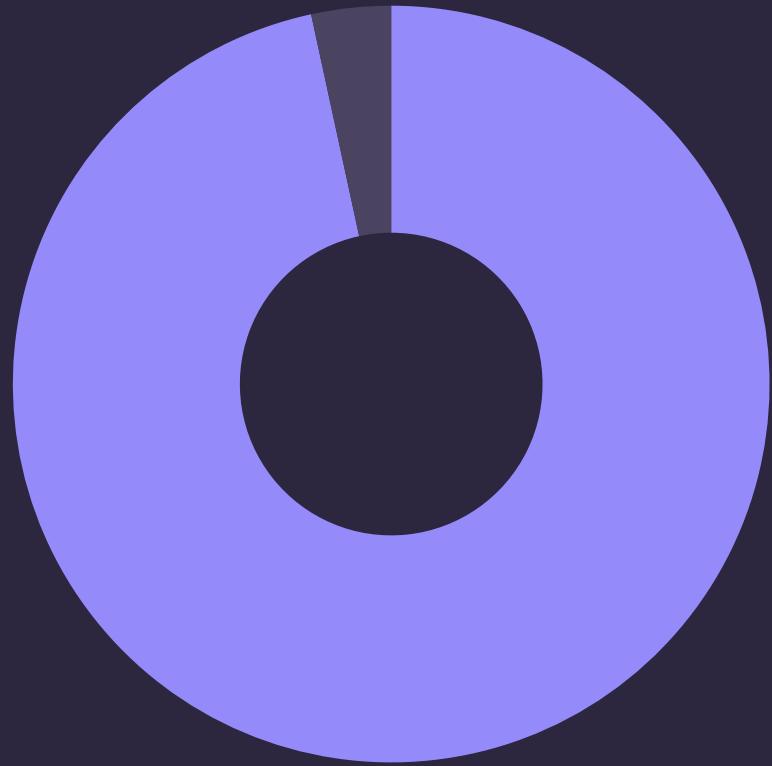
영상 데이터에
주목할 필요성 ↑



주제 선정 동기

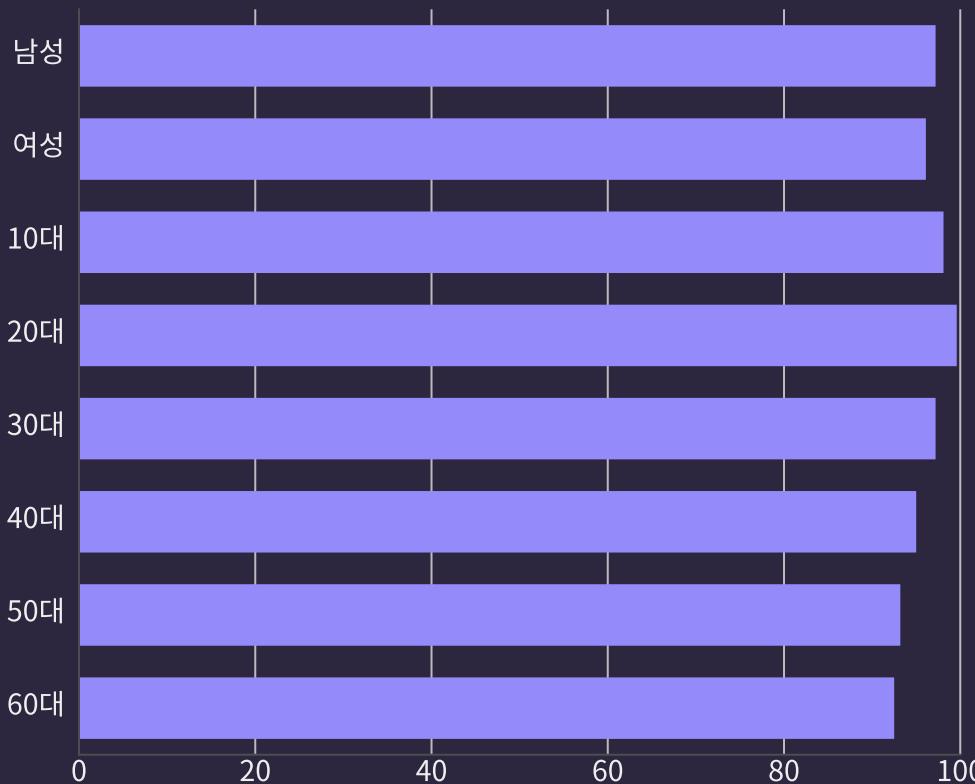
나스미디어 ‘2022 인터넷 이용자 조사’

: 최근 일주일 내 온라인 동영상을 이용한 적이 있냐



전체 이용자 분포

전체 이용자의 약 96%가 ‘시청했다’ 답변



성/연령별 분포

모든 성별과 연령대에서 90% 이상이
‘시청했다’ 답변

주제 선정 동기

영상 데이터의 중요성 증가

Problem

경험 전달의 효율성이 중요도가 높아지며
데이터 해상도가 높은 영상 데이터의 중요성 증가

Solution

영상 데이터에
주목할 필요성 ↑

텍스트 데이터와 달리 영상 데이터를 효과적으로
관리 및 검색하는 방법에 관한 연구나 서비스의 미비

영상 데이터의 관리 및 검색
방법에 관한 연구의 필요성 ↑



주제 선정 동기

영상 데이터의 중요성 증가

Problem

경험 전달의 효율성이 중요도가 높아지며
데이터 해상도가 높은 영상 데이터의 중요성 증가

텍스트 데이터와 달리 영상 데이터를 효과적으로
관리 및 검색하는 방법에 관한 연구나 서비스의 미비

데이터 관리 및 검색 기능 향상의 핵심은 해당 데이터를
나타내는 유용한 정보를 추출, 저장하는 것

Solution

영상 데이터에
주목할 필요성 ↑

영상 데이터의 관리 및 검색
방법에 관한 연구의 필요성 ↑



영상 데이터의 핵심 정보를
추출하는 서비스의 필요성 ↑

주제 선정 동기

영상 데이터의 중요성 증가

Problem

경험 전달의 효율성이 중요도가 높아지며
데이터 해상도가 높은 영상 데이터의 중요성 증가

텍스트 데이터와 달리 영상 데이터를 효과적으로
관리 및 검색하는 방법에 관한 연구나 서비스의 미비

데이터 관리 및 검색 기능 향상의 핵심은 해당 데이터를
나타내는 유용한 정보를 추출, 저장하는 것

Solution

영상 데이터에
주목할 필요성 ↑

영상 데이터의 관리 및 검색
방법에 관한 연구의 필요성 ↑

영상 데이터의 핵심 정보를
추출하는 서비스의 필요성 ↑



Raw Data가 가지는
본질을 최대한 보존하며,
영상 데이터의 핵심 정보를
추출, 데이터를 관리하는

retrievAI

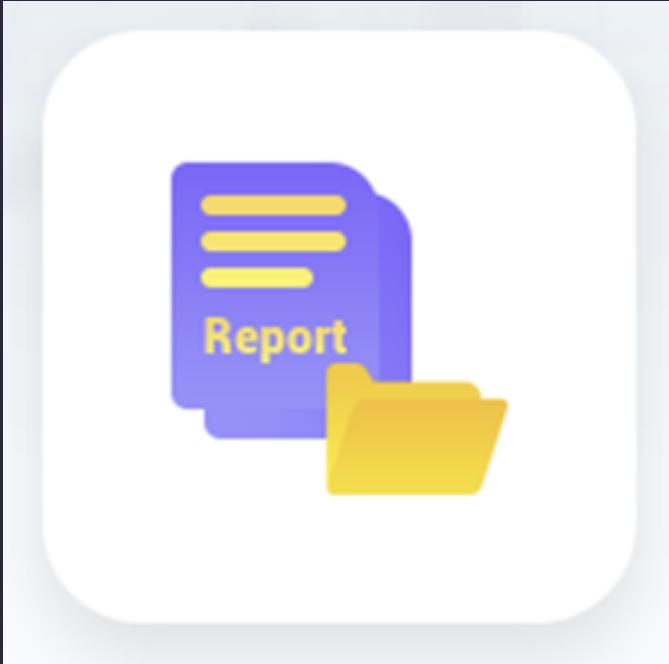
데이터 수집
데이터 출처



AI 기술 및 제품 서비스 개발에 필요한
AI 인프라를 지원함으로써 누구나 활용하고
참여하는 AI 통합 플랫폼

데이터 수집 데이터 설명

데이터 종류	원문 규모	어노테이션 규모	결과 규모		비고
			추출 요약	생성 요약	
회의록	27,000	74,800	18,700	37,400	2~3문장 추출
			18,700		20% 추출
연설문	40,000	88,000	22,000	44,000	2~3문장 추첨
			22,000		20% 추출
총계	67,000	162,800	81,400	81,400	



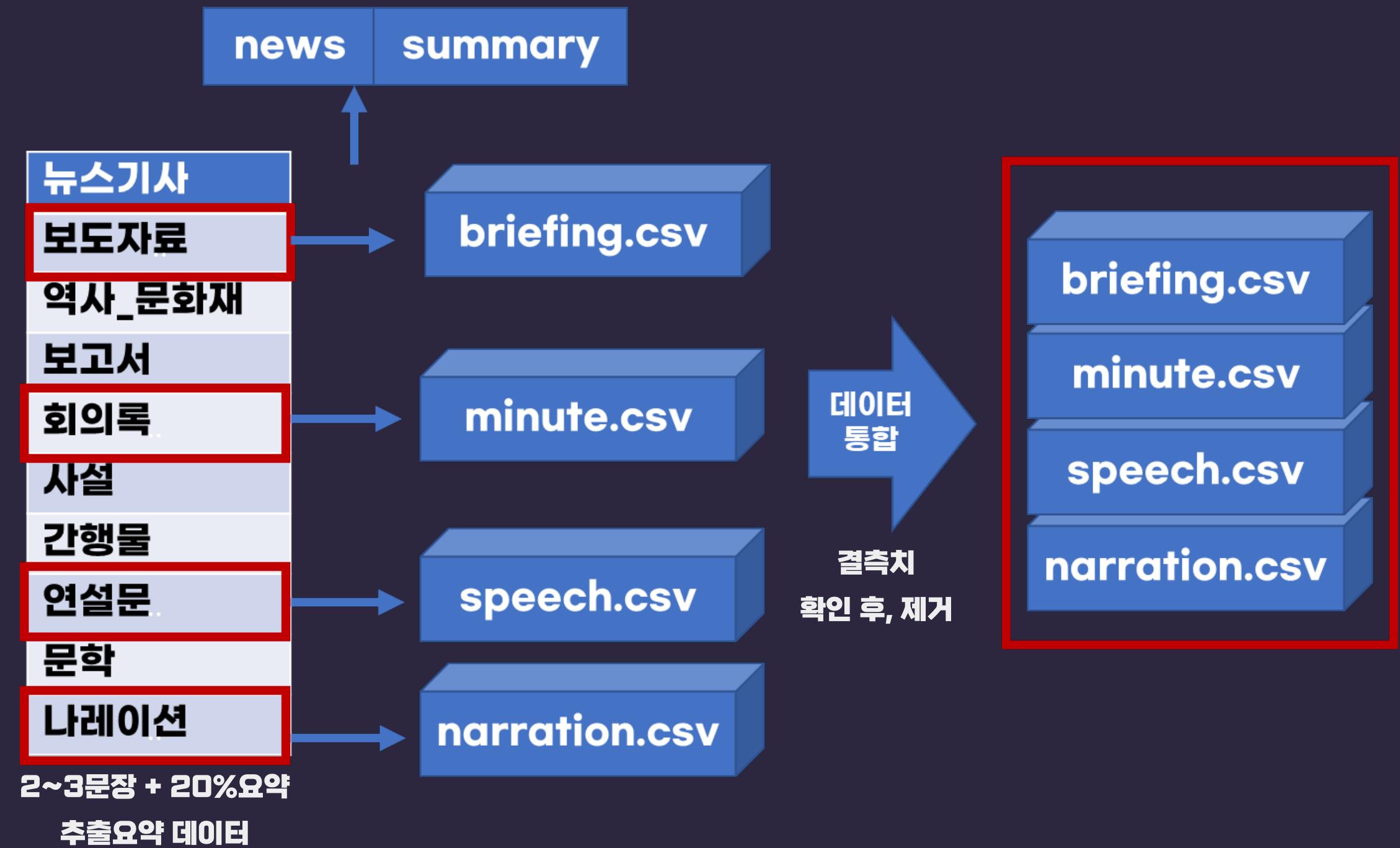
요약문 및 레포트 생성 데이터

- 다양한 한국어 원문 데이터로부터 정제된 추출하고 검증한 한국어 문서요약 AI 데이터셋
- 추출요약 포함 본문에서 중요한 문장을 하나의 새로운 요약문으로 참조하는 생성요약을 위한 데이터세트 구축

데이터 수집

데이터 전처리

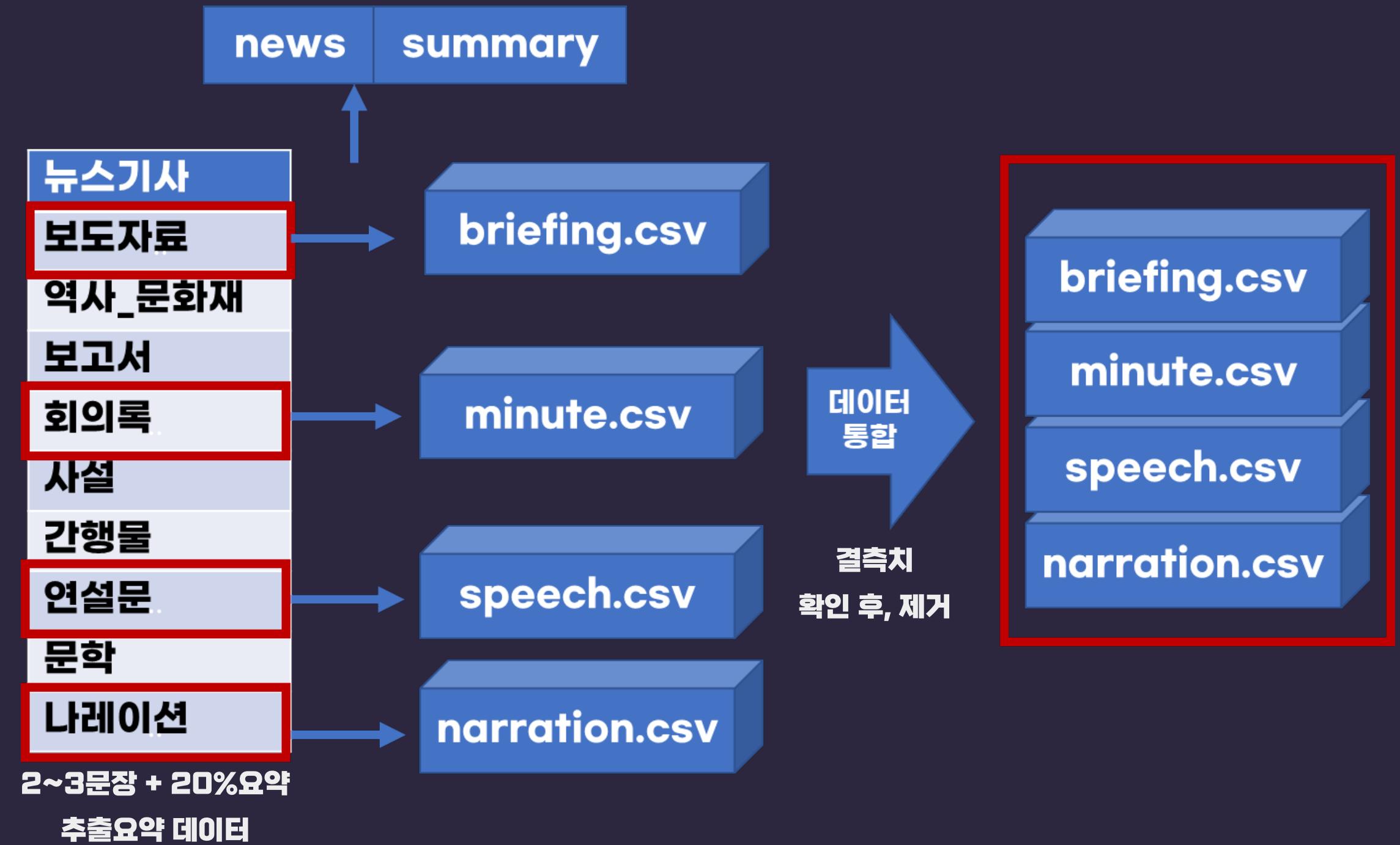
1) 훈련 데이터 (training/라벨링 데이터)



데이터 수집

데이터 전처리

2) 테스트 데이터 (Validation/라벨링 데이터)



데이터 수집

데이터 전처리 결과

1) 훈련 데이터

	news	summary
0	입지요건₩n 입지 특성에 따라 '주거상업고밀지구(역세권)', '주거산업융합지구(준...)	사업 추진 과정에서 지정권자는 사전검토기구를 구성하여 용적률 등을 사전에 검토하고 ...
1	해당 차량은 7월 23일부터 비엠더블유코리아 공식 서비스센터에서 무상으로 수리(점...	포르쉐코리아에서 수입, 판매된 타이칸이 주행 중 시동이 꺼질 가능성이 확인되어 리콜...
2	정부는 2021 P4G 서울 녹색미래 정상회의(5.30.(일)~5.31(월))에 ...	정부가 개최한 녹색미래주간 개막식은 DDP에서 20분간 개막 영상, 국회의장 축사,...
3	₩n 건축물대장상 실제 소유자와 같은데도 잘못 작성됐다면...	국민권익위는 건축물대장상 소유자가 개인 명의이나 사업자로 잘못 기재됐으니 정정해달라...
4	₩n 앞으로 예술인 복지서비스를 받기 위한 예술활동증명 ...	국민권익위원회는 코로나19로 어려움을 겪고 있는 문화예술인을 지원하기 위해 문화체육...
...
83566	대자연의 위대함과 생태계의 신비! BBC를 비롯하여, 세계 일류의 다큐멘터리 전문 ...	동물의 왕국은 세계 일류의 다큐멘터리 전문 제작사들이 제작한 고급 다큐멘터리들을 우...
83567	대자연의 위대함과 생태계의 신비! BBC를 비롯하여, 세계 일류의 다큐멘터리 전문 ...	BBC를 비롯하여 다큐멘터리 제작사들이 제작한 동물의 왕국이 우리말로 더빙되어 일반...
83568	역사의 커다란 물줄기가 바뀐 결정적 하루! 역사가 움직인 터닝 포인트를 입체적으로 ...	무령왕릉 발굴 50주년을 맞아 발굴의 생생한 현장과 종흥군주 무령왕에 대해 알아볼 ...
83569	역사의 커다란 물줄기가 바뀐 결정적 하루! 역사가 움직인 터닝 포인트를 입체적으로 ...	청기즈칸의 명령 아래 몽골 최고의 장수 제베와 수부데이는 호라즘 왕 무함마드 추격에...
83570	BBC, 내셔널 지오그래픽사, CCTV 등에서 제작한 문화, 환경, 과학, 시사 등...	부르고뉴에 도착한 스타인은 샤토에 머물며 샤토 주인이 텃밭에서 직접 딴 채소로 만든...
83571	rows × 2 columns	확인 후 제거

데이터 수집

데이터 전처리 결과

2) 테스트 데이터

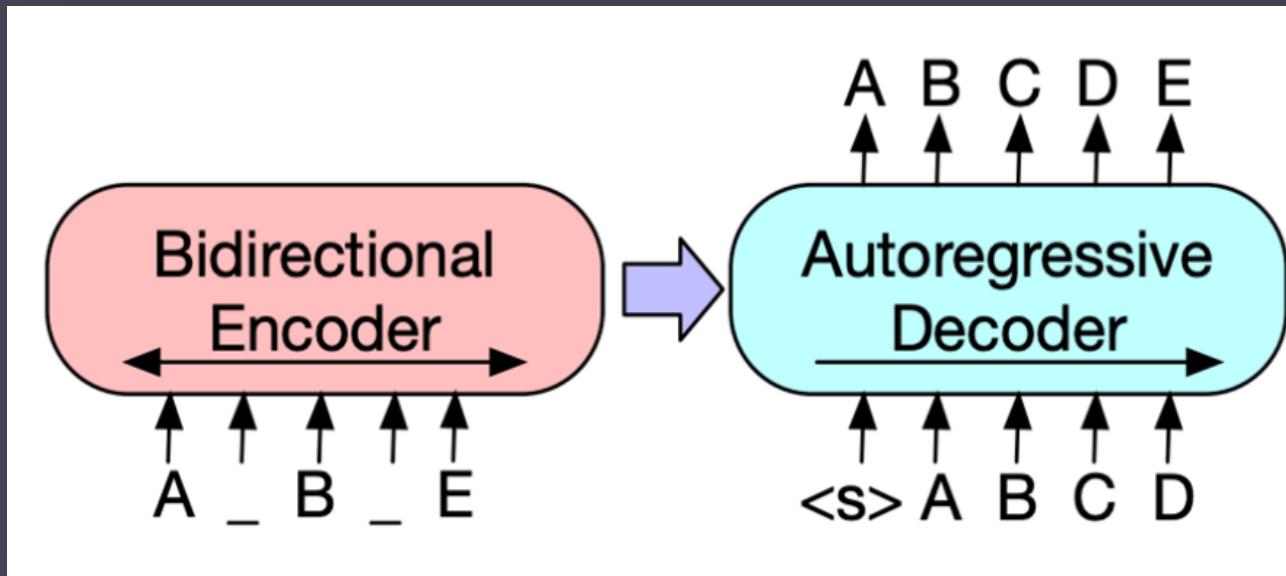
	news	summary
0	제 목 : 상호금융, 우체국, 증권사 오픈뱅킹 서비스 개시 및 입금 가능 계좌 확대 (12...)	금융위원회는 오픈뱅킹 기능을 보완할 필요가 있어 오픈뱅킹 고도화 방안을 발표하고 오...
1	기획재정부는 개도국의 코로나19 효과적 대응을 위해 10월 14일 필리핀 캄보디아에...	기획재정부가 감염병 대응 조직체계 구축, 의료진 역량 강화 등을 위해 방글라데시에 ...
2	개발 과제들에 있어서도 실제적인 진전이 있어야 합니다. 이것은 모든 각료회의에서 중...	WTO는 WTO 규정 업데이트와 경제적 의미가 있는 협정 합의, 분쟁 해결 시스템 ...
3	환경부(장관 조명래)와 한국수자원공사(사장 박재현)는 최근 내린 집중호우로 대 청댐...	환경부와 한국수자원공사는 집중호우로 전국 댐에 유입된 초목류나 생활 쓰레기 등의 부...
4	소방청(청장 정문호)은 지난 2월 소방항공 전문인력 채용 원서접수(2.17.~2.1...	소방청은 코로나19로 연기했던 실기시험 등의 시험 일정을 실시 예정일로부터 약 세 ...
...
10395	일단은 초기에 이제 유입 환자를 차단하는 부분에 있어서는 어느 정도 성과를 보고 있...	지역 사회 내에서 증상이 있는 시간이 길었던 사람들을 통해 추가적으로 2차와 3차 ...
10396	그러니까 과거의 우리나라 선거를 한 번 회고해보시면 생각보다 양당 경쟁으로 끝난 적...	양당 경쟁으로 갈 것으로 예상했는데 3당 합당에 불만을 갖던 많은 수의 유권자들이 ...
10397	네, 그런 얘기를 하는데야 아니 뭐 조금 전에 우리 대변인께서도 아니 본인 당이 집...	자유한국당은 대통령 이외에는 바뀐 것이 없는 상태에서 국회 권력과 여당의 지위를 여...
10398	우리 정부도 무인기 산업의 가능성을 높이 보고 10년 동안 국비 1000억 원을 들...	우리 정부가 국비 1000억 원을 들여 세계 두 번째로 수직이착륙 무인기를 독자 개...
10399	지난달 8일, 경찰관 2명을 흉기로 찔러 1명을 사망에 이르게 한 사건이 있었습니까...	흉기를 찔러 경찰관 1명을 사망에 이르게 한 사건의 범인은 조현병 환자였다.

10400 rows × 2 columns

모델 설명

KoBART Summarization 모델

BART (Bidirectional and Auto-Regressive Transformers)



- ▶ BART의 Encoder, Decoder 구조

BERT(Encoder) + GPT(Decoder)

인코더이기 때문에 내용 생성 Task에

대응할 수 없다는 BERT의 한계 극복

+

디코더만 존재하기 때문에 양방향으로
문맥 정보를 반영하지 못하는
GPT의 한계 극복

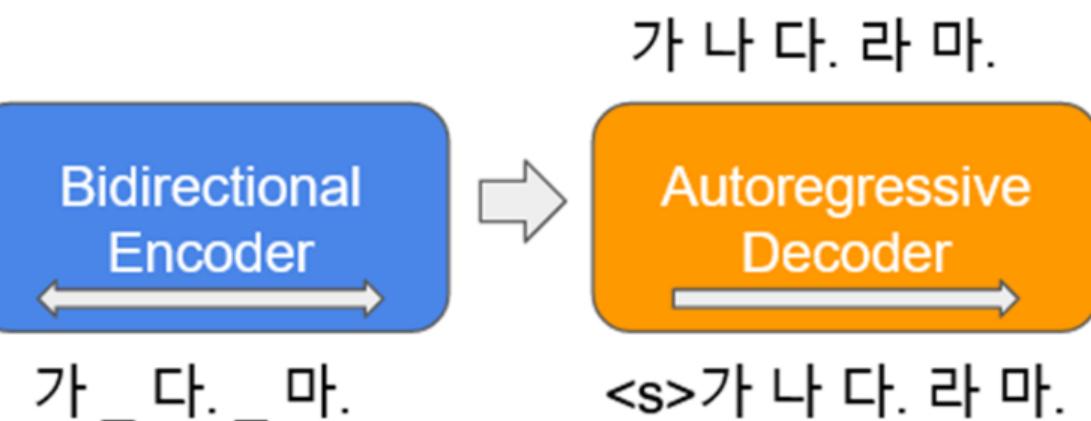
=

마스킹을 통한 텍스트 변형과 이를 복원하도록 학습시킴

모델 설명

KoBART Summarization 모델

KoBART (Korean BART)



- ▶ KoBART의 Encoder, Decoder 구조

- BART 모델을 한국어에 적용할 수 있도록 만들어진 언어 모델

- Text Infilling 노이즈 함수를 사용해 40GB 이상의 한국어 텍스트에 대해서 학습했으며, 다른 언어 모델들에 비해 뉴스 기사나 신문에 대한 요약 Task에 있어서 특히 뛰어난 성능을 보임

모델 설명

모델 요약 예시

광부 생활 기적, 기쁨 별개로…오늘 ‘책임규명’ 현장합동 감식

경찰이 경북 봉화군 아연광산 붕괴사고 현장 합동 감식에 나선다. 경북경찰청 아연광산 붕괴사고 전담수사팀은 7일 “오후 1시부터 전담수사팀, 과학수사과, 산업통상자원부 동부광산안전사무소가 광산 붕괴사고 원인과 책임을 밝히기 위한 현장 감식을 진행한다”고 밝혔다. 현장 감식에서 경찰은 사고가 난 제1수직갱도와 제2수직갱도 등 광산 구조를 파악하고, 사고 당시 갱도에 쏟아진 토사 시료를 채취해 국립과학수사연구원에 성분 분석을 의뢰할 예정이다. 붕괴사고가 발생한 갱도 안 현장조사는 안전하다고 판단되면 진행할 것으로 알려졌다. 이는 현재까지도 갱도 안이 안정적이지 못해 현장 진입이 위험할 수 있다는 전문가 진단에 따른 것이다. 경찰은 사고 발생 직후 구조된 광부 5명의 참고인 조사를 진행하고 있다. 하지만 이들도 사고로 심리적 안정이 필요한 상황이라고 보고, 조심스럽게 조사하고 있다고 한다. 지난 4일 구조된 2명은 입원치료가 끝나고 퇴원하면 일정을 조율해 마지막에 참고인으로 조사할 예정이다. 또 제1수직갱도에서는 지난 8월29일에도 사고가 발생해 광부 1명이 숨졌는데, 경찰은 지난달 26일 붕괴사고와 묶어 두 사고를 한꺼번에 조사를 진행하고 있다. 정찬익 경북경찰청 강력계장은 “참고인 조사 등 기초조사를 마친 뒤 사고 원인과 업체 쪽 안전 의무 위반 여부 등을 조사할 계획이다. 업체 말고도 관련 기관에 대해서도 관리·감독 등을 살펴볼 계획”이라고 말했다. 앞서 지난달 26일 오후 6시께 경북 봉화군 재산면 갈산리 금호광업소 제1수직갱도가 붕괴했다. 당시 지하 190m 갱도 안에서 작업하고 있던 광부 7명 가운데 5명은 자력으로 탈출하거나 구조됐지만, 조장 박아무개(62)씨와 보조작업자 박아무개(56)씨는 갱도에 갇혔다. 광산 채굴업체 성안엔엠피코리아는 자력 구조에 나섰다가 여의치 않자 다음날인 지난달 27일 아침 8시34분께 119에 신고했다. 당국은 뒤늦게 구조 작업에 나섰고, 사고 발생 221시간 만인 지난 4일 밤 11시3분께 갱도에 갇혔던 박씨 등 2명을 구조했다. 박씨 등은 안동병원 일반병실에서 입원치료를 받고 있으며, 빠르게 건강을 회복하고 있다. 병원 쪽은 며칠 안에 이들이 퇴원할 수 있을 것으로 보고 있다.



경북경찰청 아연광산 붕괴사고 전담수사팀은 7일 오후 1시부터 전담수사팀, 과학수사과, 산업통상자원부 동부광산안전사무소가 광산 붕괴사고 원인과 책임을 밝히기 위한 현장 감식을 진행한다고 밝혔다

모델 설명

모델 요약 예시

광부 생활 기적, 기쁨 별개로…오늘 ‘책임규명’ 현장합동 감식

경찰이 경북 봉화군 아연광산 붕괴사고 현장 합동 감식에 나선다. 경북경찰청 아연광산 붕괴사고 전담수사팀은 7일 “오후 1시부터 전담수사팀, 과학수사과, 산업통상자원부 등 부광산안전사무소가 광산 붕괴사고 원인과 책임을 밝히기 위한 현장 감식을 진행한다”고 밝혔다. 현장 감식에서 경찰은 사고가 난 제1수직갱도와 제2수직갱도 등 광산 구조를 파악하고, 사고 원인과 책임을 밝힐 수 있는 조사 내용을 확정해 전달수사팀과 협력해 조사할 예정이다.

인터넷 기사, 신문 기사와 같은 줄글로 작성된 문어체의 언어 데이터를 기반으로 학습

→ **발화문 요약 데이터를 통해 기존 KoBART**

요약 모델에 대한 Fine Tuning 진행

등은 안동병원 일반병실에서 입원치료를 받고 있으며, 빠르게 건강을 회복하고 있다. 병원 쪽은 며칠 안에 이들이 퇴원할 수 있을 것으로 보고 있다.

모델 설명

STT(Speech To Text) 모델

KoBART
(Korean BART)



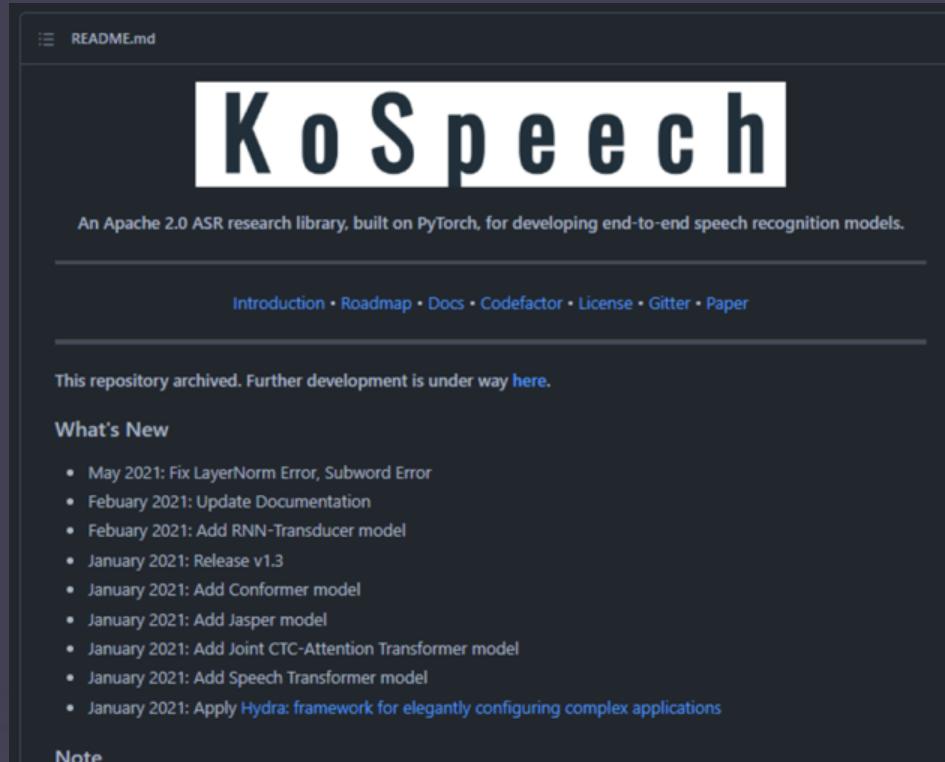
- ▶ KoBART의 Encoder, Decoder 구조

- 사람의 음성으로부터 발화한 Text를
얻어내는 기술
- 딥러닝 End-to-End 모델 사용

모델 설명

STT(Speech To Text) 모델

KoSpeech
(Korean Speech)



▶ 한국어 STT Open Source

현실적인 어려움...

- 음성데이터의 거대한 용량 (100GB 이상)
- 요약모델도 훈련시키기 바쁜 Colab

모델 설명

Python Speech Recognition

Python Speech Recognition

```
import speech_recognition as sr\n\naudio = sr.AudioFile(f"{new_name}.wav")\nwith audio as source:\n    audio = r.record(source)\ncaption_text = r.recognize_google(audio_data=audio, language="ko-KR")\nreturn caption_text
```

- ▶ Google API 활용 STT

모델 설명

Structure



spring[®]



모델 설명

Structure



spring[®]



FastAPI

데이터 분석

예시 화면

DScover Youtube Summarization

<https://www.youtube.com/watch?v=HhVRBhNIRdY> URL 검색

푸틴, "핵으로 공격하면 맞대응"...美 "무책임한 행동" / YTN

"러시아 핵으로 공격하면, 흔적 없이 사라질 것" "핵 능력을 만지 방어 수단일 뿐" "증전 위한 합의 중요하지만, 서방 맨지 못해" [영화] 블라디미르 푸틴 러시아 대통령이 러시아를 핵으로 공격하면 맞대응할 것이라며 다시 한 번 강한 어조로 경고했습니다. 푸틴의 잇따른 핵 위협에 로이드 오스틴 미 국방장관은 무책임한 행동이라며 핵보유국은 도발적인 행동을 피해야 한다고 지적했습니다. 김선희 기자가 보도합니다. [기자] "러시아를 핵무기로 공격하는 나라는 흔적도 없이 사라지게 될 것이다." 블라디미르 푸틴 러시아 대통령이 또 한 번 강하게 핵으로 맞대응할 수 있다고 경고했습니다. 러시아는 미국이 갖지 못한 국초울속 시스템이 있다는 것도 다시 한 번 상기시켰습니다. 그리면서도 러시아의 핵 능력을 방어 수단일 뿐이라고 거듭 강조했습니다. [블라디미르 푸틴 / 러시아 대통령 : 우리는 비밀이 없습니다. 우리를 겨냥한 공격에 대한 송합입니다. 만지 충돌일 뿐입니다.] 푸틴 대통령은 또 증전을 위한 외교적 합의가 중요하지만 서방을 믿을 수 없다고 비난했습니다. 2014년 우크라이나 동부 돈바스 내전 증전을 위한 민스크 협정이 무력화된 사실을 거론하며 다른 참가자들이 자신들을 속였다고 주장했습니다. [블라디미르 푸틴 / 러시아 대통령 : 그들이 우리에게 거짓말을 했다는 것이 밝혀졌고 우크라이나에 무기를 제공하고 있었습니다. 우리는 적대 행위에 대비하는 것뿐이었습니다.] 푸틴의 거듭된 핵 위협에 로이드 오스틴 미 국방 장관은 무책임한 행동이라고 우려를 표했습니다. 오스틴 장관은 핵보유국은 도발적인 행동을 피하고 핵전쟁과 핵무기 확산을 방지해야 할 종대한 책임이 있다고 강조했습니다. 세계는 푸틴의 무책임한 핵 위협을 염두하고 있다며 미국은 모든 분야에서 핵 억지력을 높이는 데 최선을 다할 것이라고 밝혔습니다. YTN 김선희입니다. YTN 김선희 (sunny@ytn.co.kr) ※ '당신의 제보가 뉴스가 됩니다' [카카오톡] YTN 검색해 채널 추가 [전화] 02-398-8585 [메일] social@ytn.co.kr ► 기사 원문 : https://www.ytn.co.kr/_ln/0104_202212100723065477 ► 제보하기 : https://mj.ytn.co.kr/mj/mj_write.php YTN 유튜브 채널 구독 : <http://goo.gl/Ytb552> © YTN 무단 전재 및 재배포금지

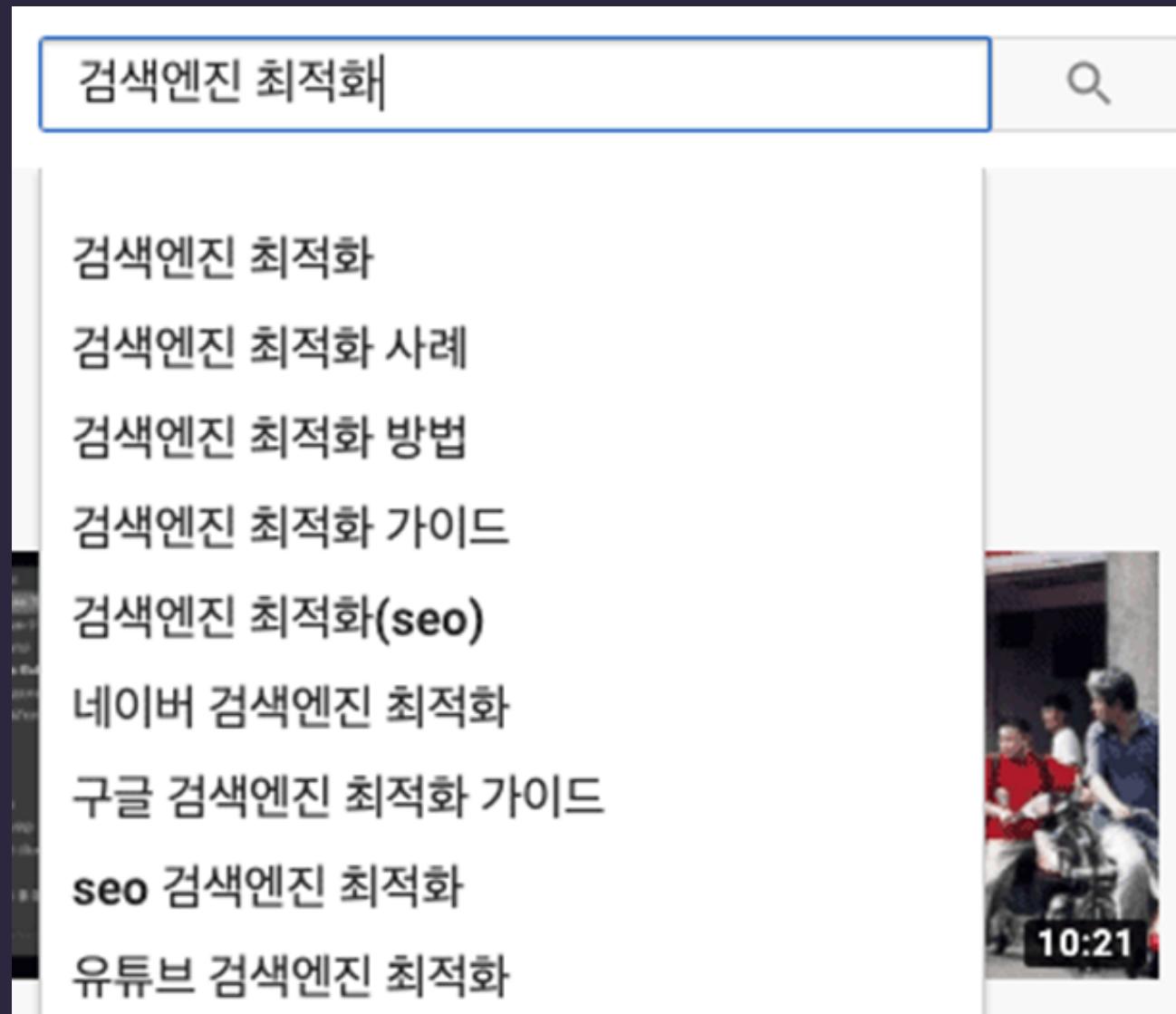
Summarize

Result

푸틴의 핵 위협에 오스틴 미국방장관은 무책임한 행동이라며 핵보유국은 도발적 행동을 피해야 한다고 지적했다.

기대 효과/한계점

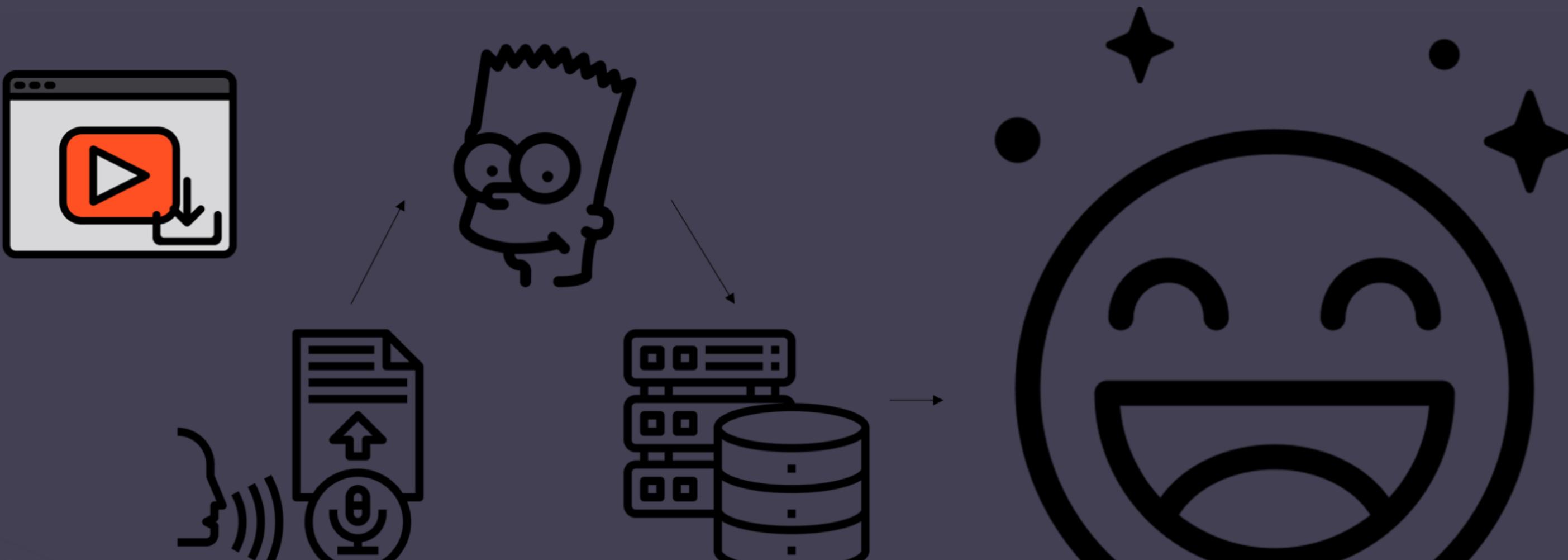
기대효과 및 활용가능성



소유한 동영상에 대한 추가적인 정보를 자동적으로 생성할 수 있는 기술로 사용 가능

기대 효과/한계점

기대효과 및 활용가능성



기대 효과/한계점

한계점

1. 오픈소스로 사전학습 모델이 잘 공개되어 있는 KoBART
모델과 달리 STT 모델은 활용이 자유롭지 않음
2. STT 모델 자체의 성능적 한계 및 동영상 데이터의 특성
상 STT 모델만으로 동영상에서 유의미한 정보를 모두 추출
하는 것은 어려움

**THANK YOU
FOR WATCHING**

프로젝트 G조