귀가노니 - 출퇴근길에 듣는 인공지능 뉴스 팟캐스트

#뉴스요약 #팟캐스트 #TTS #갓생

NLP-05

(ㅇㄱㄹㅇ팀)

boostcamp ai tech

CONTENTS

01 팀 소개

02 프로젝트 소개

03 모듈별 소개

04 시연 영상

05 향후 개선사항

06 Q&A

01

팀 소개

귀가노니 - 출퇴근길에 듣는 인공지능 뉴스 팟캐스트

01 팀 소개

권혜빈
캠퍼 ID: T2014
#Clustering #TTS

김민진
캠퍼 ID: T2029
#Summarization

김상현
캠퍼 ID: T2034
#Summarization

김성한
캠퍼 ID: T2039
#PM #Crawling #Serving

김제우
캠퍼 ID: T2052
#Crawling #Serving #TTS

이노아
캠퍼 ID: T2152
#Clustering #TTS

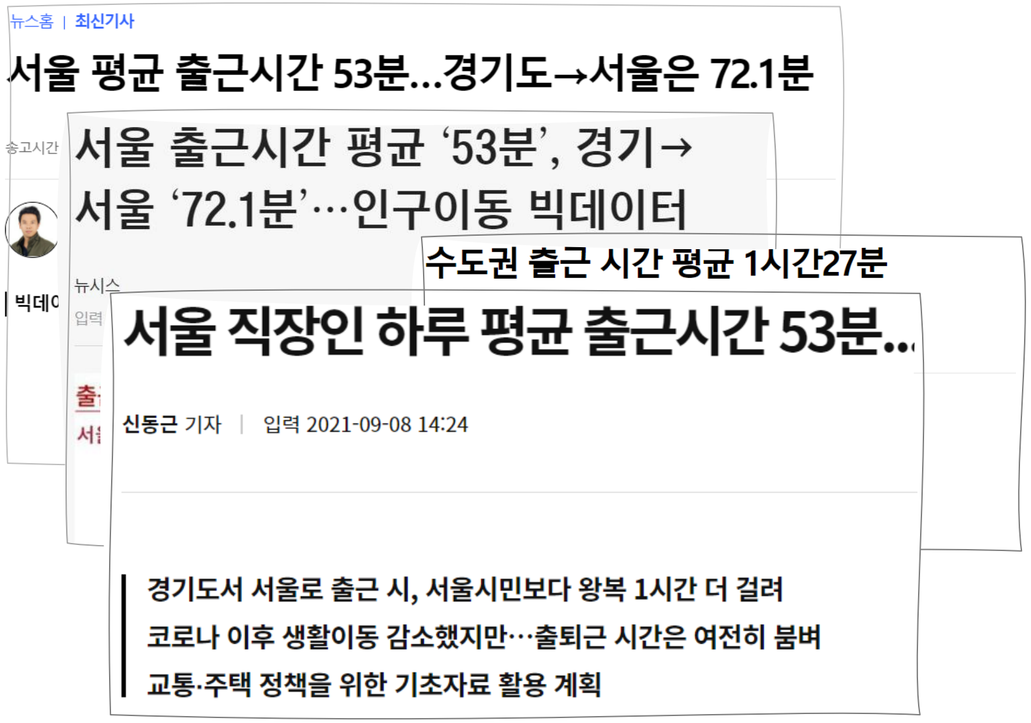
이다솔
캠퍼 ID: T2154
#PPT #Clustering #TTS

02

프로젝트 소개

귀가노니 - 출퇴근길에 듣는 인공지능 뉴스 팟캐스트

02 프로젝트 소개



서울 평균 출근시간 53분, 경기도→서울은 72분
이 시간을 유익하게 쓸 수는 없을까?

+

Spotify®

밀리의 서재, 클럽하우스 등
오디오 서비스 플랫폼의 등장

=

어제 기사들에 대한 주제별 클러스터링을 이용한
개인맞춤형 인공지능 TTS 팟캐스트, 귀가노니

03

모듈별 소개

귀가노니 - 출퇴근길에 듣는 인공지능 뉴스 팟캐스트

Clustering

Model Serving

FastAPI

TTS

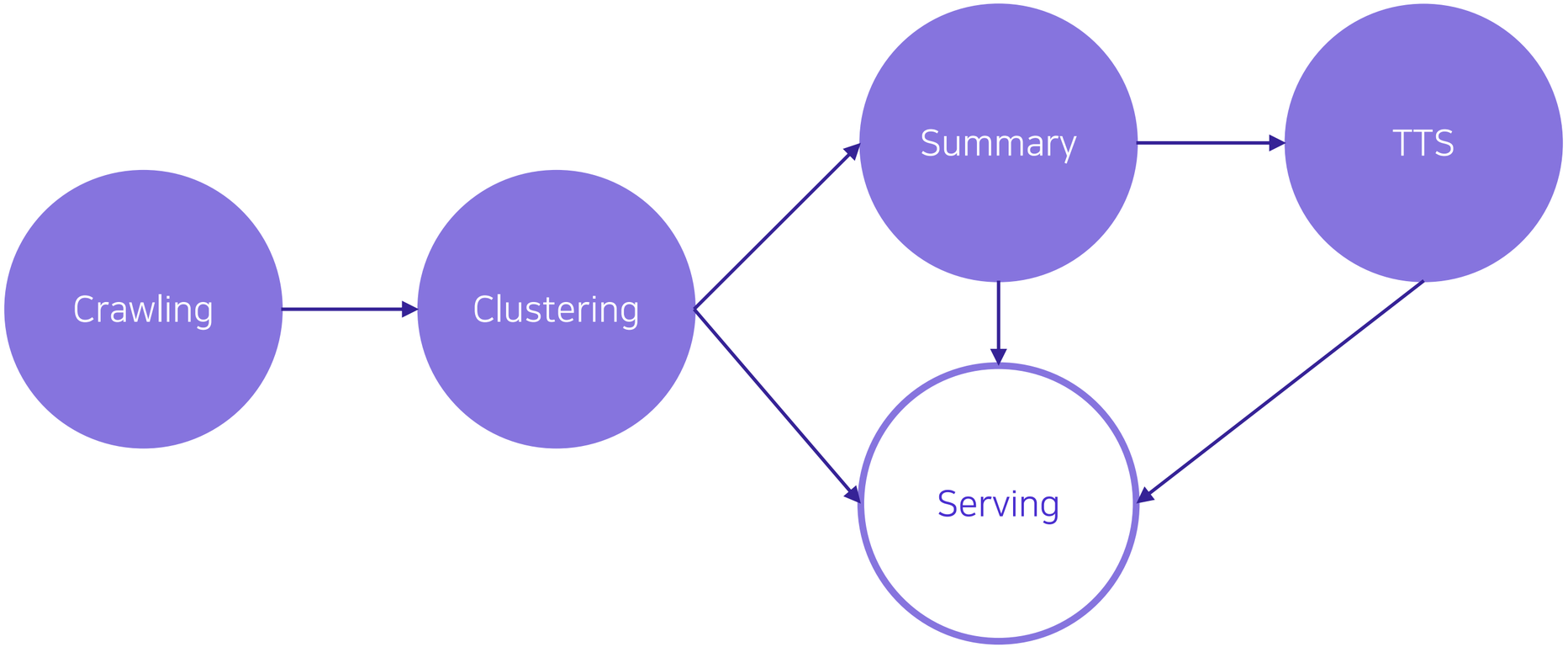
EasyBART

Crawling

03 모듈별 소개

쉘 스크립트에 각각의 모듈을 순차적으로 실행하는 파이썬 명령어 정의

Crontab을 통해 매일 밤 12시에 어제 날짜에 해당하는 데이터가 생성되도록 지정



<데이터 생성 배치 프로세싱>

03 모듈별 소개

담당: 김성한, 김제우

01 Crawling

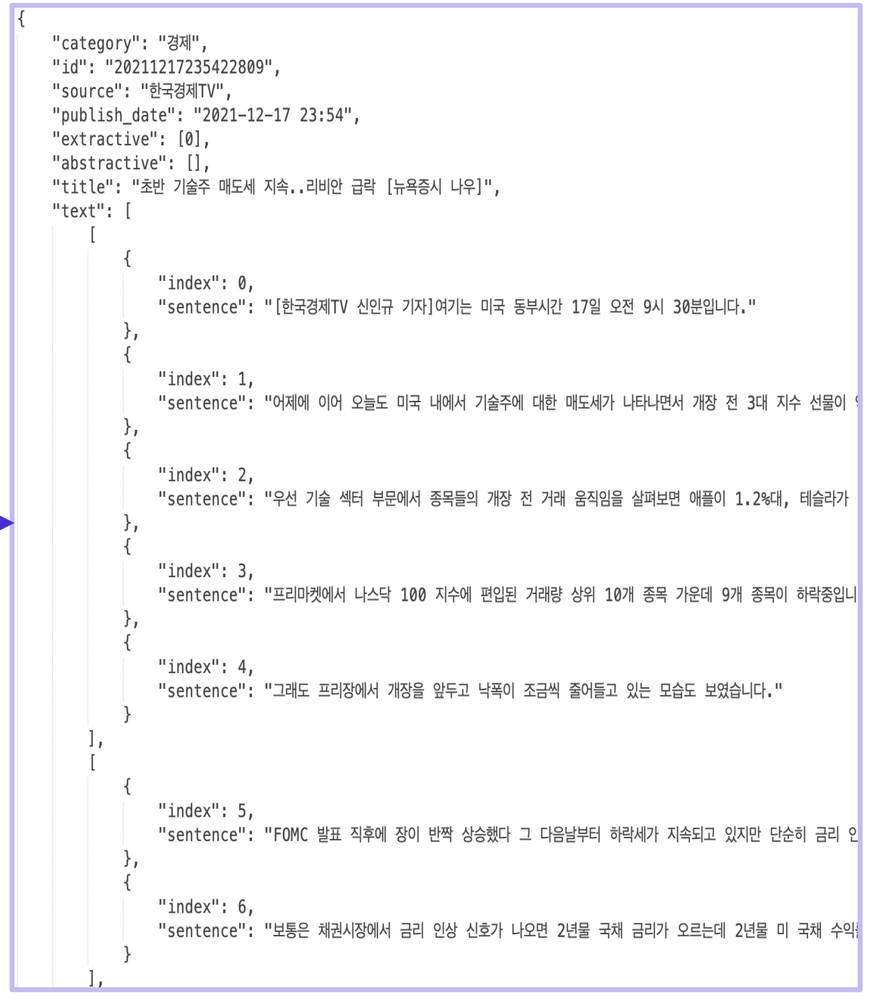
02 Clustering

03 Summarization

04 Text to Speech

05 Serving

카테고리당 약 3,000 ~ 4,000개 기사



AI Hub 포맷의 JSON 파일

03 모듈별 소개

담당: 김성한, 김제우

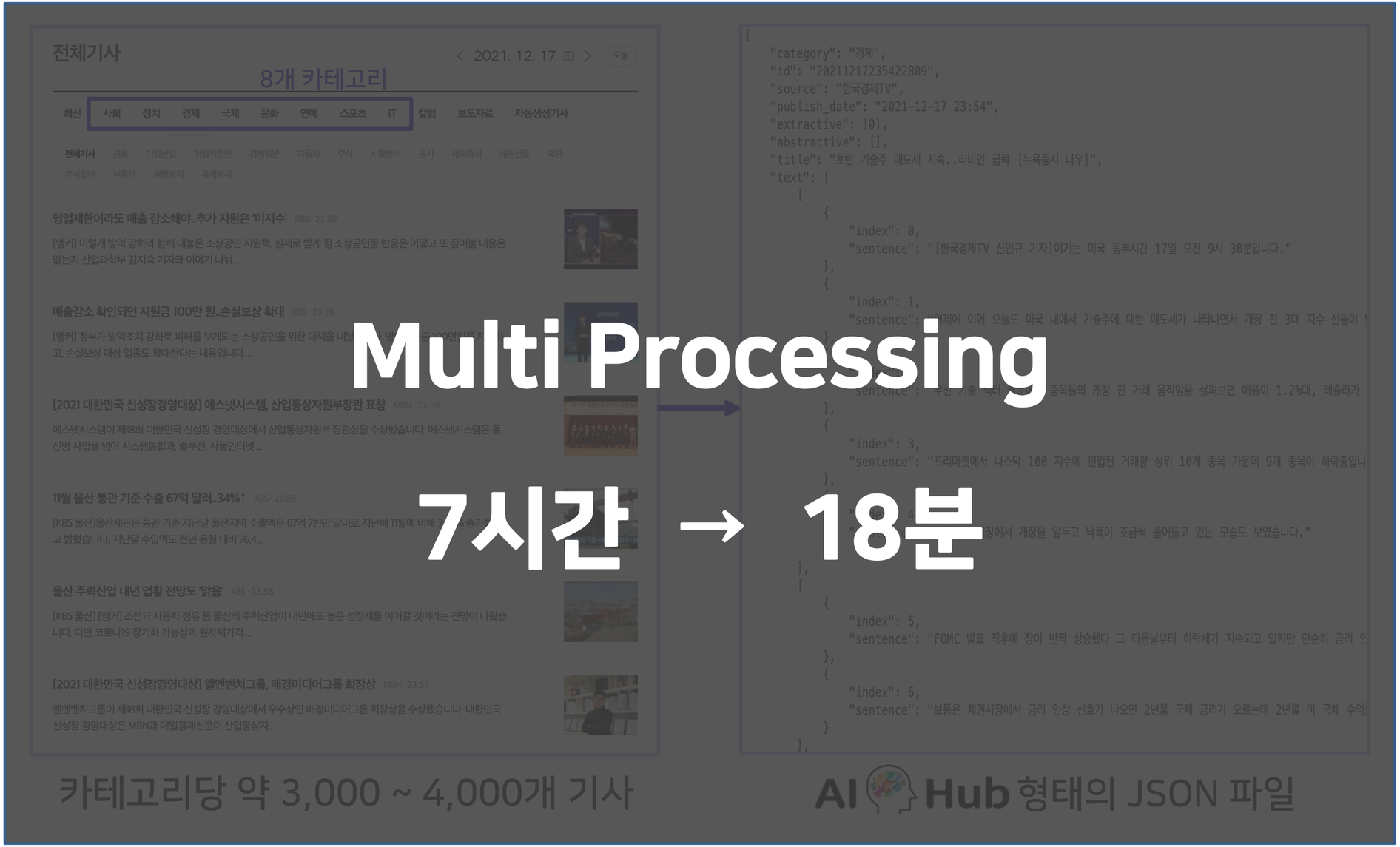
01 Crawling

02 Clustering

03 Summarization

04 Text to Speech

05 Serving



03 모듈별 소개

담당: 권혜빈, 이노아, 이다솔

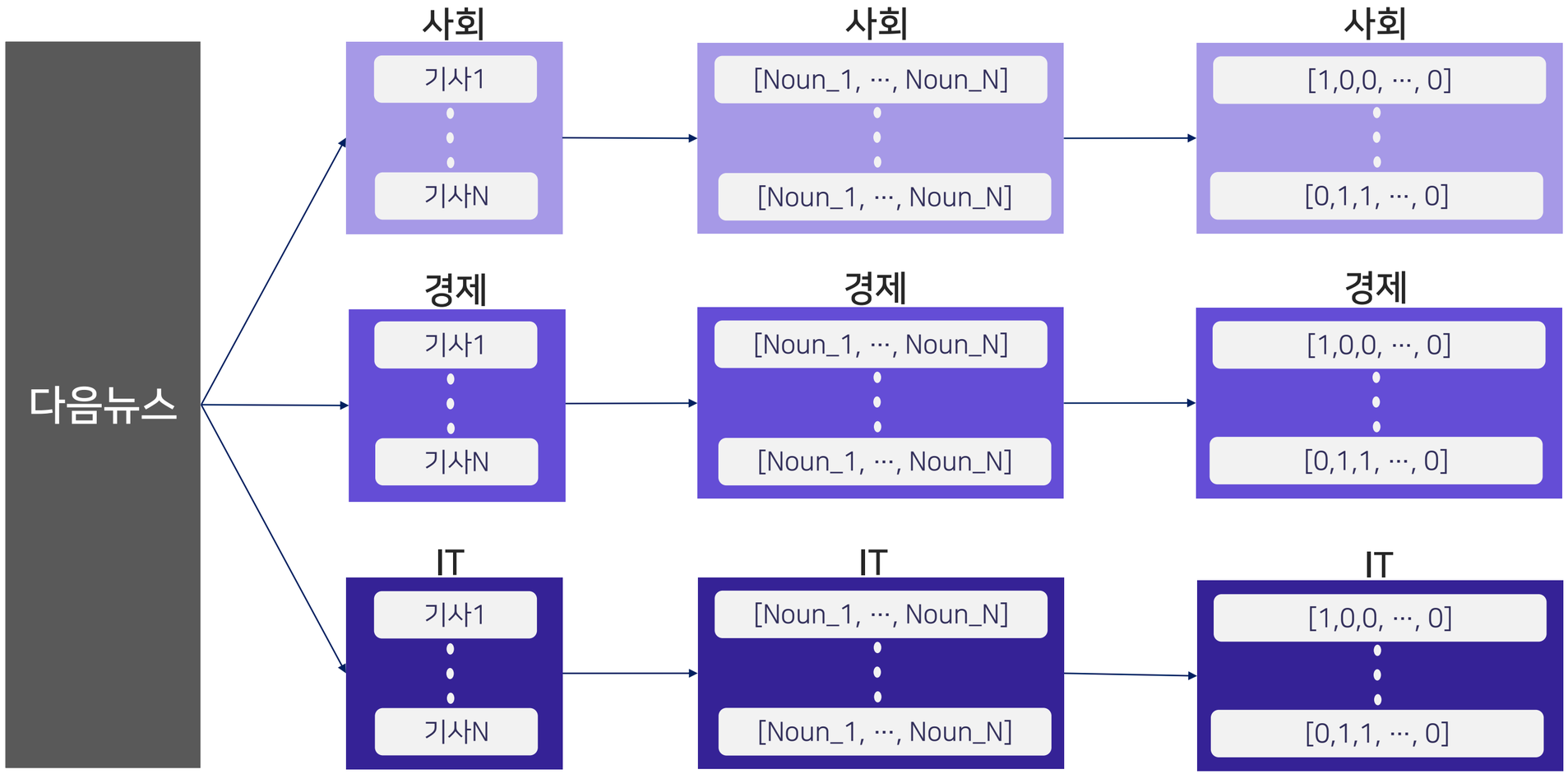
01 Crawling

02 Clustering

03 Summarization

04 Text to Speech

05 Serving



1) 전처리 & 필터링

- 특수문자 제거
- 문장 개수 3개 미만
- 본문 길이 300자 미만

2) 명사 추출

- 6개의 형태소 분석기 비교
- Hannanum[1)]

3) 벡터화

- 제목 + 본문 TF-IDF 변환

1) http://semanticweb.kaist.ac.kr/hannanum/

03 모듈별 소개

담당: 권혜빈, 이노아, 이다솔

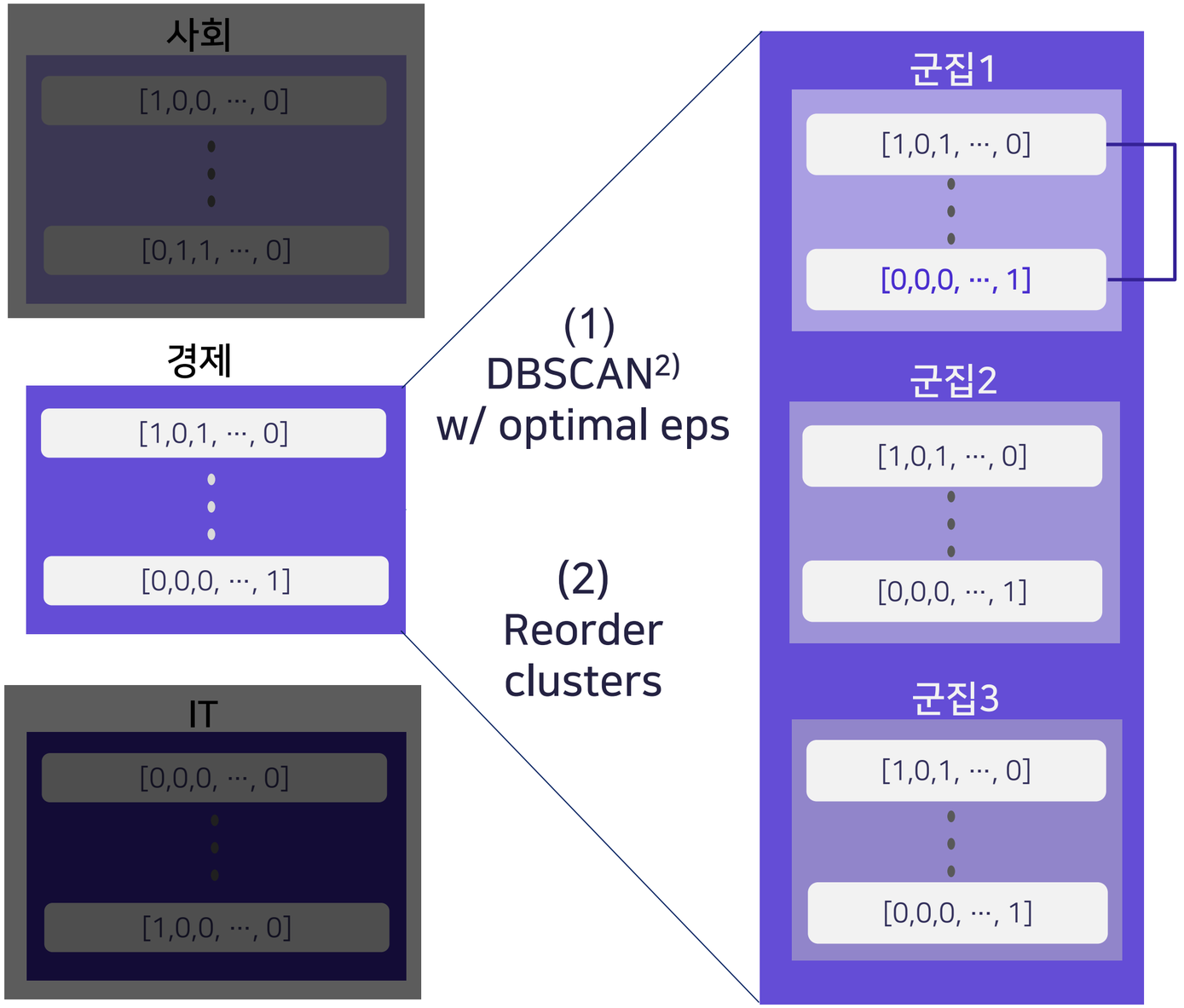
01 Crawling

02 Clustering

03 Summarization

04 Text to Speech

05 Serving



$$center\bar{V}_j$$

$$center_j = \bar{V}_j = \frac{1}{N_j}\sum_{i=1}^{N_j} V_{ij}$$

$$\boldsymbol{J} = [1, \ldots, k]$$

$$\boldsymbol{N} = [N_1, \ldots, N_k]$$

$$featured\ article_j$$

$$\underset{x \in J}{argmin}\ dist(V_{xj}, \bar{V}_j)$$

$$\Rightarrow V_{xj}$$

$$top\ k\ keywords_j$$

$$\bar{V}_j = [x_1, \ldots, x_M]$$

$$\Rightarrow [x_{top\_1}, x_{top\_2}, \ldots]$$

2) Ester, Kriegel, et al. "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise" (1996)

01 Crawling

02 Clustering

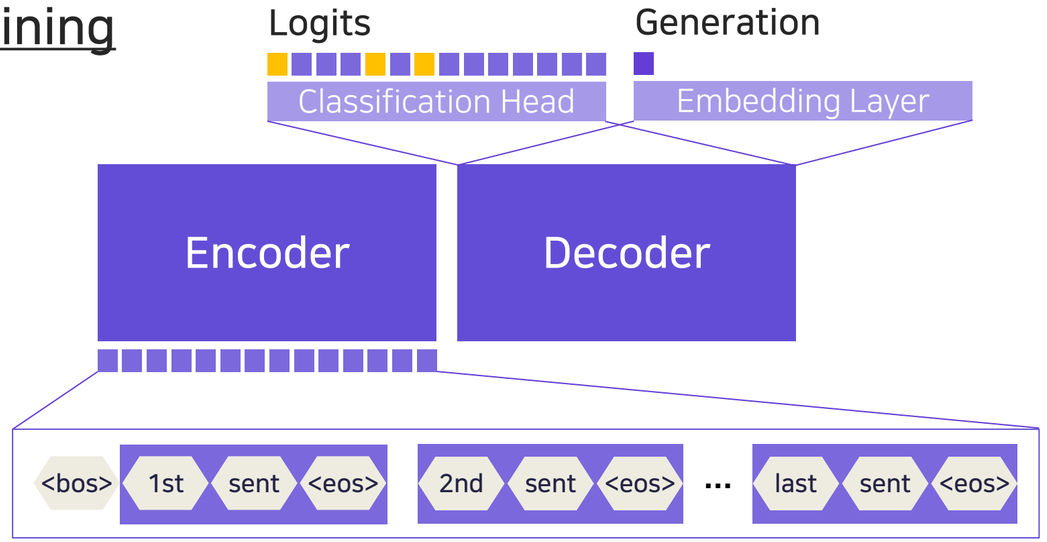
03 Summarization

04 Text to Speech

05 Serving

EasyBART – Extractive/Abstractive SummarY

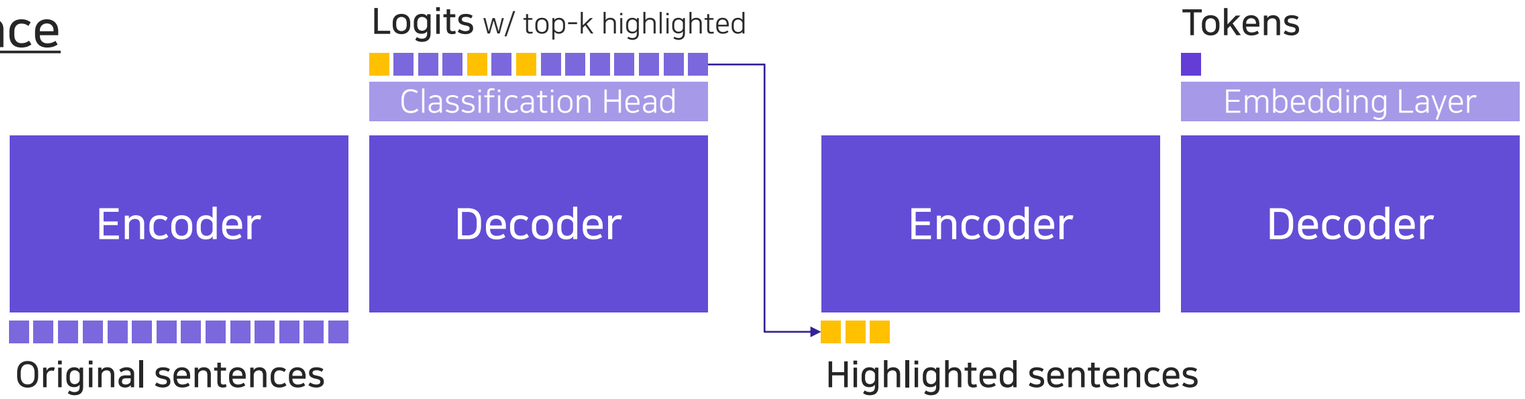
Training



> One-Step Training &
> Two-Step Inference

- BART[3]에 **custom classification head**를 추가하고, inference logic을 수정한 모델
- 추출/생성 요약 동시 진행 + 모델 구조와 훈련이 쉽다는 점에서 **"EasyBART"**로 명명
- Input 형태가 BERTSum[4]과 유사하지만, <sep> 대신 <eos> 토큰을 활용

Inference



3) Lewis, Mike, et al. "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension." *arXiv preprint arXiv:1910.13461* (2019).
4) Liu, Yang. "Fine-tune BERT for extractive summarization." *arXiv preprint arXiv:1903.10318* (2019).

담당: 김민진, 김상현

01 Crawling

02 Clustering

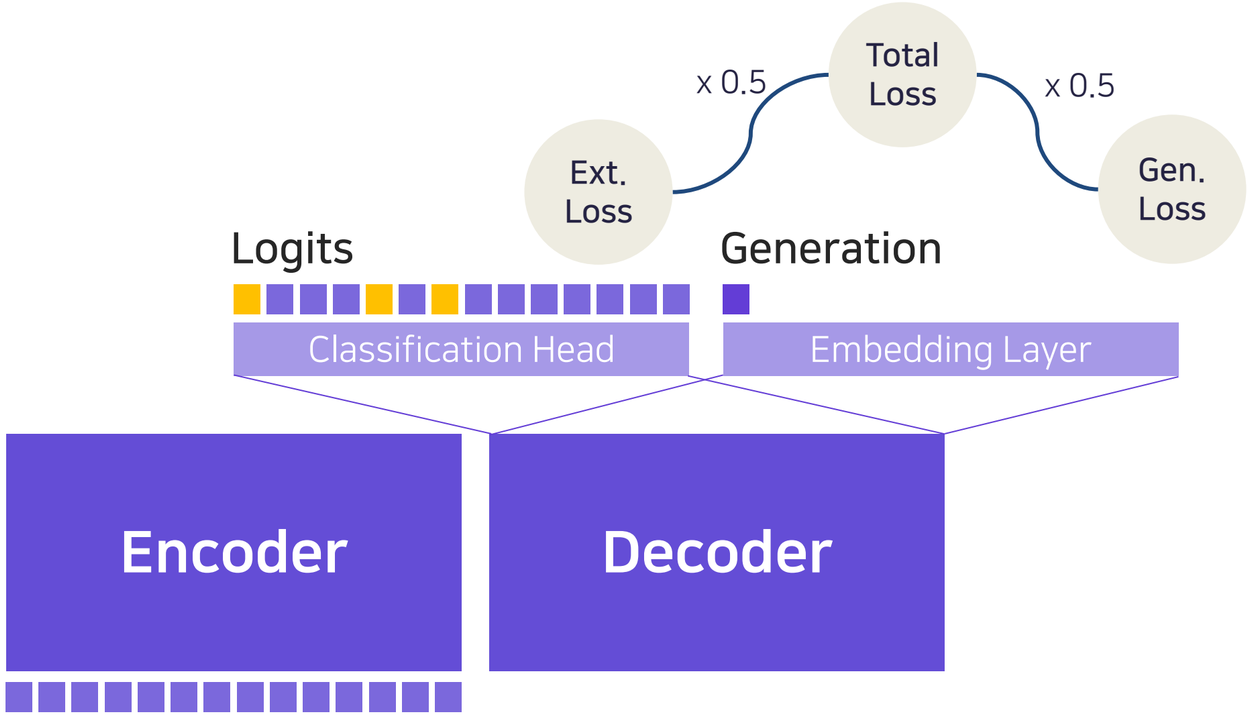
03 Summarization

04 Text to Speech

05 Serving

EasyBART – Extractive/Abstractive SummarY

Training



- 학습 데이터: **AI Hub의 "문서 요약 텍스트"**[5] 말뭉치 - 추출/생성요약 라벨 모두 존재
- 추출 요약문 기반의 생성 요약문 라벨은 존재하지 않으므로, **생성 요약 시에도 원문 그대로를 input**으로 활용

5) 본 AI 데이터는 한국지능정보사회진흥원의 사업 결과입니다. https://aihub.or.kr/aidata/8054

담당: 김민진, 김상현

01 Crawling

02 Clustering

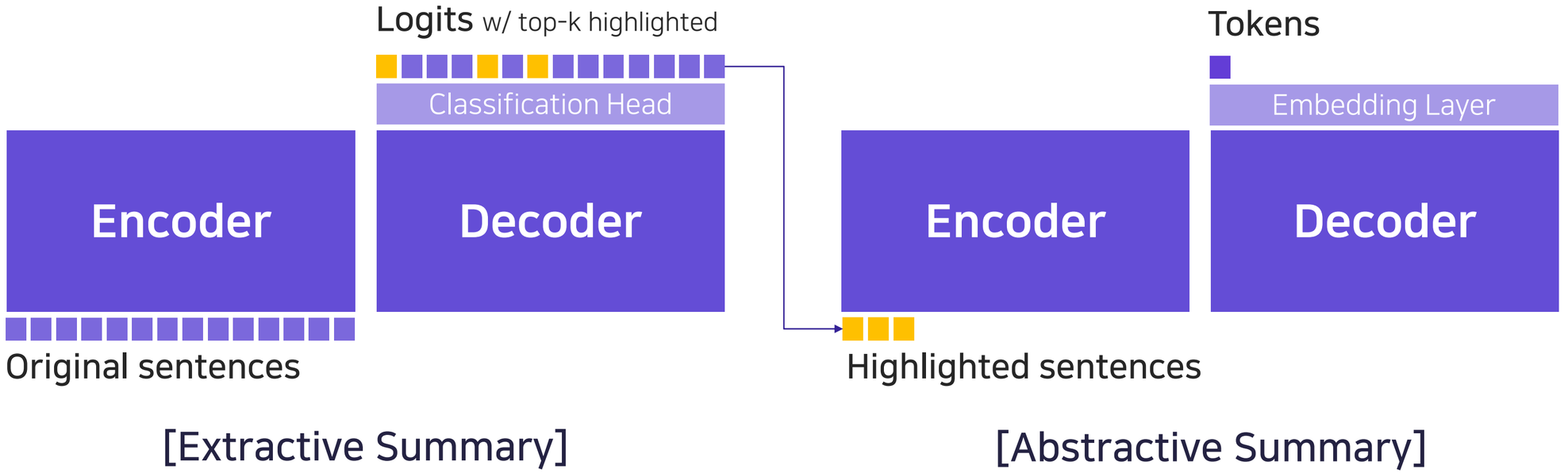
03 Summarization

04 Text to Speech

05 Serving

EasyBART – Extractive/Abstractive SummarY

Inference



- 선정된 기사 본문에서 **주요 문장 top-k개 추출** 후, 추출된 문장으로 **요약문 생성**
- 원문이 긴 경우에도 **recursive하게 top-k를 반복적으로 추출**하도록 개선 (서비스에는 미적용)

01 Crawling

02 Clustering

03 Summarization

04 Text to Speech

05 Serving

Example #1

[KoBART-summarization]
"김경남 소속사 제이알이엔티는 배우 김경남 측이 측간 소음으로 이웃에게 피해를 준 데 대해 17일 공식입장을 통해 먼저 좋지 않은 일로 심려를 끼쳐 죄송하다고 사과했다."

[EasyBART]
"김경남 소속사 제이알이엔티는 배우 김경남 측이 측간 소음으로 이웃에게 피해를 준 데 대해 17일 공식입장을 통해 먼저 좋지 않은 일로 심려를 끼쳐 죄송하다고 사과했고, 이날 저녁 김경남 배우가 당사자 분을 찾아가 이야기를 나눴다며 진심으로 사과드리고 앞으로는 더 주의하겠다고 말씀드렸다고 전했다."

담당: 김민진, 김상현

01 Crawling

02 Clustering

03 Summarization

04 Text to Speech

05 Serving

EasyBART – Extractive/Abstractive SummarY

A/B Test 결과 분석

4가지 요약문에 대한 Friedman Test

	EasyBART			BART
	Top-3	Top-5	Top-7	baseline[6]
접수합*	1755.5	1629.5	**1841.5**	1673.5
순위합*	296.0	255.5	**333.0**	265.5

- $\chi^2 = 21.452\ (df = 3)$
- $p < 0.005$

→ 4가지 요약문의 품질 차이가 존재

Top-3 vs. Baseline Sign Test

	EasyBART	BART	Total
	Top-3	baseline	
응답수*	**610**	310	920
비율	**66.3%**	33.7%	

- $\hat{p} = 0.663$
- $p < 0.005$

→ Top-3를 압도적으로 선호

"k에 따른 **readability**와 **information** 사이의 trade-off가 존재"

일종의 hyperparameter로, 도메인 및 목적에 따른 tuning 가능성

Top-5의 성능이 낮은 이유도 두 측면 모두 열등하기 때문

* 높을수록 긍정적인 평가를 의미함. 자세한 방법론은 Appendix 참조.

6) https://huggingface.co/gogamza/kobart-summarization

03 모듈별 소개

담당: 권혜빈, 김제우, 이노아, 이다솔

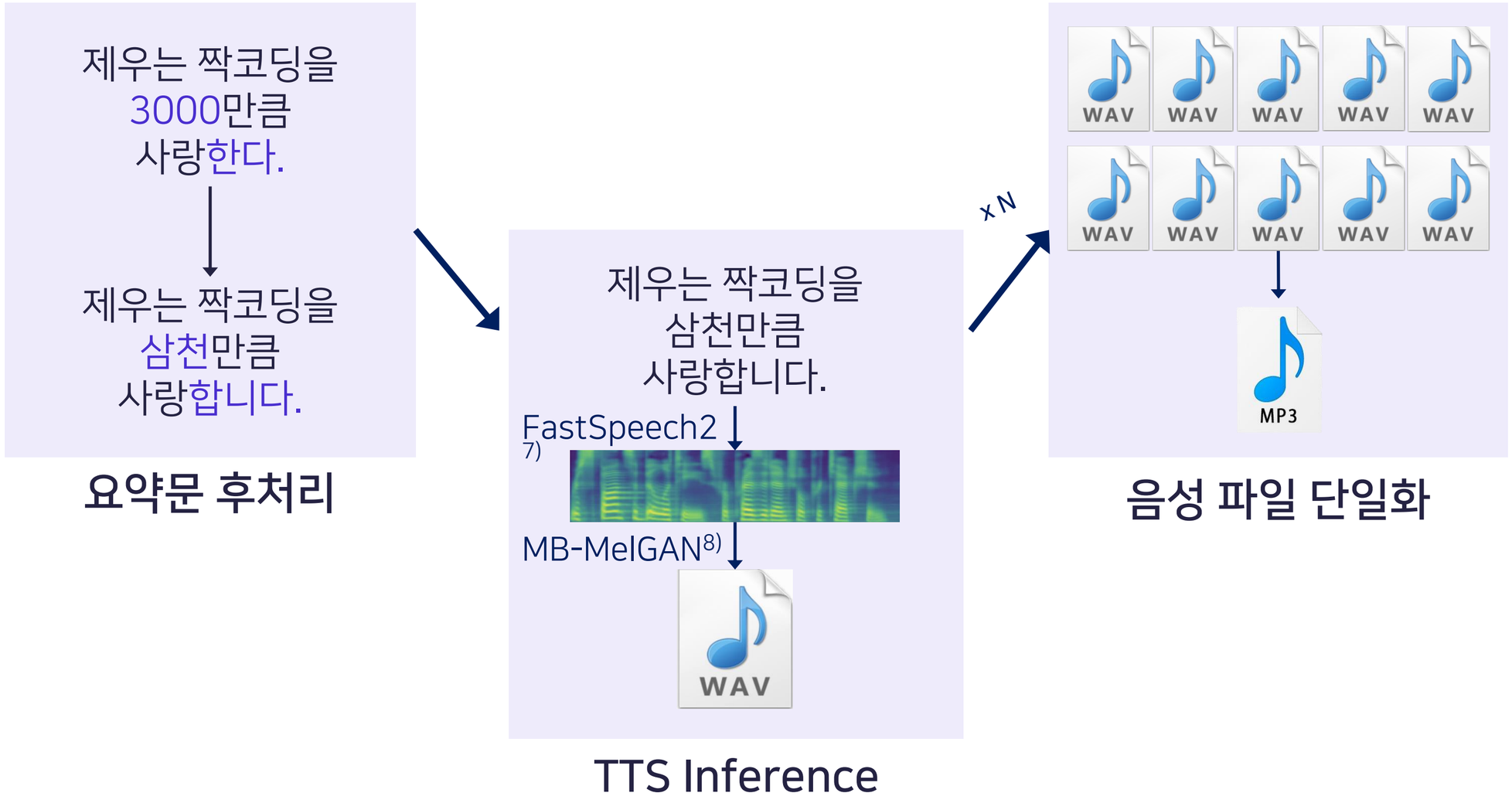
01 Crawling

02 Clustering

03 Summarization

04
Text to Speech

05 Serving



7) https://huggingface.co/tensorspeech/tts-fastspeech2-kss-ko
8) https://huggingface.co/tensorspeech/tts-mb_melgan-kss-ko

담당: 권혜빈, 김제우, 이노아, 이다솔

01 Crawling

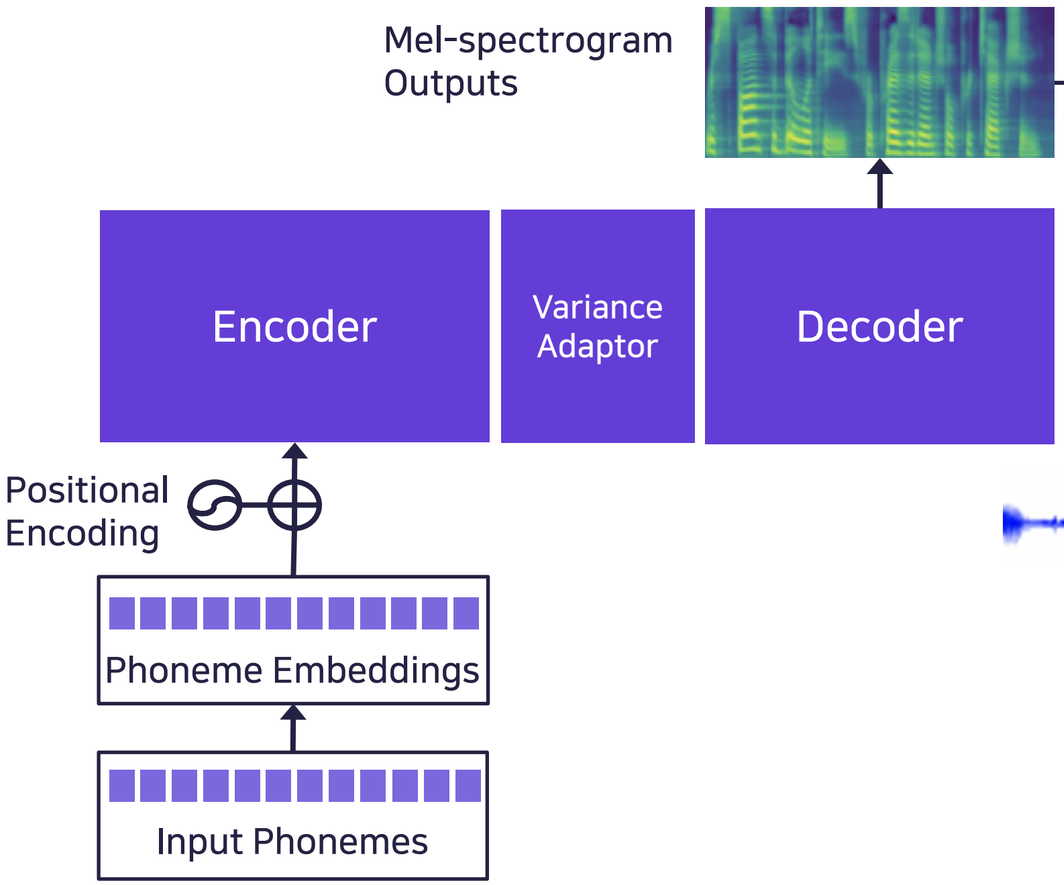
02 Clustering

03 Summarization

04
Text to Speech

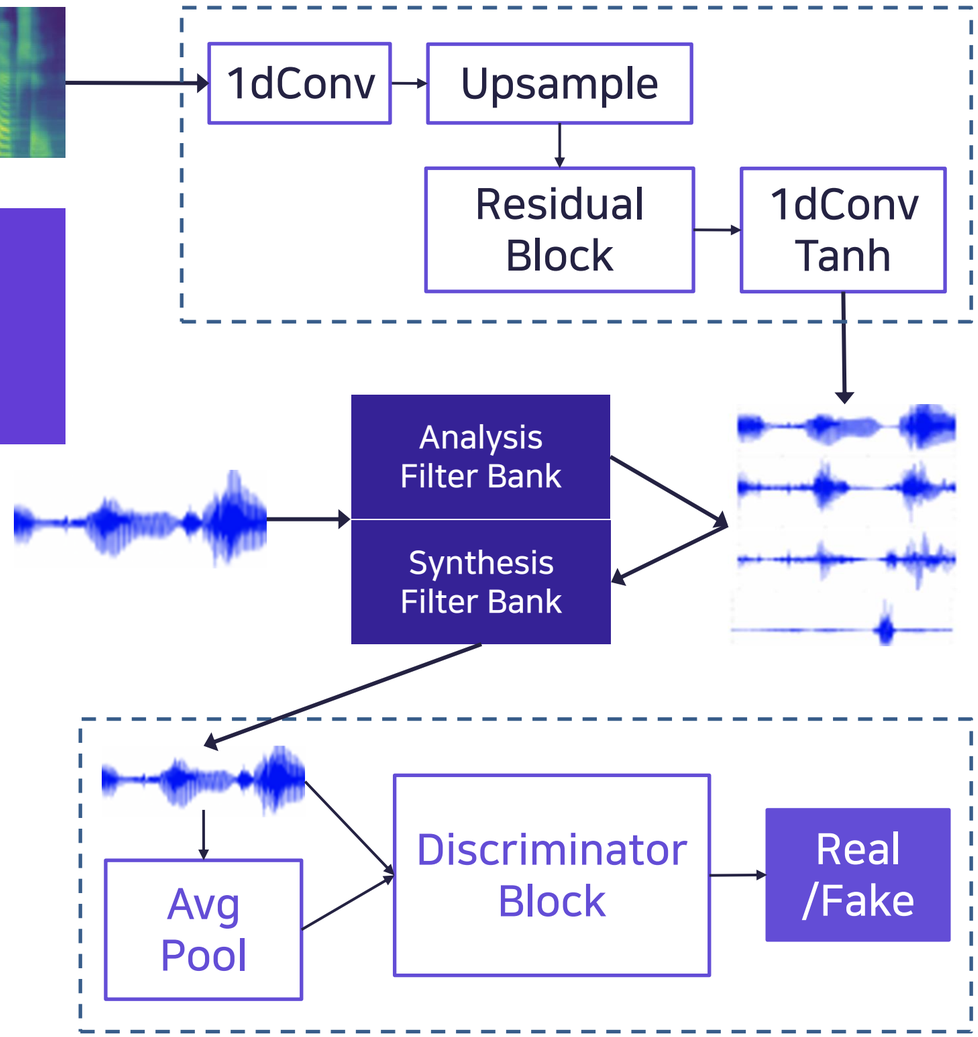
05 Serving

TTS 모델 구조



<FastSpeech2>

<Multi-Band MelGAN>



03 모듈별 소개

담당: 김성한, 김제우

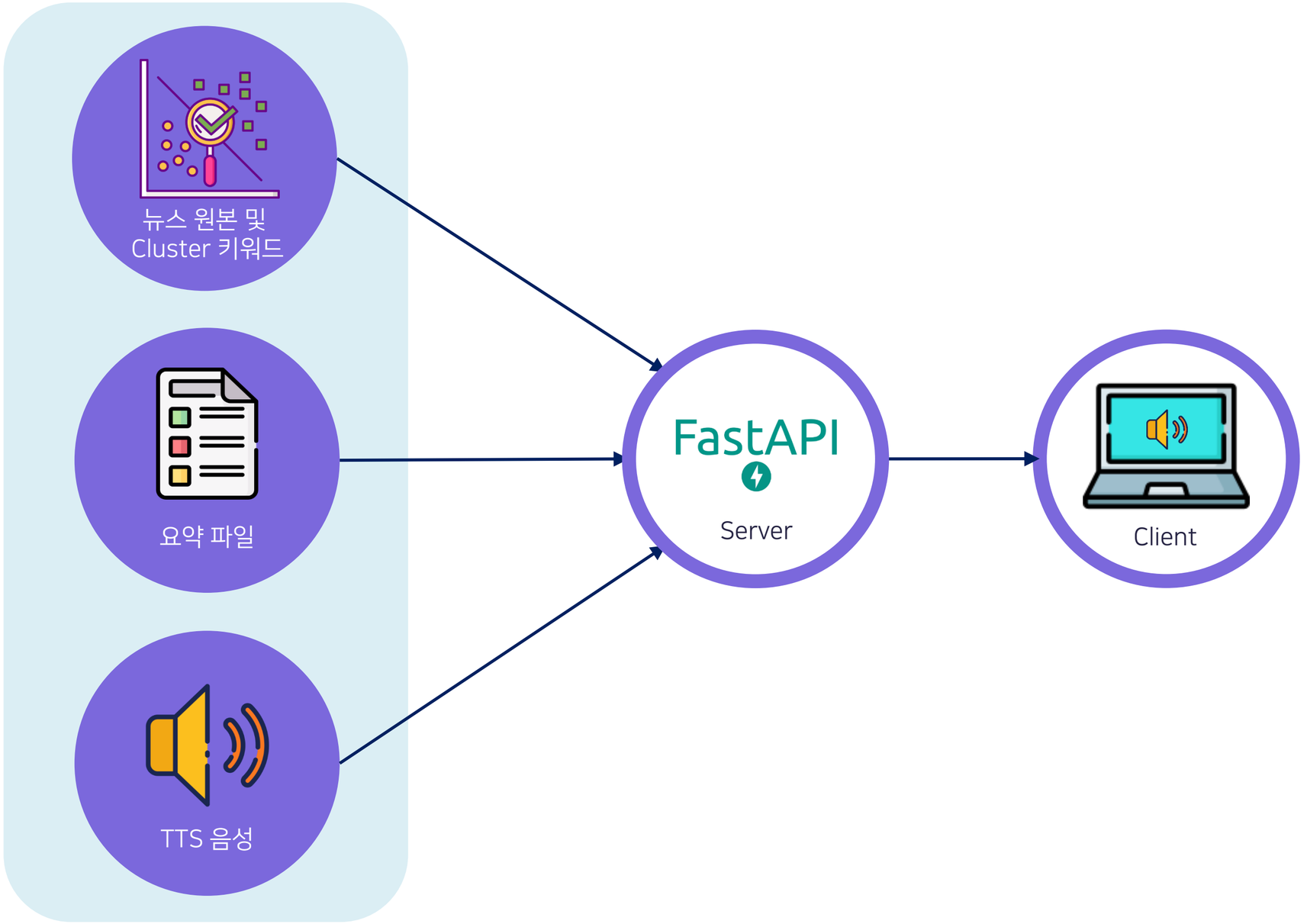
01 Crawling

02 Clustering

03 Summarization

04 Text to Speech

05 Serving



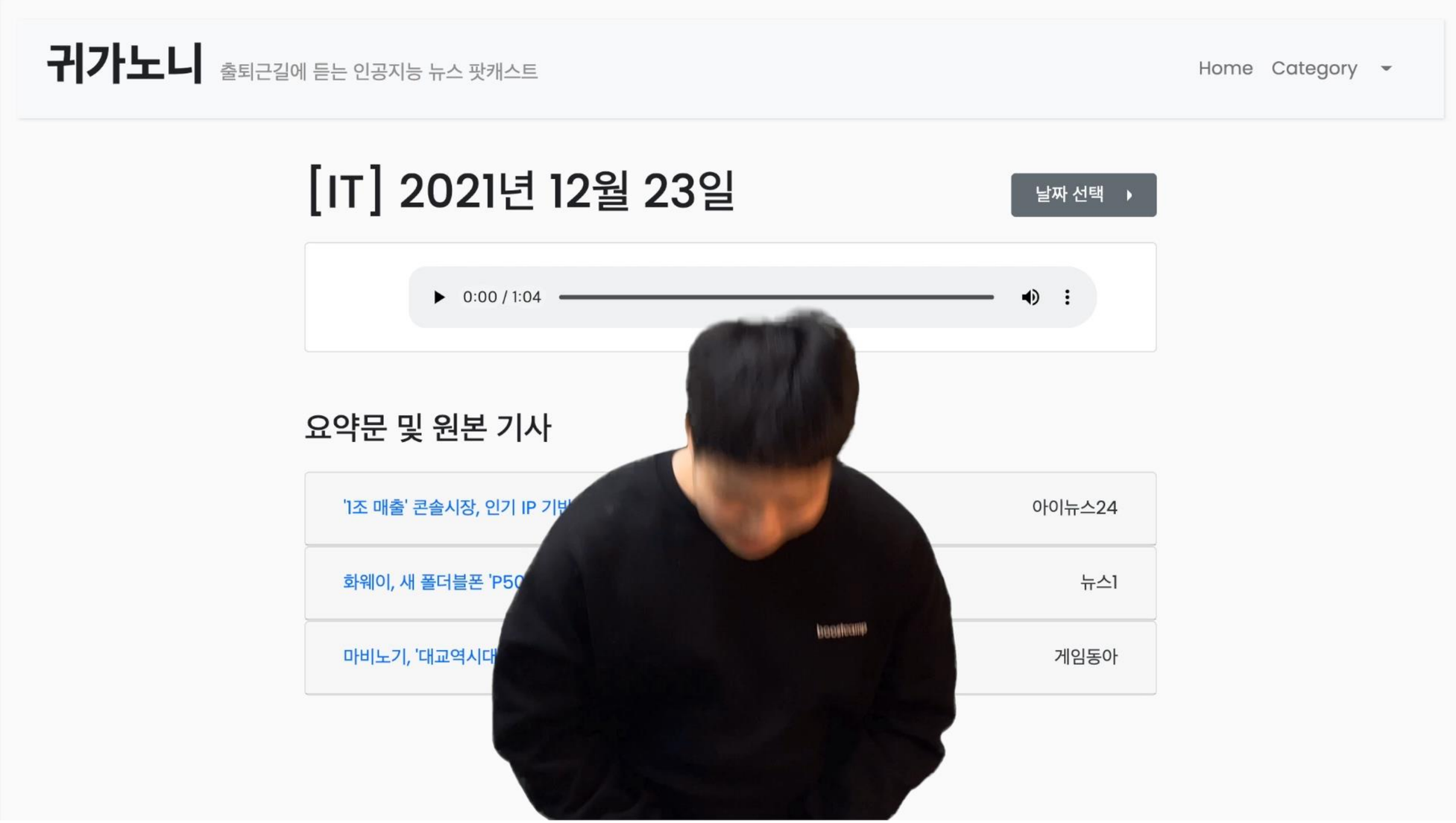
04

시연 영상

귀가노니 - 출퇴근길에 듣는 인공지능 뉴스 팟캐스트

Demo

04 시연 영상

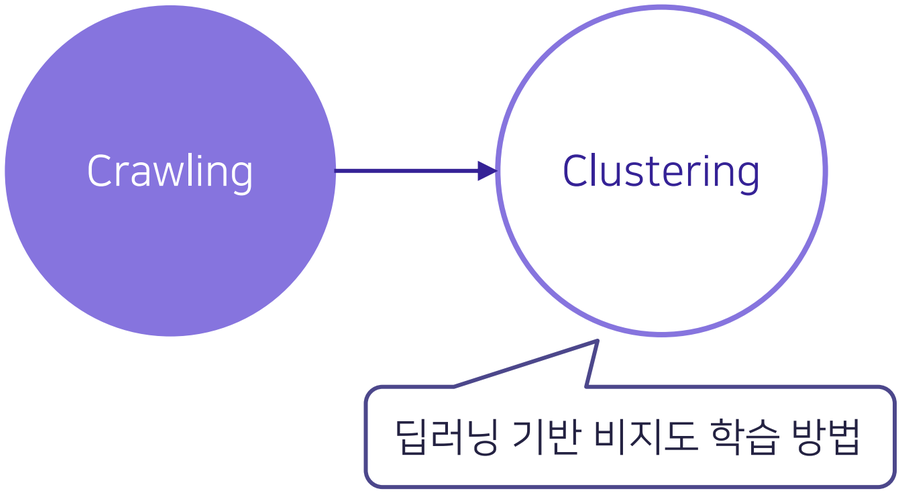

귀가노니
출퇴근길에 듣는 인공지능 뉴스 팟캐스트
Home
Category
[IT] 2021년 12월 23일
날짜 선택
0:00 / 1:04
요약문 및 원본 기사
'1조 매출' 콘솔시장, 인기 IP 기반
아이뉴스24
화웨이, 새 폴더블폰 'P50
뉴스1
마비노기, '대교역시대
게임동아

05

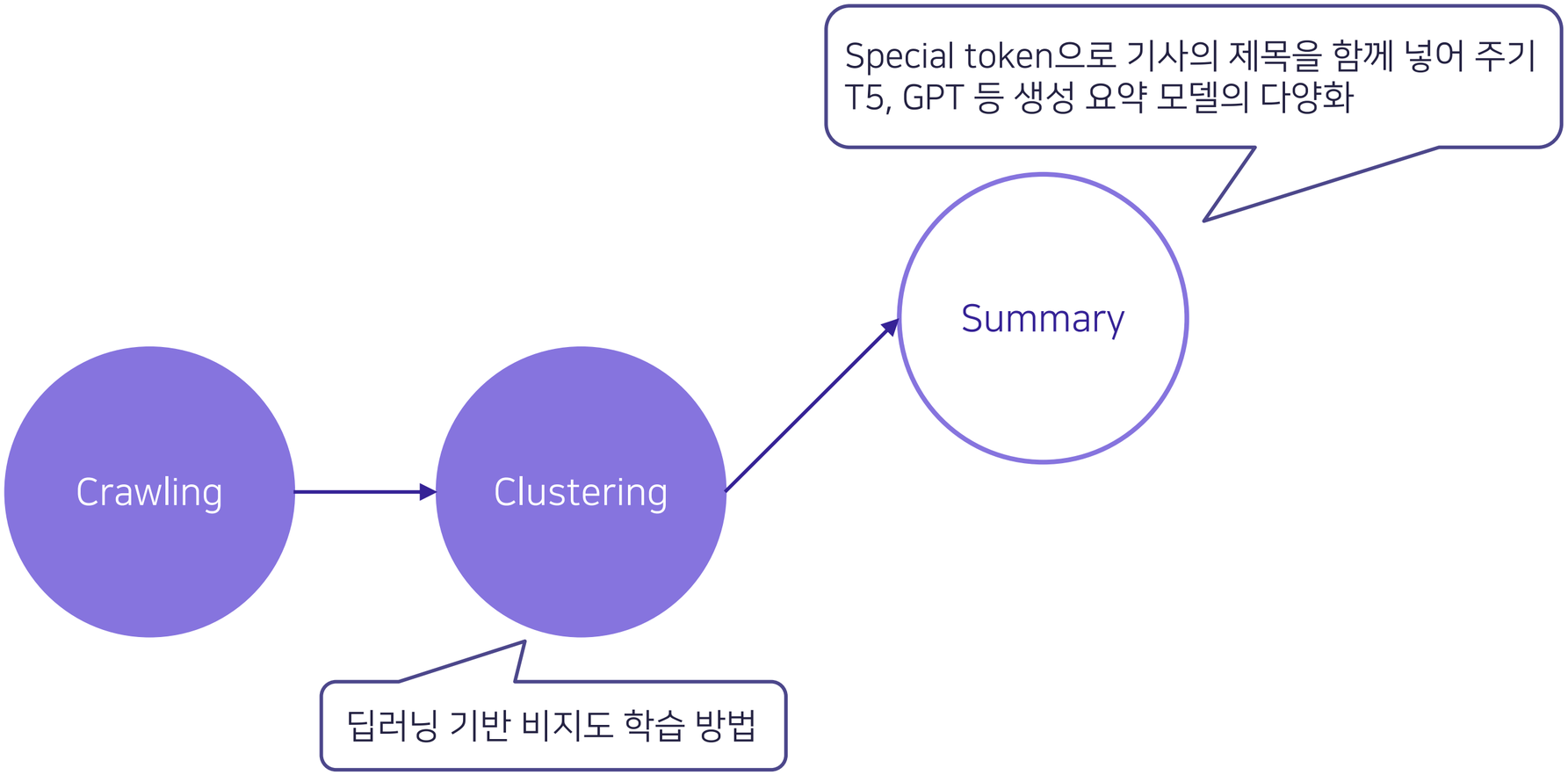
향후 개선사항

귀가노니 - 출퇴근길에 듣는 인공지능 뉴스 팟캐스트

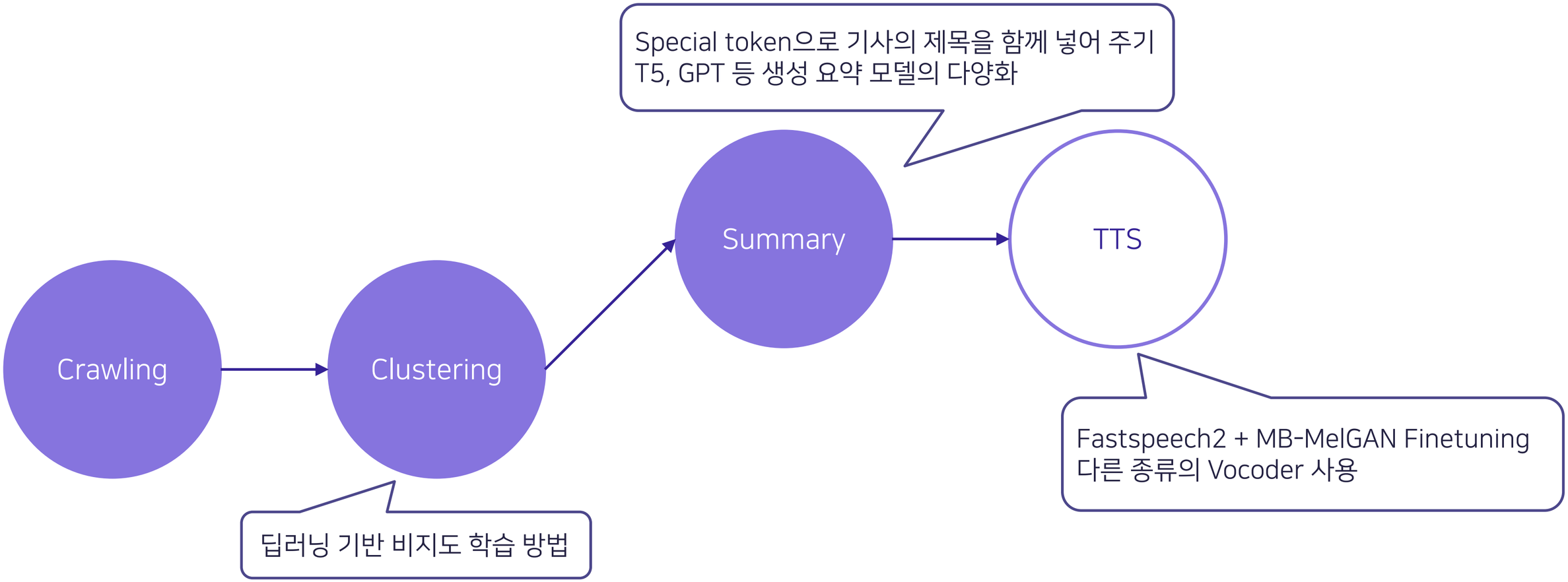
04 향후 개선 사항



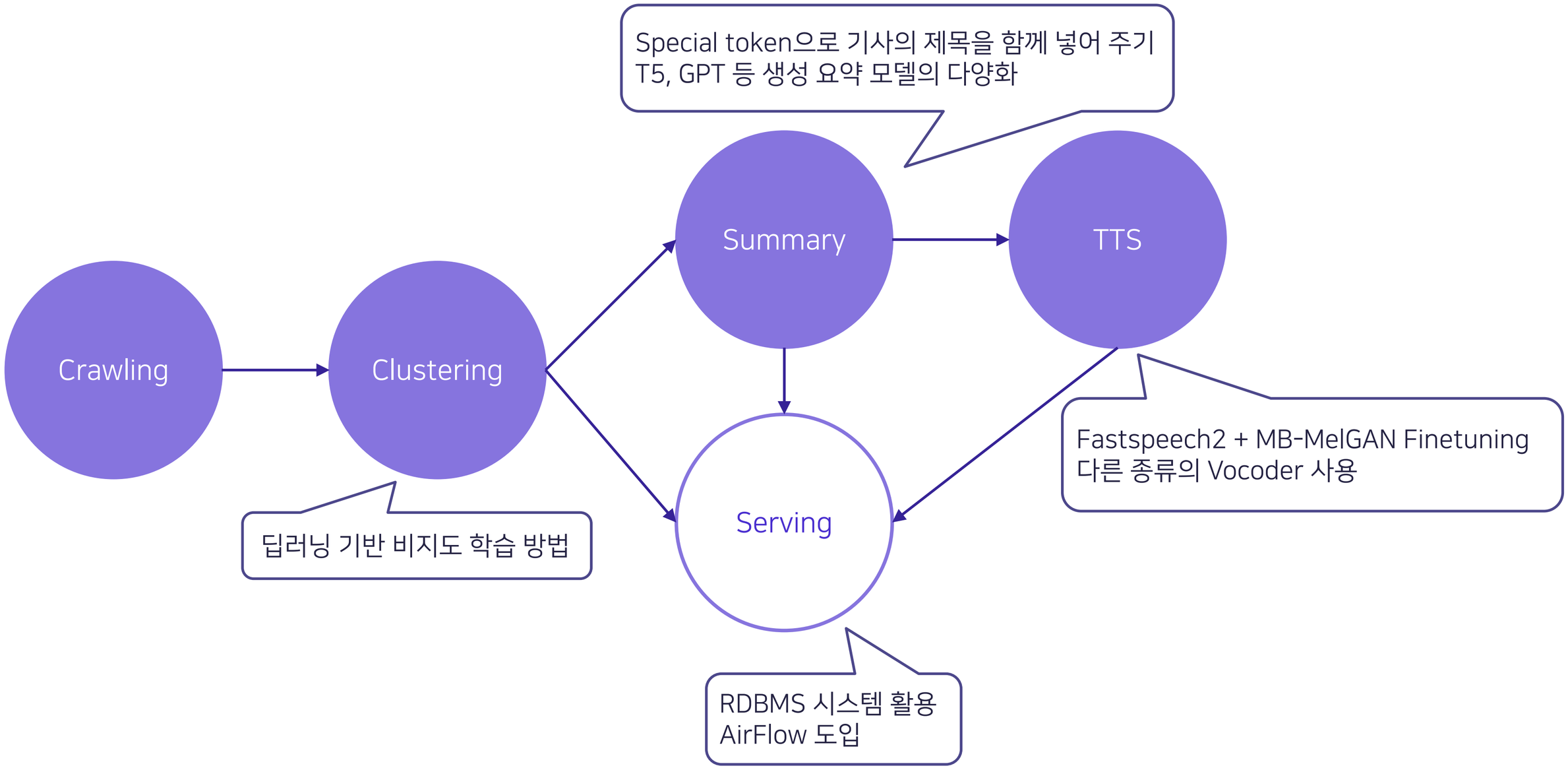
04 향후 개선 사항


Special token으로 기사의 제목을 함께 넣어 주기
T5, GPT 등 생성 요약 모델의 다양화
Summary
Crawling
Clustering
딥러닝 기반 비지도 학습 방법

04 향후 개선 사항


Special token으로 기사의 제목을 함께 넣어 주기
T5, GPT 등 생성 요약 모델의 다양화
Crawling
Clustering
Summary
TTS
딥러닝 기반 비지도 학습 방법
Fastspeech2 + MB-MelGAN Finetuning
다른 종류의 Vocoder 사용

04 향후 개선 사항


Special token으로 기사의 제목을 함께 넣어 주기
T5, GPT 등 생성 요약 모델의 다양화
Summary
TTS
Crawling
Clustering
Serving
Fastspeech2 + MB-MelGAN Finetuning
다른 종류의 Vocoder 사용
딥러닝 기반 비지도 학습 방법
RDBMS 시스템 활용
AirFlow 도입

06

Q & A

귀가노니 - 출퇴근길에 듣는 인공지능 뉴스 팟캐스트

감사합니다.

boostcamp ai tech

Example #2: 유럽 에너지 대란에 산업계도 초비상.. “금속, 비료 생산 감축” (국제)

[KoBART-summarization]

유럽 천연가스 공급난으로 가격이 급등하면서 금속 제련, 비료 제조 관련 기업들은 수익성이 낮아져 생존까지 위협받고 있으며 블룸버그통신은 "유럽 천연가스 공급난으로 가격이 급등하면서 산업계 전체가 막대한 재정적 피해를 보고 있다"며 "금속 제련소와 비료 제조업체는 이미 생산량 감축에 들어갔다"고 보도했다.

[EasyBART]

22일(현지시간) 블룸버그통신은 "유럽 천연가스 공급난으로 가격이 급등하면서 산업계 전체가 막대한 재정적 피해를 보고 있다"며 "금속 제련소와 비료 제조업체는 이미 생산량 감축에 들어갔다"고 보도했다.

✓ 주로 학습된 도메인에서 벗어나 새로운 단어들이 등장하는 경우 반복되는 표현 억제됨

담당: 김민진, 김상현

Example #3: 한은 “집값 폭등해 금융불균형 심화시 성장률 3퍼센트 추락” (사회)

[KoBART-summarization]

"한국은행은 국내외 금융불균형 상황에서 실물경제 하방리스크를 점검하기 위해 국내외 금융취약성지수를 활용한 GaR(최대성장감소율) 분석을 실시한 결과 국내 금융불균형 상황은 실물경제 하방리스크를 확대시키고, 특히 주요국 금융불균형을 감안할 경우 국내 실물경제 하방 리스크가 더 커지는 것으로 나타났다."

[EasyBART]

"한국은행이 국내외 금융불균형 상황에서 실물경제 하방리스크를 점검하기 위해 국내외 금융취약성지수를 활용한 GaR(최대성장감소율) 분석을 실시한 결과 부동산 가격 등 자산가격 상승에 따른 '금융불균형'이 심화될 경우 우리나라 경제성장률(GDP)이 최악의 경우 3퍼센트를 기록할 수 있다는 전망을 내놨다."

✓ **문법적(주술 관계)으로 더 나은 문장, 제목을 주지 않았음에도 원문의 핵심이 담김**

담당: 김민진, 김상현

Example #3: 한은 "집값 폭등해 금융불균형 심화시 성장률 3퍼센트 추락" (사회)

[원문]

한국은행이 부동산 가격 등 자산가격 상승에 따른 '금융불균형'이 심화될 경우 우리나라 경제성장률(GDP)이 최악의 경우 3퍼센트를 기록할 수 있다는 전망을 내놨다.

한국은행이 23일 발표한 '2021년 하반기 금융안정 보고서'에 따르면 한은은 국내외 금융불균형 상황에서 실물경제 하방리스크를 점검하기 위해 국내외 금융취약성지수를 활용한 GaR(최대성장감소율) 분석을 실시했다.

분석 결과 국내 금융불균형 상황은 실물경제 하방리스크를 확대시키고, 특히 주요국 금융불균형을 감안할 경우 국내 실물경제 하방 리스크가 더 커지는 것으로 나타났다.

한은 시산 결과 올 3분기 GDP는 국내 금융취약성지수만을 고려하면 10퍼센트의 확률로 1.4퍼센트를 기록하고, 여기에 주요국의 금융취약성지수를 추가하면 3.0퍼센트로 크게 낮아지는 것으로 나타났다.

국내 금융불균형은 최근 들어 일부 개선되고 있으나 과거에 비해서는 부동산 시장을 중심으로 크게 누중된 상황이다.

(…)

담당: 김민진, 김상현

Example #4: KT, 고속 양자암호통신 독자 기술 개발 (IT)

[KoBART-summarization]

"22일 KT는KT는 빛의 가장 작은 단위인 광자를 이용해 정보를 전달하는 차세대 통신 기술인 고속 양자암호통신용 독자 기술을 독자적으로 개발했다고 밝혔다."

[EasyBART]

"22일 KT는 20kbps(초당 킬로비트) 속도의 고속 양자암호통신 기술을 독자적으로 개발했다고 밝혔으며 정부의 디지털뉴딜 양자암호통신 속도 기준(5bps)에 비해 4000배 빠르다."

✓ 주로 학습된 도메인에서 벗어나 새로운 단어들이 등장하는 경우 반복되는 표현 억제와 핵심을 잘 요약함

담당: 김상현

A/B Testing 시행 방법 소개

대조군: Top-3, Top-5, Top-7 (추출 + 생성 요약문)
비교군: KoBART-Summarization (생성 요약문)

- 12월 17일 실제 기사 12건의 요약문을 이용
- 위의 4가지 case 요약문이 동일한 경우를 제외하고 A/B Test 10문항 구성
- 카테고리(정치, 경제, 사회, 연예)와 요약문 case를 고르게 반영
- n-gram (n=1, 2, 3)을 이용하여 요약문 간 중첩도를 계산 → **가장 멀리 떨어진 두 요약문을 A, B로 선정**
- 이후 n-gram 중첩도와 사람이 판단했을 때 유사한 수준을 고려 → **나머지 2개의 문장도 A, B 그룹에 속하도록 함**

담당: 김상현

Coding Scheme & 검정 방법

Coding Scheme

- 요약문들이 비슷하기 때문에 최소한의 문항 수를 통해 최대한의 정보를 얻도록 설계
- 요약문1과 요약문2가 동일한 집단에 속한다면 각각 0.5, 0.5점을 부여 (우열이 없으므로)
- 요약문1과 요약문2가 다른 집단에 속한다면 응답에 따라 1, 0점 혹은 0, 1점을 부여
- 따라서, 랜덤하게 응답했다면 각 요약문의 기댓값이 동일함

검정 방법: Friedman Test, Quade Test

- Dependent case에 사용되는 nonparametric statistics 방법론
- Friedman Test: 개인이 block으로 사용되어 순위를 매기므로 개인별 차이가 0이 됨
- Quade Test: 개인 내 평가 차이에 대해서도 순위를 매겨, 차이가 크고 적음을 반영하여 더 검정력이 큼
- 정규분포 가정이 필요 없으므로 위의 같은 데이터 형태에 적합

Coding Scheme 도식화

응답자가 응답 A를 선택했을 때,

- “응답 A”에 속한 요약문
- “응답 B”에 속한 요약문

고려사항 & 한계점

문항 작성 시 고려사항

- 설문조사가 온라인 익명으로 진행되고 긴 요약문을 읽어야하는 특성을 고려
- 응답률을 높이기 위해 문항 수를 최대한 적지만 많은 정보를 추출할 수 있도록 설계함

한계점 및 논의

- 기사 원문을 제공하지 않아 의미적/맥락적 요소를 살피기 어려움
- 많은 응답이 문법적 요소 등은 고려되지 않음
- “요약”문보다는 “생성된” 문장 자체의 품질을 평가하는 문항

담당: 김상현

Friedman Test & Quade Test

4가지 요약문에 대한 Friedman test 결과

- Friedman Chi-Squared = 21.452
- Degree of Freedom = 3
- p-value < 0.005

4가지 요약문에 대한 Quade test 결과

- Quade F = 7.9312
- df1 = 3, df2 = 342
- p-value < 0.005

따라서, 4가지 요약문에 대한 품질 차이가 존재함

	EasyBART			BART
	Top-3	Top-5	**Top-7**	baseline
접수합*	1755.5	1629.5	**1841.5**	1673.5
순위합*	296.0	255.5	**333.0**	265.5

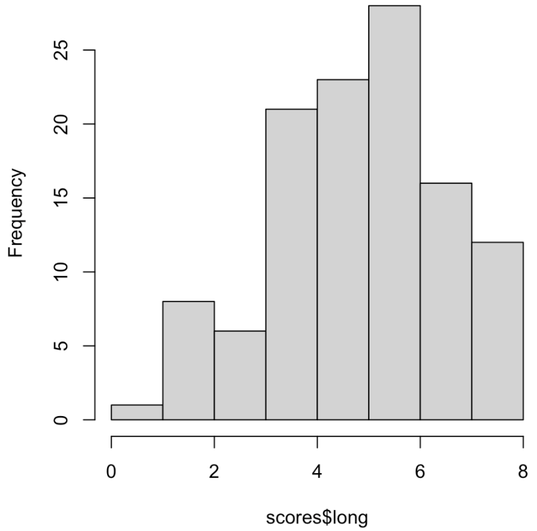
- $\chi^2 = 21.452\ (df = 3)$
- $p < 0.005$

담당: 김상현

Binomial Test

Top-3 vs. Baseline

- 8개 응답은 Top-3와 Baseline을 비교할 수 있는 문항
- 920개(115 * 8) 응답에 대해 총 610개 응답이 Top-3를 선택
  - $\hat{p}$ = 0.663
- Binomial Test 결과 p-value < 0.005으로 유의한 차이를 보임 **-> Top-3 선호**



긴 요약문 vs. 짧은 요약문

- A/B로 제시된 요약문 중 긴 것을 택한 경우를 평가
- 전체 1150개 중 526 응답
  - $\hat{p}$ = 0.457
- Binomial test 결과 p-value < 0.005 로 유의한 차이를 보임 **-> 짧은 요약문 선호**