AI를 이용한 뉴스 요약 AI비스 "New Ssum"

1조

I.O.T 김도균 노지윤 박강인 박선홍 정혜리



Saltlux





발표 순서

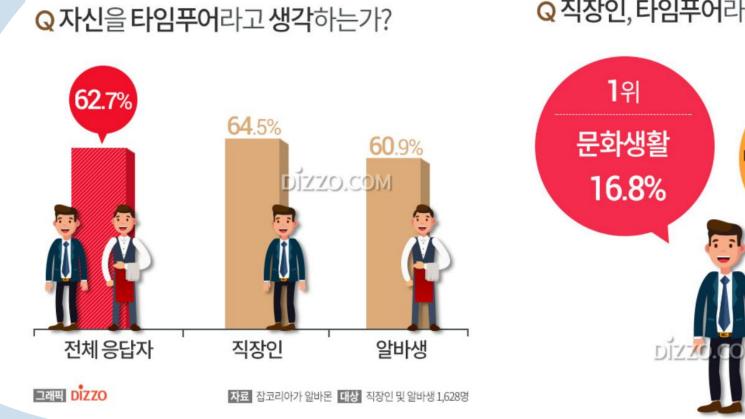
- 1 프로젝트 개요
- 2 모델링

- 3 서비스 구현
- 4 결론

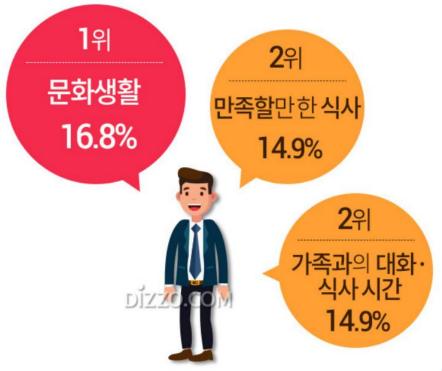


01

프로젝트 개요



Q 직장인, 타임푸어라도 포기할 수 없는 것?



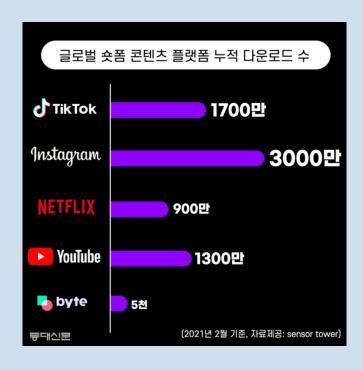
숏폼(Short-form) 및 서머리 콘텐츠(Summary contents) 유행



'타임푸어(Time-Poor)'란, 자신의 일에 쫓겨 **개인 시간이 없는** 사람들

사람들은 PMI(Please More Information)를 원하기 시작

핵심이 되는 부분만을 소비하고 필요하면 더 찾아보겠다는 의미



숏품(Short-form)

길이가 짧은 콘텐츠가 특징인 미디어



서머리 콘텐츠(Summary Contents)

도서, 영화, TV프로그램 등 콘텐츠를 요약하여 제공

긴 글과 영상에 쉽게 피로감을 느끼며 짧은 시간 내에 핵심적인 정보만을 얻고자 함.

시간이 부족한 현대인에게 숏폼과 서머리 콘텐츠는 문화생활의 대체제.







연합뉴스

- AI 뉴스 요약 서비스(2021)

→ 서머리 컨텐츠 중 뉴스 요약 서비스에 대한 수요 있음

"NewSsum" 서비스



"NewSsum"

시간은 없지만, 주요 뉴스를 알고 싶은 타임푸어족을 위해 매일 뉴스를 수집하고 **요약하여 읽어주는 서비스**

"NewSsum" 서비스 – 뉴스 요약

Summary Service : ① 주제별 뉴스 요약 서비스

언론사별 정치

경제

사회 생활/문화 IT/과학 세계 랭킹 신문보기 오피니언 TV 팩트체크

데일리안 🛨 구독

경총 "임금피크제 무력화 결사저지...정년유지형도 인정돼 야"

입력 2022.06.06. 오후 12:31 기사원문

[데일리안 = 박영국 기자] 한국경영자총협회가 최근 대법원의 '임금피크제는 무효' 판결에 따른 노동계의 임금피크제 무력화 시도에 적극 대응하기로 했다. 정년연장형 임금피크제는 물론, 정년유지형 임금피크제의 정당성도 인정돼야 한다는 주장도 내세웠다.

경총은 오는 7일 '임금피크제 대법원 판결 관련 대응방향'을 회원사에 배포할 예정이라고 6일 밝혔다.

이번 대응방향 배포는 지난달 26일 대법원의 임금피크제 판결과 관련해 실시된 '30대 기업 긴급 임금피크제 실태조사', '임금피크제 관련 30대 기업 부서장 회의'에 이어 산업현장의 혼란을 최소 화하기 위한 조치다.

경총은 "정년연장형 임금피크제는 원칙적으로 고령자고용법상 연령차별에 해당하지 않고, 정년 유지형 임금피크제라도 기존 규정상의 정년을 보장하기 위해 도입된 임금피크제라면 이번 판결을 그대로 적용해서는 안될 것"이라고 강조했다.

NEWSSUM



경총 "임금피크제 무력화 결사저지...정년유지형도 인정돼 야"

2022-06-06 12:31:00

언론사 : 데일리안

요약 기사: 한국경영자총협회가 정년연장형 임금피크제는 무효 판결에 따른 노동계의 임금피크제 무력화 시도에 적극 대응하기로 했으며, 정년연장형 임금피크제는 물론, 정년유지형 임금피크제의 정당성도 인정돼야 한다는 주장도 내세웠다.

기사 보기

"NewSsum" 서비스 – 뉴스 요약

Summary Service : ② 키워드별 뉴스 요약 서비스

▶지금 새싹 수강생 35명이 뉴스썸을 읽고 있어요

세상 돌아가는 소식, 알고 싶지만 쉴 새 없이 바쁜 게 우리 탓은 아니잖아요! 월/화/수/목/금/토/일 아침마다 세상 돌아가는 소식을 핵심만 읽어보세요.

URL을 입력해주세요

개ㅁ

요약된 뉴스 검색하기

'동학<mark>개미'</mark> 떠난다...예탁금 1년새 20조원 이상 줄어

2022.06.06. AM 6:56

언론사: 한국경제TV

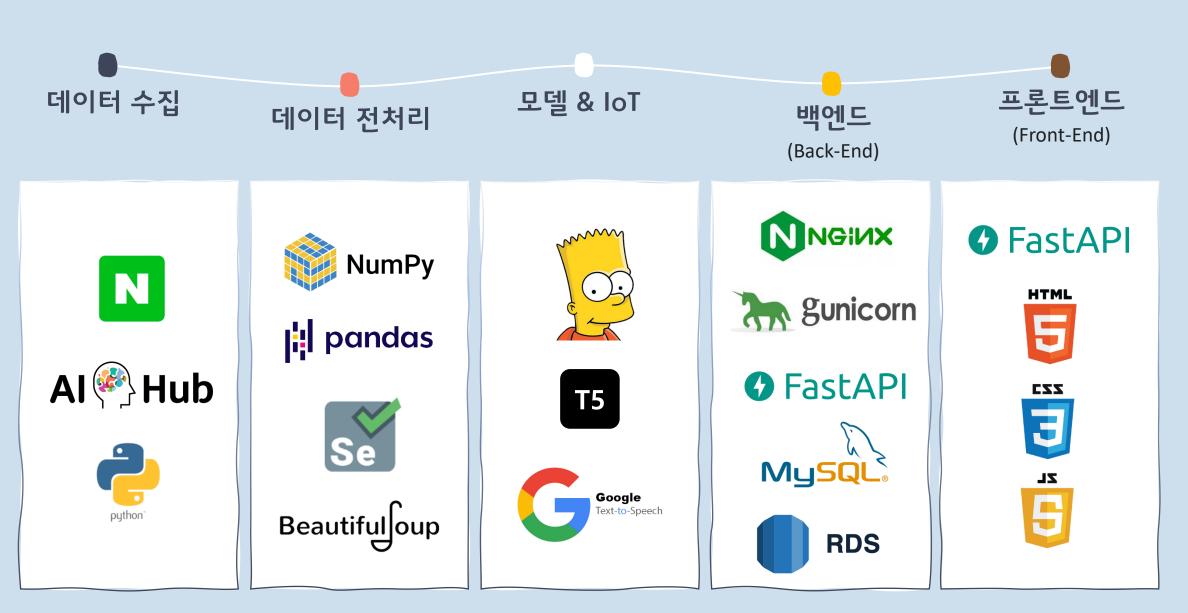
요약 기사: 지난 5월 31일 기준 투자자예탁금은 57조5천671억원으로 집계됐는데, 이는 코스피가 3,300까지 뛰어오른 작년 6월 고점을 기점으로 지금까지 1년간 지루한 하락장이 이어지면서 개미들의 주식투자 열기도 한풀 꺾였다.

기사 보기

"NewSsum" 서비스 – AI 스피커



서비스 구현 과정 – Using Stack



모델링 02_1

데이터 수집 & 전처리

데이터 수집 및 전처리 – 학습 데이터 수집



데이터 종류	데이터 형태	목표 수량
신문기사	뉴스 텍스트	원문데이터 30만 건 / 요약문 60만 건

Raw Text	Actual Text
전남드래곤즈(사장 신승재)는 지난 4일 구봉산 해맞이 행사를 통해 새해 각오를 다졌다. 신승재 사장은 "유상철 감독을 비롯한 코칭스텝, 선수단 구성이 마무리 된 만큼 구성원 모두가 하나되어 K리그 클래식 5위 이내 진입, FA컵 우승 등 ACL 진출권 획득을 목표로 최선을 다하자"고 선수들에게 신년 인사말을 전했다(중략)	전남드래곤즈 임직원과 선수단이 4일 구봉산 정상에 올라 일출을 보며 2018년 구단 목표 달성을 위한 새 해 각오를 다졌다.

신문기사 데이터 300,000건 중 30,000건의 데이터 활용

데이터 수집 및 전처리 – 서비스 데이터 선정 기준



정치, 경제, 사회, 생활/문화, 연예, 스포츠 6개의 카테고리

데이터 수집 및 전처리 – 서비스 데이터 선정기준



민주당 비대위 총사퇴..."지방선거 결과에 책임"

더블어민주당 비상대책위원회 전원이 61 지방선거 패배에 책임을 지고 전원 사퇴하기로 했습니다. 오늘(2일) 윤호중박지현 공동 비상대책위원장 등을 비롯 ...

'지선 참패' 민주 비대위 총사퇴..."의총 통해 새 지도부 논의"(상보) 뉴스네 ♀ 50+ [속보] 민주당 새 지도부, 의총・당무위・중앙위 거쳐 구성 조선비즈

[소보1 민주단 HI대의 지방성거 패배 책인지고 총사퇴 데일리안 | ◎ 50+

51 개의 관련뉴스 더보기 >



김동연 밤샘 추격끝 대역전국...경기지사 당선

61 지방선거 최대 격전지인 경기도지사 선거는 개표 과정 내내 쫓고 쫓기는 추격전

을 벌이다가 더불어민주당 김동연 후보의 막판

김동연, 역전극 배경... 화성·부천의 엇갈린 개표 시간 세계일보 경기지사 선거 '0.14%p 차이' 대역전극 원인은?[61 지방선가] 밤새 맹추격 나선 김동연, 새벽 5시부터 김은혜에 '막판 대역간 관련 뉴스 갯수를 기준으로 3개의 뉴스 토픽 선택

15] 개의 관련뉴스 더보기 >



권성동, **野**에 법사위원장 요구..."한 치하라 민심에 응답해야"

권성동 국민의힘 원내대표는 2일 61지 방선거 개표 결과와 관련해 "더불어민주당은 협치하라는 민심에 정말 응답해야한다"고 촉구했다. 권 원내대표는 이날 ...

뉴시스 | 🤛 10+

권성동 "野, 1년 전 약속대로 법사위원장 국힘에 돌려줘야" 한국경제 권성동 "민주당, 법사위원장 돌려줘야...민심에 답하라" YTN

장 독권의()반(주하) 여하뉴스 [@)

15] 개의 관련뉴스 더보기 >



데이터 수집 및 전처리 – 서비스 데이터 수집

<u> </u>	작성 시간	언론사	분야	기사 제목	기사 본문	기사 요약문	71/1\ URL
1	2022-06-02 05:01:00	파이낸셜뉴스	문화	"영화관 지키자" 박찬욱 감독 비	28개월 만에 최다 기록사회적	박찬욱 감독이 영화 헤어질 결심 으.	
2	2022-06-02 02:01:00	이데일리	정치	"한 명 살고 다 죽었다"지선 침	이석현 쇄신은 책임 큰 사람이	인천 계양을 국회의원 보궐선거에서	
3	2022-06-02 13:06:00	시사저널	사회	"헬기타고 온다더니" 재난 앞(밀양 울진서 연이어 산불 발생	지난 3월4일 대선 당시 산불 피해 현	https://n.news.naver.com/mnews/
4	2022-06-02 11:42:00	스타뉴스	스포츠	'SON 결국 동료들이 외면했다	손흥민. /AFPBBNews 뉴스1 같	잉글리시 프리미어리그에서 23골을	☑ https://sports.news.naver.com//ne
5	2022-06-02 09:19:00	부산일보	정치	'대역전' 김동연 "진심으로 감사	더불어민주당 김동연 경기.	! 더불어민주당 김동연 경기도지사 후	https://n.news.naver.com/mnews/
6	2022-06-02 08:35:00	스포츠조선	연예	'동생 하늘나라 보낸' 자두, 5살	스포츠조선닷컴 정안지 기자	가수 자두는 2일 자신의 인스타그램	☑ https://entertain.naver.com/rankir
7	2022-06-01 16:11:00	머니투데이	경제	'롯데타워' 인질된 롯데百 광복점	6월 1일 임시 휴무에 들어?	'롯데백화점 부산 광복점이 임시 휴두	https://n.news.naver.com/mnews/
8	2022-06-02 11:30:00	매일경제	경제	10일부터 열흘간 용산공원 시범	장군숙소 대통령실 남측구역	용산공원원은 용산공원을 찾은 국민	
9	2022-06-02 07:12:00	매일경제	경제	1인 다수 사업체 경영 25만개사	지난달 30일 서울 중구 신병	소상공인 1명이 여러 사업체를 경영	
10	2022-06-01 09:30:00	아시아경제	경제	5월 무역적자 17.1억달러에너	지난달 수출 615.2억달러 역대	글로벌 인플레이션(물가 상승) 우려	
11	2022-06-01 15:56:00	문화일보	경제	5월 수출 역대 두번째 규모무역	수입 32.0 증가한 632억 달러	고물가와 공급망 불안 등 대외 불확·	
12	2022-06-02 10:20:00	스포츠조선	스포츠	[RYU리뷰]58구만에 교체, 도대치	토론토 블루제이스 류현진이 2	토론토 블루제이스 류현진이 2일(이	☑ https://sports.news.naver.com//ne
13	2022-06-02 12:31:00	SBS	문화	[날씨] 낮 기온 크게 올라 '초여!	초여름 더위가 이어지고 있 _는	초여름 더위가 이어지고 있는 가운데	https://n.news.naver.com/mnews/

- 1시간 단위 수집 → 실시간 & 빠른 서비스 제공
- 동영상 뉴스 제외 → 앵커와 리포터의 대화 내용 제외

02_2

활용 모델

활용 모델 – 문서요약 기법

[톱스타뉴스 김명수]뉴시스에 따르면 1일 오후 9시까지 최소 20만3220명이 코로 나19에 신규 확진됐다. 또다시 동시간대 최다 기록으로, 사상 처음 20만명대에 진 입했다.

방역 당국과 서울시 등 각 지방자치단체에 따르면 이날 0시부터 오후 9시까지 전국 신규 확진자는 총 20만3220명으로 집계됐다.

국내 신규 확진자 수가 20만명대를 넘어선 것은 이번이 처음이다.

동시간대 최다 기록은 지난 23일 오후 9시 기준 16만1389명이었는데, 이를 무려 4만1831명이나 웃돌았다. 전날 같은 시간 기록한 13만3481명보다도 6만9739명 많다.

확진자 폭증은 3시간 전인 오후 6시 집계에서도 예견됐다.

오후 6시까지 최소 17만8603명이 신규 확진돼 동시간대 최다 기록(24일 13만8419명)을 갈아치운 데 이어 이미 직전 0시 기준 역대 최다 기록도 넘어섰다. 역대 최다 기록은 지난 23일 0시 기준 17만1451명이었다.

17개 지자체별로 보면 서울 4만6938명, 경기 6만7322명, 인천 1만985명 등 수도권이 12만5245명으로 전체의 61.6%를 차지했다. 서울과 경기는 모두 동시간대 기준최다로, 처음으로 각각 4만명과 6만명을 넘어섰다.

비수도권에서는 7만7975명(38.3%)이 발생했다. 제주를 제외한 나머지 지역에서 모두 동시간대 최다를 새로 썼다.

문서 요약 - 추출

주어진 텍스트에서 중요한 문장만 그대로 추출하여 요약

▶ "방역 당국과 서울시 등 각 지방자치단체에 따르면 이날 0시부터 오후 9시까지 전국 신규 확진자는 총 20만3220명으로 집계됐다."

→ 요약문의 <mark>응집도</mark>나 **가독성**을 확보하기 어려움

활용 모델 – 문서요약 기법

[톱스타뉴스 김명수]뉴시스에 따르면 1일 오후 9시까지 최소 20만3220명이 <mark>코로나19에 신규 확진됐다.</mark> 또다시 동시간대 최다 기록으로, 사상 처음 20만명대에 진입했다.

방역 당국과 서울시 등 각 지방자치단체에 따르면 이날 0시부터 오후 9시까지 전국 신규 확진자는 총 20만3220명으로 집계됐다.

국내 신규 확진자 수가 20만명대를 넘어선 것은 이번이 처음이다.

동시간대 최다 기록은 지난 23일 오후 9시 기준 16만1389명이었는데, 이를 무려 4만1831명이나 웃돌았다. 전날 같은 시간 기록한 13만3481명보다도 6만9739명 많다.

확진자 폭증은 3시간 전인 오후 6시 집계에서도 예견됐다.

오후 6시까지 최소 17만8603명이 신규 확진돼 동시간대 최다 기록(24일 13만8419명)을 갈아치운 데 이어 이미 직전 0시 기준 <mark>역대 최다</mark> 기록도 넘어섰다. 역대 최다 기록은 지난 23일 0시 기준 17만1451명이었다.

17개 지자체별로 보면 서울 4만6938명, 경기 6만7322명, 인천 1만985명 등 수도권이 12만5245명으로 전체의 61.6%를 차지했다. 서울과 경기는 모두 동시간대 기준최다로, 처음으로 각각 4만명과 6만명을 넘어섰다.

비수도권에서는 7만7975명(38.3%)이 발생했다. 제주를 제외한 나머지 지역에서 모두 동시간대 최다를 <mark>새로 썼다</mark>.

문서 요약 - 생성

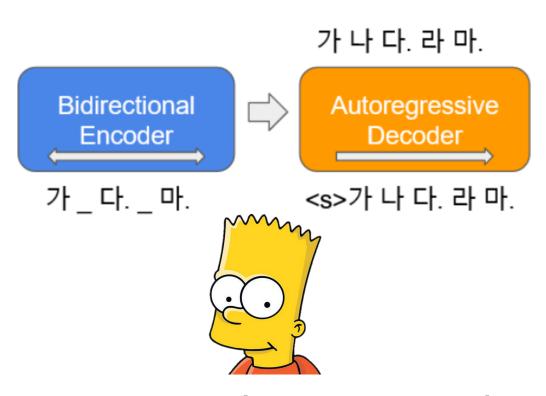
주어진 텍스트에서 중요한 의미를 가진 단어를 판별하고 문장을 새롭게 생성하여 요약



"1일 0시부터 9시까지 최소 20만3200명이 코로나19에 신규 확진되어 역대 최다 기록을 새로 썼다 "

→ 문서의 내용을 압축하여 핵심 내용을 반영

활용 모델 - KOBART & ET5



1. KoBART(한국어 BART모델)

= BERT + GPT

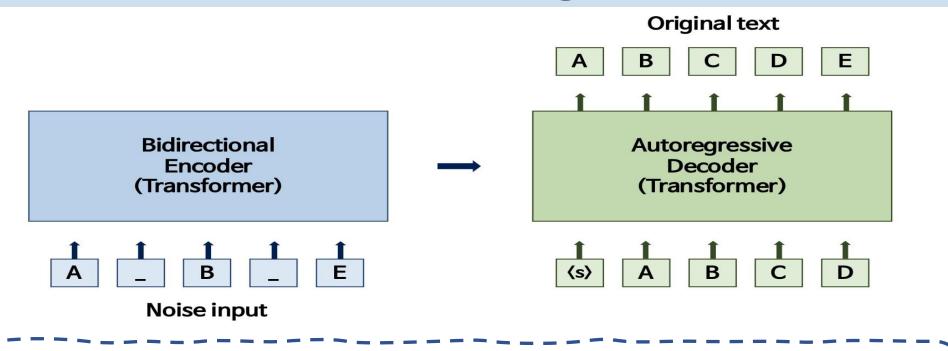


2. ET5(한국어 이해생성 언어모델)

$$= T5 + GPT$$

활용 모델 - BART

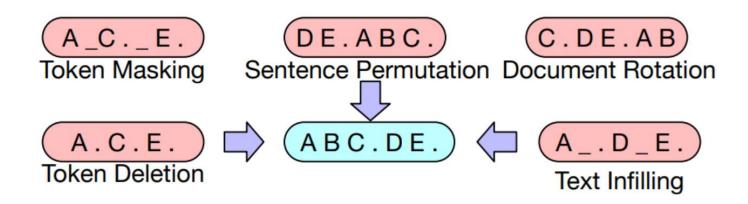
BART: Bidirectional Auto-Regressive Transformer



- 트랜스포머 모델을 새로운 Pre-training objective를 통해 학습한 모델
- Self Attention을 적용하여 양방향으로 인코딩 -> Causal Attention을 적용하여 디코딩
- 손상된 text를 BERT로 Encoding -> 복원된 text를 GPT의 Decorder가 입력값으로 받음

활용 모델 - BART

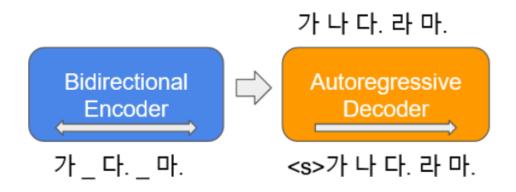
BART - Sentence Corrupt



- Token Masking : 랜덤한 토큰이 샘플링 되어 [MASK] 토큰으로 치환 → [MASK]토큰을 맞춰야 함
- **Text Infiling**: 포아송 분포를 따르는 길이의 span을 생성해서 이를 하나의 mask 토큰으로 masking →여러 토큰이 하나의 mask 토큰으로 바뀔 수 있고 길이가 0인 경우에는 mask 토큰만 추가 → **토큰의 개수를 맞춰야 함**
- Sentence Permutation: 마침표를 기준으로 문장으로 분류 후 섞임 → 원래의 순서를 맞춰야 함
- Document Rotation : 랜덤한 토큰이 문서의 시작점이 됨 → 원래 문서의 시작점을 찾아야 함
- Token Deletion : 랜덤한 토큰을 제거 → 없어진 위치와 토큰을 찾아야야 함

활용 모델 - KoBART

KoBART



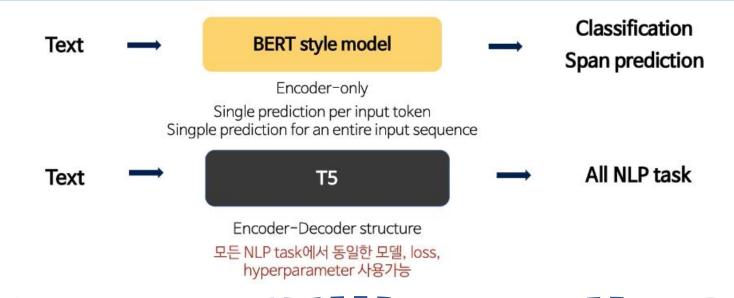
"Text Infilling" 노이즈 함수를 사용하여

40GB 이상의 한국어 텍스트에 대해 학습한

한국어 Encoder-Decoder 언어 모델

활용 모델 - T5

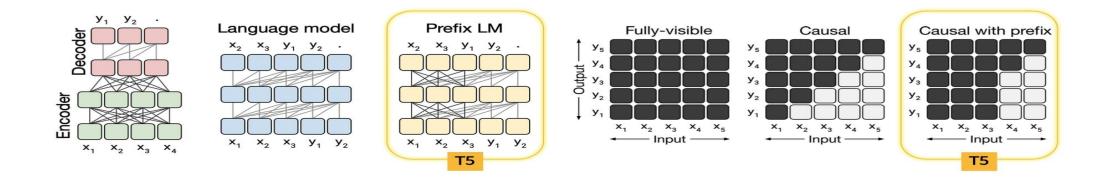
T5: Text-To-Text Transfer Trasnsformer



- Multi Task Learning: Task별 loss를 바꾸어 주지 않고 Transfer Learning이 가능함
- Prefix Language Model: Source data(양방향 어텐션), Generation Part(한방향 어텐션)
 - → 기존 NLP방법론 중 가장 좋은 점들을 모두 조합한 형태로 모델을 구현

활용 모델 - T5

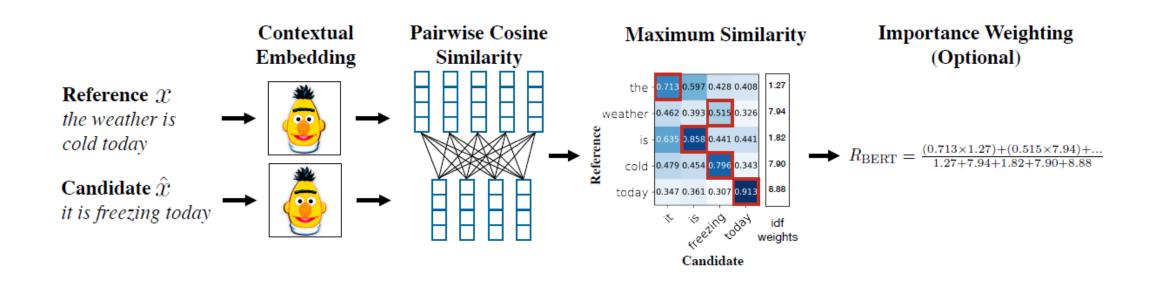
T5: Prefix Language Model



- 1. Encoder Decoder: Encoder(입력 시퀀스를 공급) / Decoder(새로운 출력 시퀀스를 생성)
- 2. Language Model: Autoregressive한 방법으로 문장을 생성
- 3. Prefix Language Model: Source data(양방향 어텐션) ,Generation Part(한방향 어텐션)

모델 선정 및 평가 – BERTScore

평가지표: BERTScore



Reference 문장과 생성 문장을 BERT를 사용해 Contextual embedding을 구한 후,

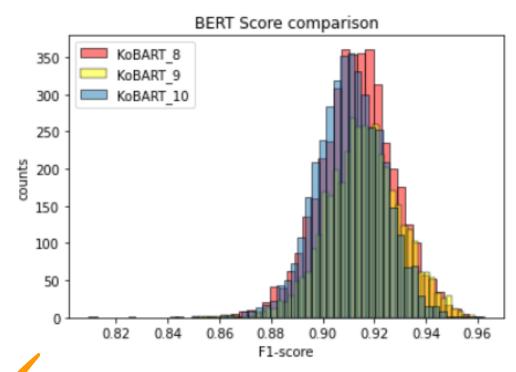
코사인 유사도를 통해 두 문맥 벡터 간의 정밀도와 **재현율**을 구함.

기존 문장 요약 평가 방법으로 많이 쓰이는 ROUGE는 Exact match로 평가해 문맥을 평가하지 못한다.

반면에 BERTScore는 문장 간의 문맥을 평가할 수 있다는 장점.

모델 선정 및 평가 – BERTScore

Bert Score 성능 비교



	Epoch 8	Epoch 9	Epoch 10
KoBART	0.913302	0.915929	0.910503
ET5	0.902948	0.901373	0.902476

모델 선정 및 평가 – 정성적 평가

ET5



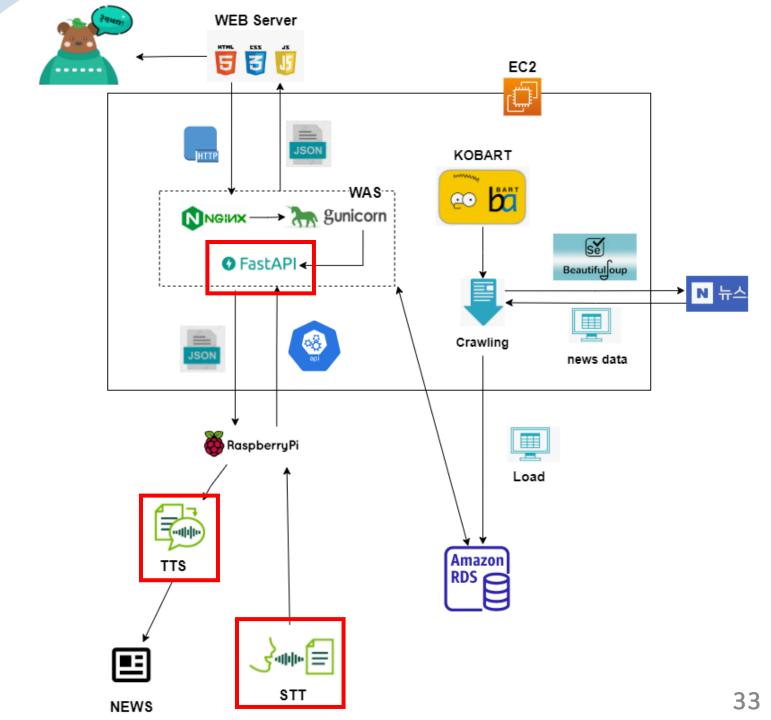
Generated Text	Actual Text	Generated Text
나 지방자치단체가 합동으로 실시한 국가안전 대진단 결과 전국에서 1만5319곳이 보수 · 보강 이 필요한 것으로 나타났는데, 행정처분을 받은 시설은 과태료 757곳, 구직자는 33곳, 금융위 는 31곳이고 보수 · 보강 조치 시 575곳은 75.9%로 가장 많았다.	중앙정부와 지방자치단체가 합동으로 실시한 국가안전대진단 결과 전국 생활밀접시설, 사회 기반시설 등 16만1588곳 중 10%에 해당하는 1 만5319곳이 보수 첫보강이 필요한 것으로 나타 났으며 현장 시정 조치가 9218곳, 과태료 등 행 정처분이 2263곳, 정밀안전진단이 필요한 곳도 190곳으로 나타났다.	중앙정부와 지방자치단체가 합동으로 실시한 국가안전대진단 결과 전국에서 1만5319곳이 보 수ㆍ보강이 필요한 것으로 나타났다.
가 기대되는 차량용 프로세서를 개발하여 출시할 예정인 '엑시노스 오토 8890'을 시작으로 차세대 모델에도 '엑시노스 오토' 제품을 공급하며 차량용 프로세서 기술 경쟁력을 강화해 나갈 계획이다.	삼성전자가 아우디 신형 A4 모델에 탑재되는 엑 시노스 오토 8890(메인프로세서)을 시작으로 차세대 모델에도 엑시노스 오토 제품을 공급하 며 차량용 프로세서 기술 경쟁력을 강화해 나갈 계획이다.	삼성전자가 아우디의 신형 차량 A4 모델에 탑재 되는 '엑시노스 오토(Exynos Auto) 8890'을 시 작으로 차량용 프로세서 시장 공략에 본격적으 로 나선다.

'주어'가 생략됨

'문법'에 맞는 문장 생성

03 서비스 구현

Service Architecture



백엔드



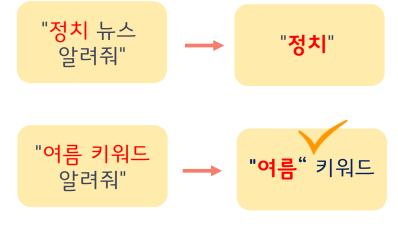
마이크에 입력된 소리를 파일로 저장 (5초로 제한)

Speech Recognition을 통해 소리를 텍스트로 변환

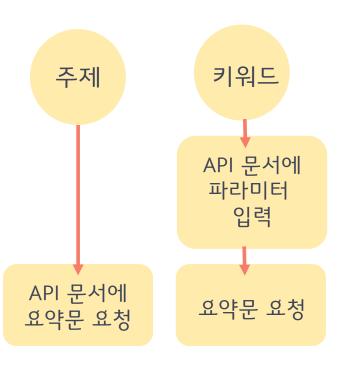


인식된 명령어를 통해 주제와 귀워드 구분

변환된 텍스트에서 주제 혹은 키워드 찾음



인식된 명령어를 APIM버에 요청





4

요약문을 사운드 파일로 저장

요청받은 요약문을

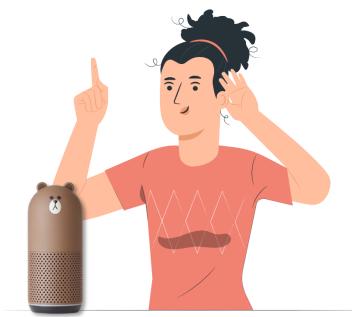
GoogleTTS를 이용하여 사운드파일로 저장



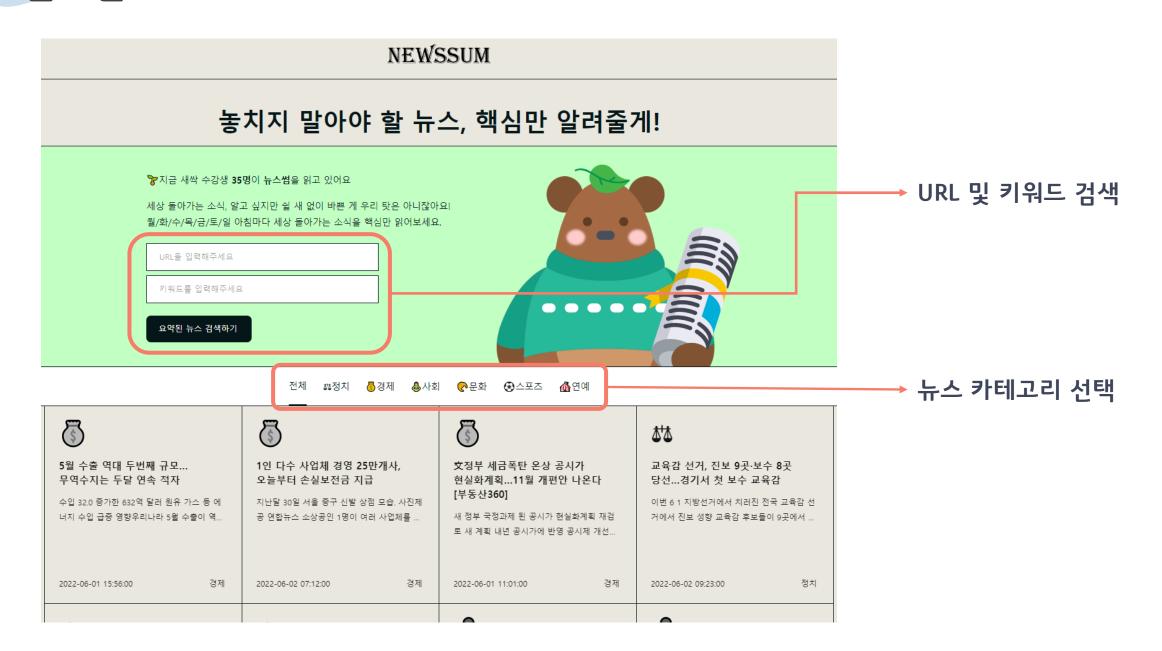
5

Pygame 이용하여 사운드로 호출

Pygame을 이용하여 사운드 출력



프론트엔드





서비스 시연 - A.I. 스피커



서비스 시연 - Web



보완 및 개선할 점



데이터 분석을 위한 사용자의 검색 키워드 데이터 베이스에 저장



카테고리 늘리기



키워드 검색 시 로딩 창 생성



키워드 검색 속도 개선



특정 날짜나 요일 뉴스 요약 구현



음성 생성 모델을 접목시킨 스피커 구현





