

한눈에 파악하는 기업 뉴스

NEWS.tar # 뉴스 토픽 분류 # 뉴스 요약

NLP - 05






바스버거 | 김진호, 신혜진, 이효정, 이상문, 정지훈

1. Introduction
2. Working Process
3. Model Optimization
4. Service & Product
5. Conclusion
6. Appendix

1. Introduction

1.1 팀 소개

 **BASBURGER** Team member

김진호	신혜진	이효정	이상문	정지훈
				
토픽 모델링	추출 요약 생성 요약	감정 분석 Frontend & Model 서버 구축	데이터 수집 데이터 전처리 DB 서버 구축	생성 요약 유사도 분류



NEWS.tar 는 뉴스 데이터를 주제 별로 분류하고
기사 내용을 요약하여 보여줌으로써
사용자들이 짧은 시간에 주요 뉴스 내용을 파악할 수 있도록 도와줍니다.

너무 많은 뉴스들



뉴스1 | 9시간 전 | 네이버뉴스

美 카즈닷컴 '올해 최고의 차'에 현대차·기아 4종..."그룹 기준 최다"

카니발, 현대차의 아이오닉 5, 제네시스 G90 등 4개 차종이 수상했다고 8일 밝혔다. 2023 최고의 차 어워즈는... 현대차 아이오닉5는 '2023 최고의 전기차'로 선정됐다. 넓고 쾌적한 실...

현대차·기아-제네시스, 美 카즈닷컴 '최고의 차' 4개 부... 더팩트 | 9시간 전 | 네이버뉴스
현대차·기아-제네시스, 美 카즈닷컴 '2023 최고의 차'... 이데일리 | 8시간 전 | 네이버뉴스
현대차·기아-제네시스, 美 카즈닷컴 '2023 최고의 차' 4개 부문 석권 매일일보 | 5시간 전

머니S PICK | 8시간 전 | 네이버뉴스

美서 또 상품가치 입증한 현대차·기아

카니발 ▲현대차 아이오닉5 ▲제네시스 G90 등 4개 차종이 수상했다고 8일 밝혔다. 2023 최고의 차... 현대차그룹은 총 6개 부문 중 4개 부문에 선정돼 미국에서 그룹 기준 최다 수상을 달...

현대차 데일리카 | 6시간 전

현대차·기아-제네시스, 美 카즈닷컴 '최고의 차' 4부문... 데일리인 | 9시간 전 | 네이버뉴스
현대차·기아·제네시스, 美 '2023 최고의 차 어워즈'... 노컷뉴스 | 8시간 전 | 네이버뉴스
현대차·기아-제네시스, 美 카즈닷컴 '2023 최고의 차'... 아이뉴스24 | 8시간 전 | 네이버뉴스

관련뉴스 9건 전체보기 >

연합뉴스 PICK | 13시간 전

현대차 아이오닉5, 美 카즈닷컴 선정 '최고의 전기차'

현대차 미국판매법인(HMA)에 따르면 카즈닷컴은 미국 뉴저지 뉴포트 노프-150-리브링 등 결선에 진출한 전기차 3대 가운데 아이오닉5를 1위로 최종 뽑았다. 제니 뉴먼 카즈닷컴 편...

현대차 아이오닉5, 美 카즈닷컴 선정 '최고의 차'... 한국경제 PICK | 6시간 전 | 네이버뉴스
현대차·기아·제네시스, 미국서 '2023 최고의 차 어워즈'... 동아일보 | 8시간 전 | 네이버뉴스
현대차·기아-제네시스, 美 카즈닷컴 '2023 최고의 차'... 스포츠동아 | 3시간 전 | 네이버뉴스
현대차·기아-제네시스, 美 카즈닷컴 '2023 최고의 차'... 스포츠경향 | 4시간 전 | 네이버뉴스

관련뉴스 14건 전체보기 >



뉴스 서비스에서의 AI



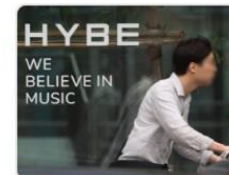
단일 뉴스에 대한 긍정/부정 분류
추출 요약물 통한 뉴스 3줄 요약

➡ 단순 모델 사용 수준

실시간 뉴스 마이종목 뉴스

중립 11분전 | 한국경제

게임·가상인간·NFT까지...
하이브 '광폭 행보' 이유는



애플페이, 다음달 초부터 사용 가능... "NFC 가맹점부터" (종합)

송고시간 | 2023-02-03 14:59



조성미 기자
기자페이지



오규진 기자
기자페이지

| 국내 간편결제 방식·스마트폰 점유율 판도에 변화 부를까 촉각

금융당국이 애플사의 비접촉식 간편결제 시스템 애플페이의 국내 서비스가 가능하다는 해석을 내린 가운데 다음 달 초부터 애플페이를 사용할 수 있을 것으로 3일 알려졌다.

연합뉴스 취재를 종합하면 애플페이 국내 서비스 개시일은 다음 달 초가 될 것으로 파악됐다.

애플페이 결제에 필요한 NFC(근거리 무선 통신) 단말기를 갖춘 곳부터 서비스가 시작될 것으로 보인다.

엔씨소프트 ▲1.29% 테마 #게임

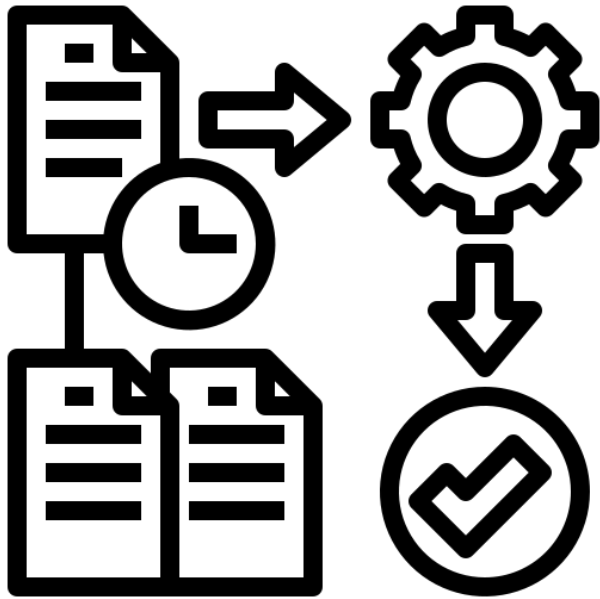
중립 15분전 | 비즈니스워치

네오위즈, 대만 게임 시장
'도전'...TGS 참가



네오위즈 ▼0.12% 테마 #게임

NEWS.tar의 새로운 기능

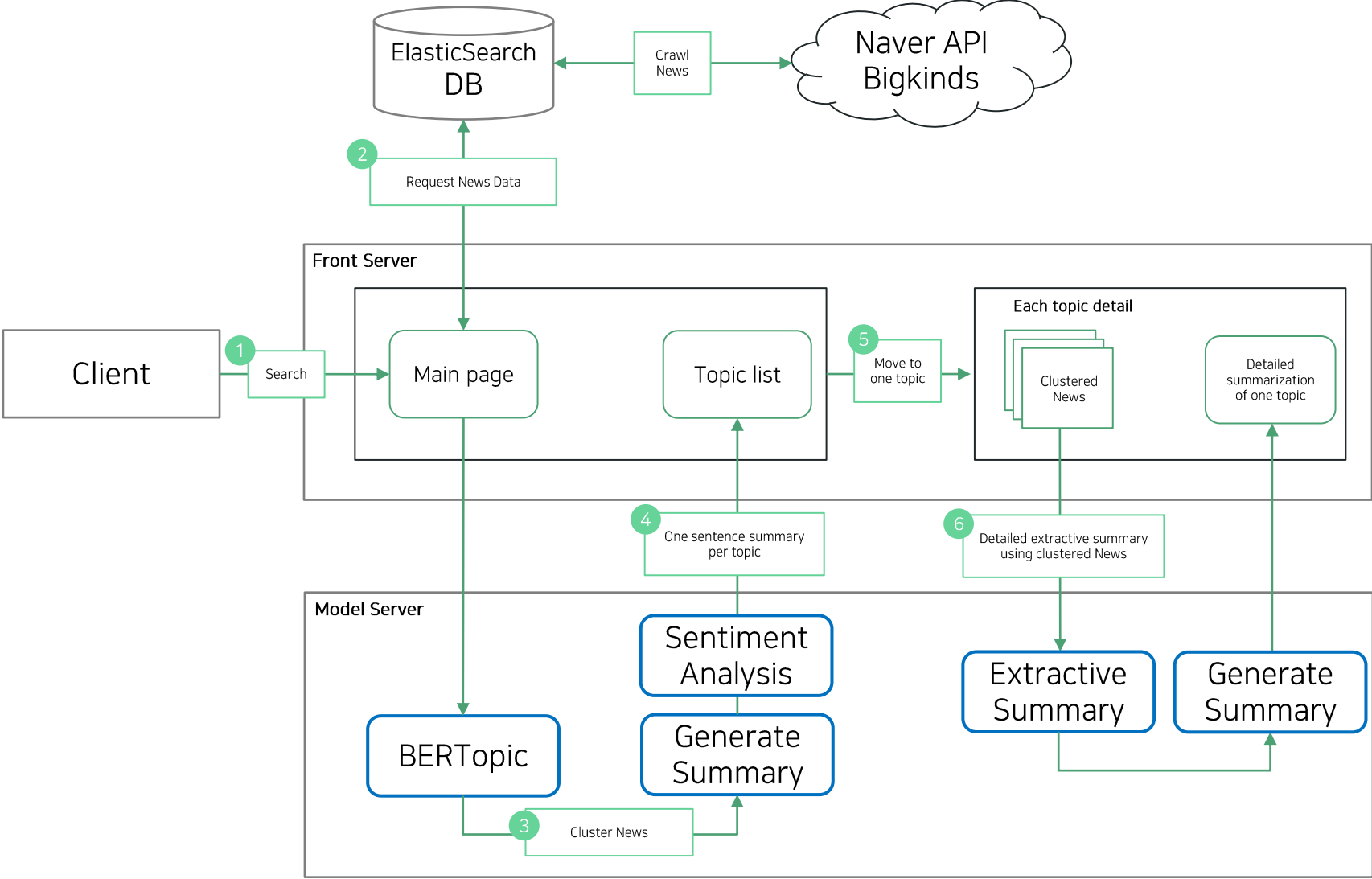


1. 비슷한 주제의 뉴스를 모아서 제공
2. 각 주제의 기사들을 하나의 문장으로 요약
3. 해당 주제에 대한 감정 분석 제공
4. 같은 주제로 묶인 기사들의 전반적인 요약 문단 제공

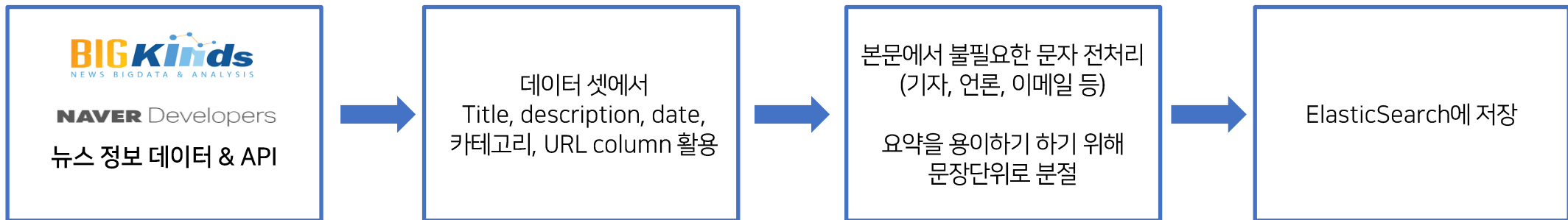
➡ 특정 기업에 대한 뉴스를 주제 별로 빠르게 파악

2. Working Process

2.1 Flow Chart



2.2 데이터 수집 및 전처리



- Multiprocessing을 이용하여 데이터 처리 작업 병렬화 (1000개 기사 처리 15 ~ 20초)
- Airflow를 이용하여 데이터 갱신 1일 단위 배치화
- 전체 작업시간 (10 ~ 15분 소요)
기사 평균 개수 2,000 ~ 12,000개

삼성전자, 역대급 혜택의 '삼성전자 세일 페스타' 개최

A 김현우 | © 입력 2022.12.26 16:23 | © 수정 2022.12.26 19:07 | ■ 댓글 0

삼성전자가 2023년 1월 1일부터 2월 12일까지 '삼성전자 세일 페스타'를 연다고 26일 밝혔다.

국민 모두가 새해를 더욱 희망차게 시작하길 응원하는 취지에서 지난 2021년 시작한 삼성 세일 페스타는 다양한 인기 모델을 풍성한 혜택과 함께 판매해 완판 행렬을 이어왔다.

3회째를 맞는 2023년 행사는 가전·모바일 등 대상 모델과 구매 혜택을 확대해 온·오프라인 매장에서 동시에 진행된다.

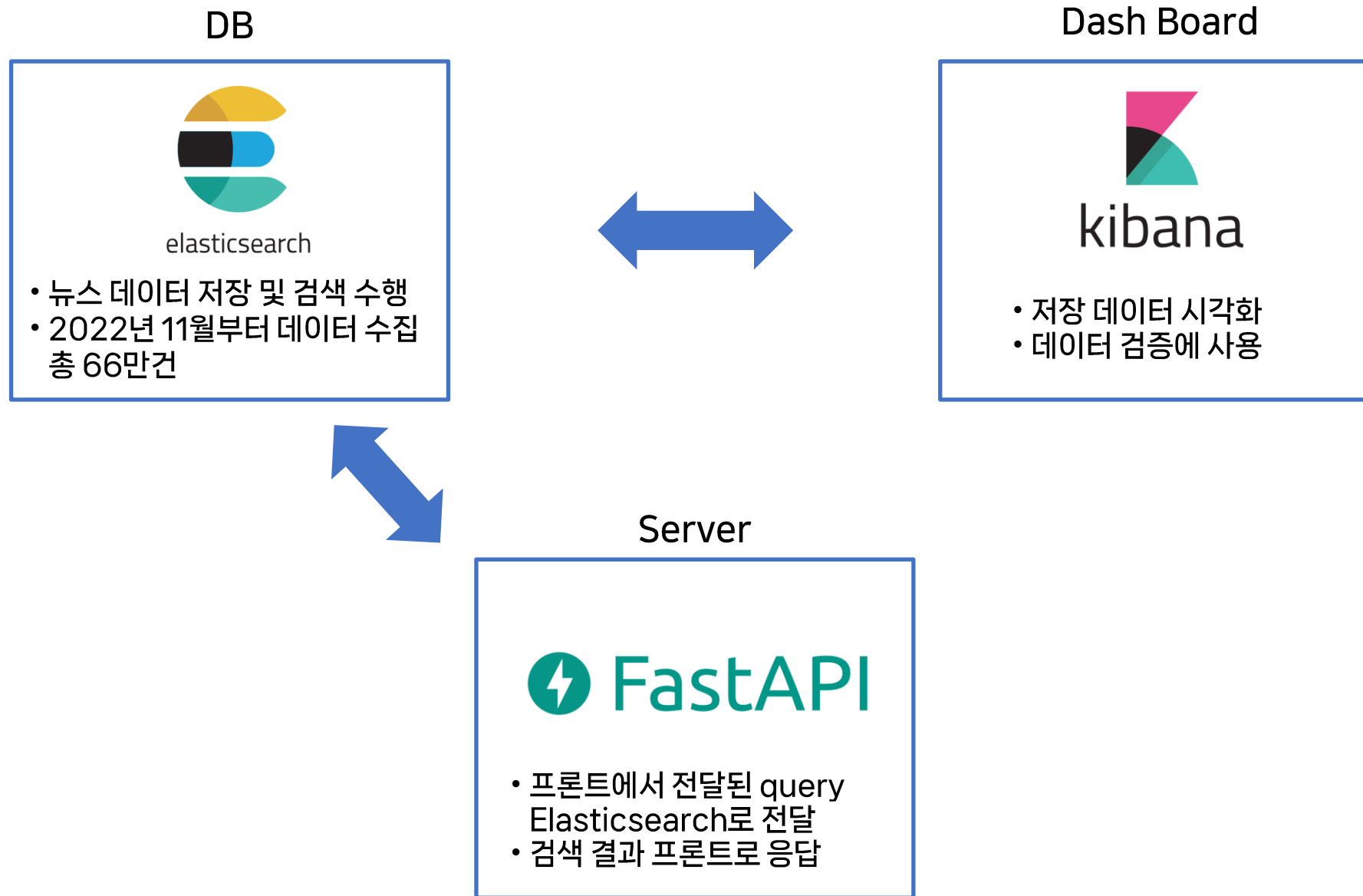
특히, 삼성전자는 그 동안의 행사에 대한 고객들의 성원에 보답하고자 90만 원대 특별가 한정 판매 모델을 늘렸다.

QLED TV(138cm, 55형), 비스포크 그랑데 AI 세탁기(24Kg)-건조기(20Kg), 양문형 냉장고 등을 90만 원대에 판매한다.

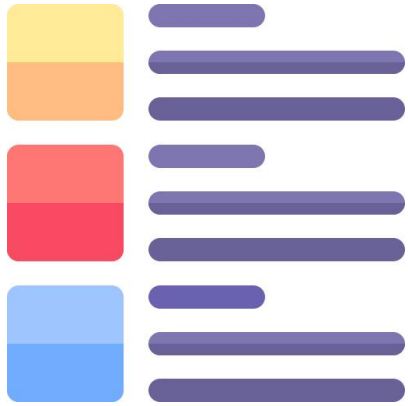
이 밖에도 ▶비스포크 냉장고부터 무풍에어컨, 에어드레서, 제트, 큐커, 식기세척기 등 다양한 비스포크 가전 ▶Neo QLED, The Serif 등 TV ▶스마트 모니터 ▶갤럭시 Z 폴드·Z 플립, 갤럭시 시 북 등 모바일 기기 ▶하만카돈, JBL 등 오디오 기기 등 총 150여개 모델을 할인가로 판매한다.

['삼성전자 모델이 삼성 디지털프라자 서초본점에서 '삼성전자 세일 페스타'를 소개하고 있다.', '삼성전자 삼성전자가 2023년 1월 1일부터 2월 12일까지 '삼성전자 세일 페스타'를 연다고 26일 밝혔다.', '국민 모두가 새해를 더욱 희망차게 시작하길 응원하는 취지에서 지난 2021년 시작한 삼성 세일 페스타는 다양한 인기 모델 행렬을 이어왔다.', '3회째를 맞는 2023년 행사는 가전·모바일 등 대상 모델과 구매 혜택을 확대해 온·오프라인 매장에서 동시에 진행된다.', '특히, 삼성전자는 그 동안의 행사에 대한 고객들의 성원에 보답하고자 90만 원대 특별가 한정 판매 모델을 늘렸다.', 'QLED TV, 비스포크 그랑데 AI 세탁기-건조기, 양문형 냉장고 등을 90만 원대에 판매한다.', '이 밖에도 비스포크 냉장고부터 무풍에어컨, 에어드레서, 제트, 큐커, 식기세척기 등 다양한 비스포크 가전 Neo QLED, The Serif 등 TV ▶스마트 모니터 ▶갤럭시 Z 폴드·Z 플립, 갤럭시 북 등 모바일 기기 ▶하만카돈, JBL 등 오디오 기기 등 총 150여개 모델을 할인가로 판매한다.', '또한, 행사 기간에 추첨을 통해 구매 금액의 최대 3배를 삼성전자 멤버십 포인트로 제공하며, 제품 구매 후 이벤트 응모 멤버십 포인트와 기프트콘 등 풍성한 경품을 제공한다.', '향대한 삼성전자 한국총괄 부사장은 "삼성전자의 다양한 제품을 큰 혜택으로 만나볼 수 있는 '삼성전자 세일 페스타'를 통해 하길 바란다"고 말했다.']

2.2 데이터 저장 및 검색



사용한 Task 종류



1. 토픽 모델링



2. 한 줄 요약

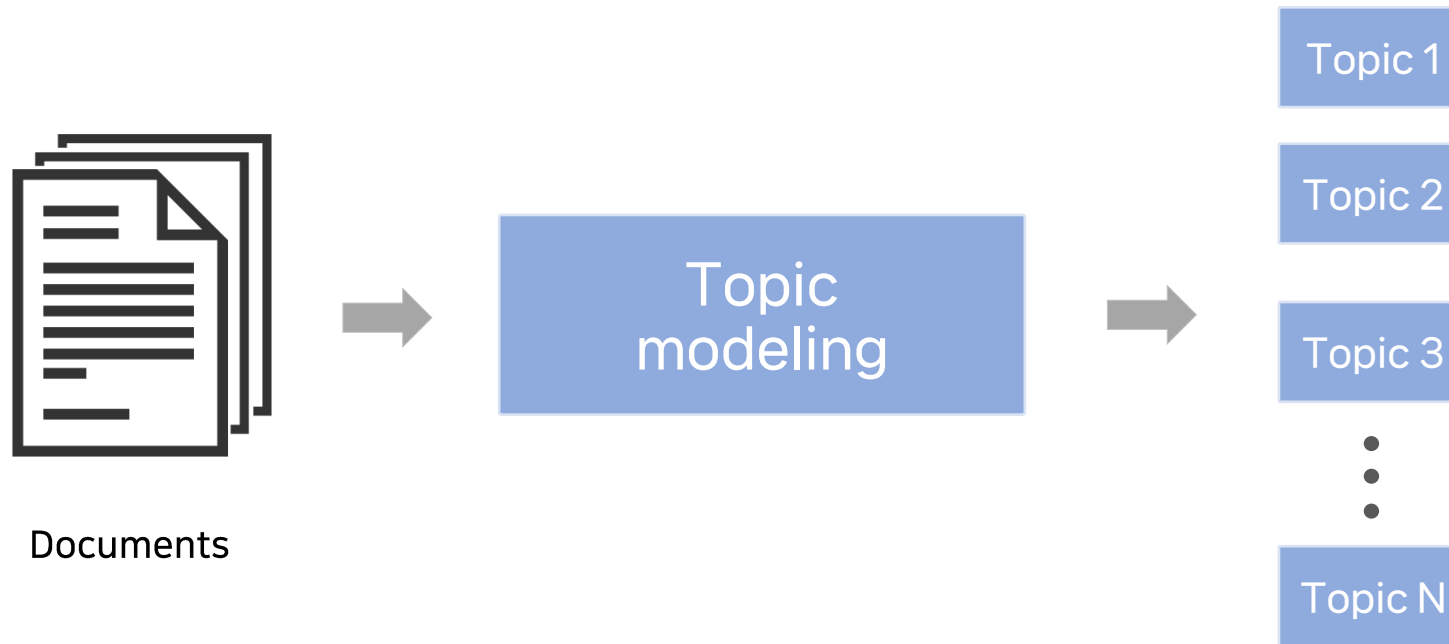


3. 감정 분석



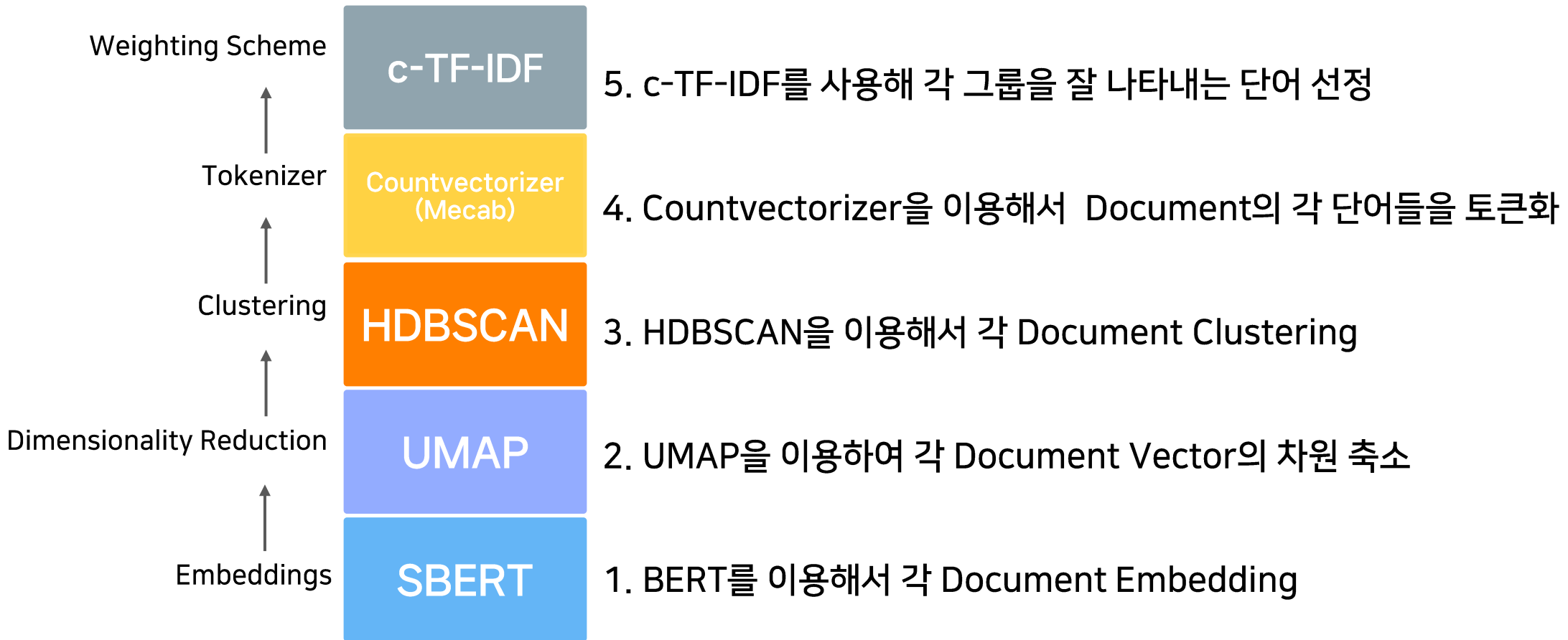
4. 문단 요약

토픽 모델링(Topic Modeling)

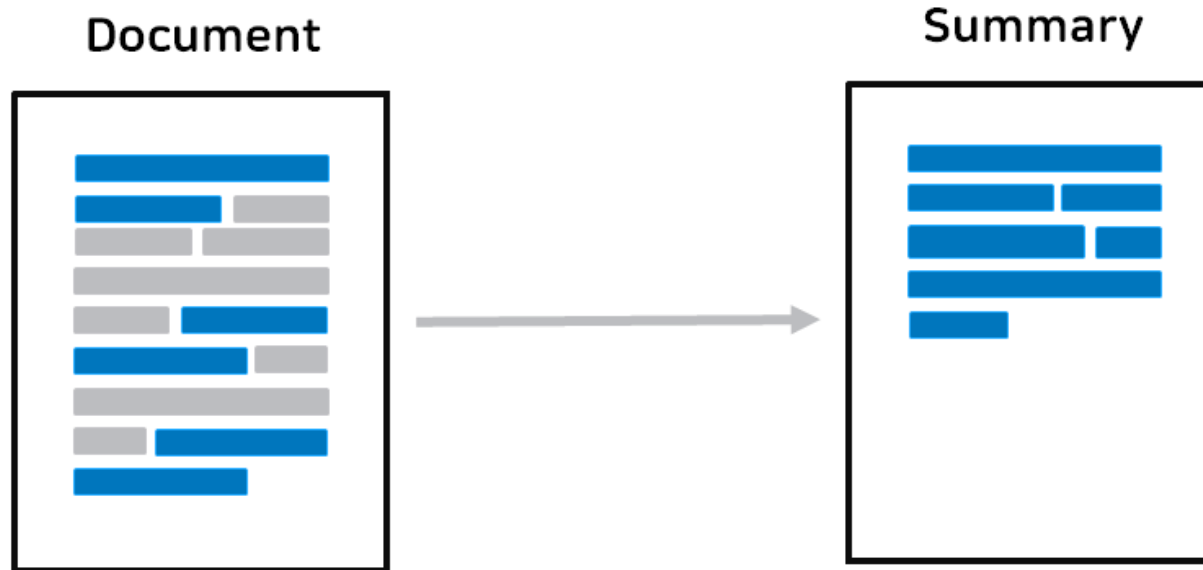


각 단어들이 통계적으로 특정 토픽에 포함될 **확률을 추출**하여
문서의 **주제를 파악**하는 기법

버토픽(BERTopic)

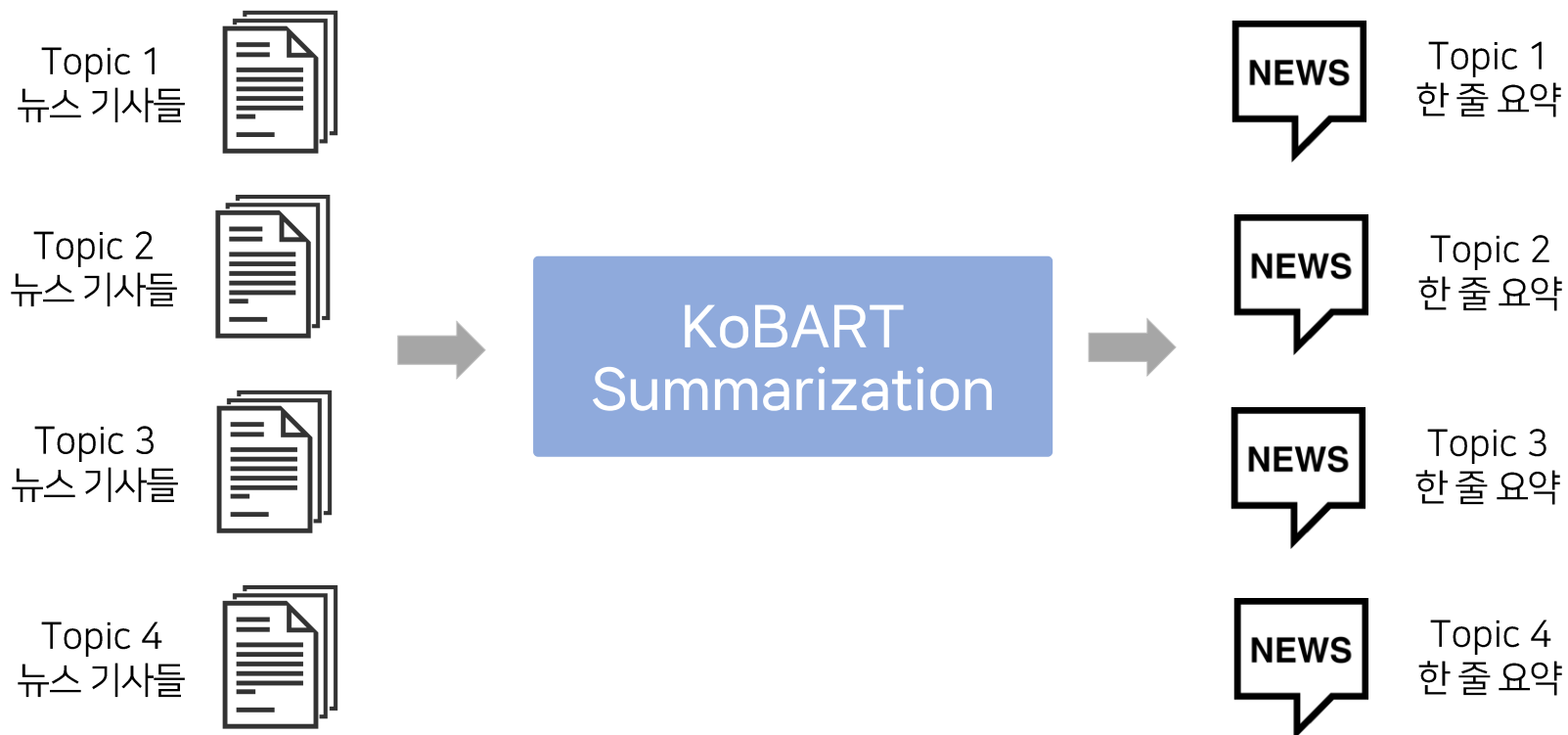


토픽 한 줄 요약(Topic Summary)



1. 기사의 **제목과 본문 앞 2문장**을 concat
2. 같은 주제로 분류된 기사들의 concat 문장을 합쳐서 모델에 입력
3. 하나의 주제에 대해서 하나의 한 줄 요약문 생성

토픽 한 줄 요약(Topic Summary)



감성 분석(Sentiment Analysis)



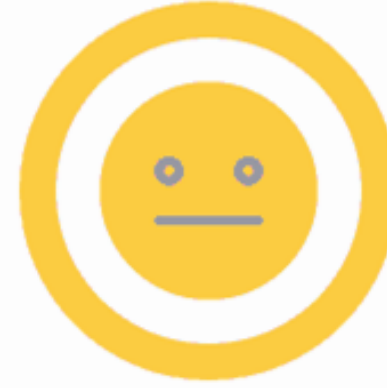
Positive

매출이 작년에
비해 증가함



Negative

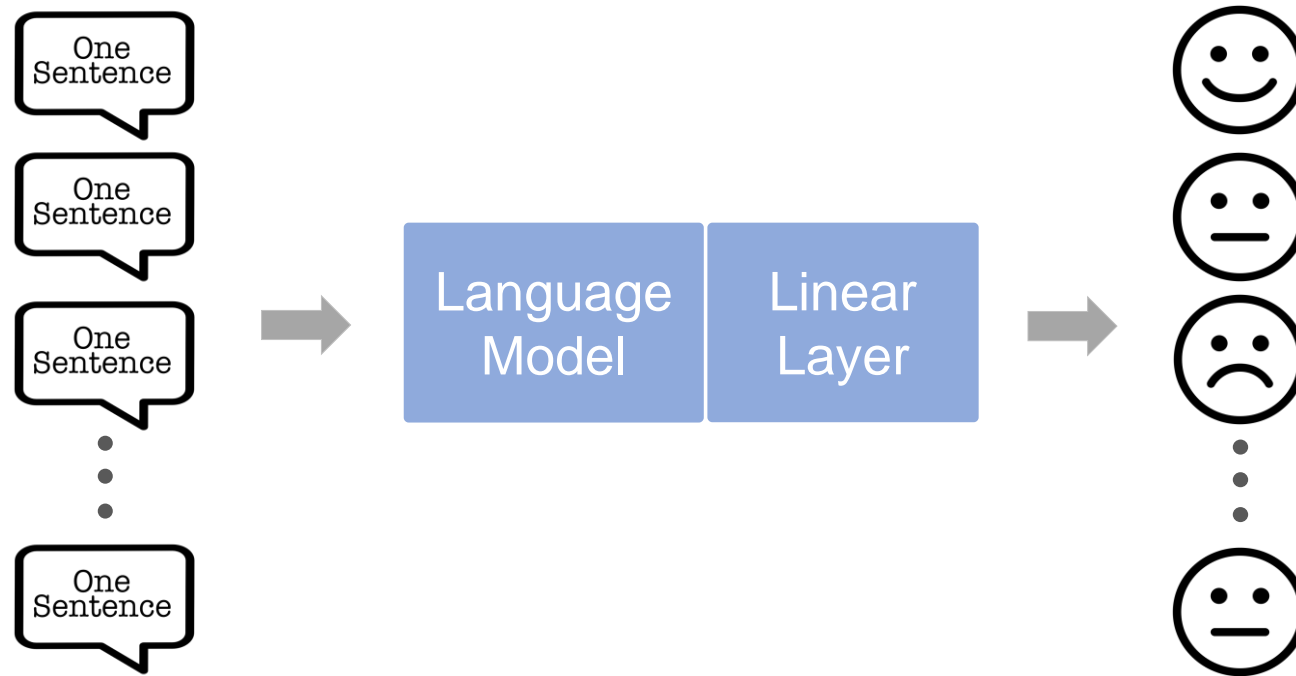
매출이 작년에
비해 하락함



Neutral

매출이 발생함
(비교군 없음)

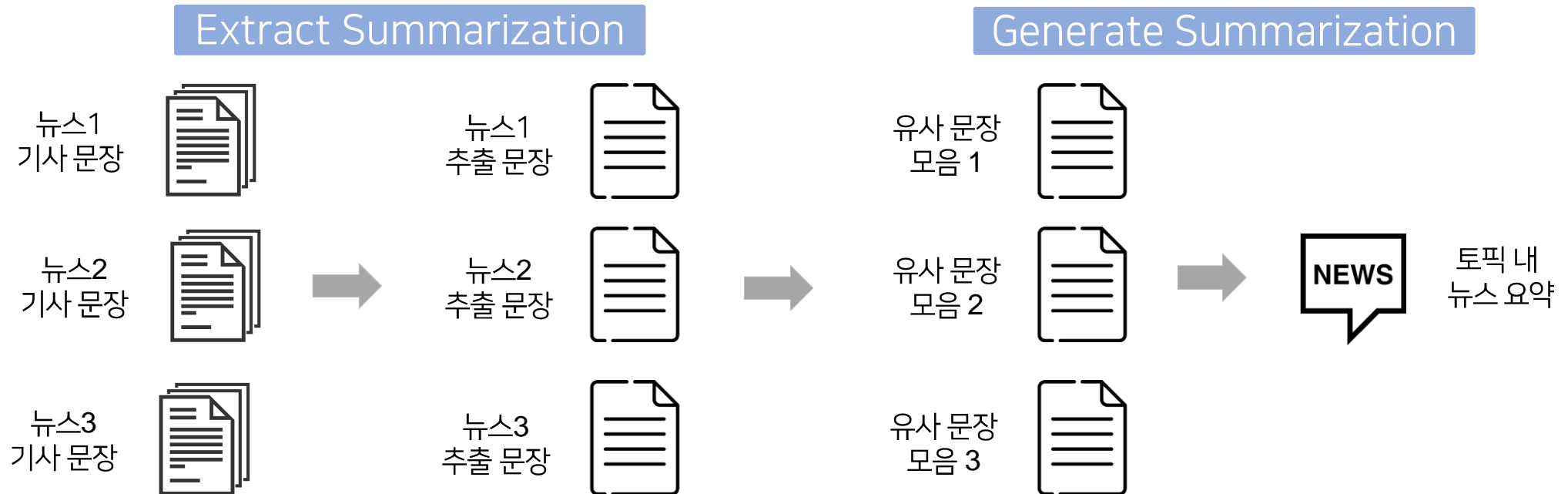
감성 분석(Sentiment Analysis)



한 문장을 Sequence Classification Model에 넣어
Positive, Neutral, Negative 세 가지 class로 분류

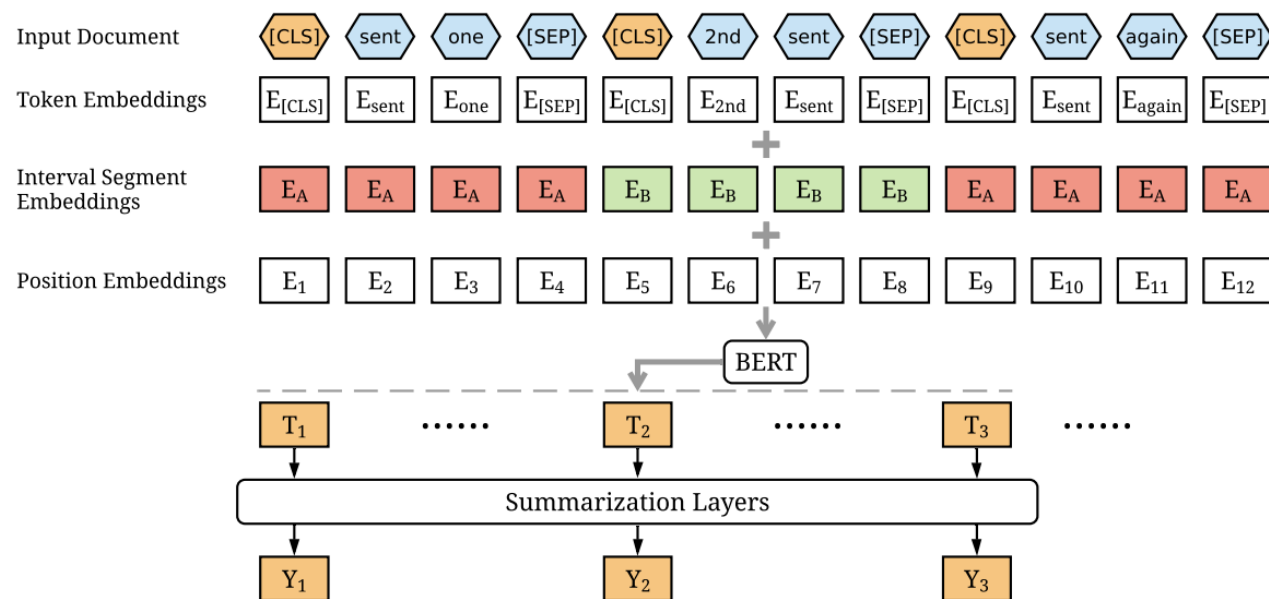
토픽 내 뉴스 요약(News Context Summary)

한 토픽 내에 포함된 여러 개의 뉴스들을 한눈에 볼 수 있도록 요약



토픽 내 뉴스 요약(News Context Summary)

추출 요약 : KorBertSum을 사용해 전처리된 문장을 중요도 순으로 정렬



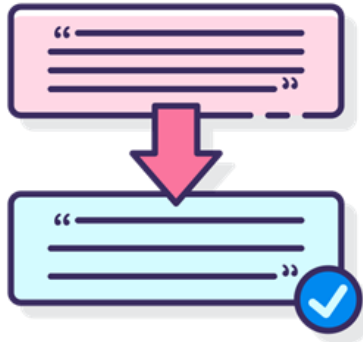
토픽 내 뉴스 요약(News Context Summary)

생성 요약 : 유사한 문장들을 concat해서 한 줄로 요약

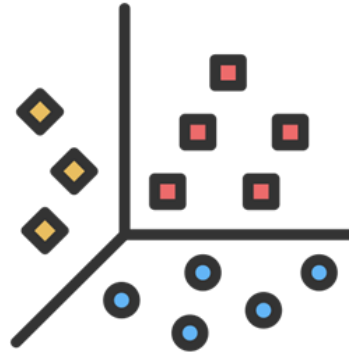
1. 추출 요약된 문장을 입력



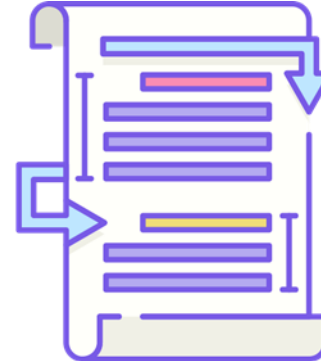
2. 비슷한 문장 제거



3. 문장 주제별 클러스터링



4. 생성 요약



5. 요약문 후처리



3. Model Development

토픽 모델링(Topic Modeling)

Embedding 모델을 기준으로 실험 진행

Embedding Model	Silhouette Score	Speed(sec)
Paraphrase mpnet ¹⁾	0.7585	7.34
KR-SBERT ²⁾	0.7439	6.68
DistillBERT ³⁾	0.7012	7.88
Paraphrase MiniLM ⁴⁾	0.6994	5.81
QA mpnet ⁵⁾	0.6927	11.16

Dataset : 2022.12.01 ~ 2022.12.15 '삼성전자'를 검색어로 한 717개 기사

- 1) paraphrase-multilingual-mpnet-base-v2 모델(다중 언어) <https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2>
- 2) KR-SBERT-V40K-klueNLI-augSTS(한국어), <https://huggingface.co/snunlp/KR-SBERT-V40K-klueNLI-augSTS>
- 3) multi-qa-distilbert-cos-v1(다중 언어), <https://huggingface.co/sentence-transformers/multi-qa-distilbert-cos-v1>
- 4) paraphrase-multilingual-MiniLM-L12-v2(다중 언어), <https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2>
- 5) multi-qa-mpnet-base-dot-v1(다중 언어), <https://huggingface.co/sentence-transformers/multi-qa-mpnet-base-dot-v1>

토픽 한 줄 요약(Topic Summary)

Model	Rouge-1(F1)	Rouge-2(F1)	Rouge-3(F1)	Length	Speed(sec)
kobart-summarization⁶⁾	0.495	0.339	0.413	115.83	0.46
KoT5_news_summarization ⁷⁾	0.495	0.329	0.385	201.49	3.19
kobart-news ⁸⁾	0.488	0.324	0.394	180.29	0.64

Dataset : Alhub 뉴스 기사 생성 요약 평가 데이터 18,000 rows

6) 한국어 wiki 데이터, DAICON 한국어 문서 생성요약 AI 경진대회 데이터로 학습한 모델 (<https://github.com/seujung/KoBART-summarization>)

7) AlHub 요약문 및 레포트 생성 데이터, Huggingface 'naver-news-summarization-ko' 데이터로 학습한 모델 (https://huggingface.co/noahkim/KoT5_news_summarization)

8) 한국어 wiki 데이터, AlHub 문서요약 텍스트/신문기사 데이터로 학습한 모델 (<https://huggingface.co/ainize/kobart-news>)

감성 분석(Sentiment Analysis)

한국어로 번역한 Finance Phrase Bank 데이터를 학습한 모델

Model	Loss	AUPRC	Micro F1	Speed(sec)	Easy data (#48)	Medium data(#22)	Hard data (#23)	Total data (#93)
roberta-large	0.4667	88.1713	82.7956	0.7371	43	18	16	77
roberta-base 1	0.9074	87.4126	76.3440	0.2793	42	17	12	71
roberta-base 2	0.5078	88.6208	78.4946	0.2668	42	14	17	73
KorFinASC-XLM-RoBERTa ⁹⁾	4.3266	29.8050	32.2580	0.8201	14	7	7	28

Dataset : 2022.12.01 ~ 2022.12.31 삼성전자, 하이닉스, 네이버, 카카오 한 줄 요약

9) KorFin-ASC, [Ko-FinSA](#), [Ko-ABSA](#) and [ModuABSA](#) 데이터로 학습한 모델 (<https://huggingface.co/amphora/KorFinASC-XLM-RoBERTa>)

토픽 내 뉴스 요약(News Context Summary)

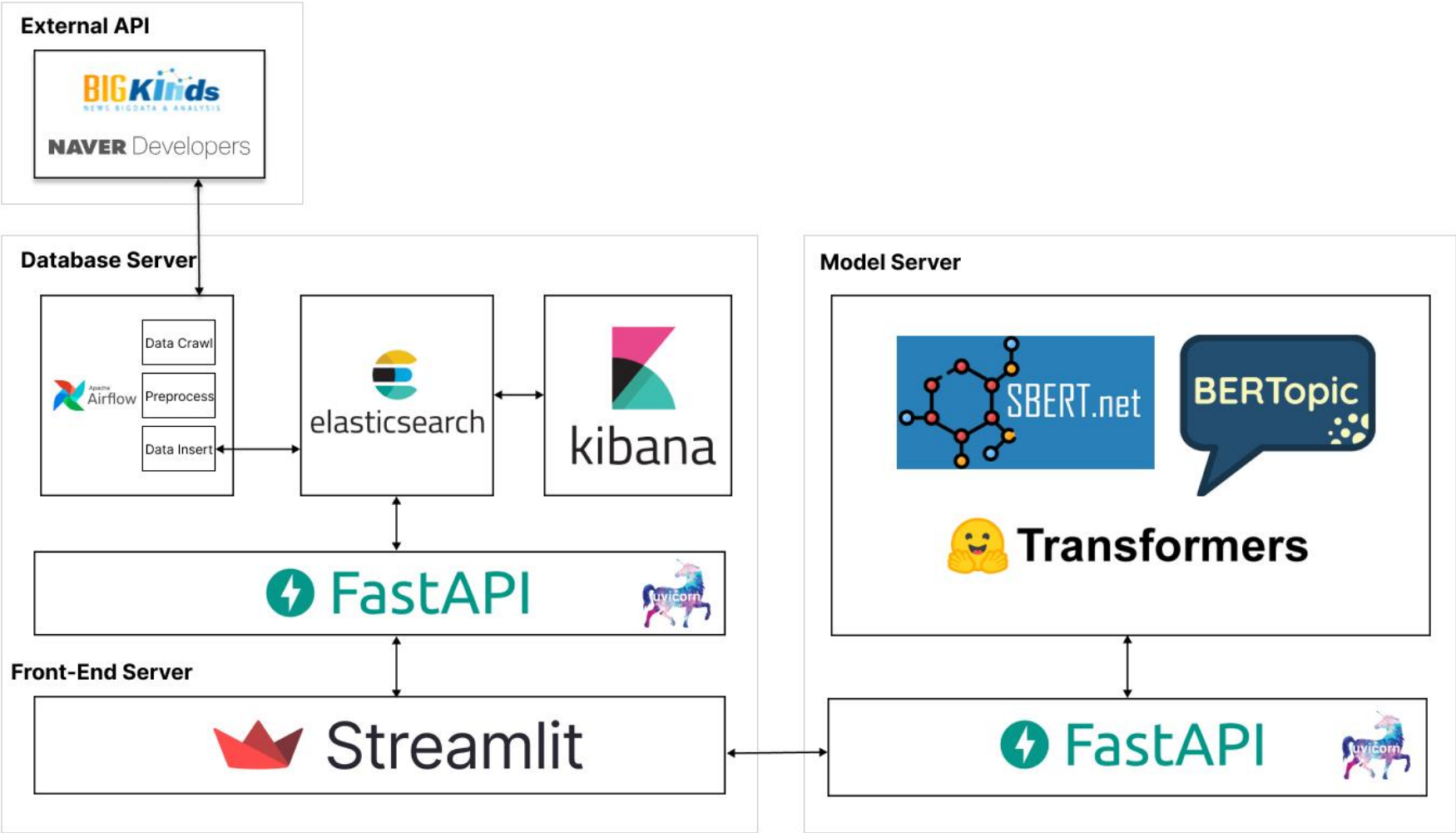
Etri pretrained 한국어 BERT 모델을 AIHub 추출 요약 데이터셋으로 추가 학습

Model	Rouge-1(F1)	Rouge-2(F1)	Rouge-3(F1)	Rouge-1 (Recall)	Rouge-2 (Recall)	Rouge-3 (Recall)
Etri pretrained model ¹⁰⁾	0.7550	0.5944	0.7045	0.7213	0.5661	0.6714
AIHub data fine-tuned model	0.7834	0.6365	0.7295	0.7969	0.6467	0.7421

10) 한국어 뉴스/백과사전 데이터로 학습한 모델(<https://github.com/BM-K/KoSentenceBERT-ETRI>)

4. Service & Product

4.1 아키텍처



4.2 시연 영상

NEWSUMMARY

삼성전자

기사 검색 선택

경제

기사 카테고리 선택

정치 경제 국제 지역 스포츠

2022/12/01 - 2022/12/15

검색된 뉴스 170개
추출 토픽 수 26개

경제(6)

경제 CATEGORY

솔루션 반도체 메모리

삼성전자와 네이버가 AI 반도체 솔루션 개발 협력을 위한 업무협약 체결하고 실무 태스크포스를 발족하여 초거대 AI에 최적화된 반도체 솔루션 개발에 협력하기로 했다.

경제 CATEGORY

강병일 해린 리조트

삼성물산은 7일 정해진 삼성전자 사업지원 태스크포스 부사장을 삼성물산 리조트부문 이사 사장 겸 삼성웰스토리 이사로 승진 내정하고 강병일 삼성물산 건설부문 경영지원도 각각 사장으로 승진 내정했다고 밝혔다.

경제 CATEGORY

철도 이달 벤치기업

중소벤처기업부는 기업의 상생 협력 활동을 격려하고 동반성장 문화를 확산하기 위해 기업의 상생협력 활동 우수사례에 대해 포상하는 행사인 '이달의 상생볼'에 삼성전자와 SKT 그리고 포스코, 국가철도공단, 국민은행 등 5개사가 10월 '이달의 상생볼' 대상으로 선정됐다고 5일 밝혔다.

경제 CATEGORY

중동 바라카 아부다비

이재용 삼성전자 회장이 취임 후 처음으로 정보 UAE 아부다비 알 다프라주에 있는 바라카 원자력 발전소 주를 방문하여 중동 '신시장 개척' 본격화

경제 CATEGORY

무역의 날 산업 서울

산업통상자원부와 한국무역협회가 5일 서울 삼성동 코엑스에서 올 한 해 우리 무역의 확대를 위해 노력한 수출기업과 유공자를 격려하기 위한 '제59회 무역의 날' 기념식을 개최하고 1780개사에 '수출의 탑'을 수여한다.

경제 CATEGORY

하수 반도체 화성

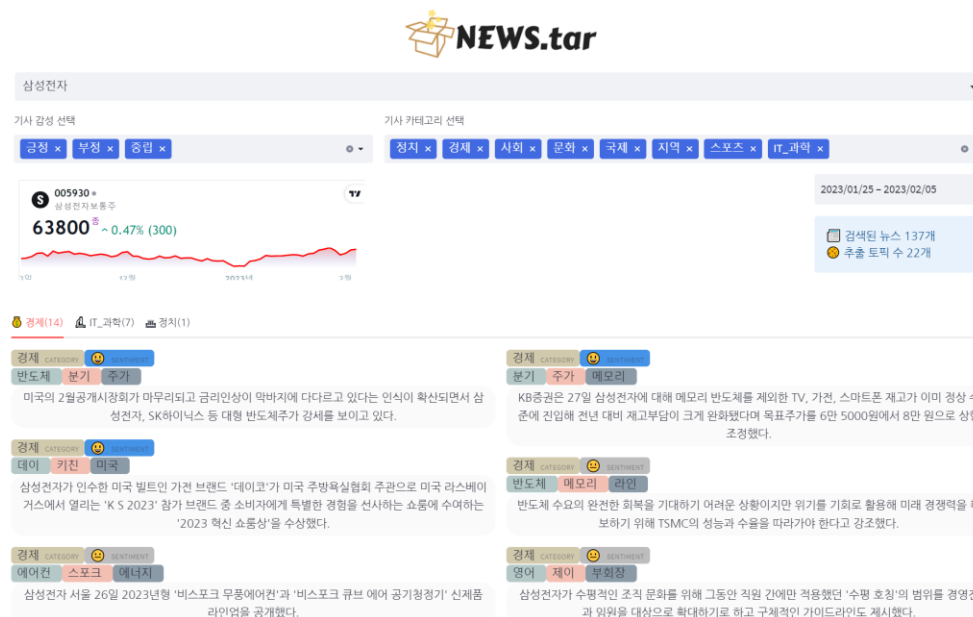
30일 삼성전자는 화성·수원·용인·화성·오산시 공공하수처리장의 방류수를 반도체 사업장에서 필요한 공업용수 수준으로 처리해 기흥·화성·평택 사업장에서 공급하는 업무협약을 체결했다.

KAPWING

5. Conclusion

5.1 개선사항

① UI/UX



❑ 각 기능에 대한 부연 설명 탭 추가

5.1 개선사항

② 속도

검색 기간	주제별 분류	한 줄 요약 생성	추출 요약 (토픽 당)	문단 요약 생성 (토픽 당)
3일	9.5257	3.8788	6.4556	2.4388
7일	5.2232	9.8479	5.9897	1.4234
15일	9.1015	19.3342	6.0331	1.3325
30일	16.3753	40.4939	6.2538	1.8158

검색어 : 삼성전자

- ❑ 더 작은 사이즈의 Embedding 모델을 사용
- ❑ 기사 별 추출 요약을 DB에서 미리 처리한 상태로 호출
- ❑ 데이터 입출력을 배치 단위로 묶어서 처리

5.1 개선사항

③ 뉴스 토픽 모델링 & 키워드

현재 생성된 토픽 그룹 및 키워드

'5만전자' 되자 자사주 담는 삼성 임원들
'5만전자' 되자 다시 자사주 사들이는 삼성전자 임원들
삼성전자 주가 6만원대로...'삼성생명법' 재논의 영향
삼성전자 다시 6만1000원대로 '뚝' 코스피도 1% 하락 중

코스피_분기_주가_주식

예상되는 토픽 그룹

'5만전자' 되자 자사주 담는 삼성 임원들
'5만전자' 되자 다시 자사주 사들이는 삼성전자 임원들
자사주_5만전자_임원

삼성전자 주가 6만원대로...'삼성생명법' 재논의 영향
삼성전자 다시 6만1000원대로 '뚝' 코스피도 1% 하락 중
주가_하락_6만원_코스피

- ❑ 단어 그래프나 랭킹을 통하여 중요한 단어와 중요하지 않은 단어들 가중치 표시
- ❑ 더 정교한 토픽 모델링을 통하여 뉴스 문서 클러스터링

5.1 개선사항

④ 생성된 요약문 품질

생성된 한 줄 요약문

삼성물산은 삼성물산 리조트부문 삼성웰스토리 이사에 정해린 삼성전자 사업지원 태스크포스 부사장을 리조트부문 이사 사장 겸 삼성웰스토리 이사로 승진 내정했다고 7일 밝혔다.

기대하는 한 줄 요약문

삼성물산은 정해린 삼성전자 사업지원 태스크포스 부사장을 삼성물산 리조트부문 대표이사 사장 겸 삼성웰스토리 대표이사로 승진 내정했다고 7일 밝혔다.

검색어 : 삼성전자

- ❑ 더 정교한 전처리를 통해 향상된 품질의 인풋 데이터 입력
- ❑ 추가적인 뉴스 요약문 데이터를 활용하여 모델 추가학습 진행
- ❑ 자체적인 Test set을 구축하여 정량적인 요약문 품질 평가

5.1 개선사항

⑤ 감정 분석 정교화

금융과 직접적인 연관이 없는 문장

동서와 동서식품은 연말을 맞아 사회복지공동모금회·초록우산어린이재단·한국여성재단·대한적십자사·대한적십자사·따뜻한동행·한국소아암재단·네이버 해피빈 등 총 7억6000만원의 이웃돕기 성금을 기탁했다고 27일 밝혔다

공정거래회가 김범수 카카오 미래이니셔티브센터장이 지분을 100% 보유한 개인회사 케이큐브홀딩스를 검찰에 고발하기로 했다.

감정 분류가 어려운 문장

현대그룹 광고회사 이노션에 2005년 창립 이후 최초로 여성 부사장으로 승진한 김정아 전무는 1996년 광고계에 입문한 이래 26년 동안 현대자동차그룹, 한화그룹, SKT, 신세계, KT, KT, CJ, 카카오 등 대한민국 기업들의 브랜드 캠페인을 제작, 책임, 총괄 진행해 왔다.

네이버가 부동산 매물 정보 제공 업체에 대한 계약을 통해 카카오의 시장 진입을 막았다는 혐의로 진행 중인 재판에서 관련 혐의를 전면 부인했다.

- ❑ Task 특화된 학습 데이터 구축
- ❑ 긍정/중립/부정 세 가지 분류보다 더 세부적인 단계로 분류

5.2 후속개발

① 데이터 크롤링 1일단위 배치 -> 실시간

- ❑ 현재 방식 : 1일 단위로 전날 뉴스 전체를 받아와서 크롤링 및 전처리 하여 저장
- ❑ 후속 개발 방식 : 더 작은 단위의 시간으로 API를 호출하여 크롤링 및 전처리 하여 저장
=> 새로운 뉴스 즉각 반영
- ❑ Airflow를 이용하여 해당 작업 자동화 (5분 or 그 이상)



Apache
Airflow

6. Appendix

예상 Q&A

1. Transformers를 활용한 생성 요약 모델은 잘못된 문장 종종 있는데, 이러한 출력을 목격한 경우가 있는지, 어떻게 해결하려고 했는지?
2. 같은 주제로 분류된 기사들을 모두 합쳐서 입력했다고 했는데, 긴 시퀀스를 어떻게 처리해서 입력했는지?
3. Non-AI 방식을 사용하는 방법도 생각해 봤는지?

6.2 참고자료

- Grootendorst, Maarten. "BERTopic: Neural topic modeling with a class-based TF-IDF procedure." *arXiv preprint arXiv:2203.05794* (2022).
- Malo, Pekka, et al. "Good debt or bad debt: Detecting semantic orientations in economic texts." *Journal of the Association for Information Science and Technology* 65.4 (2014): 782-796.
- Lewis, Mike, et al. "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension." *arXiv preprint arXiv:1910.13461* (2019).
- Lee, Dongyub, et al. "Reference and document aware semantic evaluation methods for Korean language summarization." *arXiv preprint arXiv:2005.03510* (2020).
- Liu, Yang, and Mirella Lapata. "Text summarization with pretrained encoders." *arXiv preprint arXiv:1908.08345* (2019).

감성 분석(Sentiment Analysis)

Baseline(한국어 뉴스 감성분석 모델)

- [amphora/KorFinASC-XLM-RoBERTa](#) trained on KorFin-ASC, [Ko-FinSA](#), [Ko-ABSA](#) and [ModuABSA](#)

문제점

- 입력값이 "감성분석 대상이 포함된 문장</s>감성분석 대상" 으로 구성되어야 함.
- 실제 문장에 대해 감성분석 결과가 좋지 않음.
- 하지만 학습 데이터의 문장에 회사명이 들어가지 않는 경우가 많아 학습이 어려움

따라서 입력값이 문장만 들어가도록 Sentiment Analysis 모델 학습이 필요함.

Model	Loss	AUPRC	Micro F1
KorFinASC-XLM-RoBERTa	4.3266	29.8050	32.2580

Dataset : 2022.12.01 ~ 2022.12.31 삼성전자, 하이닉스, 네이버, 카카오 한 줄 요약

감성 분석(Sentiment Analysis)

Train/valid dataset

- [Finance Phrase Bank 번역 데이터^{11\)}](https://github.com/ukairia777/finance_sentiment_corpus)
- Positive, Netural, Negative 3개의 카테고리로 분류
- Train/valid 구성
 - Train dataset : 4361개
 - Valid dataset : 485개

Test dataset

- 2022.12.01 ~ 2022.12.31 삼성전자, 하이닉스, 네이버, 카카오 한 줄 요약 93개의 문장에 대해 라벨링
- Test dataset level : 학습데이터와 실제 데이터와의 비교를 위해 다음과 같이 난이도를 부과함
 - 난이도 하(0) - 상승, 하락 등 쉬운단어 또는 데이터셋에 비슷한 문장이 존재
 - 난이도 중(1) - 데이터셋과 비슷한 듯 하면서 다른 문장
혹은 데이터셋에는 등장하지 않지만 라벨을 유추할 수 있는 경우
 - 난이도 상(2) - 데이터셋에 비슷한 문장이 없거나 금융관련 문장이 아닌 경우

11) https://github.com/ukairia777/finance_sentiment_corpus

감성 분석(Sentiment Analysis)

난이도에 따른 Dataset 예시

level	Train dataset	Train dataset label	Test dataset	Test dataset label
0	ADP 뉴스 - 2009년 2월 13일 - 핀란드 소매업체 Kesko Oyj HEL: KEBV는 오늘 부가가치세 부가세를 제외한 총 매출이 2009년 1월 6억 6,130만 유로로 전년 동기 대비 15.2% 하락했다고 밝혔다.	negative	반도체 업황 악화로 삼성전자의 올해 4분기 실적이 전년 대비 '반토막' 날 것이라는 전망이 나오자 삼성전자 주가가 전다 1.49% 하락한 5만9500원에 거래를 마쳤다.	negative
1	카메코는 탈비바라와의 이륙 협약 조건에 따라 우라늄 추출 회로 건설 비용을 충당하기 위해 최대 6000만 달러를 선불로 투자하기로 했다.	positive	카카오엔터테인먼트가 사우디아라비아 국부펀드 등 해외 유수의 국부펀드로부터 약 1조2000억원 규모의 투자를 유치했다고 12일 공시했다.	positive
2	-	-	삼성전자, SKC, 포스코, 국가철도공단, 국민은행 총 5개사가 10월달 '이달의 상생불'로 삼성전자, SKC, 포스코, 국가철도공단, 국민은행 총 5개사의 상생협력 활동을 선정했다고 5일 밝혔다.	positive

토픽 내 뉴스 요약(News Context Summary)

전처리되어 문장으로 나누어진 context를 추출 요약 모델을 이용해 중요도순으로 정렬
사용한 추출 요약 모델 : KorBertSum

["반도체 업황 악화로 삼성전자의 올해 4분기 실적이 전년 대비 '반토막' 날 것이라는 전망이 나오자 삼성전자 주가가 12일 다시 5만원대로 하락했다. ",
'이날 유가증권시장에서 삼성전자는 전 거래일보다 1.49% 하락한 5만9500원에 거래를 마쳤다. ',
'삼성전자는 최근 6만원 안팎에서 등락하며 '5만전자'와 '6만전자'를 오가고 있다. ',
"삼성전자는 이미 지난 3분기 실적이 '어닝 쇼크'를 기록했다. ",
'연결 기준 3분기 영업이익은 10조8520억원으로 지난해 동기보다 31.39% 감소했다. ',
'증권가는 삼성전자의 실적 부진이 4분기에도 지속될 것으로 보고 있다. ',
'DB금융투자 어규진 연구원은 이날 보고서를 통해 삼성전자의 올해 4분기 매출액은 전년 동기 대비 3.8% 감소한 73조7000억원, 영업이익은 49.9% 감소한 6조9000억원을 기록할 것으로 전망했다. ',
'사업부별 영업이익 기대치는 반도체 2조원, 디스플레이 1조8000억원, 스마트폰(MX) 2조6000억원, 소비자가전(CE) 5000억원으로 각각 추정됐다. ']



[6, 8, 10, 4, 2, 0, 9, 7, 5, 1, 3, 11]

[(6,
'DB금융투자 어규진 연구원은 이날 보고서를 통해 삼성전자의 올해 4분기 매출액은 전년 동기 대비 3.8% 감소한 73조7000억원, 영업이익은 49.9% 감소한 6조9000억원을 기록할 것으로 전망했다. '),
(4, '연결 기준 3분기 영업이익은 10조8520억원으로 지난해 동기보다 31.39% 감소했다. '),
(2, "삼성전자는 최근 6만원 안팎에서 등락하며 '5만전자'와 '6만전자'를 오가고 있다. "),
(0,
"반도체 업황 악화로 삼성전자의 올해 4분기 실적이 전년 대비 '반토막' 날 것이라는 전망이 나오자 삼성전자 주가가 12일 다시 5만원대로 하락했다. "),
(7,
'사업부별 영업이익 기대치는 반도체 2조원, 디스플레이 1조8000억원, 스마트폰(MX) 2조6000억원, 소비자가전(CE) 5000억원으로 각각 추정됐다. '),
(5, '증권가는 삼성전자의 실적 부진이 4분기에도 지속될 것으로 보고 있다. '),
(1, '이날 유가증권시장에서 삼성전자는 전 거래일보다 1.49% 하락한 5만9500원에 거래를 마쳤다. '),
(3, "삼성전자는 이미 지난 3분기 실적이 '어닝 쇼크'를 기록했다. ")]

토픽 내 뉴스 요약(News Context Summary)

추출 요약을 이용해 중요도 순으로 정렬한 문장 배열에서 처음 n 퍼센트를 가져온 후, k-means 알고리즘을 이용해 비슷한 주제의 문장끼리 묶는다.

각 주제의 문장들을 concat해서 생성 요약하고, 문장들이 뉴스에서 몇 번째 문장이었는지 평균을 계산해 요약문에서의 순서를 정한다.

문장 번호	(6, '그는 갤럭시S 시리즈, 폴더블폰 개발을 주도한 공로를 인정받아 40대 부사장으로 승진했다.')
	(18, '다양성과 포용성에 기반한 혁신적 조직문화를 구축하고 지속 가능한 기업으로서의 경쟁력 강화를 위함이다.')
	(8, '여성 및 외국인 발탁도 이어졌다.')
	(6, '이번에 승진한 DX부문 MX사업부 전략제품개발1그룹장인 문성훈 부사장의 경우 갤럭시 S 시리즈와 폴더블폰 등 주력 스마트폰의 하드웨어 개발을 주도하며 신규 기술발굴에 기여하는 등 모바일 비즈니스 성장을 견인한 점을 인정받았다.')
	(8, '이정원 DS부문 S.LSI사업부 모뎀개발팀장은 올해 45세로 가장 젊은 부사장이 됐다.')
	(4, '지난해에 이어 올해도 40대 부사장(17명)과 30대 상무(3명) 등 젊은 리더들을 중용했다.')

문장번호 평균	[8.333333333333334, ['문성훈 부사장의 경우 갤럭시 S 시리즈, 폴더블폰 개발을 주도하며 신규 기술발굴에 기여하는 등 모바일 비즈니스 성장을 견인한 점을 인정받아 40대 부사장으로 승진했다.'])]
------------	---

토픽 내 뉴스 요약(News Context Summary)

후처리로 요약문에서 반복되는 구절을 삭제한다.

deleted : 27 | 삼성전자를 비롯해 삼성디스플레이, 삼성SDI, 삼성전기

deleted : 72 | 등 23개 계열사가

사회복지공동모금회에 1999년부터 24년간 연말 이웃사랑 성금을 기탁한 삼성전자를 비롯해 삼성디스플레이, 삼성SDI, 삼성전기 등 23개 계열사가 연말 이웃사랑 성금 모금에는 삼성전자를 비롯해 삼성디스플레이, 삼성SDI, 삼성전기, 삼성SDS, 삼성생명, 삼성화재, 삼성카드, 삼성증권, 삼성물산, 삼성엔지니어링, 제일기획, 에스원 등 23개 계열사가 참여했다.

사회복지공동모금회에 1999년부터 24년간 연말 이웃사랑 성금을 기탁한 삼성전자를 비롯해 삼성디스플레이, 삼성SDI, 삼성전기 등 23개 계열사가 연말 이웃사랑 성금 모금에는 삼성SDS, 삼성생명, 삼성화재, 삼성카드, 삼성증권, 삼성물산, 삼성엔지니어링, 제일기획, 에스원 참여했다.

deleted : 18 | 고물가·고금리·고환율 등 이른바 ‘3고 시대’를 극복하기 위한 위기 대응책이 될 것으로

고물가·고금리·고환율 등 이른바 ‘3고 시대’를 극복하기 위한 위기 대응책이 될 것으로 예상되는 이번 회의 주요 안건은 고물가·고금리·고환율 등 이른바 ‘3고 시대’를 극복하기 위한 위기 대응책이 될 것으로 예상된다.

고물가·고금리·고환율 등 이른바 ‘3고 시대’를 극복하기 위한 위기 대응책이 될 것으로 예상되는 이번 회의 주요 안건은 예상된다.

deleted : 44 | 지난 5월 방한 때도 이 회장을 만나 차세대 메모리, 팹리스 시스템 반도체,

deleted : 6 | 팹리스 시스템 반도체, 파운드리(위탁생산), PC 및 모바일 등 다양한 분야에서 협력방안을

지난 5월 방한 때도 이 회장을 만나 차세대 메모리, 팹리스 시스템 반도체, 파운드리(위탁생산), PC 및 모바일 등 다양한 분야에서 협력방안을 논의했으며 지난 5월 방한 때도 이 회장을 만나 차세대 메모리, 팹리스 시스템 반도체, 팹리스 시스템 반도체, 파운드리(위탁생산), PC 및 모바일 등 다양한 분야에서 협력방안을 논의했다.

지난 5월 방한 때도 이 회장을 만나 차세대 메모리, 팹리스 시스템 반도체, 파운드리(위탁생산), PC 및 모바일 등 다양한 분야에서 협력방안을 논의했으며 논의했다.

토픽 내 뉴스 요약(News Context Summary)

요약문의 길이가 부족하다면 다음 n개의 문장들을 가져와 반복해서 수행한다.

```
.....448<570 get additional topk 22 .....
```

```
before delete similar: 12
```

```
겔싱어 CEO는 2021년 1월 인텔의 여덟 번째 CEO로 선임됐다.
```

```
겔싱어 CEO는 2021년 1월 인텔의 여덟 번째 CEO로 선임됐다.
```

```
delete_similar: 1.0
```

유사 문장 제거

```
after delete similar: 11
```

클러스터링

```
1
```

```
(10, 'TV·스마트폰·가전을 담당하는 삼성전자 DX(디바이스경험) 부문은 15~16일, 반도체를 담당하는 DS(디바이스솔루션) 부문은 22일 회의를 연다. ')
```

```
(6, '겔싱어 CEO 방한은 5월에 이어 올해에만 두 번째다. ')
```

```
(6, '겔싱어 CEO는 2021년 1월 인텔의 여덟 번째 CEO로 선임됐다. ')
```

```
(8, '경쟁 관계이자 밀접한 협력 관계이기도 하다. ')
```

```
(4, '김 사장도 만나 5세대(5G) 통신 관련 협의를 진행한 것으로 알려졌다. ')
```

```
(8, '당시 차세대 메모리, 팹리스(반도체 설계), 파운드리(반도체 위탁생산), PC 등 반도체와 세트 부문에 걸친 전방위적인 협력 방안을 논의했다. ')
```

```
(5, '앞서 5월 방한 때도 겔싱어 CEO는 이 회장을 만나 차세대 메모리, 팹리스 시스템 반도체, 파운드리(반도체 위탁생산), PC 및 모바일 등 여러 분야에서의 협력방안을 논의했다. ')
```

```
(8, '이 때문에 이 회장이 방한 중인 겔싱어 CEO를 만날지도 관심사다. ')
```

```
(4, '이번 방한 기간 겔싱어 CEO는 이재용 삼성전자 회장 등을 만날 것으로 예상된다. ')
```

```
(16, '이번 회의 주요 안건은 고물가·고금리·고환율 등 이른바 '3고 시대'를 극복하기 위한 위기 대응책이 될 것으로 예상된다. ')
```

```
(6, '이에 삼성전자와 인텔은 글로벌 반도체에서 1, 2위를 다투는 라이벌인 동시에 메모리와 중앙처리장치를 선도하는 동반자 관계를 형성하고 있다. ')
```

다음 n개 주요문장으로
문장 생성 후 요약문에 추가

End of Document

Thank You.