SAMSUNG SW AI ACADEMY FOR YOUTH

함께가요 미래로!
Enabling People

삼성청년 SW·AI아카데미

스트리밍 데이터 파이프라인 구축 (Kafka + Flink)

SAMSUNG SW AI ACADEMY FOR YOUTH

함께가요 미래로!
Enabling People

스트리밍 데이터 파이프라인 구축 (Kafka + Flink)

챕터의 포인트

- 관통 프로젝트 안내

SAMSUNG SW AI ACADEMY FOR YOUTH

함께가요 미래로!
Enabling People

관통 프로젝트 안내

관통 프로젝트 안내

Confidential

목표

- 이번 프로젝트의 목표는 실시간 스트리밍 데이터를 효율적으로 처리하고 적절히 저장·분석하는 파이프라인을 구축하는 것입니다.
- 이를 위해 Kafka와 Flink를 이용하여 스트리밍 환경을 구성하고, 실시간으로 데이터를 수집·처리할 수 있는 토대를 마련합니다.
- 이전 주차에서 구축한 데이터베이스와 데이터를 적극 활용하여, 기존 RSS 피드에서 새로운 데이터를 지속적으로 가져와 분석 가능하도록 하는 과정을 포함합니다.
- Kafka + Flink 구조를 이해하고 직접 구현해볼 수 있다.
- 이전 주차에서 획득한 데이터 처리 구조를 재활용하여 전체 파이프라인의 연속성을 확보한다.
- 실시간 데이터 처리 형태의 환경을 경험한다.
- 데이터 분류, 키워드 추출, 임베딩 등 다양한 전처리/분석 기술을 적용할 수 있다.

준비사항

- 사용 데이터
  - PostgreSQL DB
  - 이전까지 수집한 데이터

- 개발언어/프로그램
  - Python: Kafka Producer, Flink Consumer(파이썬 API) 구현
  - Kafka: 스트리밍 데이터 큐(브로커) 역할
  - Flink: 실시간 스트리밍 데이터 처리 엔진
  - PostgreSQL: 처리된 데이터 저장 (DB)
  - Requests, BeautifulSoup: 웹 크롤링 및 파싱

구현 방법

1) 데이터 수집 및 Kafka Producer 구현

- 실시간의 최신 데이터를 주기적으로 수집
- BeautifulSoup을 활용해 수집하는 데이터를 파싱
- 중복 데이터 처리를 위한 방식 활용
- Kafka Producer로 수집된 데이터를 JSON 형태로 직렬화하여 특정 토픽(TOPIC)에 전송

구현 방법

2) Flink를 통한 스트리밍 처리

- Flink Environment를 설정하고, Kafka Consumer를 사용하여 토픽의 실시간 메시지 수신
- 수신된 JSON 데이터를 Pydantic 모델로 매핑하여 구조화
- 예시로 확인할 transform_classify_category, transform_extract_keywords, transform_to_embedding 등 본인의 태스크에 맞는 사용자 정의 처리 과정을 통해 내용 전처리 및 분석

구현 방법

3) DB 저장

- Flink의 sink 형태로 PostgreSQL에 데이터(ex. 제목, 작성자, 본문, 카테고리, 키워드, 임베딩 등) 저장
- 실시간으로 저장된 데이터를 기반으로 후속 분석 가능

관통 프로젝트 가이드

1) 데이터 수집:

- 수집한 데이터 가공하고 취합하여 처리할 수 있는 Flink 코드 구성
- 크롤링을 통해 데이터 추가 확보
- Kafka Producer를 이용해 실시간 전송

관통 프로젝트 가이드

2) 데이터 처리:

- Flink로 스트리밍 데이터를 실시간 소비
- 구조화 및 전처리
- 카테고리 분류, 키워드 추출, 임베딩 처리

관통 프로젝트 가이드

3) 데이터 처리 이후 저장

- 분류된 데이터를 PostgreSQL에 설계한 DB의 스키마에 맞게 저장
- 중복으로 삽입되는 형태가 발생하지 않도록 처리 (카프카에서 처리되긴 하겠지만, 처리를 최대한 줄이기 위함)

관통 프로젝트 가이드

```
SELECT * FROM mynews_article ORDER BY "write_date" desc LIMIT 100
```

Search Results | Export | Cost: 594ms | 1 2 3 4 ··· 33 | Total 3295

id inte	title varchar(200)	writer varchar(255)	write_date timestamp w	cate varch	content text	url varchar(200)	keywords jsonb	embedding vector
3744	고층빌딩·골프장·테마파크 ᄌ	박준철 기자 terryus@kyung	2025-03-24 11:22	경제	인천 송도국제도시에 초고ᄎ	https://www.khan.co.kr/artic	["송도랜드마크시티", " 초	[0.005244136,0.012104424,(
3743	우원식 의장, 연금개혁 '미래	허진무 기자 imagine@kyun	2025-03-24 11:21	정치	우원식 국회의장은 최근 여(	https://www.khan.co.kr/artic	["국민연금", " 개혁안", " ᄋ	[0.014015239,0.01575779,0.
3742	혁신당 "군수선거서 민주당	이유진 기자 yjleee@kyungh	2025-03-24 11:19	정치	조국혁신당은 24일 이재명 ᄃ	https://www.khan.co.kr/artic	["조국혁신당", " 이재명", "	[0.008693117,0.027707772,-
3741	"전통시장 점포도 상세주소	김은성 기자 kes@kyunghya	2025-03-24 11:16	사회일반	빌딩과 아파트처럼 동·층·호	https://www.khan.co.kr/artic	["전통시장", " 3D 입체 주ᄉ	[-0.024739787,-0.02308054,
3740	6·25전쟁 당시 실종된 남아공	곽희양 기자 huiyang@kyun	2025-03-24 11:14	국제	6·25전쟁에 참전했다 실종된	https://www.khan.co.kr/artic	["유해 발굴", " 남아프리카	[0.0192715,0.05665059,0.03
3739	이제 안경 없이 3D 즐겨요···	최민지 기자 ming@kyunghy	2025-03-24 11:13	IT_과학	이제 별도의 안경 없이도 3D	https://www.khan.co.kr/artic	["3D 게이밍", " 오디세이 ᄉ	[-0.0060794456,-0.0134353
3738	'윤석열 탄핵심판 결정문' 미	박용하 기자 yong14h@kyur	2025-03-24 11:09	정치	조국혁신당이 24일 예상되는	https://www.khan.co.kr/artic	["조국혁신당", " 윤석열 다	[0.019564822,0.04293365,0.
3737	SKB, 아시아-미국 잇는 해저	배문규 기자 sobbell@kyung	2025-03-24 11:06	IT_과학	SK브로드밴드가 아시아와 미	https://www.khan.co.kr/artic	["SK브로드밴드", " 국제 해	[0.0022665525,-0.01785444
3736	미 정보수장도 한국 '패싱'···	윤기은 기자 energyeun@ky	2025-03-24 11:05	국제	미국 정보기관을 총괄하는 탐	https://www.khan.co.kr/artic	["털시 개버드", " 국가정보	[-0.027561024,-0.017770896
3735	[속보] 법원, 윤석열 첫 재판	김정화 기자 clean@kyunghy	2025-03-24 11:03	정치	12·3 비상계엄 관련 내란 우ᄉ	https://www.khan.co.kr/artic	["윤석열", " 비상계엄", " ᄂ	[0.032831922,0.028163875,-
3733	광진구청 직원 생일인 달 하·	김은성 기자 kes@kyunghya	2025-03-24 11:02	사회일반	서울 광진구청 직원은 자신(	https://www.khan.co.kr/artic	["서울 광진구청", " 생일 ᄐ	[0.0087944465,0.03951189,(
3734	셀트리온 스테키마, 미 유통·	이진주 기자 jinju@kyunghy·	2025-03-24 11:02	미분류	셀트리온의 자가면역질환 치	https://www.khan.co.kr/artic	["셀트리온", " 스테키마", "	[0.0030780577,-0.00994867·
3732	'비싼 아파트가 왜 이래?'···초	김지혜 기자 kimg@kyunghy	2025-03-24 11:00	사건사고	아파트 분양가가 매년 비싸ᄌ	https://www.khan.co.kr/artic	["아파트", " 하자", " 분양ᄌ	[-0.012741898,0.003917167,
3730	민주당, 한덕수 탄핵 기각 유	손우성 기자 applepie@kyur	2025-03-24 10:57	정치	더불어민주당은 24일 헌법재	https://www.khan.co.kr/artic	["더불어민주당", " 헌법재	[-0.008961228,0.01167937,-
3731	부산시, 건설업 등 고용유지	권기정 기자 kwon@kyungh	2025-03-24 10:57	경제	부산시는 중소기업의 고용·ᄀ	https://www.khan.co.kr/artic	["부산시", " 중소기업", " ᄀ	[-0.011828643,0.07418075,0
3729	국힘, 한덕수 탄핵 기각에 "C	문광호 기자 moonlit@kyun	2025-03-24 10:54	정치	국민의힘은 24일 헌법재판소	https://www.khan.co.kr/artic	["국민의힘", " 헌법재판소	[-0.0023872056,0.02554068

요구사항

- 기본 기능

- Kafka + Flink를 통한 실시간 데이터 파이프라인 구축
  - 카프카 토픽 생성 및 실시간으로 Producer가 데이터 전송
  - Flink Consumer가 메시지를 수신하여 DB에 저장
  - DB에는 이전에 구상한 동일한 스키마(또는 확장)를 사용하여 데이터 적재
- 요청조건
  - 이전 과정의 데이터 및 DB 사용을 하되, kafka와 flink를 거쳐 적재되는 형태로 구성
  - 적재 과정에서 키워드 및 카테고리 분류 등의 데이터 특성에 맞는 전처리 과정 포함

관통 프로젝트 안내

Confidential

요구사항

- AI 활용을 위한 데이터 가공 (임베딩(Embedding) 및 심화 분석)
- 요청 조건
  - 키워드 추출, 임베딩 로직을 코드로 구현
  - 임베딩 결과를 DB에 json 문자열로 저장
- 결과

  DB에 각 요소의 임베딩 필드가 추가로 저장되어야 함

  고급 분석(기사 간 유사도 분석 등) 가능

결과

- Kafka 토픽에 적재된 데이터
- Flink 환경에서의 Consumer 및 DB 적재 로그
- DB 테이블에 (준)실시간으로 쌓이는 데이터

관통 프로젝트 안내

Confidential

산출물

- Gitlab에 올라온 de-project 코드 기반의 레포지토리 지속적 커밋
- 이후 PJT도 해당 레포지토리에 이어 나가면서 개발

정리

구현 기능	체크 포인트
데이터 수집	Kafka Producer/Consumer 구성 Flink 처리
DB 적재	DB 스키마 맞게 중복 처리 예외 처리
AI 활용을 위한 데이터 가공	임베딩 기반 및 AI 활용 로직 처리를 위한 아이디어 유사도 분석 등의 처리를 위한 임베딩 적재 등 다양한 방향의 AI 활용에 필요한 형태로 데이터 가공

SAMSUNG
SW
AI ACADEMY
FOR
YOUTH

함께가요 미래로!
Enabling People

내일 방송에서 만나요!

삼성청년SW·AI아카데미