웹 크롤링 기반 쇼핑몰(쇼핑창고)

팀명:데미안
팀원:유진영, 임지수, 신지호, 트란트룽하우

프로젝트 개요

구글과 네이버와 같이 통합 검색을 제공하는 사이트에서도 쇼핑 탭에 개인 쇼핑몰의 상품은 표시되기 어렵다.

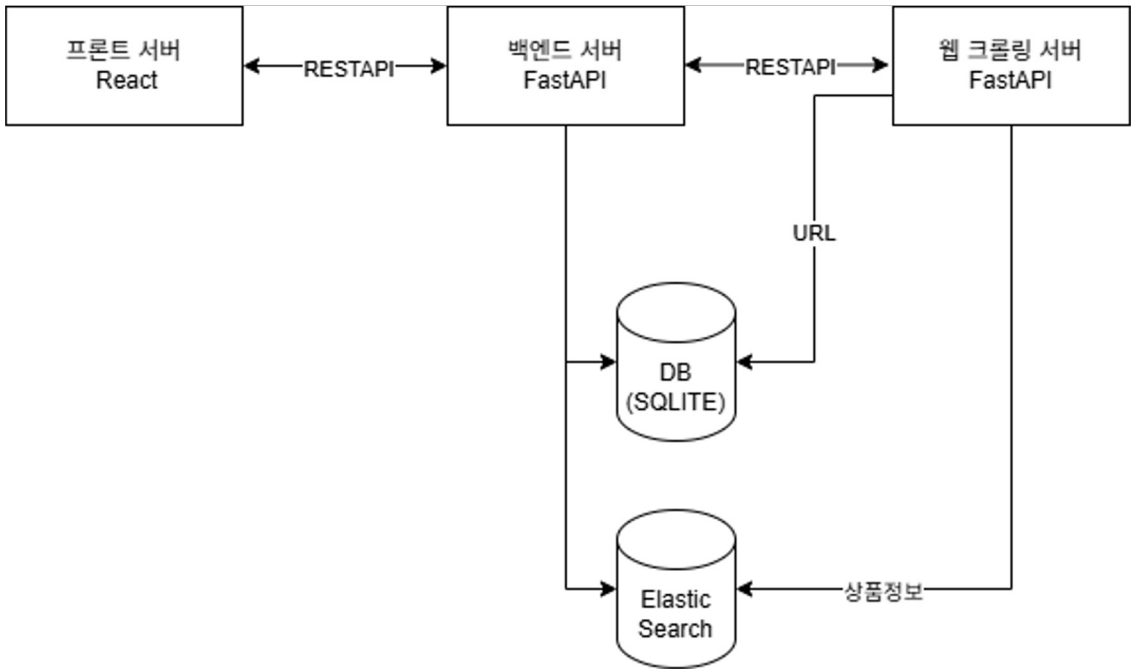
일부 제휴된 사이트에 등록하거나 구글에서 요구하는 정형적인 구조의 html을 맞춰야되는 한계가 있다.

이를 해결하기 위해 웹 크롤링 기반 쇼핑몰을 구성한다.

비정형적인 구조의 쇼핑몰에서 상품데이터를 추출하기 위해 직접 만든 쇼핑몰과 개인 쇼핑몰을 대상으로 한다.

VIDE(Vision-based Data Extraction)기반으로 하되 경량화하여 크롤링 주기를 줄여 정보의 신뢰성을 높임.

구조도


프론트 서버
React
RESTAPI
백엔드 서버
FastAPI
RESTAPI
웹 크롤링 서버
FastAPI
URL
DB
(SQLITE)
Elastic
Search
상품정보

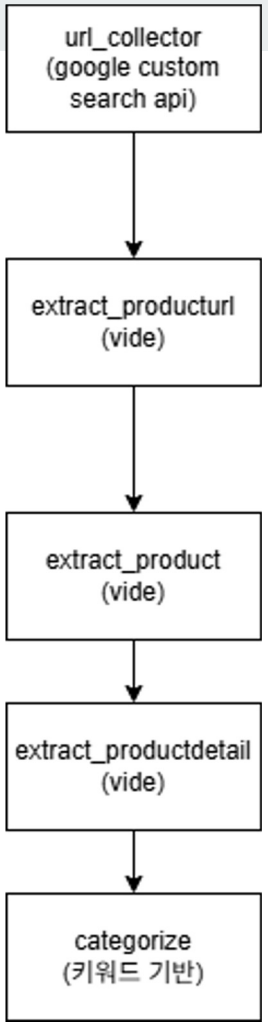
웹 크롤링 서버

-구글 custom search api에 키워드를 통해 root url 수집이후 vide로 상품 목록 url 수집

-수집된 상품 목록 url에서 상품 데이터 추출(vide)

-수집된 상품 url에서 상품 상세정보 추출(vide)

-키워드 기반 상품 카테고리 분류

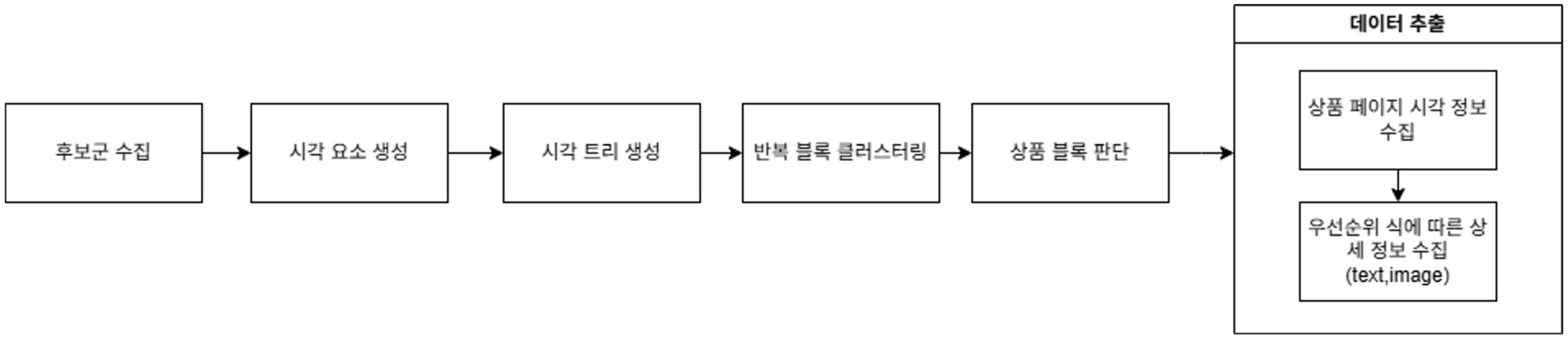


상품 데이터 추출

-이미지 텍스트 url 요소가 있는 후보군 수집

-포함관계에 따른 시각 트리 생성 및 비슷한 블록끼리 클러스터링

-상품 블록 판단(자식 요소 유무,가격 유무)



프론트 및 백엔드 서버

-ES에서 상품 이름, 카테고리, 텍스트 상세설명에 검색 단어가 포함된 경우 검색

-마지막 상품 클릭한 상품 기준으로 상세설명과 이름 기준으로TF-IDF 기반 유사 상품 추천

해결하지 못한 문제

-검색 엔진을 통한 url수집이라 개인 쇼핑몰 전체를 커버한다고 할 수는 없음

-키워드 룰 기반 카테고리 분류라 당연하게도 오류가 있고 완벽하지 못함

-상품 블록에서 데이터를 추출하는 부분이 hearistic한 방법이기에 완벽한 처리라고 보기에는 어려움

-상품 상세 데이터 추출은 시간이 너무 오래걸림(vide의 한계)

-관리자 기능이 없음