뉴스 기사 요약 자동화를 위한 서비스 개발

Ver 1.2

팀명: 대체불가

Table of Contents

1. 프로젝트 제목

뉴스 기사 요약 자동화를 위한 서비스 개발

2. 필요성 및 배경

최근 정보의 양이 폭발적으로 증가하면서 매일 수많은 뉴스 기사가 생산되고 있다. 기업, 투자자, 연구자 등 다양한 이해관계자들은 자신과 관련된 뉴스를 빠르게 파악하고 인사이트를 도출해야 하나, 수많은 기사 중 원하는 정보를 수작업으로 선별하고 요약하는 데에는 큰 시간과 비용이 소모되고 있다.

특히 산업별, 기업별 뉴스를 정기적으로 모니터링해야 하는 직군에게 정보 탐색과 요약의 자동화는 생산성 향상을 위한 핵심 과제이다.

현재 미디어 분야의 주요 문제는 다음과 같다.

1. 방대한 뉴스 양으로 인한 정보 과부하
2. 관심 있는 기업 또는 분야에 대한 기사만 골라내는 데 드는 시간 소모
3. 기사의 핵심 내용을 빠르게 파악하기 어려움
4. 기존 뉴스 알림 서비스의 낮은 개인화 및 실시간성 부족

3. 목표

본 프로젝트는 사용자가 설정한 언론사 또는 관심 분야에 맞는 뉴스를 자동으로 수집, 요약하고, 맞춤형 뉴스 기사를 추천해주는 서비스를 구축하는 것을 목표로 한다.

- 뉴스 내용 검색(RAG)
- 뉴스 요약(LLM)
- 관심 분야 뉴스 알림(push 알림)

4. 팀별 역할

김기훈	조장, UX 설계, 프론트엔드 개발	UX 설계도, 개발 코드
안영민	백엔드 개발, 논문 작성	개발 코드, 논문

지은주	백엔드 개발	개발 코드, 보고서
강민지	데이터 처리, 서비스 개념도 생성	데이터 처리 코드, 데이터 결과물

5. 요구사항 분석 및 정의

a. 자료 조사

i. LLM 기반 뉴스 추천 시스템

최근 연구에서는 LLM을 활용하여 뉴스 콘텐츠 분석, 사용자 관심사 예측, 추천 기능을 구현한 시스템이 활발히 연구되고 있다. → 출처: https://arxiv.org/abs/2502.09797

ii. RAG 시스템에서의 뉴스 통합

RAG 기반 뉴스 요약 시스템에서는 HTML, PDF, CSV 등 다양한 형식의 비정형 데이터를 통합하고 요약하는 기능이 중요하게 다뤄지고 있다.

iii. 실시간 뉴스 요약 오픈소스 프로젝트

GitHub 등에서는 실시간 뉴스 수집 및 요약을 위한 오픈소스 프로젝트가 다수 공유되고 있다. 예를 들어 itnews_qna_rag는 SQLite 및 FAISS를 이용해 IT 뉴스 요약 및 Q&A 기능까지 제공하는 프로젝트이다. → 출처: https://github.com/taejongK/itnews_qna_rag

b. 경쟁사 벤치마킹

Bigkinds

- **주요 기능**: 뉴스 빅데이터 분석, 기사 검색 및 필터링, 트렌드 분석 제공
- **차별화 포인트**: 사용자가 능동적으로 검색해야 하며, 자동 요약은 X, 맞춤형 이메일 발송 기능도 X
- 링크: https://www.bigkinds.or.kr/

Perplexity AI

- **주요 기능:** 사용자 질문 기반 기사 요약 제공
- **차별화 포인트**: 특정 회사에 대해 지속적으로 요약/구성된 뉴스 제공 기능은 없음
- 링크: https://www.perplexity.ai/

c. 요구사항 정의

주 사용자(고객)

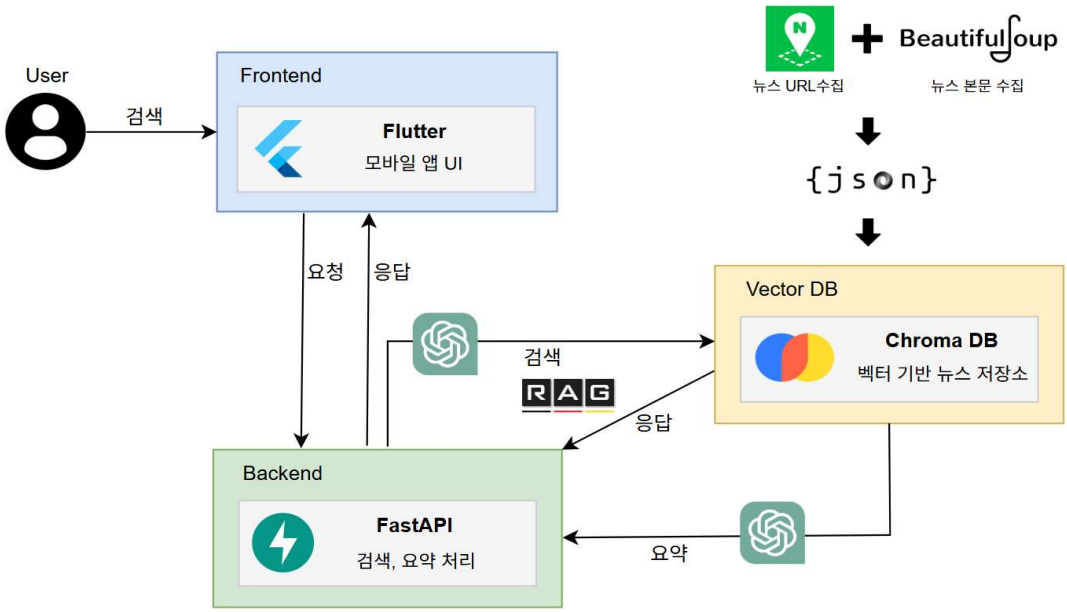
1. 기업 정보 담당자: 자사 또는 경쟁사에 대한 뉴스를 신속하게 파악하고, 전략 수립 및 보고서 작성에 활용하고자 하는 실무자
2. 투자자 및 애널리스트: 산업 동향과 특정 기업의 이슈를 실시간으로 파악하여 투자 전략 수립 및 분석 보고서에 활용하고자 하는 개인 및 기관 투자자
3. 관심 산업에 대한 뉴스를 원하는 일반 사용자: 특정 산업(예: 반도체, 친환경 에너지 등)에 관심이 있으며, 관련 뉴스를 효율적으로 모니터링하고 요약된 정보를 통해 빠르게 이슈를 파악하고자 하는 일반 사용자

요구사항

1. 기업/키워드/산업 기반 뉴스 수집: 사용자가 입력한 언론사, 키워드, 또는 산업 분야를 기반으로 관련 뉴스를 자동으로 검색하고 수집
2. 뉴스 요약 자동화: 수집된 뉴스 기사를 최신 언어모델을 활용하여 관련성 높은 뉴스를 선별하고 핵심 내용을 중심으로 간결하게 요약
3. 맞춤형 뉴스 전달: 사용자별 관심사에 맞는 뉴스를 자동으로 선별하여 특정 시간에 알림

6. 서비스 설계(예시)

a. 서비스 개념 및 구조도



b. UX 설계


뉴스 검색 챗봇
뉴스 검색
저장
채팅 내역
검색 및 언론사 선택
검색어를 입력하세요
언론사 선택
검색
요약
설정
뉴스 검색
관련 답변
관련 기사
검색어를 입력하세요
전체
검색
요약
설정


키워드 기반 검색 및 요약
키워드 선택
환영합니다!
관심 있는 키워드를 선택해주세요
처음 시작 화면
정치
경제
사회
국제
문화
스포츠
IT
과학
환경
교육
시작하기
관심 키워드 선택 후 시작
설정
관심 키워드 관리
관심 키워드 관리
선택된 키워드 (2개)
추가할 키워드
키워드 기반 검색 및 정보 제공
키워드 요약
키워드
뉴스 정보

c. 데이터 설계

(1) 데이터 수집

- 출처: Naver News API
- 수집 기준: 키워드 기반 최신 뉴스, 하루 100건 내외
- 포맷: title, content, link, date, journal 등의 JSON 구조
- 저장 위치: data/news.json 형태로 정형화


```
{
    "lastBuildDate": "Wed, 07 May 2025 21:15:34 +0900",
    "total": 6054094,
    "start": 1,
    "display": 10,
    "items": [
        {
            "title": "CJ그룹 <b>IT</b> 인프라 관리 'CJ올리브네트웍스' 해킹 의심",
            "originallink": "https://www.ytn.co.kr/_ln/0102_202505071627332256",
            "link": "https://n.news.naver.com/mnews/article/052/0002189932?
                sid=101",
            "description": "SK텔레콤의 유심 해킹 사태 파장이 가라앉지 않는 가운데
                CJ그룹의 <b>IT</b> 인프라를 관리하는 'CJ 올리브네트웍스'의 인증서 파일도
                해킹으로 유출된 것으로 의심되고 있습니다. 중국 보안 기업으로 알려진 '레드
                드립팀'은... ",
            "pubDate": "Wed, 07 May 2025 16:27:00 +0900"
        },
```


(2) 전처리 및 필터링

- url을 매핑하여 언론사 데이터 추가
- 날짜 포맷을 변경하고, url을 활용해 BeautifulSoup으로 기사 본문 크롤링 후 json 파일로 저장

```
news_media_mapping = {
    "yonhapnews.co.kr": "연합뉴스",
    "yonhapnewstv.co.kr": "연합뉴스TV",
    "news1.kr": "뉴스1",
    "edu.donga.com": "동아일보",
    "biz.heraldcorp.com": "헤럴드경제",
    "daily.hankooki.com": "한국일보",
    "kmib.co.kr": "기독교일보",
    "kbs.co.kr": "KBS",
    "munhwa.com": "문화일보",
    "sports.naver.com": "네이버 스포츠",
```


```
[
    {
        "title": "체코 원전 계약 제동…국내 건설업계 '초긴장'",
        "date": "2025-05-07 10:20",
        "link": "https://n.news.naver.com/mnews/article/003/0013225746?sid=101",
        "content": "두산에너빌리티·대우건설 수조원 수주 여부 달려[두코바니(체코)=AP/뉴시스]체코 두코바니에 있는 두코바니 원자력
        "journal": "뉴시스"
    },
    {
        "title": "이창용 &amp;quot;최상목 사퇴 해명에 곤혹…환율 예단 어려워&amp;quot;",
        "date": "2025-05-06 15:07",
        "link": "https://n.news.naver.com/mnews/article/422/0000737805?sid=101",
        "content": "연합뉴스 제공이창용 한국은행 총재가 국내 정치 불확실성이 경제를 가라앉히는 주요 요인 중 하나가 될 수 있다고
        "journal": "연합뉴스TV"
    },
```


(3) 임베딩 및 벡터화

- 임베딩 모델: OpenAI text-embedding-ada-002, KoSBERT (jhgan/ko-sbert-nli)
- 기사별로 title + content를 하나의 문장으로 연결 후 임베딩 수행
- 벡터 크기: 768차원 (KoSBERT 기준)

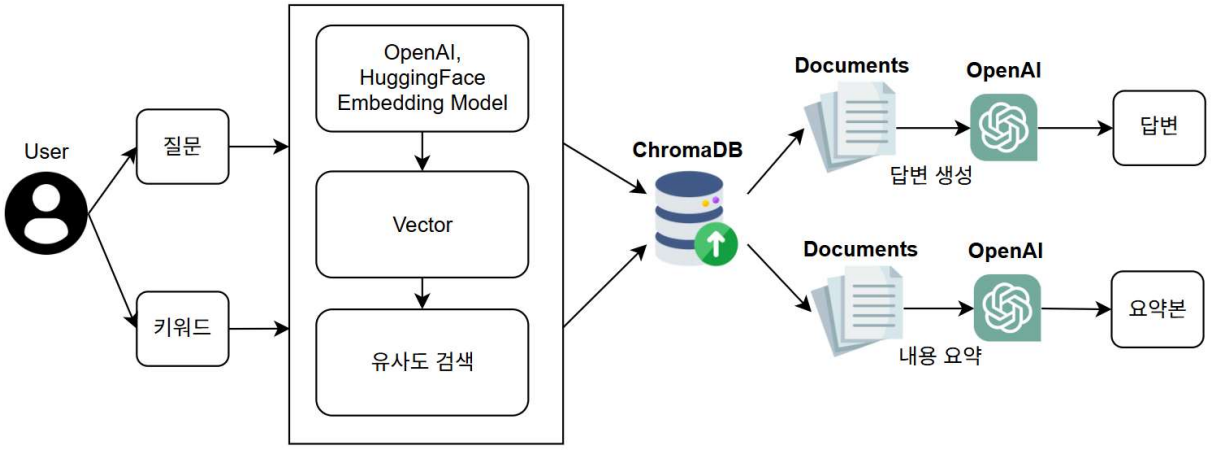
- 벡터 객체 구성: Document(page_content=..., metadata={title, date, link})

(4) 벡터 DB 저장 (ChromaDB)

- 저장 경로: ./chroma_news_db
- ChromaDB는 빠른 유사도 검색을 위해 cosine similarity 기반 인덱싱 제공
- chroma DB 구조

content	metadata	ids
검색 및 유사도 분석 대상인 뉴스 본문	부가정보 (기사 제목, 날짜, 링크, 언론사)	식별자

d. 프로그램 모듈 구조, 모델 및 알고리즘 설계



7. 테스트 계획

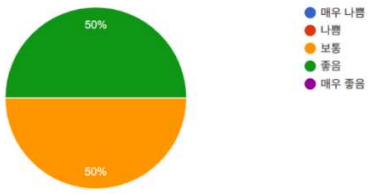
정확도: 벡터 검색(RAG)을 통해 제공되는 뉴스 콘텐츠가 사용자의 키워드 또는 관심 주제와 얼마나 일치하는지를 평가

생산성: 수동 뉴스 탐색 대비 시간 절약 효과를 정량적으로 비교

정보 리드타임: 뉴스 수집부터 요약 완료까지 소요되는 전체 시간 측정

생성된 뉴스의 **전반적인 품질**은 어땠나요?

응답 2개



생성된 뉴스의 **사실관계나 정보 정확성**은 얼마나 신뢰할 수 있었나요?

응답 2개

문장이 **간결**하고 **자연스럽게** 구성되었나요?

응답 2개

원문에 비해 **핵심 내용을 잘 요약**하고 있나요?

응답 2개

응답 **속도**는 **적절**했나요?

응답 2개

생성된 뉴스에서 **좋았던 점**과 **개선할 점**을 자유롭게 적어주세요.

응답 2개

처음 카테고리를 지정하는 페이지로 돌아갈 수 있게끔 뒤로가기 버튼이 있었으면 좋겠습니다.

설정창에서 카테고리를 재선택할 수 있게 되어있는데 메인페이지에서 선택할 수 있게끔 하면 더 좋을 것 같습니다.

생성된 뉴스의 양이 조금 더 많았으면 좋겠음

8. 구현 및 결과

- **구현 개요**

본 시스템은 사용자의 뉴스 키워드 또는 질문에 대해 관련 뉴스를 요약하거나 질의응답을 수행하는 LLM, RAG 기반 아키텍처로 구현되었다.

기능	설명
키워드 기반 요약	사용자가 입력한 키워드에 대해 벡터 유사도를 기반으로 유사 기사 4개를 추출하고, GPT-4 Turbo로 요약 생성
질의응답 (RAG)	질문을 임베딩 후 ChromaDB에서 관련 기사 검색 → GPT-4 Turbo로 응답 생성
뉴스 수집	naver news API를 사용해 하루에 한 번(오전 9시) 자동 수집하여 데이터 업데이트

1. 기능별 응답 예시

기능	예시 입력	제목	응답 요약
키워드 요약	해킹	"CJ그룹 'SW 인증서'도 해킹에 유출"	"CJ그룹의 SW 인증서가 해킹에 유출되었다. CJ올리브네트웍스는 CJ그룹의 IT 인프라를 관리하는 기업이다. SK텔레콤과 마찬가지로 CJ그룹도 해킹 문제에..."
질문 응답	해킹된 회사를 알려줘.	"최태원 SKT 해킹에 결국 고개 숙였지만...위약금 면제에는 '회의적'"	"해킹된 회사로는 SKT와 CJ그룹이 있습니다. SKT의 최태원 회장과 CJ그룹의 CJ올리브네트웍스에서 SW 인증서가 해킹에 유출되었습니다.... "

2. 속도

항목	json	ChromaDB
요약 생성 시간	2분 내외	20초 내외

3. 추후 개선 방안

항목	개선 방향	설명
1. 모델 비용 최적화	HuggingFace 기반 요약 모델로 전환	OpenAI API 호출 비용 절감, 속도 향상 가능
2. 벡터 DB 최적화	Chroma → FAISS + SQLite 이중 저장	검색 속도 향상, 배포 편의성 개선
3. 캐싱 도입	Redis 캐시로 질문-응답 결과 저장	동일 질의에 대한 중복 API 호출 방지
5. 데이터 수집 늘리기	뉴스 수집량 늘리기	더 많은 키워드에 대한 답변 개선

9. 감사의 글

본 과제 수행에 있어 유익한 피드백과 실질적인 조언을 아끼지 않으신 강사님께 깊은 감사를 드립니다. 프로젝트의 방향 설정과 기술적 문제 해결에 큰 도움이 되었으며, 과제를 성실히 완수할 수 있는 원동력이 되었습니다.

또한, 다양한 기술적 자료와 연구 기반을 바탕으로 한 강의 내용은 본 과제의 수행에 있어 중요한 참고가 되었습니다. 이에 다시 한 번 감사의 마음을 전합니다.

10. 레퍼런스

본 과제 할 때 다음 도구 및 자료들을 참고하였다.

1. [API] 네이버 뉴스 검색 API
2. [논문] LLM 기반 전세계 영어 및 한국어 뉴스 실시간 번역 및 요약 시스템 개발
3. [논문] **RAG 기반 한국어 뉴스기사 질의응답 및 팩트체크 시스템**
4. [도구] Chroma DB, Open AI LLM, Flutter, FastAPI, BeautifulSoup, RAG
5. [Github] https://github.com/dhivyeshrk/Retrieval-Augmented-Generation-for-news?utm_source=chatgpt.com