

**SK네트웍스 Family AI과정 6기  
 데이터 전처리 학습된 인공지능 모델**



**□ 개요**

* 산출물 단계 : 데이터 전처리
* 평가 산출물 : 학습된 인공지능 모델
* 제출 일자 : 2025.03.07
* 깃허브 경로 : https://github.com/SKNETWORKS-FAMILY-AICAMP/SKN06-FINAL-1TEAM.git
* 작성 팀원 : 백하은

| **사용 데이터 및 기술** | [사용 데이터 및 전처리]사업보고서 데이터  * **텍스트 분할(Text Split)**: LLM 기반 자연어 처리를 위해 문서를 소제목 단위로 분할 * **임베딩(Embedding)**: 텍스트를 벡터화하여 유사도 검색 최적화 (OpenAI Embeddings) * **벡터 데이터베이스(Vector DB)**: Chroma DB를 활용하여 벡터 검색 기능 제공  재무제표 데이터  * **API 활용**: OpenDart API를 통해 기업 재무제표 데이터 수집 * **데이터 전처리**:   + 단위 변환 (예: 백만 원 → 억 원)   + **Pivot 변환**을 통한 데이터 정렬 및 분석 최적화 * **관계형 데이터베이스(RDBMS)**: MySQL을 활용하여 정형 데이터 저장  뉴스 데이터  * **웹 크롤링(Web Crawling)**:   + 10분마다 **"국내자동차"** 키워드로 네이버 뉴스 크롤링   + Selenium, BeautifulSoup 등 크롤링 라이브러리 활용 * **BERT 모델 기반 NLP 처리**:   + **키워드 추출**: 사전 학습된 **BERT 모델**을 사용하여 뉴스에서 주요 키워드 추출   + **요약(Summarization)**: Transformer 기반 모델로 뉴스 내용을 요약   + **감정 분석(Sentiment Analysis)**: 긍정/부정 감정 분석 적용 * **MySQL 저장**: 크롤링 및 NLP 분석 결과를 관계형 DB에 저장하여 데이터 관리 최적화  [기술]데이터 저장 및 검색  * **Chroma DB**: 사업보고서 벡터 검색을 위한 **Vector DB** * **MySQL**: 재무제표 및 뉴스 데이터를 저장하는 **관계형 데이터베이스**  AI 및 데이터 분석  * **LLM 기반 문서 요약**: GPT, Llama 등 대형 언어 모델 활용 * **BERT 기반 뉴스 분석**: 키워드 추출, 요약, 감정 분석 수행 * **데이터 기반 예측 모델:**   + 선형회귀(Linear Regression): 전년도 매출 데이터를 기반으로 다음년도 매출 예측   + Train/Test Split(80:20) 활용  웹 크롤링 및 데이터 수집  * **크롤링 라이브러리**: Selenium, BeautifulSoup 활용 * **API 활용**: OpenDart API, 네이버 뉴스 검색 API 등  서버 및 백엔드  * **Django**: API 및 서버 구축 |
| --- | --- |
| **상황** | 1. 현재 모델 구조  * **Router 기반 Multi-Agent System 구현 완료**   + 사용자의 입력을 분석하여 적절한 Agent로 라우팅 가능 * **4개의 주요 Agent 구성**   + **요약 Agent**: 사업보고서 및 뉴스 요약   + **검색 Agent**: 기업 정보 및 데이터 검색   + **예측 Agent**: 기업 재무 데이터 및 트렌드 분석 기반 예측 수행   + **뉴스 Agent**: 뉴스 크롤링 및 분석 (현재 미완성)  2. 현재 실행 가능 상태  * **요약, 검색, 예측 Agent는 정상적으로 작동 중** * **뉴스 Agent는 분석 기능 구현 중**  3. 고도화 작업 진행 사항  * **답변의 폭과 질을 넓히기 위한 최적화 진행 중**   + 보다 정교한 검색 및 요약 기능 강화   + 예측 모델의 정확도 향상 * **Agent 간 협업 및 최적화**   + 다중 Agent 협업을 통한 복합 질의 처리 강화 * **RAG 최적화**   + 검색 성능 향상을 위해 벡터 DB(Chroma DB) 최적화   + LLM 기반 정보 추출 및 응답의 질 개선 |
| **수행 결과** | 1. 사업보고서 요약 모델  * LLM(GPT, Llama 등) 기반으로 사업보고서 핵심 내용 자동 요약 * Chroma DB를 활용한 벡터 검색 적용 완료 * 질의응답(Q&A) 기능 정상 작동  2. 뉴스 분석 모델  * 10분마다 "국내자동차" 관련 뉴스 크롤링 진행 중 * BERT 모델을 활용한 키워드 추출 완료 * 분석 결과 MySQL 데이터베이스에 저장  3. 예측 모델 (매출액 예측)  * 선형 회귀(Linear Regression) 모델을 활용하여 다음년도 매출 예측 가능 * 전년도 매출 데이터를 기반으로 예측 수행 * Train/Test Split (80:20 비율) 및 MSE, R² Score 평가 완료 * 사용자 입력 시 매출 예측 가능  4. 검색 기능  * Django API 기반으로 사업보고서 검색 구현 * Chroma DB를 활용한 벡터 검색 정상 작동 * 사용자의 질의에 맞춰 최적화된 검색 결과 제공  5. 전체 시스템 현황  * Router를 통한 Agent 분기 가능   + 요약, 검색, 예측 Agent 정상 작동 |
| **보완점** | 1. 뉴스 분석 및 감정 분석 고도화  * BERT 기반 감정 분석 모델의 정밀도 개선 필요 * 뉴스 데이터의 감정 분석 결과를 예측 모델과 연계하여 활용  2. 뉴스 Agent 구현  * 현재 크롤링 및 감정 분석은 진행 중이나, 뉴스 Agent는 미완성 상태 * Agent 구조를 구축하여 실시간 뉴스 분석 및 요약 자동화 필요 * 다른 Agent와 연계하여 뉴스 분석 결과를 검색 및 예측 모델에 반영  3. 예측 모델 고도화  * 다중 회귀(Multiple Linear Regression) 모델 적용 검토 * 비재무 데이터(뉴스 감정 분석 결과, 업계 동향 등) 반영한 예측 모델 개발 * LSTM 등 딥러닝 기반 시계열 분석 모델 도입 가능성 평가  4. 검색 및 문서 요약 최적화  * Chroma DB 벡터 검색 성능 개선 (쿼리 최적화, 인덱싱 방식 조정) * 사업보고서 요약의 품질 개선 (LLM 프롬프트 엔지니어링)  5. API 및 시스템 안정성 강화  * **Django API의 속도 및 응답 성능 최적화** |