

**SK네트웍스 Family AI과정 6기  
 모델링 및 평가 수집된 데이터 및 전처리 문서**



**□ 개요**

* 산출물 단계 : 모델링 및 평가
* 평가 산출물 : 수집된 데이터 및 전처리 문서
* 제출 일자 : 2025-03-07
* 깃허브 경로 : <https://github.com/SKNETWORKS-FAMILY-AICAMP/SKN06-FINAL-1TEAM>
* 작성 팀원 : 김지영

| **개요** | 사업보고서, 재무제표, 실시간 뉴스 데이터를 분담하여 데이터 수집 및 전처리 진행함.   1. **데이터 설명**  * **사업 보고서** * 임직원 정보(csv), 사업보고서 원본(xml), 회사 개요 및 사업 개요(csv), 재무제표(txt), 기업개황(csv), 임원현황(csv), 직원현황(csv), 최대주주현황(csv) * OpendDART 제공 API를 통해 수집 진행 * **재무제표** * 대차대조표,현금흐름표,손익계산서 * **실시간 뉴스** * 네이버에서 ‘국내 자동차’와 기업명을 검색했을 시 최신순으로 업데이트되는 뉴스 기사를 2025년 1월 1일부터 현재시점까지 실시간으로 크롤링 * API endpoint를 호출하여 requests 기반 크롤링, newspaper3k 패키지를 이용하여 본문 수집  1. **데이터 수집 목적**  * **사업 보고서** * 기업 개요, 사업 개요, 재무제표 정보를 분석하여 기업 건전성 및 주요 사업 내용 파악 * 재무제표 정보 시각화 * 재무제표 정보를 기반으로 한 기업 관련 추이를 추론 * **실시간 뉴스** * 자동차 산업 관련 실시간 트렌드와 기업 관련 이슈를 분석하여 긍/부정 전망 추론 * 산업 및 기업 관련 키워드로 시각화 자료 제공 |
| --- | --- |
| **데이터 자동화 및 검증** | * **실시간 뉴스 데이터 수집 자동화 프로세스 (설계 중)**  1. 8개의 언론사를 정하여 해당 언론사들의 뉴스 데이터를 수집. 기사 제목(title), 언론사(newspaper), 기사 작성일(pub\_date), URL(link) 수집. URL을 기준으로 하여 중복되는 기사는 미수집. 2. 1번에서 저장된 URL(link)을 통해 상세페이지 접속 후 기사 본문(content) 수집 3. 첫 번째 데이터베이스(news\_data)에 저장 4. 수집된 기사 본문(content)을 로드하여 기사 특성 키워드 추출(keyword\_20)을 포함한 전처리 수행 5. 특성 키워드(keyword\_20) 컬럼이 추가된 두 번째 데이터베이스(cleaned\_news\_data)에 저장 |
| **데이터 저장 및 관리** | MySQL과 HeidiSQL 툴을 이용하여 데이터베이스 관리 중임.   * **사업 보고서 데이터 저장 방식** * 사업보고서: ChromaDB * 재무제표: MySQL * **재무회계 개념체계 데이터 저장 방식** * MySQL * **실시간 뉴스 데이터 저장 방식** * MySQL |
| **데이터 전처리 과정** | * **사업보고서 데이터 전처리 단계 및 방법** * 사용 도구   : python, pandas, xml.etree(XML parsing), json(JSON 변환), csv(저장), zipfile(압축 해제)   * 데이터 추출 방식  1. API 요청으로 압축파일(.zip) 다운로드 2. 압축파일(.zip) 해제 후 XML 파일 추출 및 확인 3. XML 데이터에서 불필요한 태그를 정규표현식을 통해 제거 후 태그 전체 제거 - CSV 4. XML 데이터를 로마 숫자와 테이블 형태를 기준으로 JSON으로 변환  * 불필요한 데이터 기준(제거)   : 중복 데이터, 사용하지 않는 컬럼, 공백   * 정제 방법  1. 기업 고유번호(8자리)와 종목코드(6자리) 데이터를 문자열로 변환 후 생략된 0개수만큼 다시 추가 2. Y, K, N, E 데이터를 한글로 변경하여 법인을 구분 3. 결측치 데이터는 ‘미제공’으로 변경 4. 섹션별 데이터 병합(ex. 회사 개요, 사업 개요, 주요 서비스)  * **실시간 뉴스 데이터 전처리 단계 및 방법** * 사용 도구   : python, BeautifulSoup, pandas   * 전처리 단계  1. ‘기사 제목, 언론사, 기사 작성일, 기사 본문, URL’이 수집된 데이터베이스(news\_data)에서 데이터 로드 2. 날짜 형식 변환 - datetime 형식으로 변환 3. 기사 작성일(pub\_date), 언론사(newspaper), 기사 제목(title), 기사 본문(content), URL(link) 중 하나라도 결측값이 있는 행 제거 4. 텍스트 전처리 - 기사 제목(title), 기사 본문(content)에서 불필요한 기호, 특수문자, 연속된 공백, 이모지 제거 5. 기사 본문(content)에서 특성 추출하여 가중치순 상위 20개 키워드(keyword\_20) 추출 |
| **데이터 전처리 결과** | * **사업보고서 데이터 전처리 결과 및 향후 사용계획** * 기업 고유번호 파일명(XML → JSON)   : 기업 고유번호.xml, 기업   * 기업 고유번호 파일명(XML → CSV)   : 고유번호.json, 기업 고유번호.csv   * 기업개황 파일명   : cleaned\_all\_company.csv   * 임원현황 파일명   : 임원현황\_cleaned.csv   * 직원현황 파일명   : 직원현황\_cleaned.csv   * 최대주주현황 파일명   : 최대주주현황.csv   * 향후 사용계획  1. 기업 고유번호와 개황 데이터를 활용하여 상세 기업 리포트 작성 및 데이터 시각화(Google Gemini 2.0 Flash API를 통해 요약 데이터를 HTML 보고서로 변환) 2. 투자 분석 및 내부 보고서에 적용  * **실시간 뉴스 데이터 전처리 결과 및 향후 사용계획** * 뉴스 데이터 전처리는 진행 중에 있음. 그러나, 데이터 전처리 후 MySQL의 두 번째 뉴스 데이터베이스(cleaned\_news\_data)에 저장할 예정임 * 향후 사용계획  1. 전처리된 데이터를 활용하여 자동차 산업 트렌드 분석 2. 머신러닝 모델을 활용한 자동 키워드 분류 진행 3. BERT Embedding을 기반으로 한 기업 긍/부정 전망 추론 |