**SK네트웍스 Family AI과정 6기  
 데이터 전처리 인공지능 데이터 전처리 결과서**

**□ 개요**

* 산출물 단계 : 데이터 전처리
* 평가 산출물 : 인공지능 데이터 전처리 결과서
* 제출 일자 : 2025-02-07
* 깃허브 경로 : <https://github.com/SKNETWORKS-FAMILY-AICAMP/SKN06-FINAL-1TEAM>
* 작성 팀원 : 김지영, 노원재, 박창규(팀장), 백하은

**□ 내용**

* 뉴스 / 사업보고서 / 재무제표 용어 및 개념체계 데이터를 각각 분담하여 진행하였습니다.
* **(사업보고서)**

|  |  |
| --- | --- |
| **데이터 전처리 개요** | **문서유형**:   * API를 통해 수집한 사업보고서 공시 서류 데이터 (XML 형식)   **데이터 수집일자**:   * 2025.02.03   **데이터 양**:   * 2,780개의 XML 파일 * 기업개황 CSV 파일 * XML 데이터 총 레코드 수: 20,000 |
| **전처리 과정** | **전처리 도구**:   * Python, pandas, xml.etree(XML 파싱), json(JSON 변환), csv(CSV 저장), zipfile(압축 해제)   **데이터 추출 방식**:   1. API 요청으로 ZIP 파일 다운로드 2. ZIP 파일 해제 후 XML 파일 추출 및 확인 3. XML 데이터에서 불필요한 태그를 정규표현식으로 제거 후 태그 전체 제거 - CSV 4. XML 데이터를 로마 숫자와 테이블 형태를 기준으로 JSON으로 변환   **불필요한 데이터 제거 기준**:   * 중복 데이터 제거 * 사용하지 않는 컬럼 제거   **정제 방법**:   * **기업 고유번호(8자리), 종목코드(6자리)**: 문자열로 변환 후 생략된 0 개수만큼 다시 추가 * **법인 구분**: Y, K, N, E 데이터를 한글로 변경 * **결측치 데이터**: '미제공'으로 변경 |
| **데이터 전처리 결과** | * **결과**:   + 기업 고유번호(XML → JSON): 기업 고유번호.xml, 기업   + 기업 고유번호(XML → CSV)   + 고유번호.json, 기업 고유번호.csv   + 기업개황: cleaned\_all\_company.csv |

* **(재무제표)**

|  |  |
| --- | --- |
| **데이터 전처리 개요** | **문서유형:**   * API를 통해 수집한 기업의 재무제표   **데이터 수집일자:**   * 2025.02.03   **데이터 양:**   * 단일재무제표CFS csv : 약 32,000개 관측치. 11개 변수 * 단일재무제표OFS csv : 약 40,000개 관측치, 11개 변수수 |
| **전처리 과정** | **전처리 도구:**   * Python, csv라이브러리 활용   **데이터 추출 방식:**   1. CSV 파일 로드 및 데이터 리스트 생성 2. 필요 없는 컬럼 제거 3. 데이터 정제   **불필요한 데이터 제거 기준:**   * 분석 목적과 무관한 변수 삭제   **정제 방법:**   * 기업 고유번호(8자리), : 문자열 변환 후 형식 통일법인 구분: 코드 값을 한글로 변경 * 연도 구분을 위한 컬럼명 변경 |
| **데이터 전처리 결과** | * **결과:** * 컬럼명이 추가된 정제된 CSV 파일 생성 * 불필요한 변수가 제거되어 분석 및 시각화 작업에 적합한 형태로 변환됨 |

* **(최신 뉴스)**

|  |  |
| --- | --- |
| **데이터 전처리 개요** | **문서유형**:   * BIGKINDS에서 제공하는 최근 3개월 자동차 관련 기사(2024/11/03-2025/02/03) (xlsx 형식) * 네이버 자동차 관련 최신 뉴스(2025-02-07) (JSON 형식)     **데이터 수집일자**:   * 2025.02.03   **데이터 양**:   * BIGKINDS 기사 31,217개 * 네이버 뉴스 실시간으로 업데이트 되는 기사 |
| **전처리 과정** | **전처리 도구**:   * Python * BeautifulSoup * 자연어 처리 라이브러리 (KoNLpy, Mecab) * BERT   **데이터 추출 방식**:   * Naver에서 크롤링을 통한 뉴스 수집 * API 엔드포인트 호출하여 requests 기반 네이버 뉴스 크롤링   **불필요한 데이터 제거 기준**:   * 중복 기사(동일 제목 및 본문) 제거 * 결측값이 포함된 데이터(기사 제목 또는 본문이 없는 경우) 제거 * 기사 본문이 100자 미만인 경우 제거   **정제 방법**:   * 크롤링 과정에서 포함된 HTML 태그 및 특수 문자 제거 * 불필요한 공백 제거(연속된 공백을 단일 공백으로 변환) * 특수문자 및 숫자 제거(기사 내용에서 분석에 불필요한 문자 제거) * 한국어 자연어 처리를 위해 Mecab 또는 KoNLPy 사용하여 형태소 분석 및 불용어 제거 * WordPiece Tokenizer를 사용하여 문장을 BERT 토큰으로 변환 * BERT Embedding을 생성하여 문서 간 유사도 분석 수행 * Attention 메커니즘을 활용하여 기사 내 핵심 키워드 추출 * 문맥 기반 불용어 제거 및 중요 단어 강조 |
| **데이터 전처리 결과** | **결과**:   * 아직 데이터 수집 과정에 있어 상세 결과 작성 불 * 총 X개의 기사 중 Y개의 유효한 기사 데이터 확보 * 중복 기사 Z개 제거 * 결측치 처리 후 유효 데이터 비율 A% 확보 |