

**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.
* **박창규(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 * **향후 사용계획**:   + 기업 고유번호와 개황 데이터를 활용하여 상세 기업 리포트 작성 및 데이터 시각화   + 투자 분석 및 내부 보고서에 적용 |

* **노원재 (사업보고서, 재무제표 용어/개념체계)**

| **데이터 전처리 개요** | **문서유형**:   * OpenDART에서 제공하는 재무제표 데이터 (txt 파일 형식) * OpenDART에서 제공하는 사업보고서 원본 파일(Api 호출, xml 형식) * 재무제표 용어 (PDF 형식)   **데이터 수집일자**:   * 2025-02-03   **데이터 양**:   * 약 2,800개 기업의 재무제표 및 사업보고서 데이터 * 157개의 재무제표 용어 및 개념체계 해설 데이터 |
| --- | --- |
| **전처리 과정** | **전처리 도구**:   * Python (pandas, re, google generative AI API 등)   **데이터 추출 방식**:   * OpenDART에서 txt 파일로 데이터 다운로드 * Python을 활용하여 txt 데이터를 csv 파일로 변환   **불필요한 데이터 제거 기준**:   * 회사 개요 및 사업 개요에 포함되지 않은 섹션 필터링 * 불완전한 데이터(누락된 값) 제거   **정제 방법**:   * 텍스트 정규화 (HTML 태그, 공백 제거) * 결측치 0으로 대체체 * 섹션별 데이터 병합 (예: 회사 개요, 사업 개요, 주요 서비스) * API 호출 전 데이터 길이 제한 |
| **데이터 전처리 결과** | **결과**:   * 기업 개요 및 사업 개요 정보를 포함한 요약 csv 파일 생성 * 주요 항목(회사명, 사업 내용, 주요 서비스)을 추출하여 가독성 있는 요약 데이터 생성   **향후 사용계획**:   * Google Gemini 2.0 Flash API를 통해 요약 데이터를 HTML 보고서로 변환 * 요약 데이터를 활용하여 투자 분석, 기업 리포트 작성 등 다양한 활용 가능 |

* **백하은(사업보고서 일부)**

| **데이터 전처리 개요** | **문서유형:**   * API를 통해 수집한 기업 관련 데이터 (CSV 형식)   **데이터 수집일자:**   * 2025.02.03   **데이터 양:**   * 기업개황.csv: 3,827개 관측치, 19개 변수 (987 KB) * 임원현황.csv: 33,955개 관측치, 18개 변수 (10.2 MB) * 직원현황.csv: 9,104개 관측치, 22개 변수 (1.2 MB) * 최대주주현황.csv: 24,519개 관측치, 16개 변수 (3.1 MB) |
| --- | --- |
| **전처리 과정** | **전처리 도구:**   * Python, csv라이브러리 활용   **데이터 추출 방식:**   1. CSV 파일 로드 및 데이터 리스트 생성 2. 필요 없는 컬럼 제거 3. 컬럼명 추가 및 데이터 정제   **불필요한 데이터 제거 기준:**   * 분석 목적과 무관한 변수 삭제   **정제 방법:**   * 기업 고유번호(8자리), 종목코드(6자리): 문자열 변환 후 형식 통일 * 법인 구분: 코드 값을 한글로 변경 * 결측값: '미제공'으로 대체 |
| **데이터 전처리 결과** | * **결과:** * 컬럼명이 추가된 정제된 CSV 파일 생성 * 불필요한 변수가 제거되어 분석 및 시각화 작업에 적합한 형태로 변환됨 * 생성된 파일:  1. 기업개황.csv: 3,827개 관측치, 19개 변수 (987 KB) 2. 임원현황\_cleaned.csv: 3,827개 관측치, 19개 변수 (3.8 MB) 3. 직원현황\_cleaned.csv: 3,827개 관측치, 19개 변수 (755 KB) 4. 최대주주현황.csv: 24,519개 관측치, 16개 변수 (3.1 MB)   **향후 사용계획:**  기업 고유번호와 개황 데이터를 활용하여 상세 기업 리포트 작성 및 데이터 시각화 |

* **김지영 (BIGKINDS뉴스 최신 뉴스, 네이버 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**:   * BIGKINDS에서 엑셀 파일로 제공해주는 뉴스 데이터 직접 다운받기 * 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   **향후 사용계획**:   * 전처리된 데이터를 활용하여 뉴스 트렌드 분석 수행 * 머신러닝 모델을 활용한 자동 키워드 분류 진행 * BERT Embedding을 기반으로 문서 유사도 분석 및 핵심 문장 요약 기능 추가 검토 |