**SK네트웍스 Family AI과정 3기  
 데이터 전처리 인공지능 데이터 전처리 결과서**

**□ 개요**

* 산출물 단계 : 데이터 전처리
* 평가 산출물 : 인공지능 데이터 전처리 결과서
* 제출 일자 : 2024-12-31
* 깃허브 경로 : <https://github.com/SKNETWORKS-FAMILY-AICAMP/SKN03-FINAL-6Team>
* 작성 팀원 : 최연규

|  |  |
| --- | --- |
| **데이터 전처리 개요** | * 문서유형 : pdf * 데이터 수집일자 : 2024-11-30 * 데이터 양 : 자동차 매뉴얼 pdf 1개 |
| **전처리 과정** | * **전처리 도구**: Llama Index, OpenAI * **데이터 추출 방식**:  1. Llama Index로 문서 로드 2. 로드된 텍스트를 256 청크 사이즈로 분할하여 Parquet로 저장 3. OpenAI 모델(gpt-40-mini)을 통해 질문-답변 생성 |
| **데이터 전처리 결과** | * 결과:   문서와 관련된 질문과 답변 샘플 추출     * 향후 사용계획:   작성된 매뉴얼 기반 Q&A챗봇의 평가 데이터 셋으로 활용 |