

**SK네트웍스 Family AI 과정 13기  
 데이터 전처리 인공지능 데이터 전처리 결과서**



| **산출물 단계** | 데이터 전처리 |
| --- | --- |
| **평가 산출물** | 인공지능 데이터 전처리 결과서 |
| **제출 일자** | 2025.08.10 |
| **깃허브 경로** | <https://github.com/SKNETWORKS-FAMILY-AICAMP/SKN13-FINAL-6Team> |
| **작성 팀원** | 김동욱, 안수민 |

1. **문서 개요**

* 프로젝트명: 사내 업무 가이드 챗봇
* 전처리 목적: 문서의 내용 정제를 통한 임베딩 품질 향상
* 문제 정의:  
  RAG 시스템에서 사용하는 벡터 데이터베이스 저장용 데이터셋 구축

1. **데이터셋 개요**

* 데이터 출처 및 수집 방법:  
   한국인터넷진흥원의 사내 내규 문서를 다운로드함. (<https://www.kisa.or.kr/609>)
* 데이터 구성:

pdf형식의 사내 내규 문서 122건이다.

* 원본 데이터 샘플:

| 1\_01\_한국인터넷진흥원 정관(240715)의 일부: 제정, 개정 날짜 |
| --- |
| 텍스트, 스크린샷, 문서, 폰트이(가) 표시된 사진  AI 생성 콘텐츠는 정확하지 않을 수 있습니다. |
| 1\_01\_한국인터넷진흥원 정관(240715)의 일부: 본문 |
| 텍스트, 편지, 폰트, 종이이(가) 표시된 사진  AI 생성 콘텐츠는 정확하지 않을 수 있습니다. |
| 1\_01\_한국인터넷진흥원 정관(240715)의 일부: 부칙 |
| 텍스트, 문서, 폰트, 편지이(가) 표시된 사진  AI 생성 콘텐츠는 정확하지 않을 수 있습니다. |
| 2\_01\_이사회 운영규정(221222)의 일부: 별지 |
| 텍스트, 스크린샷, 폰트, 영수증이(가) 표시된 사진  AI 생성 콘텐츠는 정확하지 않을 수 있습니다. |

1. **전처리 프로세스 개요**

* 전체 흐름도:  
    
   ① 임베딩 옵션 설정 → ② 임베딩
* 전처리 파이프라인 요약:

| **단계** | **목적** | **수행 작업** | **사용 도구/라이브러리** |
| --- | --- | --- | --- |
| 임베딩 옵션 설정 | 청크 사이즈 등의 요소 고려 | 청크 사이즈, 청크 오버랩, 임베딩 모델, 임베딩 벡터 차원 결정 | PyPDF2, RecursiveCharacterTextSplitter |
| 임베딩 | 임베딩 벡터 저장 | 임베딩 벡터 저장 | QdrantClient |

1. **세부 전처리 단계**

4.1 결측치 처리

* 결측치 존재 여부: 없음

4.2 이상치 처리

* 이상치 존재 여부: 없음

4.3 정규화 및 표준화

* 정규화 및 표준화를 진행하지 않음.

4.4 데이터 변환 및 생성

1. **학습/검증 데이터 분리**
2. **전처리 결과 요약 및 평가**
3. **변경 이력**