

**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

* PDF 원문에서 텍스트 추출 후 전체 데이터셋에 대해 결측 여부를 확인함.
* 모든 문서가 완전한 텍스트 데이터를 포함하고 있어 결측치는 발견되지않음.

4.2 이상치 처리

* 문서 특성상 수치형 데이터 같은 이상치 개념은 해당 사항이 없음.
* 텍스트 추출 과정에서 별지, 개정이력, 특수문자 등 검색 품질에 영향을 줄 수 있는 텍스트는 확인 후 제거하였음.

4.3 정규화 및 표준화

* nlpai-lab/KoE5 모델은 사전학습된 언어모델로 텍스트 전처리 시 별도의 정규화/표준화가 필요하지 않음.

4.4 데이터 변환 및 생성

* 텍스트 추출 : PyPDF2를 이용하여 PDF문서의 본문 텍스트를 추출하고, 불필요한 특수문자, 개정이력, 별지 제거
* 청크 분할
  + RecursiveCharacterTextSplitter를 사용하여 문단->문장->단어 순으로 분리
  + Chunk\_size : 1000 토큰
  + Chunk\_overlap : 200 토큰
* 임베딩 :
  + nlpai-lab/KoE5를 사용하여 chunk를 1024차원 벡터로 변환
  + 변환된 벡터는 Qdrant 에 저장
* 메타데이터 생성
  + 각 Chunk에서 문서명, 페이지 번호, 카테고리 등의 메타데이터를 부착
  + 검색 시 필터링 조건으로 활용 가능
* 점검 사항
  + 검색 풀질 저하를 유발하는 불필요한 청크는 재검토 후 삭제 요망.

1. **학습/검증 데이터 분리**
   * 본 데이터는 텍스트 임베딩을 통한 검색 용도로만 사용되며, 모델 학습이 목적이 아니므로 별도의 학습/검증 데이터 분리를 수행하지 않음.
   * 데이터 전처리 후 전체 Chunk를 벡터DB(Qdrant)에 저장하여 RAG 시스템 검색에 바로 활용.
2. **전처리 결과 요약 및 평가**
   * 데이터정제 결과 :
     + 결측치, 이상치 없음.
     + 불필요한 목차, 별지 제거
   * 임베딩 파라미터
     + 청크사이즈 : 1000 토큰
     + 청크 오버랩 : 200 토큰
     + 모델 : nlpai-lab/KoE5 (1024차원)
   * 저장소 : Qdrant 벡터 데이터베이스에 저장
3. **변경 이력**

* 없음.