**데이터 수집 및 전처리 정의서 항목**

**1. 문서 개요**

* 목적: RAG 기반 QA 챗봇 구축을 위한 데이터 수집 및 전처리 절차 명세
* 범위: 내부 규정, 인사 정보, 조직도

**2. 원천 데이터 개요**

| **문서명** | **형식** | **내용 요약** | **주요 특성** |
| --- | --- | --- | --- |
| DM\_rules.docx | Word | 사내 규정  ( 8개 챕터로 구성 ) | 텍스트 + 표 + 숫자 혼합형 |
| HR\_information.xlsx | Excel | 인사 정보  ( 직원별 정보 ) | 문장 기반 시트 구성 |
| org\_chart.docx | Word | 조직도 문서  (상하위 구조 포함) | 특수문자/도식 포함, 구조화 필요 |
|  |  |  |  |

**3. 데이터 수집 방식**

* 내부 문서: 수작업 수집 (Word/Excel)

**4. 전처리 절차 정의**

**1) 사내 규정 데이터 (Word/Excel)**

| **단계** | **설명** | **사용 기술** |
| --- | --- | --- |
| 1 | 문서 내 목차, 불필요한 행 제거 | python-docx, pandas |
| 2 | 조문 단위 텍스트 추출 | 정규표현식 + 제목/내용 구분 |
| 3 | 표/숫자 항목 별도 추출 | 표 → 텍스트, 단위 추출 |
| 4 | 청크 작업 | 조문 단위로 청크 설정값 / 결과 |
| 5 | 청크 보완작업 | 슬라이딩 윈도우방식 설정값/결과 |
| 6 | 특수기호 제거 및 한글화 정리 | re.sub, textnormalizer |

**2) 조직도 데이터**

* 계층 구조 파싱 → 상하위 노드로 분리
* └, ├, ─등의 특수문자 제거

**5. 청크 전략**

**1) 작업과정**

|  |  |  |
| --- | --- | --- |
| **대상 문서** | **전략 구분** | **설명** |
| DM\_rules.docx | **기본 청크 ( 기본 조문 단위 청크 )** | 제n조 단위로 분리 |
| DM\_rules.docx | **청크 보완 ( 슬라이딩 윈도우 청크 )** | 일정 토큰 길이만큼 겹치며 생성  (stride 포함) |
| DM\_rules.docx  (표/숫자) | **정량 정보 청크** | 테이블 기반 정보만 별도로 추출하여 별도 청크 구성 |
| HR\_information.xlsx | **문장 단위 청크** | 엑셀 시트를 문장 단위로 파싱, 항목별 정보 묶음 처리 |
| org\_chart.docx | **계층 구조 청크** | 상위 → 하위 구조를 기반으로 문장화하여 청크 생성  (예: CEO → 팀장 → 사원) |
| org\_chart.docx | **특수기호 정제 청크** | 도식/구조 기호( └, ├, ─) 제거 후 계층 기반 텍스트화 처리 |

**2) 청크 결과**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 총 청크 수 | 평균길이 | 최대길이 | 최소길이 |
| 기본 청크 | 42 | 38 | 515 | 8 |
| 청크 보완 | 88 | 303 | 527 | 180 |

**6. 임베딩 설정**

* 모델: nlpai-lab/kure-v1 (Ko-SBERT 기반)
* 방식: 텍스트 normalization → 슬라이딩 청크 → 벡터화
* 저장: FAISS DB (2개)
  + faiss\_win: 사내 규정
  + faiss\_org\_hr: 인사 정보

**7. 검증 질문지 생성**

* LLM: GPT-4o-mini 사용
* 질문 유형:
  + 사내 규정 : 일반 질문 (정책 이해) / 정량 질문 (숫자/표 기반)\
  + 인사 정보 :
* 생성 수:
  + 규정 문서: 챕터별 15~20개 (총 137개)
  + 인사 정보: 직원별 메타데이터 기반 속성 질문 (총 60개 )

**8. 최종 검증 및 평가**

* 사용 지표: F1 / EM (Exact Match)
* 내부 규정
  + 청크보완방식 : 슬라이딩 윈도우
  + 임베딩모델 : KURE-V1
  + 리랭커모델 : BGE-RERANKER-KO
  + 최종 구성 : 리랭커(topk:3)+GPT-4o-mini+금지형프롬프트
  + 평가 점수 : F1/EM score : 55.04/10.61
* 인사 정보
  + 임베딩모델 : KURE-V1
  + 리랭커모델 : BGE-RERANKER-KO
  + 최종 구성 : 리랭커(topk:3)+GPT-4o-mini
  + 평가 점수 : 48.09/38.30:

**9. 산출물 정리**

**1) 임베딩 코드**

**- embedding\_win.ipynb : 내부규정 임베딩 코드**

**- embedding\_hr.ipynb : 인사정보 임베딩 코드**

**2) 청크 파일**

**- dm\_chunks\_window.json : 내부규정 청크보완 파일**

**- combined\_chunks.json : 인사정보 청크 파일**

**3) 질문지 파일**

**- eval\_questions\_window.jsonl : 내부규정 (청크보완) 질문지**

**- eval\_questions\_gpt.jsonl : 인사정보 질문지지**

**4) 테스트 파일**

**- test\_win.ipynb : 내부 규정 평가 코드**

**- test\_hr.ipynb : 인사 정보 평가 코드**

**- rag\_eval.py : 테스트 함수**

**5) 벡터DB**

**- faiss\_win : 내부규정 벡터DB폴더**

**- faiss\_org\_hr : 인사정보 벡터DB 폴더**

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |