SK네트웍스 **Family AI**과정 **3**기

데이터 수집 및 저장 데이터 수집 보고서

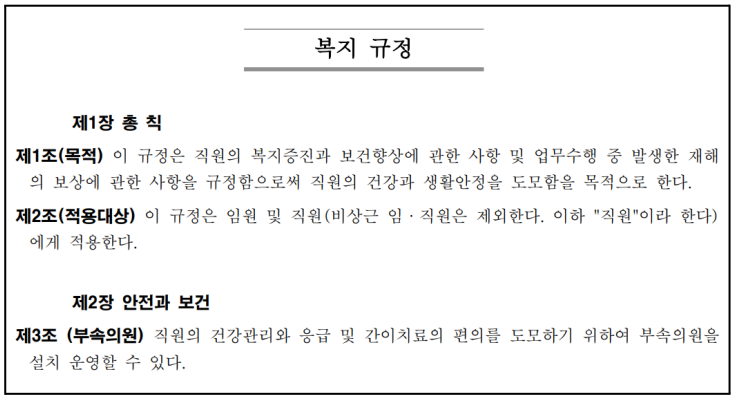
☐ 개요

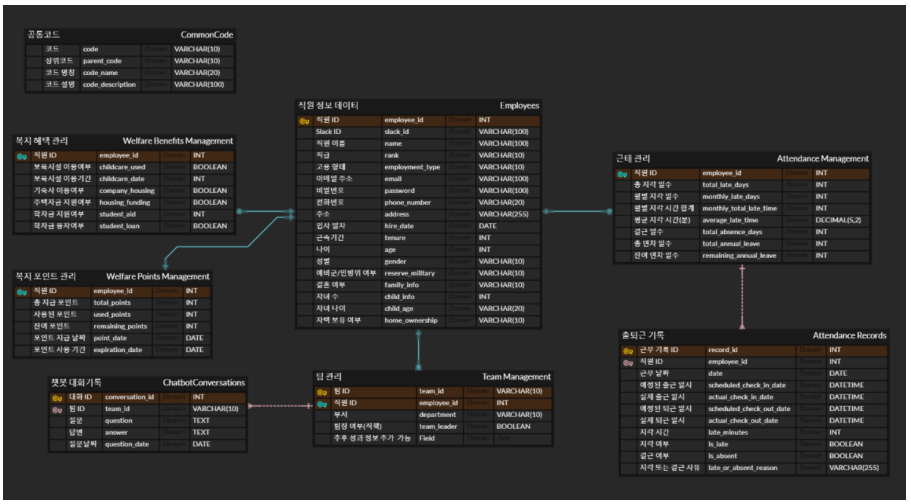
- 산출물 단계 : 데이터 수집 및 저장
- 평가 산출물 : 데이터 수집 보고서
- 제출 일자 : 2024.11.11
- 깃허브 경로 : https://github.com/SKNETWORKS-FAMILY-AICAMP/SKN03-FINAL-3Team
- 작성 팀원 : 송영빈, 송명신

데이터 수집 목적	• 사내 sLLM 시스템 개발을 위한 정보 수집 • HR 관련 질의응답, 정책 안내, 복지 정보 제공 등을 자동화 함으로써 업무 효율성을 높이고 직원 만족도를 개선

데이터 수집 방법	**1.** 웹 데이터 수집 ● 노동 관련 규정 및 정책 문서**:** 노동OK 홈페이지 자료실 (출처: https://www.nodong.kr) ● 활용 목적**:** 규정 데이터를 전처리 후 sLLM 모델이 학습할 수 있는 FAQ 데이터를 제작함으로써 정확한 답변을 제공하는 시스템 구축 **2. HR** 데이터베이스 자체 생성 ● 특정 사내 **HR** 데이터베이스 접근이 불가능해 자체 생성 ● 데이터 소스**(**참고**):** ○ **kaggle HR Analytics Dataset** ○ **IBM HR Analytics Dataset** ● 사용도구 및 라이브러리: Python (pandas, random, datetime, faker) ● 컬럼 별로 특정 로직을 적용하여 자체 생성**:** 이름, 이메일, 주소 등 faker 혹은 random 라이브러리 활용하여 약 700명 정도의 HR 직원 정보 자체 생성 ● 활용 목적**:** ○ HRDB를 참조하여 직원의 개별적인 상황(예: 근태 관리,사용 가능한 복지 포인트 등)에 맞는 개인화된 답변 생성 ○ 데이터를 분석 및 활용하여 HR 정책 개선이나 직원 만족도 향상을 위한 의사결정에 기여하는 관리자 대시보드 개발 **3. sLLM** 학습용 **FAQ** 데이터 ● 데이터 소스: 전처리된 규정 데이터 활용 ● 생성 모델: OpenAI GPT 기반 생성형 AI ● 질문 생성 방법: 1. 규정 문서 조항 별로 구분 2. 생성형 AI 모델 활용하여 조항과 관련된 질문, 카테고리 생성 3. EXCEL 형식으로 데이터 저장

	• 답변 생성 방법: 1. 규정 텍스트 추출 2. 텍스트 임베딩 생성 3. FAISS 인덱스 생성 및 임베딩 추가 4. 질문(EXCEL) 데이터 읽기 5. FAISS를 사용해 유사한 규정 검색 6. 검색된 규정을 기반으로 Open AI API 사용해 답변 생성 7. 질문, 답변 데이터 EXCEL 형식으로 제작 8. EXCEL → JSONL 형식 변환환 • 활용 목적: ○ 효율적인 모델 학습 데이터 제공 ○ 사용자 질문에 최적화된 정보 제공 ○ 신뢰정과 일관성 있는 답변 생성

수집 데이터	**1.** 웹 데이터 **(**노동 관련 규정 및 정책 문서**)** ● 원본 데이터 예시**:**  복지 규정 **제1장 총 칙** **제1조(목적)** 이 규정은 직원의 복지증진과 보건향상에 관한 사항 및 업무수행 중 발생한 재해의 보상에 관한 사항을 규정함으로써 직원의 건강과 생활안정을 도모함을 목적으로 한다. **제2조(적용대상)** 이 규정은 임원 및 직원(비상근 임 · 직원은 제외한다. 이하 "직원"이라 한다)에게 적용한다. **제2장 안전과 보건** **제3조 (부속의원)** 직원의 건강관리와 응급 및 간이치료의 편의를 도모하기 위하여 부속의원을 설치 운영할 수 있다. ● 데이터 유형**:** 비정형 데이터 (PDF문서) → 정형 데이터 (JSONL) ● 데이터 구조**:** 장, 조, 주제, 내용 구분된 JSONL 형식 파일 (ex: 제1장/제1조/주제/내용) ● 특징**:** ○ 특정 혜택 세부내용 추가 (ex: 학자금 신청시 필요한 문서) ○ 비정형 데이터를 텍스트로 변환 및 분할 ○ jsonl 형태로 저장 ● 데이터 예시**:** `{"장": "1", "조": "1", "주제": "목적", "내용": "이규정은직원의복지증진과보건향상에관한사항및업무수행중발생한재해의보상에관한사항을규정함으로써직원의건강과생활안정을도모함을목적으로한다."}` **2.** 자체 생성된 **HR** 데이터베이스 ● 데이터 유형**:** 정형 데이터(CSV) ● 데이터 구조**:** ERD 참고 (https://www.erdcloud.com/d/2WqHTgjaCkqpH44iE)



- 특징:
  - Python 코드 기반으로 데이터 생성
  - 근태, 복지 관리 등 데이터 관리를 세분화
  - 실무에서 인사이트를 얻을만한 요소 추가(근태, 복지 등)
  - 공통코드 시스템을 활용하여 유연한확장성과 중복 방지

- 데이터 예시:

employee_id	slack_id	name	rank	employment_type
352	351	김성현	RANK06	TYPE01

3. sLLM 학습용 FAQ 데이터

- 데이터 유형: 정형 데이터(JSONL)

- 데이터 구조: source, category, question, answer로 이루어진 JSONL

- 특징:
  - 전처리된 규정 문서를 기반으로 GPT 모델을 사용하여 생성
  - 질문과 답변 생성시 각각 파라미터를 다르게 설정
  - 수작업 수정 및 검토를 통해 데이터 검증
  - JSONL 형태로 구조화

	● 데이터 예시: {**'source'**: '직원의건강관리와응급및간이치료의편의를도모하기위하여의료실을설치운영할수있다. 의료실운영시간: 평일9:00~18:00, 공휴일휴무 의료실위치: 본사1층F동', **'category'**: '공휴일 의료 이용', **'instruction'**: '공휴일에는 의료실을 이용할 수 없나요?', **'response'**: '공휴일에는 의료실을 이용할 수 없습니다. 의료실은 평일 9:00부터 18:00까지 운영되며, 공휴일에는 휴무입니다. 따라서 공휴일에는 의료실을 이용할 수 없음을 참고해 주시기 바랍니다'}