SK네트웍스 **Family AI** 과정 **15**기

데이터 수집 및 저장 데이터 조회 프로그램

산출물 단계	데이터 수집 및 저장
평가 산출물	데이터 조회 프로그램
제출 일자	2025. 10. 17
깃허브 경로	https://github.com/SKNETWORKS-FAMILY-AICAMP/SKN15-FINAL-3TEAM
작성 팀원	조솔찬

원본 데이터 (텍스트)	**[특허 정보 수집]** 1. 자료 출처 : KIPRIS(특허정보, 특허거절결정서,의견제출통지서), aihub(법령), arxiv(논문), 지식 재산처(특허 -실용신안 심사기준) 2. 수집 방식 : API 연동 및 웹 크롤링(Web Crawling) 3. 수집 항목 : KIPIRS CPC A63 특허 정보, 의견제출통지서, 특허거절결정서, 특허-실용신안 심사기준, 관련 논문 4. Output 파일 : 특허정보.csv, 의견제출통지서.pdf, 특허거절결정서.pdf, 특허-실용신안 심사기준.pdf, 논문.pdf
데이터 전처리 과정	**[수집데이터 구조 및 전처리 대상]** 1. **전처리 대상** : KIPIRIS CPC A63특허, 특허 - 실용신안특허 심사기준,의견제출통지서, 법령 2. **이상치 탐색** 논문 : 발행일 오류(미래 시점), DOI 형식 불일치, 동일 논문 중복 수집 특허 정보 : 존재하지 않는 출원번호, 등록/거절 상태 불일치, 출원인명 비정형 패턴 거절 결정서 : 동일 출원번호 내 중복 기록, 심사일자 역전(등록일보다 빠름 입력 누락 3. **처리 방식** **단순 오류** : 입력단 타이포, 포맷 깨짐 → 정규식 변환 및 포맷 교정(OCR, replace rule) **중복 데이터**: 동일 키(ID·제목·출원번호 등) → 유사도·발행일 기준 최신본만 유지 **결측치 발생** : DOI/심사일 등 누락 → 외부 API(학술·특허 DB) 재조회 또는 '결측' 플래그 설정 **모호값(다중기록)** : 동일 키에 상이한 값 존재 → 검증 신뢰도가 높은 출처

	우선 선택(공식 DB → 내부 DB → 사용자 입력 순) **법령** : 전체 법령 중 특허법 데이터만 추출, 추출 후 청구항 판단 관련 조문만 정제 **전처리 된 최종 파일 : reject_documents.csv, unique_links.csv, with_application_number.csv, 심사기준.csv, 특허법(청구항).csv**
DB 사용 용도	**[DB 구축 및 활용 목적]** **1. 사용 목적** **1-1. 특허 정보 관리** 특허 출원번호, 제목, 법적 상태, 분류코드 등 특허 관련 메타데이터를 체계적으로 저장. 거절결정서 및 거절사유 데이터를 연동해 **거절 원인 분석** 및 **모델 학습용 데이터셋 생성** **1-2 법률·심사기준 정보 관리** 「특허법」, 「특허심사기준」 등 법률·기준 데이터를 구조화하여 AI가 청구항 판단 시 관련 조항을 **근거로 자동 매칭**할 수 있도록 설계. **1-3 사용자 관리 및 접근 제어** 사용자 계정, 역할(Role), 권한(Permission)을 분리하여 **관리자/연구자/일반 사용자별 접근권한을 제어**. 데이터 무결성과 보안을 동시에 확보. **1-4 질의 응답 및 로그 기록 관리** 사용자 세션, 질의(query), 챗봇 응답, 검색기록 등을 모두 로그화하여 **AI 성능평가 및 사용자 행동분석에 활용**.

2. 활용 구조

구분	주요 엔터티	주요 기능 및 역할
사용자 관리	users, roles, permissions, user_role_map, role_permission_map	사용자 계정, 역할(Role), 권한(Permission) 관리 및 접근 제어 기능 제공
특허 데이터	patents, rejection_decisions, rejection_reasons	특허 기본정보, 거절결정 및 거절사유 저장 및 분석. AI 학습용 데이터셋 생성 기반
법률 및 심사기준	patent_laws, examination_criteria	특허법, 심사기준 등의 법률적 근거를 체계적으로 저장하고 특허 거절 판단 시 근거 데이터로 활용
질의/세션 관리	user_sessions, user_queries	사용자 질의 및 세션 이력 관리. 챗봇 및 질의응답 로그 기록 저장
로그 및 기타	chatbot_logs, keyword_search_logs, admin_page_logs, lunch_suggestions	검색·챗봇·관리자 활동 로그 관리 및 운영 데이터 분석용