SK네트웍스 **Family AI**과정 **15**기

데이터 수집 및 저장 수집 데이터 보고서

산출물 단계	데이터 수집 및 저장
평가 산출물	수집 데이터 보고서
제출 일자	2025.10.02
깃허브 경로	https://github.com/SKNETWORKS-FAMILY-AICAMP/SKN15-FINAL-3TEAM.git
작성 팀원	김주형, 이소정

1. 특허 데이터

1.1 수집 데이터 개요

데이터명	수집 대상	수집 목적	사용 예정 기능	출처/저작권
특허 데이터	특허청에 올라온 특허 정보	RAG 검색 시 사용될 데이터 수집, 거절된 특허의 경우 거절 원인 패턴 학습	키워드 기반 특허 검색, 아이디어 기반 유사 특허 검색	kipris

1.2 수집 방법 및 자동화 절차

- 수집 방식 (해당 항목에 체크)
  - ‘특허로’에서 엑셀 파일로 다운로드 가능
- 수집 도구 또는 스크립트 설명:
  - 사용한 언어/라이브러리: Python
  - 자동화 여부 및 주기: 일회성
- 예시 스크립트 또는 흐름도 첨부: (이미지, 순서도 또는 코드)

  ①’특허로’에서 데이터 다운로드 → ②파일 병합 및 ‘특허번호’ 기준으로 중복 제거

1.3 파일 및 필드 설명

파일명 또는 테이블명	필드명	데이터 타입	설명	예시
patents.csv	invention_title	string	발명의 명칭	“물리적 영역에서 고객과 상호작용하는 디지털 캐릭터”
patents.csv	invention_title_eng	string	발명의 명칭(영문)	“digital character interacting with customer in physical realm”
patents.csv	ipc_class	string	IPC 분류	“A63F 13/67(2014.01)”
patents.csv	cpc_class	string	CPC 분류	“A63F 13/67(2014.09)”
patents.csv	application_number	string	출원번호	“1020227000900”
patents.csv	application_date	datetime	출원일자	2025-03-15
patents.csv	applicant	string	출원인	“유니버셜 시티 스튜디오스 엘엘씨”
patents.csv	translation_date	datetime	번역문 제출일자	2025-04-01
patents.csv	registration_number	string	등록번호	“1021650760000”
patents.csv	registration_date	datetime	등록일자	2026-01-10
patents.csv	publication_number	string	공개번호	“1020220019282”
patents.csv	publication_date	datetime	공개일자	2025-09-20
patents.csv	announcement_number	string	공고번호	“1020220019282”
patents.csv	announcement_date	datetime	공고일자	2025-11-01
patents.csv	pct_application_num	string	국제출원번호	“PCT/US2020/037314”

patents.csv	pct_application_date	datetime	국제출원일자	2025-05-20
patents.csv	pct_publication_num	string	국제공개번호	“WO2020252210”
patents.csv	pct_publication_date	datetime	국제공개일자	2025-07-01
patents.csv	priority_info	string	우선권정보	“62/860,188 (2019.06.11) 미국US|15/931,377 (2020.05.13) 미국US”
patents.csv	legal_status	string	법적상태	“등록”
patents.csv	examination_status	string	심사진행상태	“등록결정(일반)”
patents.csv	original_app_num	string	원출원번호	“1020157012011”
patents.csv	original_app_date	datetime	원출원일자	2024-12-15
patents.csv	related_app_num	string	관련출원번호	“1020157012011”
patents.csv	examination_request	string	심사청구여부(일자)	Y(2023.06.12)
patents.csv	claim_count	int	심사청구항수	15
patents.csv	abstract	string	요약	“디스플레이 장치에 묘사되는…”
patents.csv	inventor	string	발명자	“사이알리 사라”
patents.csv	agent	string	대리인	“제일특허법인”
patents.csv	assignee	string	최종권리자	“(주) 엑스골프”
patents.csv	designated_country	string	지정국	“유럽특허청”
patents.csv	citation_count	int	피인용 횟수	24
patents.csv	claims	text	청구항	“[청구항1]디스플레이 장치에서 묘사되는”

1.4 데이터 양

- 전체 수집 데이터 건수: 77292건
- 추출된 고품질 데이터 건수 (필터링 후 기준): 61499건

1.5 저장 위치 및 포맷

- 저장 포맷: CSV
- 인코딩: UTF-8

1.6 데이터 품질 및 정합성 관리 방안

- 중복 제거 기준: 출원번호 기준 중복 제거
- Null 처리 및 결측치 전략: '요약'과 '청구항'에 널이 있는 경우 행 제거

2. 거절 결정서

2.1 수집 데이터 개요

데이터명	수집 대상	수집 목적	사용 예정 기능	출처/저작권
거절결정서 데이터	특허청에서 발송된 거절결정서 문서 (PDF → 텍스트)	거절 사유 패턴 학습 및 모델 판단 근거 확보	모델 파인튜닝, 거절 사유 예측	특허청 공개 문서 기반 / 자체 생성 및 내부 테스트 목적

2.2 방법 및 자동화 절차

- 수집 방식 (해당 항목에 체크)
    - API 호출
    - 문서 파일 처리(로컬/업로드된 PDF)
- 수집 도구 또는 스크립트 설명:
    - 사용한 언어/라이브러리: Python(pdfplumber, json)
    - 자동화 여부 및 주기: 일회성
- 예시 스크립트 또는 흐름도 첨부: (이미지, 순서도 또는 코드)

  ① PDF 수집 → ② 텍스트 레이어 추출(pdfplumber) → ③ JSONL 저장(페이지별 레코드)

2.3 파일 및 필드 설명

파일명 또는 테이블명	필드명	데이터 타입	설명	예시
reject.jsonl	doc_id	string	사용자 질문	2020210001809_952023068810958.pdf
reject.jsonll	page	int	페이지 번호	32

reject.jsonl	text	string	해당 페이지 내용	"발송번호: 9-5-2021-083847504 수신 :\n발송일자: 2021.10.25.\n제출기일\nYOUR INVENTION PARTNER\n특 허 청\n- - - 거절결정서 - - -\n출 원 인 성 명 이준형 (특허고객번호: 420200122423)\n주 소\n대 리 인 성 명\n주 소\n출 원 번 호 20-2020-0000633 \n고 안 의 명 칭 배드민턴 매직 라인\n지정기간 2021.06.20.까지 의견서 또는 보정서 제출이 없었으며 …”

2.4 데이터 양

- 전체 수집 데이터 건수:1122건
- 추출된 고품질 데이터 건수 (필터링 후 기준):1114건

2.5 저장 위치 및 포맷

- 저장 포맷:JSONL
- 인코딩: UTF-8

3. 법령

3.1 수집 데이터 개요

데이터명	수집 대상	수집 목적	사용 예정 기능	출처/저작권
특허 법령	특허법	특허법 위반 사항이 없는지 확인	특허 첨삭	AI Hub

3.2 수집 방법 및 자동화 절차

- 수집 방식 (해당 항목에 체크)
    - AI Hub에서 다운로드
- 수집 도구 또는 스크립트 설명:
    - 사용한 언어/라이브러리: Python
    - 자동화 여부 및 주기: 일회성
- 예시 스크립트 또는 흐름도 첨부: (이미지, 순서도 또는 코드)

  ① AI Hub에서 데이터 다운로드 → ② 데이터 병합

3.3 파일 및 필드 설명

파일명 또는 테이블명	필드명	데이터 타입	설명	예시
law.csv	statute_name	string	법령의 정식 명칭	"특허법·실용신안법·의장법및상표법에의한특허료·등록료와수수료의징수규정"
law.csv	effective_date	datetime	법령이 시행된 날짜	"1979-08-27 00:00:00"

law.csv	proclamation_date	datetime	법령이 공포된 날짜	"1979-08-27 00:00:00"
law.csv	statute_type	string	법령의 형식 (예: 법률, 대통령령, 총리령, 부령 등)	"대통령령"
law.csv	statute_abbrv	string	법령의 약칭	"특허법·실용신안법·의장법및상표법에의한특허료·등록료와수수료의징수규정"
law.csv	statute_category	string	법령이 속하는 분야 카테고리	"행정일반"
law.csv	sentences	string	법령 조문 전체 텍스트	"제1조 (특허료)\n 특허법 제76조의 규정에 의한 특허료는 다음과 같다.\n …"
law.csv	data_class	int	데이터 분류 코드 (예: 라벨링/카테고리용 숫자값)	2

3.4 데이터 양

- 전체 수집 데이터 건수: 45건
- 추출된 고품질 데이터 건수 (필터링 후 기준): 45건

3.5 저장 위치 및 포맷

- 저장 포맷: JSON
- 인코딩: UTF-8

4. 심사 판단 기준

4.1 수집 데이터 개요

데이터명	수집 대상	수집 목적	사용 예정 기능	출처/저작권권
특허 심사 판단 기준 통합 데이터셋	표준 심사기준, 기술별 심사기준	SLLM 학습 및 RAG 검색 시, 심사 판단·거절 사유·법적 근거·최신 기술 트렌드를 종합적으로 판단할 수 있도록 지원	심사 판단 자동화 모델 파인튜닝, 거절 사유 자동 분류 및 근거 매핑, 사용자 질의 시 법령·판례·기술 가이드 근거 제공	특허청 공개 문서

4.2 수집 방법 및 자동화 절차

- 수집 방식 (해당 항목에 체크)
    - 문서 파일 처리(로컬/업로드된 PDF)
- 수집 도구 또는 스크립트 설명:
    - 사용한 언어/라이브러리: Python(pdfplumber, json)
    - 자동화 여부 및 주기:일회성
- 예시 스크립트 또는 흐름도 첨부: (이미지, 순서도 또는 코드)

  ① PDF 수집 → ② 텍스트 레이어 추출(pdfplumber) → ③ JSONL 저장(페이지별 레코드)

4.3 파일 및 필드 설명

파일명 또는 테이블명	필드명	데이터 타입	설명	예시
stand_pages.jsonl	doc_id	string	사용자 질문	stand_part.pdf
stand_pages.jsonl	page	int	페이지 번호	32

stand_pages.jsonl	text	string	해당 페이지 내용	"제1부 인공지능 분야 심사실무가이드\n다만, 거절이유에 대한 출원인 대응의 편의를 …"

4.4 데이터 양

- 전체 수집 데이터 건수: 표준심사기준(844건) + 기술별심사기준(1166건)
- 추출된 고품질 데이터 건수 (필터링 후 기준): 2010건

4.5 저장 위치 및 포맷

- 저장 포맷: JSONL
- 인코딩: UTF-8

5. 관련 논문

5.1 수집 데이터 개요

데이터명	수집 대상	수집 목적	사용 예정 기능	출처/저작권
논문	배경 지식 관련 논문	트랜드 분석	트랜드 시각화	arxiv

5.2 수집 방법 및 자동화 절차

- 수집 방식 (해당 항목에 체크)
    - 웹 크롤링
- 수집 도구 또는 스크립트 설명:
    - 사용한 언어/라이브러리: Python(BeautifulSoup, requests)
    - 자동화 여부 및 주기: 일회성
- 예시 스크립트 또는 흐름도 첨부: (이미지, 순서도 또는 코드)

  ①키워드 기반 url 변경 → ②크롤링 → ③데이터 병합 및 link 기준 중복 제거

5.3 파일 및 필드 설명

파일명 또는 테이블명	필드명	데이터 타입	설명	예시
chat_logs.csv	title	string	논문 제목	“Optimal estimation for regression discontinuity design with binary outcomes”
chat_logs.csv	authors	string	저자	“Takuya Ishihara”
chat_logs.csv	abstract	string	초록	“We develop a finite-sample….”

chat_logs.csv	link	string	논문 링크	“https://arxiv.org/abs/2509.18857”

5.4 데이터 양

- 전체 수집 데이터 건수: 200
- 추출된 고품질 데이터 건수 (필터링 후 기준): 196

5.5 저장 위치 및 포맷

- 저장 포맷: CSV
- 인코딩: UTF-8

5.6 데이터 품질 및 정합성 관리 방안

- 중복 제거 기준: link 기준 중복 제거
- Null 처리 및 결측치 전략: 널 존재 시 해당 행을 삭제