## **1. 문서 개요**

### **1.1 작성 목적**

본 문서는 특허 관련 LLM 학습 및 RAG 기반 DB 검색 시스템 구축을 위해 활용된 법령 및 특허 문서의 전처리 절차와 그 결과를 정리한 자료입니다.

### **1.2 문서 범위**

* 특허 관련 법령 PDF 문서 (공개된 전문)
* 등록 및 공개 특허 명세서 PDF 문서 (API 요청에 따른 명세서)
* 인공지능 분야 심사 실무 가이드 PDF 문서
* 의견 제출 통지서 PDF 문서 (특허 출원 거절 이유)
* 거절결정서 PDF 문서 (특허 출원 거절 결정)

## **2. 원본 데이터 개요**

### **2.1 데이터 유형**

* **법령 데이터**: 특허법 및 시행령 등 관련 조항 (PDF 형식)
* **특허 데이터**: 등록·공개된 특허 명세서 문서 (PDF 형식)
* **인공지능 분야 심사 실무 가이드 데이터:** 특허청 자료 공유실 문서 (PDF 형식)
* **의견 제출 통지서 데이터**: 특허 출원에 대한 거절 이유 통지 문서 (PDF 형식)
* **거절결정서 데이터**: 특허 출원에 대한 최종 거절 결정 문서 (PDF 형식)

### **2.2 수집 경로 및 방식**

* laws/ 국가법령 정보센터(<https://www.law.go.kr/>) 크롤링
* patents/kipris 특허 정보 서비스(<https://plus.kipris.or.kr/portal/main/apiStatus.do?menuNo=210157>) api 사용
* judge\_guide/특허청 자료 공유실 PDF 다운로드(<https://www.kipo.go.kr/ko/kpoContentView.do?menuCd=SCD0200147>)
* opinions/kipris 특허 정보 서비스(<https://plus.kipris.or.kr/portal/main/apiStatus.do?menuNo=210157>) api 사용
* rejects/kipris 특허 정보 서비스(<https://plus.kipris.or.kr/portal/main/apiStatus.do?menuNo=210157>) api 사용
* Python 기반 스크립트를 통해 텍스트 추출 진행

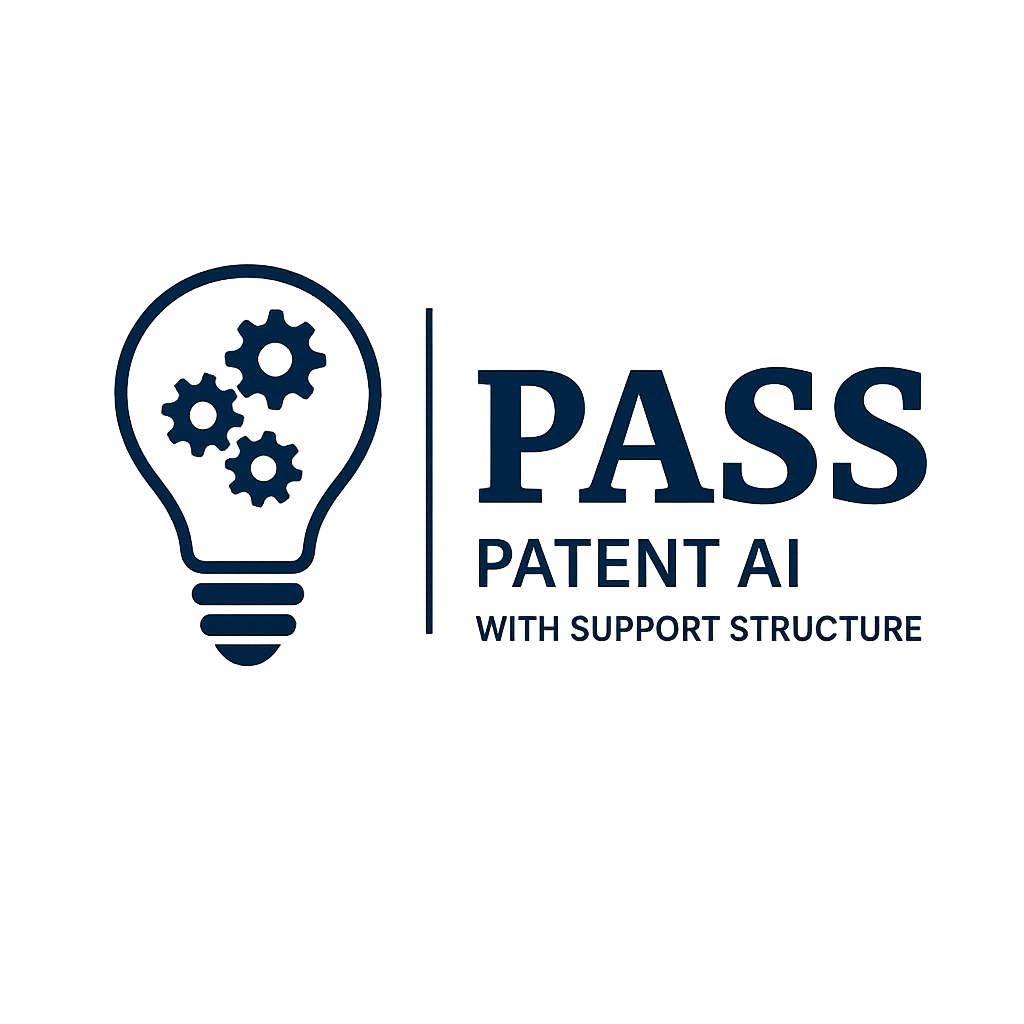
## **3. 전처리 개요**

### **3.1 전처리 목적**

* 법령 데이터는 조문 단위로 분리하여 RAG 검색을 위한 벡터화 목적
* 특허 데이터는 항목별(특허ID, 특허정보, 청구항, 요약, 기술분야 등) 분리하여 RAG 및 fine-tuning 데이터 생성 목적
* 의견 제출 통지서 데이터는 거절 이유별 분석 및 fine-tuning 데이터 생성 목적
* 거절결정서 데이터는 최종 거절 결정의 법적 근거 및 사유를 체계적으로 분석 후 fine-tuning 데이터 생성 목적

### **3.2 전처리 흐름 요약**

1. PDF 텍스트 추출
2. 불필요한 텍스트(헤더/푸터) 제거
3. 조항·항목 구분 및 구조화
4. JSONL 저장 및 검증



## **4. 전처리 세부 내용**

### **4.1 법령 데이터(patent\_laws\_preprocessing.py)**

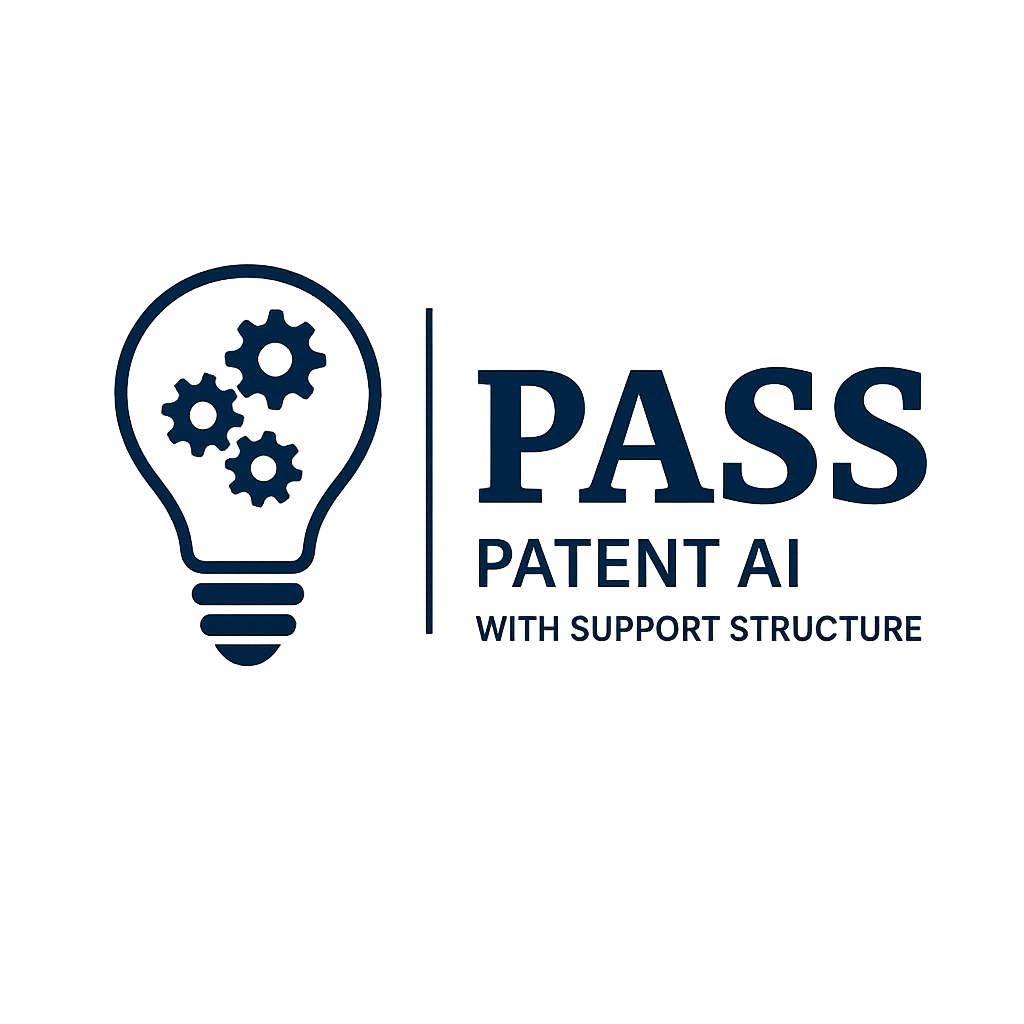
* **PDF 읽기**: pypdf 사용하여 페이지별 텍스트 추출
* **불필요한 정보 제거**: "국가법령정보센터", "법제처" 등 헤더/푸터 텍스트 삭제
* **문장 정제 및 병합**: 조문 단위로 정리하여 input, output 형식으로 구성
* **저장 포맷**: .tsv 중간 저장 후 .jsonl 최종 저장

### **4.2 특허 데이터(patent\_report\_preprocessing.py)**

* **PDF 텍스트 추출**: PyPDF2를 통해 전체 페이지 처리
* **기본 정보 추출**: 출원번호, 등록번호, 발명명 등 정규식 파싱
* **청구항/요약/기술분야 구분**: 패턴 기반 텍스트 분리 및 저장
* **항목 정렬 및 구조화**: section → subsection → paragraph 단위로 구조화
* **저장 포맷**: JSONL 구조로 각 항목 개별 저장

### **4.3 AI 분야 특허 심사 가이드 데이터(ai\_section\_judge\_guide.py)**

* **PDF 텍스트 추출**: fitz 라이브러리를 활용하여 전체 PDF 페이지의 텍스트를 추출함.
* **필요 분야 페이지 추출**: 인공지능 특허 심사 가이드가 포함된 페이지를 선별하여 추출함.
* **불용어 제거**: 개정 정보, 관리 부서 안내, 머리말 및 꼬리글 등 불필요한 텍스트를 제거함.
* **항목 정렬 및 구조화**: 소제목 및 부제목을 기준으로 내용을 분리하여 구조화함.
* **저장 포맷**: 각 항목을 JSONL 포맷으로 개별 저장함.



### **4.4 의견 제출 통지서 데이터(opinion\_report\_preprocessing.py)**

* **PDF를 XML로 변환**: PDF파일 xml 파일로 변환
* **XML 텍스트 추출**: ElementTree 라이브러리를 활용하여 전체 PDF 페이지의 텍스트를 추출함
* **기본 정보 추출**: 출원번호, 등록번호, 발명명 등 정규식 파싱
* **상세 거절 이유 구조화**: 법조항별 세부 거절 사유를 단락 단위로 분리
* **출원 정보 정규화**: 출원번호, 출원일, 발명명 등 기본 정보 표준화
* **저장 포맷**: JSONL 구조로 각 항목 개별 저장

### **4.5 거절결정서 데이터(reject\_report\_preprocessing.py)**

* **PDF 텍스트 추출**: PyPDF2를 통해 전체 페이지 처리
* **거절결정 사유 분석**: 최종 거절 결정의 법적 근거 및 세부 사유 구조화
* **보정 및 의견서 검토 내용 분석**: 출원인의 보정 내용과 심사관의 검토 의견 분리
* **인용발명 정보 추출**: 거절 근거가 된 선행기술 문헌 정보 정리
* **저장 포맷**: JSONL 구조로 각 항목 개별 저장

## **5. 전처리 결과 요약**

| **항목** | **법령 데이터** | **특허 데이터** | **AI분야 특허심사 가이드 데이터** | **의견 제출 통지서 데이터** | **거절결정서 데이터** |
| --- | --- | --- | --- | --- | --- |
| 전처리 대상 | 12개의 PDF 파일 | 약 6만 건 | 1개의 PDF파일 | **2,300개의 JSON 레코드** | **2,300개의 JSON 레코드**ChatGPT_Image_2025_4_29_03_15_03.png |
| 전처리 후 샘플 수 | 6024 rows의 JSONL 파일 | 약 6만 건 | 34 rows의 JSONL 파일 | **약 2,300 rows의 JSONL 파일** | **약 2,300 rows의 JSONL 파일** |
| 구조화 항목 | 장/조/항 | 청구항/기술분야/특허내용 등 | 고유id/source/text/gubun 등 | **출원번호/거절법조항/세부사유** | **출원번호/거절결정사유/법적근거** |
| 출력 형식 | input/output JSONL | section 기반 JSONL | gubun 기반 JSONL | **law\_article 기반 JSONL** | **decision\_reason 기반 JSONL** |

### 

### **5.1 법령 데이터**

| **필드명** | **데이터 타입** | **설명** | **구조** |
| --- | --- | --- | --- |
| input | 문자열 | 조문 제목 및 번호 | 장/조/항 |
| output | 문자열 | 조문 내용 | 해당 법령 전문 텍스트 |

### 

### **5.2 특허 데이터**ChatGPT_Image_2025_4_29_03_15_03.png

| **필드명** | **데이터 타입** | **설명** | **예시** |
| --- | --- | --- | --- |
| patent\_id | 문자열 | 특허 고유 식별자 | 출원번호/등록번호 |
| section | 문자열 | 문자 섹션 | 특허\_기본정보, 청구범위, 발명의 설명, 도면 등 |
| subsection | 문자열 | 섹션 내 하위 구분 | 특허 정보, 청구항, 요약, 기술분야, 배경 기술 등 |
| paragraph | 정수 | 단락 번호 | 1, 2, 3 ... |
| claim\_number | 문자열 | 청구항 번호 | 청구항\_1, 청구항\_2 |
| content | 객체/문자열 | 특허 내용 | 발명의 설명, 청구 범위, 특허 정보 등 |

### 

### **5.3 심사 가이드 데이터**

| **필드명** | **데이터 타입** | **설명** | **예시** |
| --- | --- | --- | --- |
| id | 문자열 | 특허 고유 식별자 | 인공지능분야심사실무가이드\_0001 |
| source | 문자열 | 문자 섹션 | 인공지능분야심사실무가이드 |
| text | 문자열 | 해당 가이드 설명 | 각 경우에 따른 심사 가이드라인ChatGPT_Image_2025_4_29_03_15_03.png |
| gubun | 문자열 | 소제목 | 3.1.2 판단의 구체적인 방법 |

### 

### 

### 

### **5.4 의견 제출 통지서 데이터**

| **필드명** | **데이터 타입** | **설명** | **예시** |
| --- | --- | --- | --- |
| application\_number | 문자열 | 출원번호 | 10-2022-0126528 |
| application\_date | 문자열 | 출원일 | 2022.10.04. |
| invention\_title | 문자열 | 발명의 명칭 | 배터리 셀 상태 판단 장치 및그것의 동작 방법 |
| rejection\_laws | 배열 | 거절 근거 법조항 | ["특허법 제29조제2항", "특허법 제42조제3항제2호"] |
| detailed\_reasons\_by\_law | 객체 | 법조항별 상세 거절 이유 | 법조항을 키로 하는 상세 설명 |

### 

### **5.5 거절결정서 데이터**

| **필드명** | **데이터 타입** | **설명** | **예시**ChatGPT_Image_2025_4_29_03_15_03.png |
| --- | --- | --- | --- |
| 출원번호 | 문자열 | 출원번호 | 10-2022-0175106 |
| 발명의명칭 | 문자열 | 발명의 명칭 | 블록체인 및 인공지능 기반 농산물 저장유통 스마트 유닛로드 시스템 |
| 거절결정의이유 | 문자열 | 거절 결정의 상세 이유 | 진보성 부정, 보정 미흡 등의 구체적 사유 |

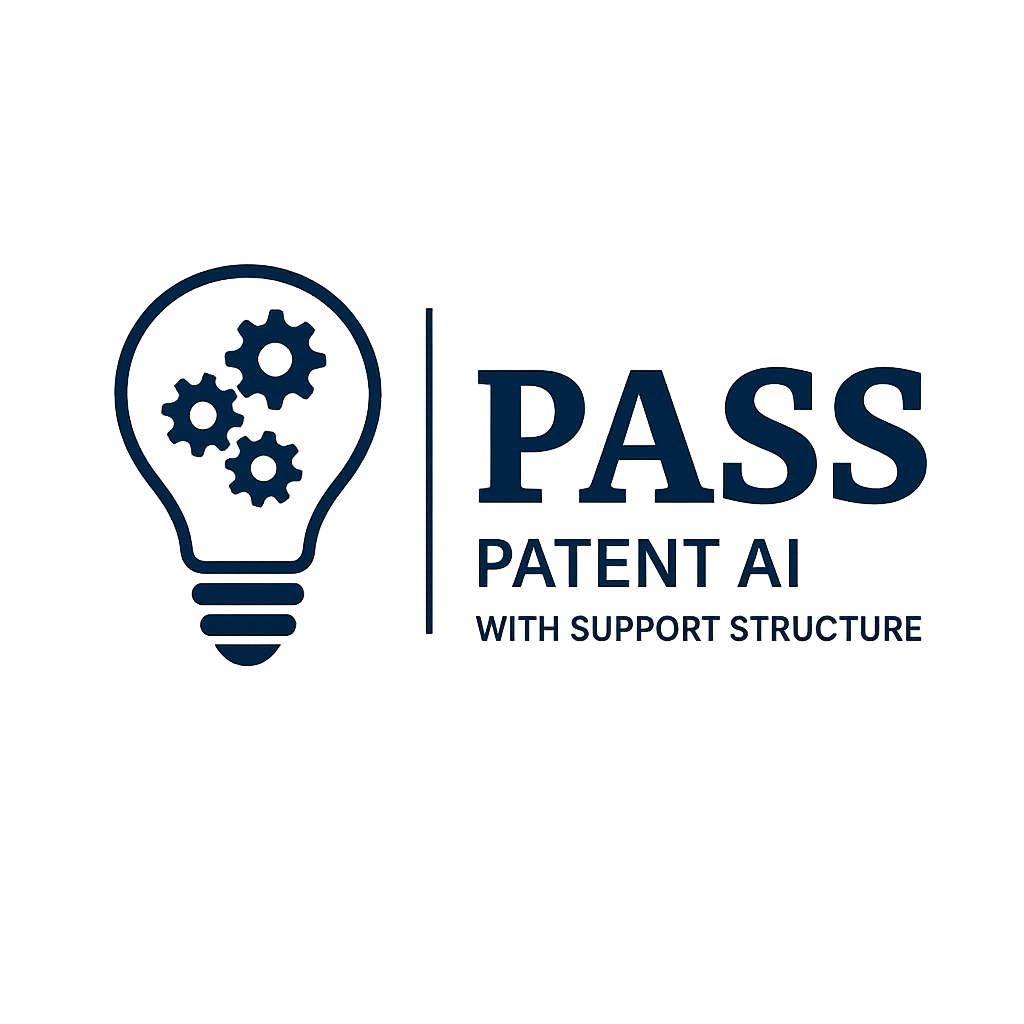
## **6. 품질 검증 및 처리 방식**

### **6.1 주요 오류 유형**

* 조문 번호 인식 오류
* 청구항 누락
* 특허 설명에 관련된 데이터 로드 불가
* 거절 이유 텍스트의 불완전한 파싱
* 법조항 참조 정보의 불일치
* 인용발명 정보의 누락

### **6.2 대응 방법**

* 법령을 명시 후 하위 조항 통합
* 청구항 누락에 대한 전처리 규칙 강화 및 필터링
* 설명에 관련된 데이터 파싱 형식 변경
* 거절 이유 텍스트의 구조화된 파싱 규칙 적용
* 법조항 매핑 테이블을 통한 일관성 검증
* 인용발명 정보 추출을 위한 정규식 패턴 강화



## **7. 활용 계획**

* **법령 데이터**: RAG 시스템 구축을 위한 벡터 DB화
* **특허 데이터**: RAG 시스템 구축을 위한 벡터 DB화, Fine-tuning 데이터
* **AI분야 특허심사 가이드 데이터**: RAG 시스템 구축을 위한 벡터 DB화, Fine-tuning 데이터
* **의견 제출 통지서 데이터**: Fine-tuning 데이터
* **거절결정서 데이터**: Fine-tuning 데이터

