SK네트웍스 **Family AI** 과정 **15**기

데이터 전처리 인공지능 데이터 전처리 결과서

산출물 단계	데이터 전처리
평가 산출물	인공지능 데이터 전처리 결과서
제출 일자	2025.10.17
깃허브 경로	https://github.com/SKNETWORKS-FAMILY-AICAMP/SKN15-FINAL-3TEAM
작성 팀원	김주형

1. 문서 개요

- 프로젝트명: 특허정보, 관련 법안, 문서 등을 통합 관리·자동화하여 팀 간 특허 관련 소통과 업무 생산성을 높이는 사내 전용 AI 허브
- 전처리 목적: 특허 등록 가능/불가능 분류 모델 학습용 데이터 정제
- 문제 정의: 특허 아이디어 및 청구항을 기반으로 특허 등록 가능성을 판단하는 모델을 학습하기 위한 데이셋 구축

2. 데이터셋 개요

- 데이터 출처 및 수집 방법: API 호출, 문서 파일 처리(로컬/업로드된 PDF), Python(pdfplumber, json)
- 데이터 구성:

항목명	설명	예시
doc_id	파일 식별자	1019950049190_952004009275519.pdf
발송번호	발송 번호	9-5-2004-009275519
발송일자	발송된 날짜	2004.03.10
출원인코드	출원인 식별자	519980961336
출원인	출원인 이름	가부시키가이샤 산요붓산
대리인	대리인 이름	황의만
출원번호	서류 식별자	10-1995-0049190
발명의_명칭	발명 명칭	파친코기
심사기관	심사기관 이름	기계금속심사국 제어기계심사담당관실
심사관	심사관 이름	이성섭
text	문서 내용	10-1995-0049190\n\nYOUR INVENTION PARTNER\n특 허...
tables_raw	파일 내부 표	[{""table_id"": 1, ""columns"": [""순번"", ""거절이유가 해소되지 않은 부분"", ""관련 법조항""]

- 원본 데이터 샘플(5~10건 첨부):
  (스크린샷 또는 테이블 형태)

	doc_id	발송번호	발송일자	출원인코드	출원인	대리인	출원번호	발명의_명칭	심사기관	심사관	text	tables_raw
0	1019950049190_952004009275519.pdf	9-5-2004-009275519	2004.03.10	519980961336	가부시키가이샤 산요붓산	황의만	10-1995-0049190	파친코기	기계금속심사국 제어기계심사담당관실	이성섭	10-1995-0049190\n\nYOUR INVENTION PARTNER\n특 허...	NaN
1	1019970011220_952006022243218.pdf	9-5-2006-022243218 수신 :	2006.04.20	519980961394	가부시키가이샤 세가	장수길 외 1 명	10-1997-0011220	태블릿유니트	기계금속건설심사본부 20 제어기계심사팀	조영길	제출기일\n특 허 청\n- - - 특허거절결정서 - - -\n\n주 소\n\n주 소...	NaN
2	1019970702450_952003008835047.pdf	9-5-2003-008835047	2003.03.10	519980594621	그라코 칠드런스 프로덕츠 인 크.	김용인 외 1 명	10-1997-0702450	상부개방형그네및그제어장치및방법	심사2국 제어기계심사담당관실	여원현	10-1997-0702450\n\nYOUR INVENTION PARTNER\n특 허...	NaN
3	1019980009800_952003012043234.pdf	9-5-2003-012043234	2003.03.31	419986012078	오충효	NaN	10-1998-0009800	전자오락기용본체프레임	심사2국 제어기계심사담당관실	이성섭	10-1998-0009800\n\nYOUR INVENTION PARTNER\n특 허...	NaN
4	1019980019945_952005011001790.pdf	9-5-2005-011001790 수신 :	2005.03.11	419980443366	강효동	NaN	10-1998-0019945	완력을조절하는 핸드그립	기계금속건설심사국 19 제어기계심사담당관실	이상선	제출기일\n특 허 청\n특허거절결정서\n\n주 소\n주 소\n\n이 출원은 지정기간...	NaN

3. 전처리 프로세스 개요

- 전체 흐름도:

  ① 수집 → ② 결측치 처리 → ③ 이상치 탐지 → ④ 데이터 분리

- 전처리 파이프라인 요약:

단계	목적	수행 작업	사용 도구/라이브러리
결측치 처리	누락값 제거	Null 행 제거, 특수값 대체	pandas
이상치 처리	비정상 데이터 제거	단어 수 기준, 이상시간 필터링	numpy
정규화	텍스트 전처리	소문자 변환, 불용어 제거	nltk
분리	학습/검증 분할	train:test = 8:2	train_test_split

4. 세부 전처리 단계

4.1 결측치 처리

- 결측치 존재 여부: 있음
- 결측 컬럼 및 비율:

컬럼명	결측률	처리 방법
대리인	50%	해당 열 제거
tables_raw	60%	해당 열 제거

- 코드 예시: df = df.drop(columns=['대리인', tables_raw])

4.3 정규화 및 표준화

- 텍스트 정규화:

항목	기준	처리 방식	제거 수
출원인코드	519980961336	- 제거	2274건
출원인	오충효	텍스트 추가	1035건
text	2026년 초과	불필요한 텍스트 제거	2274건

- 수치형 표준화: z-score, Min-Max
- 사용 라이브러리: re, nltk, sklearn.preprocessing

4.4 데이터 변환 및 생성

- 레이블 인코딩: tag 컬럼 → 숫자 레이블

  ```
  from sklearn.preprocessing import LabelEncoder
  encoder = LabelEncoder()
  df['tag_id'] = encoder.fit_transform(df['tag'])
  ```

- 파생 변수 생성 (예: 메시지 길이):

  ```
  df['msg_length'] = df['message'].apply(len)
  ```

5. 학습/검증 데이터 분리
   `dom_state=42)

- 분리 후 건수:

구분	데이터 수
학습 데이터	1819건
테스트 데이터	455건

6. 전처리 결과 요약 및 평가

- 전처리 후 전체 건수: 2274건
- 품질 향상 지표:
  - 결측값 제거: 1035건 정제
  - 컬럼 10개 제거
- 향후 활용 방안:
  - 특허 등록 가능/불가능 분류 모델 학습용으로 사용