SK네트웍스 Family AI과정 10기

데이터 수집 및 저장 수집 데이터 보고서

산출물 단계	데이터 수집 및 저장
평가 산출물	수집 데이터 보고서
제출 일자	2025-05-30
깃허브 경로	https://github.com/SKNETWORKS-FAMILY-AICAMP/SKN10-FINAL-3Team
작성 팀원	김현수

1. 수집 데이터 개요

데이터명	수집 대상	수집 목적	사용 예정 기능	출처/저작권
판례문	국가법령정보 공동활용	파인 튜닝 및 RAG	판례 조회 및 파인튜닝 데이터 생성 참조	국가법령정보 공동활용
용어	웹 크롤링	파인 튜닝	판례 / 문서 요약 및 분석	국가법령정보 공동활용 및
문서	자체 생성	RAG	문서 조회 및 요약, 분석	자체 생성

2. 수집 방법 및 자동화 절차

- 판례문
  - 수집 방식
    - API 호출
    - 웹 크롤링
  - 수집 도구 또는 스크립트 설명:
    - Python + requests, BeautifulSoup, Selenium
    - 자동화 여부 및 주기:
      - 서버 구축 이후 자동화 예정
        - 매주 일요일마다 판례 목록 조회
        - 이후 저장되어 있지 않은 데이터 존재 시 판례 상세 내용 수집
      - 자동화 예정

- 예시 스크립트

```python
import os
import requests
import json
import pandas as pd
from dotenv import load_dotenv

load_dotenv()

def page_presearch(page, display = 100):
  precsearch_url = "/DRF/lawSearch.do?"

  params = {
    "OC": OC,
    "target": "prec",
    "type": "json",
    "display" : display,
    "JO": "민법",
    "page": page
  }

  response = requests.get(LAW_URL + precsearch_url, params=params)
  data = response.json()

  df = pd.DataFrame(data["PrecSearch"]['prec']).iloc[:, 1:]
  return df

def page_presearch_logic(df):
  if os.path.exists(os.path.join(DATA_PATH, "precsearch_df.csv")):
    saved_presearch_df = pd.read_csv(os.path.join(DATA_PATH, "precsearch_df.csv"))
    new_presearch_df = page_presearch(1, 10)
    for row in new_presearch_df.iterrows():
      if row["사건명"] not in saved_presearch_df["사건명"].values:
        saved_presearch_df = pd.concat([row, saved_presearch_df])
    saved_presearch_df.reset_index(drop=True, inplace=True)
    saved_presearch_df.to_csv(os.path.join(DATA_PATH, "precsearch_df.csv"), index=False)
```

- 데이터 파일 예시


```
{
    "판시사항": "[1] 불법행위로 인한 손해배상 청구소송에서 피해자의 사고 당시 수입이 자백의
    "참조판례": "[1] 대법원 1980. 3. 25. 선고 80다68 판결(공1980, 12742), 대법원 1982. 5. 
    "사건종류명": "민사",
    "판결요지": "  [1] 타인의 불법행위로 인하여 피해자가 상해를 입게 되거나 사망하게 된 경
    "참조조문": "[1] 민사소송법 제261조 / [2] 민법 제393조, 제763조 / [3] 민법 제393조, 제
    "선고일자": "19980515",
    "법원명": "대법원",
    "사건명": "손해배상(자)",
    "판례내용": "【 원고, 피상고인 】 원고 (소송대리인 변호사 정장훈)<br/>【 피고, 상고인 】
    "사건번호": "96다24668",
    "사건종류코드": "400101",
    "판례정보일련번호": "116679",
    "선고": "선고",
    "판결유형": "판결",
    "법원종류코드": "400201"
}
```

- 용어
  - 수집 방식
    - tavily 기반 웹 크롤링
  - 수집 도구 또는 스크립트 설명
    - Python + tavily, langchain_openai
    - 자동화 여부 및 주기
      - 도메인 지식 학습용이므로 자동화 불필요
  - 예시 스크립트

```python
llm = ChatOpenAI(model="gpt-4o-mini", temperature=0)

tool = TavilySearchResults(
    max_results=5,
    include_answer=True,
    include_raw_content=True,
    search_depth="advanced",
    include_domains=[
      "ko.wikipedia.org/wiki/",
      "naver.com",
      "google.com",
      "sldongbu.scourt.go.kr/word/new/WordList.work"
      ]
)

llm_with_tools = llm.bind_tools([tool])

llm_chain = prompt | llm_with_tools

@chain
def tool_chain(user_input: str, config: RunnableConfig):
```

  - 데이터 파일 예시:

```
"대습상속": {
  "title": "대습상속",
  "definition": "대습상속은 상속인이 상속 개시 전에 사망한 경우, 그 상속인의 자녀가 그 상속인의 지
},
"텔레뱅킹": {
  "title": "텔레뱅킹",
  "definition": "텔레뱅킹은 전화 또는 인터넷을 통해 은행 서비스에 접근하고 거래를 수행할 수 있는
},
"부제": {
  "title": "부제",
  "definition": "부제는 법률 문서나 계약서에서 주제를 보충하거나 구체화하기 위해 사용되는 제목이나
},
```

- 내부 문서

  - 수집 방식
    - 자체 제작 - 판례문의 기초 사실을 기반으로 제작
  - 수집 도구 또는 스크립트 설명
    - Python + openai
    - 자동화 여부 및 주기:
      - 사용 시 사용자의 로컬 파일 기반이므로 자동화 불필요
  - 예시 스크립트

```python
import json
import os
from openai import OpenAI
from tqdm.auto import tqdm
from dotenv import load_dotenv

load_dotenv()

system_prompt = """
너는 기초사실을 기반으로 소장을 작성해주는 어시스턴트야.

행동강령:
- 이름, 주소, 전화번호와 관련된 내용은 임의로 작성해줘.
- 단 이름과 전화번호 각각은 중복되지 않게 작성해줘.
- 청구취지와 청구원인은 기초사실을 기반으로 작성해줘.
```

  - 데이터 파일 예시

```json
{
  "title": "소장",
  "사건번호": "2023가합12345",
  "사건명": "손해배상(자)",
  "원고": {
    "원고 1": {
      "이름": "김영희",
      "주소": "부산광역시 해운대구 해운대해변로 123",
      "전화번호": "010-1234-5678"
    },
    "원고 2": {
      "이름": "이민수",
      "주소": "부산광역시 사하구 하단동 456",
      "전화번호": "010-2345-6789"
    },
    "원고 3": {
      "이름": "박지훈",
      "주소": "부산광역시 동래구 온천동 789",
      "전화번호": "010-3456-7890"
    },
    "원고 4": {
      "이름": "최수진",
      "주소": "부산광역시 부산진구 부전동 101",
      "전화번호": "010-4567-8901"
    }
  },
  "피고": {
    "피고 1": {
      "이름": "이상훈",
      "주소": "경상남도 창원시 의창구 사림동 202",
```

3. 데이터 설명 및 구성

3.1 판례문

3.1.1 파일 및 필드 설명

필드명	데이터 타입	설명	예시
판시사항	string	법률적 쟁점 혹은 판단 기준 요약 정리	[1] 불법행위로 인한 손해배상 청구소송에서…
참조판례	string	판단 근거로 삼은 기존 판례 목록	[1] 대법원 1980. 3. 25. 선고…
사건종류명	string	사건의 종류	민사
판결요지	string	법원의 법적 판단 이유	[1] 타인의 불법행위로 인하여 피해자가 상해를 …
참조조문	string	판결에 적용되거나 언급된 법률 조항	[1] 민사소송법 제261조 …
선고일자	string	판결 선고일자	19980515
법원명	string	판결을 내린 법원 명칭	대법원
사건명	string	사건의 간략한 명칭	손해배상(자)
판례내용	string	주요 내용 전체	원고, 피고, 원심판결, 주문에 관한 내용 등
사건번호	string	사건 식별 고유 번호	96다24668
사건종류코드	string	사건 종류 코드	400101

판례정보 일련번호	string	고유 식별번호	116679
선고	string	판결 결과 요약	선고
판결유형	string	판결의 절차	판결
법원종류 코드	string	법원 종류 구분용 코드	400201

3.1.2 데이터 양

- 전체 수집 데이터 건수: 341
- 추출된 고품질 데이터 건수 (필터링 후 기준): 320

3.1.3 저장 위치 및 포맷

- 저장 경로: data/case
- 저장 포맷: JSON
- 인코딩: UTF-8

3.2 용어

3.2.1 파일 및 필드 설명

필드명	데이터 타입	설명	예시
용어	string	전문 도메인 용어	과실책임
설명	string	용어에 대한 정의 및 부연 설명	과실책임은 타인의 권리를 침해하거나 손해를 발생시킨 경우 …

3.2.2 데이터 양

- 전체 수집 데이터 건수: 2493
- 추출된 고품질 데이터 건수 (필터링 후 기준): 예정

3.2.3 저장 위치 및 포맷

- 저장 경로: data/law_word
- 저장 포맷: JSON
- 인코딩: UTF-8

3.3 사내 문서

3.3.1 파일 및 필드 설명

파일명 또는 테이블명	필드명	데이터 타입	설명
소장	본문 내용	json	판례의 기초 사실을 통해 만들어진 소장 데이터
진술서	본문 내용	json	판례의 기초 사실을 통해 만들어진 진술서 데이터
합의서	본문 내용	json	판례의 기초 사실을 통해 만들어진 합의서 데이터

3.3.2 데이터 양

- 전체 생성 데이터 건수
    - 소장 : 100개
    - 진술서 : 100개
    - 합의서 : 100개

3.3.3 저장 위치 및 포맷

- 저장 경로: data/doc
- 저장 포맷: JSON
- 인코딩: UTF-8

4. 데이터 품질 및 정합성 관리 방안

- 중복 제거 기준
    - 판례 데이터 : 중복되지 않는 스키마(사건번호)를 기준으로 중복을 제거
- 정합성 검증 방법
    - 판례 데이터
        - 선고일자를 YYYYMMDD로 일정한지 검증
- 표준화 전략
    - 텍스트 전처리
        - 특수문자
            - 【
            - 】
        - 텍스트
            - </br>

5. 법적·윤리적 검토

- 개인정보 포함 여부:
  - 포함: 변호사 이름, 법관 이름
- 비식별화 조치 여부:
  - 조회 시 아래 국판결서 등의 '열람 및 복사를 위한 비실명 처리 기준'에 의거하여 비식별화 조치 필요성 없음.
  - 학습 시 개인정보가 포함된 내용 제거 예정
- 출처 및 사용권:
  - 공개 여부: [ ] 공개 [ ] 내부사용 한정 [v] 일부 공개
  - 라이선스 또는 약관 검토 여부:
    - 상업적 이용 가능
      - 국가법령정보 공동활용의 저작권 정책
      - 참조링크 : https://open.law.go.kr/LSO/information/guide.do
    - 저작권
      - 저작권법 제7조 :
        보호받지 못하는 저작물로 분류됨
      - 참조링크 :
        https://www.law.go.kr/%EB%B2%95%EB%A0%B9/%EC%A0%80%EC%9E%91%EA%B6%8C%EB%B2%95/%EC%A0%9C7%EC%A1%B0
    - 판결서 등의 열람 및 복사를 위한 비실명 처리 기준(재일 2014-2)
      - 제4조 1항에 의거하여 변호인, 변호사, 법관 및 국가 기관은 비실명 처리
      - 참조링크 : https://jeju.scourt.go.kr/img/wjs/4.pdf
    - 공공데이터의 제공 및 이용 활성화에 따른 법률
      - 판례는 제1조에 의거하여 이용권 보장됨
        제2조를 참조하여 판례 데이터는 공공기관, 공공데이터에 포함
      - 참조링크 :
        https://www.law.go.kr/%EB%B2%95%EB%A0%B9/%EA%B3%B5%EA%B3%B5%EB%8D%B0%EC%9D%B4%ED%84%B0%EC%9D%98%EC%A0%9C%EA%B3%B5%EB%B0%8F%EC%9D%B4%EC%9A%A9%ED%99%9C%EC%84%B1%ED%99%94%EC%97%90%EA%B4%80%ED%95%9C%EB%B2%95%EB%A5%A0
    - 공공기록물 관리에 관한 법률

        - 참조링크 : https://www.law.go.kr/LSW//lsLinkProc.do?lsNm=%EA%B3%B5%EA%B3%B5%EA%B8%B0%EB%A1%9D%EB%AC%BC+%EA%B4%80%EB%A6%AC%EC%97%90+%EA%B4%80%ED%95%9C+%EB%B2%95%EB%A5%A0&chrClsCd=010202&mode=20&ancYnChk=0#

- 검토자 및 검토 일자:
    - 김현수 - 2025-05-30

6. 변경 이력 및 보완 내역

변경일	변경자	변경 내용	비고