**SK네트웍스 Family AI과정 13기  
 모델링 및 평가 수집된 데이터 및 전처리 문서**

**□ 개요**

* 산출물 단계 : 모델링 및 평가
* 평가 산출물 : 수집된 데이터 및 전처리 문서
* 제출 일자 : 2025.08.22
* 깃허브 경로 : https://github.com/SKNETWORKS-FAMILY-AICAMP/SKN13-FINAL-1TEAM
* 작성 팀원 : 남궁건우, 이명인

## 1. 개요

### 1.1 데이터 설명

**ClickA 프로젝트에서 RAG 기반 LLM 애플리케이션의 검색·생성 정확도를 높이기 위해 필요한 데이터 수집 및 전처리 과정을 정리한다. 벡터 검색에 최적화된 문장 단위 청크화를 목표로 한다. 문서의 범위에는 데이터 출처, 수집 방식, 전처리 단계, 품질 관리 절차, 저장 방식이 포함된다.**

1. **대상: 내부 배포 PDF, HWP(감사 결과보고, 회계감사보고서, 입찰공고문, 재무제표 등)**
2. **방식:**
3. **HWP 파일 PDF로 변환.**
4. **PyMuPDF로 본문을 추출**
5. **Camelot/Tabula로 표를 감지·추출**
6. **표의 행·열 구조를 이용해 단문 형태로 변환하는 방법을 상세히 기술**

**- 본문은 의미를 변경하지 않는 범위에서 줄바꿈·공백·하이픈을 정규화**

**- 장식 요소·페이지 바·불필요 메타 문자열 제거**

**- 이미지의 경우, 정보전달 목적의 차트·도표에 한해 사실 서술형 설명을 부여**

**- 개인정보 및 민감 정보는 정규식 기반 마스킹 규칙을 적용**

1. **최종 산출물은 인덱싱용 Markdown으로 구성하며, 문장·표 단문·이미지 설명을 청크로 분할해 임베딩 후 버전 관리와 함께 S3에 저장.**

## 1.2 데이터 수집 목적

1. **내부 고객 업무 효율성 향상  
    사내 문서·회의·일정·메일을 단일 인터페이스에서 의미 기반으로 검색·활용할 수 있게 하려면, 질의와 직접적으로 매칭되는 고품질 텍스트가 필요하다. 데이터 수집은 이 텍스트의 토대(문서 원문, 회의/업무 산출물)를 체계적으로 확보해 “한 번의 클릭으로 업무 완결”이라는 목표를 실현하기 위한 것이다.**
2. **RAG 정밀도와 재현성 확보  
    RAG 검색 품질을 좌우하는 것은 인덱스 데이터의 충실도와 일관성이다. 의미검색과 요약 기반 QnA에서 코사인 유사도 임계, recall@k, 응답 지연(P95) 등의 지표를 달성하려면, 출처와 형식이 다른 자료를 동일한 단위로 정규화해 누락 없이 수집해야 한다. 수집 단계의 목표는 모델이 아닌 데이터로 정밀도를 끌어올릴 수 있는 기반을 만드는 데 있다.**
3. **비정형 PDF의 벡터화 가능한 단위화  
    표·이미지·본문이 혼재된 PDF를 벡터 DB에 적합한 문장 단위로 변환하려면, 수집 단계에서 시각 요소까지 포함한 원본을 확보해야 한다. 이후 전처리에서 표는 행·열을 단문으로, 정보 전달 목적의 이미지는 객관 서술로 치환할 수 있도록 원천 데이터를 빠짐없이 모으는 것이 목적이다. 정보 손실을 최소화해 검색 적합성을 보장한다.**
4. **데이터 접근 제한에 대한 대안 확보  
    내부 문서 직접 수집에 제약이 있는 상황에서, 유사 도메인 공개 자료(예: 한국방송광고진흥공사 관련 공개 문서)를 합법적 범위에서 수집·활용해 프로토타입과 평가를 진행할 수 있는 기반을 마련하는 것이 목적이다. 이는 실제 도입 전 품질 검증과 파이프라인 안정화에 기여한다.**

## 2. 데이터 수집

### 2.1 데이터 출처

* **‘Kobaco’ 경영정보 공개시스템**
* **공공데이터 포털**

### 2.2 데이터 유형

* **text**
* **표(table)**
* **image**

## 3. 데이터 전처리

### 전처리 개요

* **전체 pdf 문서 : 347개 / 전체 6960pages (hwp로부터 변환된 pdf 포함)**
* **PDF → 페이지 단위 스캔 → 표 추출 → 텍스트 추출 → 이미지 설명(옵션) → MD 생성 → 품질 점검 → S3 업로드**

### 3.2 페이지 분류

* **표적 페이지, 본문 텍스트 페이지, 이미지 중심 페이지를 휴리스틱으로 구분(선분·격자 수, 공백 패턴 등)**

### 3.3 표 추출과 문장화

* **도구: Camelot(Tabula 보조). 추출 실패 시 Stream/Lattice 교차 시도**
* **기본 정제**
  + **공백·개행 정규화, 전부 문자열화, 완전 공란 행·열 제거**
  + **첫 행이 헤더 후보면 헤더 승격**
* **단위 추정: “단위: …” 또는 헤더·본문 단위 토큰(원, 억원, %, USD 등) 감지 → 값에 단위 부가**
* **문장형 변환 규칙  
   행키, 열헤더 값{단위}.  
   예) 자본총계, 제1(당)기 반기말 284,708,054,131.  
   실제 산출물에 수치 문장들이 생성됨을 확인.**
* **합계/총계/소계 행은 검색 노이즈를 줄이기 위해 기본 제외**

### 3.4 본문 텍스트 정제

* **페이지별 블록 추출 → 표 영역 제외 텍스트만 수집**
* **라인브레이크·하이픈 연결·인코딩 깨짐 최소 교정(내용 변형 금지)**
* **헤딩/페이지 마커/장식 구분선 등 인덱싱 방해 요소는 인덱싱용에서 제거**

### 3.5 이미지 설명

* **목적: 문서 흐름 보존 및 조회 가능성 확보(차트/도표는 검색 가치 높음)**
* **정책:**
  + **정보전달 목적(차트/도표/스크린샷/양식)인 경우에만 캡션 생성**
  + **장식용(로고/워터마크/배경)인 경우 표기 X**
* **가이드(요지)**
  + **그래프/차트: 제목·종류·축/단위·범례·데이터 포인트를 사실 서술**
  + **해석·추론·평가는 금지**

### 3.6 Markdown 구성

* **페이지 헤더 삭제: ## 페이지 N**
* **섹션 블록**
  + **표 단문 블록: 위 규칙으로 생성된 문장들**
  + **본문 블록: 페이지별 정제 텍스트**
  + **이미지 설명 블록(옵션)**
* **Markdown 파일에서 <NA>, Name: n, dtype: object 같은 판다스 잔여물은 제거 대상(실제 변환물에 잔존 사례 확인됨).**
* **ClickA는 jsonl/txt 별도 산출 없이 MD 자체를 인덱싱 소스로 사용**
* **인덱싱 제거 규칙(라인 필터)**
  + **문서 헤더/페이지 마커/구분선/빈 불릿/이미지 캡션 헤더는 제외**
  + **<NA>, 타입표기 잔여물은 삭제**

### 3.7 품질 검증

* **LLM이 100% 정확성을 보장하지 못하기 때문에 파일 직접 검수**
* **내용이 틀리거나 누락된 부분이 있을 경우 검수 과정에서 수**

**4. 저장 방식**

1. **로컬 저장**

* **형식: PDF당 마크다운 파일 1개(.md), UTF-8 인코딩.**
* **파일명: 원본 파일명의 글자와 숫자, 공백, 밑줄만 유지해 정제하여 사용.**
* **경로: 입력 루트 폴더 하위 \_markdown\_output\_재무성과/에 저장.**
* **내용: 벡터 인덱싱용으로 선형화된 본문만 저장.**
  + **포함: 표 단문 문장, 교정된 일반 텍스트, 이미지 설명의 본문.**
  + **제외: 헤딩(## 등), 페이지 라벨, 구분선(---), 빈 불릿, “### 이미지 설명” 헤더.**

1. **S3 업로드**

* **조건: .env에 AWS\_S3\_BUCKET이 설정된 경우에만 업로드 수행.**

## 5. 향후 사용 계획

* **전처리 파이프라인을 스케줄링해 PDF를 MD로 변환하고 S3에 저장한 뒤, 벡터 DB와 자동 동기화**
* **표는 행 단문, 본문은 최소 교정 텍스트, 정보 전달용 이미지 설명만 인덱싱하며 문서명, 페이지, 섹션, 단위 같은 메타데이터를 함께 저장**
* **검색 품질은 임베딩·하이브리드 검색과 chunk 크기, overlap, 임계값을 실험으로 최적화하고, 내부 질의응답 평가셋과 운영 로그를 주기적으로 반영해 성능을 개선**