SK네트웍스 **Family AI 17**기 **3Team**

데이터 전처리 RAG 데이터 전처리 결과서

1. 문서 개요

- 프로젝트명: ⚾BAIS⚾
- 전처리 목적: 야구 경기의 해설에 할루시네이션을 줄이기 위한 RAG 데이터 구축
- 문제 정의: AI 모델의 도메인 지식 함양과 RAG(검색 증강 생성) 시스템 구축을 위해 야구 규정 및 실제 경기 기록 데이터 수집을 필요로 함

2. 데이터셋 개요

- 데이터 출처 : 2023~2025 KBO 야구 경기 규칙 문서, 리그 규칙 문서, WBSC 국제 야구 규정 문서, 경기 기록 데이터(포스트시즌 및 국가대표)
- 데이터 수집 방법 및 명세

구분	데이터 주제	원천 소스	수집 방법	파일 포맷
규정	KBO 리그 규정 및 야구 규칙	KBO 공식 홈페이지 (PDF)	LlamaParse 활용 PDF 파싱	Markdown(.md)
규정	WBSC 국제 야구 규정	WBSC 공식 홈페이지 (PDF)	Google 문서 번역 후 LlamaParse 파싱	Markdown(.md)
기록	경기 기록 및 선수 정보	나무위키 (2021-2025 PS/국대)	수동 크롤링(Hand-Crawling) 및 구조화	Markdown(.md)

3. 전처리 프로세스 개요

- 비정형 데이터(PDF, 웹 텍스트)를 LLM이 이해하기 쉬운 마크다운(Markdown) 포맷으로 통일하고, RAG 검색 효율을 높이기 위해 구조화하는 것을 목표로 함
- 전처리 파이프라인 요약:

단계	수행 작업	목적	사용 도구/라이브러리
1. 파싱 (Parsing)	PDF 문서를 텍스트/마크다운으로 변환	비정형 문서의 텍스트 추출 및 표(Table) 구조 보존	LlamaIndex (LlamaParse)
2. 번역 (Translation)	영문 규정집을 국문으로 변환	국내 서비스 환경에 맞는 도메인 지식 확보	Google Docs Translation
3. 구조화 (Structuring)	경기 기록 텍스트를 계층적 구조로 작성	벡터 DB 청킹(Chunking) 및 검색 정확도 향상	VSCode, Markdown
4. 정제 (Cleaning)	불필요한 특수문자, 헤더/푸터 제거	노이즈 제거를 통한 모델 환각(Hallucination) 방지	Python(Regex), 수동 검수

4. 세부 전처리 단계

- 각 데이터 소스의 특성에 맞춰 차별화된 전처리 전략을 수행함

4-1. KBO 리그 규정 및 야구 규칙 (PDF → MD)

- 파싱: LlamaParse API를 활용하여 PDF 내 복잡한 표(Table)와 다단 구성을 마크다운 문법으로 변환
- 노이즈 제거: 페이지마다 반복되는 머리말, 꼬리말, 페이지 번호를 정규표현식(Regex)을 통해 일괄 제거
- 구조 보정: 파싱 과정에서 깨진 문단 구조나 들여쓰기를 수동으로 재정렬하여 가독성 확보

4-2. WBSC 국제 야구 규정 (번역 → PDF → MD)

- 번역: 영문으로 된 원본 문서를 Google 문서의 전체 번역 기능을 사용하여 1차적으로 한글화
- 노이즈 제거: 페이지마다 반복되는 머리말, 꼬리말, 페이지 번호를 정규표현식(Regex)을 통해 일괄 제거
- 파싱: 번역된 문서를 다시 PDF로 저장한 후, KBO 규정과 동일하게 LlamaParse를 통해 마크다운으로 변환하여 포맷 통일

4-3. 경기 기록 (웹 → 수기 작성 MD)

- 수집 범위: 2021~2025년 포스트시즌 전 경기 및 주요 국제대회(올림픽, WBC 등) 경기
- 계층적 구조화:
    - **Level 1(**경기 개요**)**: 대진, 날짜 (하이라이트 영상의 파일 제목과 동일)
    - **Level 2(**경기 정보**,** 출전 선수 라인업**,** 기록**)**: 구장, 승패 결과 요약, 이닝별 주요 타석 결과 및 득점 상황 상세 기술
- 식별자 관리:별도의 복잡한 ID 부여 대신, 하이라이트 영상의 파일 제목과 동일한 개요 형식을 일관되게 사용하여 텍스트 검색만으로도 식별되도록 처리

5. 전처리 요약 및 평가

5-1. 전처리 결과 데이터

데이터셋	처리 전 **(**페이지**/**건수**)**	처리 후 **(**용량**)**	비고
규정 데이터 (KBO+WBSC)	PDF 약 600 페이지 내외	1.01MB	표/리스트 구조 보존 완료
경기 기록 데이터	97 경기	531KB	경기 개요- 경기 기록 계층 구조화 완료
총계	-	1.53MB	RAG 구축용 지식 베이스 확보

5-2. 향후 활용 방안

- **RAG** 벡터 **DB** 구축 **(Knowledge Base)**
    - 전처리된 마크다운 파일은 구조 단위로 청킹하여 Vector DB에 적재
    - 캐스터가 규정이나 과거 경기 기록을 질문했을 때, 해당 지식을 검색하여 답변의 근거로 활용
- 데이터 확장성
    - 현재 구축된 마크다운 템플릿을 활용하여, 추후 2026년 시즌 데이터나 다른 리그 데이터도 손쉽게 추가 적재할 수 있는 기반 마련