Q. 도면 검색 지원 가능한지?

A. OCR 및 키워드 기반 검색, 유사도 기반 검색의 조합으로 일정 수준의 도면 텍스트 정보 검색은 가능할 수 있습니다. 다만 전문적인 도면 정보에 대한 검색은 특수 파서 및 비젼 AI 기술을 보유한 업체와의 협업이 필요합니다.

Q. 이미지 기반 벡터 검색은 어떻게 이루어지며, 표/이미지가 혼합된 문서도 검색 가능한가?

A. 현재 완벽한 수준의 표/이미지 파싱을 찾아보기는 어렵습니다. 표의 경우 OCR 기반 단순 파싱은 가능하지만, 복잡한 도면이나 인포그래픽은 비전 기술이 병행되어야 합니다.
현재는 멀티모달 LLM을 활용해 이미지를 설명 텍스트로 전환 후 임베딩하는 방식이 가장 현실적입니다다.
도면 해석이나 의미 파악은 어려우며, 라벨링이나 설명 텍스트 보강이 필요합니다.

5.2 데이터 수집 및 전처리

#전처리 #보안연계 #임베딩 #청킹

Q. 데이터 수집 전처리에서 스케쥴 관리란?

데이터 파이프라인의 스케쥴 관리란 수집·전처리 작업의 실행 시점과 주기를 관리하는 기능을 말합니다.

예를 들어, 매시간 또는 매일 자동으로 데이터 수집과 변환 작업을 실행하도록 예약하거나, 실시간 파이프라인의 트리거 조건을 설정하는 방식입니다.

이를 통해 데이터 수집 및 적재 작업이 계획된 일정에 따라 안정적으로 수행되도록 함으로써 데이터 최신성을 유지하고 운영 편의성을 높일 수 있습니다.

Q. 통합 수집 프로세스에서 정확도 점검은 어떻게 수행되는가?

A. 수집된 데이터는 자동화된 검증 로직을 통해 오류나 누락 여부를 사전에 점검합니다.

예정된 형식, 값의 범위, 필수 항목 여부 등을 기준으로 검토하며,일부 샘플은 원본과 비교해 데이터의 신뢰성을 확인하기도 합니다.