SK네트웍스 **Family AI**과정 **3**기

데이터 전처리 인공지능 데이터 전처리 결과서

□ 개요

- 산출물 단계 : 데이터 전처리
- 평가 산출물 : 인공지능 데이터 전처리 결과서
- 제출 일자 : 2025.04.18
- 깃허브경로:https://github.com/SKNETWORKS-FAMILY-AICAMP/SKN07-FINAL-5Team/tree/main/docs
- 작성 팀원 : 이재철

데이터 전처리 개요	● 문서유형 채용공고 데이터 - CSV ● 데이터 수집일자 2025/03/21 ~ 2025/03/26 ● 데이터 양 - 약 21,000건의 채용공고 - 약 26,000건의 기업별 면접후기 - 약 3,000건의 모집요강 이미지

전처리 과정	● 전처리 도구 – 의미없는 컬럼 제거 : pandas – 이미지 텍스트 추출 : multimodal model (gpt-4o-2024-08-06) – ● 데이터 추출 방식 사람인 채용공고 – 제목, 본문, 기업명, 기업정보, 회사 위치, 이미지 URL 등 크롤링 – 직무 대/중/소 분류는 API 활용하여 수집 캐치 면접후기 – 기업별 면접질문 등 크롤링 ● 불필요한 데이터 제거 기준 – 중복 채용공고 및 후기 제거 – 기업명 누락, 텍스트 미포함 공고 제거 – 너무 짧은 텍스트 or 의미 없는 이미지 제외 ● 정제 방법 – RecursiveCharacterTextSplitter 로 채용공고 본문 분할 (1000 토큰 기준) – 이미지형 모집요강은 OCR 기반 멀티모달 추출 후 텍스트화 – 채용공고 본문 임베딩 text-embedding_02_ada 사용

데이터 전처리 결과	● 결과 – 채용공고 약 13,000개 정제 완료 – 이미지형 모집요강에서 텍스트 약 3,000건 추출 – 모든 데이터는 메타정보 포함된 구조화된 형태로 저장 및 인덱싱 완료 – 데이터 저장 – 텍스트 데이터 ● Postgersql ● Elasticsearch – 이미지 데이터 ● AWS S3 – 임베딩 데이터 ● Chroma ● 향후 사용계획 – 사용자의 이력서를 입력 받아 유사 채용공고 추천 – 추천된 채용공고를 면접 질문 생성 참고 자료로 사용 예정