SK네트웍스 Family AI과정 7기

모델링 및 평가 수집된 데이터 및 전처리 문서

◘ 개요

- 산출물 단계 : 모델링 및 평가
- 평가 산출물 : 수집된 데이터 및 전처리 문서
- 제출 일자 : 2025-04-04
- 깃허브경로: https://github.com/SKNETWORKS-FAMILY-AICAMP/SKN07-FINAL-5Team/tree/main/docs
- 작성 팀원 : 김나예

개요	• **목적** ○ 면접 활용 데이터를 수집해 AI 기반의 분석 및 추천 서비스를 제공함으로써 유저의 면접 합격률 향상을 도움 • **사람인 - 채용 공고 데이터** ○ 출처 : 사람인(구직 정보 플랫폼) ○ 수집 방법 : 웹 크롤링 ○ 수집 기간 : 25.03.18 ~ 25.03.21 ○ **채용 공고 데이터** ■ 채용 공고 텍스트 및 기업 정보 전체 데이터 ■ 크기 : 약 **22,000건** ■ 주요 컬럼 : 직무코드(대/중/소), 채용공고 ID, 채용공고 제목, 공고 내용 텍스트, 공고 이미지 경로, 채용공고 URL, 기업명, 기업정보, 관련직무, 근무지역, 경력구분, 학력구분 ■ 이미지형 채용 공고(약 1,200건)는 수집 후 텍스트 변환하여 스토리지 적재 ○ **직무 데이터** ■ 직무 구분 별 이름, ID(Code) 데이터 ■ 크기 **: 2,178건** ■ 주요 컬럼 : 직무 대분류명, 대분류 코드, 직무 중분류명, 중분류 코드, 직무 소분류명, 소분류 코드

개요	● **캐치 - 면접 후기 데이터** ○ 출처 : 캐치(구직 정보 플랫폼) ○ 수집 방법 : 웹 크롤링 ○ 수집 기간 : 25.03.14 ~ 25.03.18 ○ **면접 후기 데이터** ■ 캐치에 등록된 대·중견 기업의 전체 면접 후기 데이터 ■ 크기 : **26,666건** ■ 주요 컬럼 : 기업명, 기업코드, 면접 평가, 면접 팁, 면접 질문, 면접ID, 지원 직무, 난이도, 어필역량, 합격여부, 면접구분, 경력여부 ○ **기업 정보 데이터** ■ 기업명, 기업ID(Code) 데이터 ■ 크기 : **7,000건** ■ 주요 컬럼 : 기업명, 기업 코드 ■ 면접 후기 수집에 활용
데이터 저장 및 관리	● **DB** ○ PostgreSQL(RDB) ○ ChromaDB(VectorDB) ○ ElasticSearch(검색 엔진) ● **Storage** ○ AWS S3 ● 수집 데이터 원본을 DB 3종에 각각 동일하게 적재 ○ 질문 생성 시 문서 검색과 top_k 공고 ID반환을 동시에 수행하기 위함 ● 이미지형 채용 공고 원본 파일과 유저의 이력서, 포트폴리오, 자소서 PDF 원본 파일은 스토리지에 적재

데이터 전처리 과정	● **사람인 - 채용 공고 데이터** ○ 데이터 중복 해소 ■ 하나의 채용 공고에 다수의 직무가 기입 되어 중복 발생 ■ 직무(대분류)를 기준으로 채용공고 ID를 분류한 테이블 생성 ○ 이미지 데이터 처리 ■ OpenAI 멀티모달 API 이용해 이미지에서 텍스트 추출 후 DB적재 ● **캐치 - 면접 후기 데이터** ○ **컬럼 필터링** ■ 삭제 컬럼 : 연도, 상/하반기 구분, 분위기점수, 난이도 텍스트, 분위기 텍스트, 처리과정, 면접타입, 면접타입명 ■ 삭제 이유 : 면접 질문 생성 시 참고할만한 데이터로써의 중요도가 낮은 컬럼 ○ **텍스트 정제** ■ 정제 컬럼 : 면접평가, 면접팁, 면접질문 ■ 정제 내용 : 특수문자 및 기호 (‘, [] 등) 제거, 텍스트 병합
데이터 전처리 결과	● 처리 결과 ○ 요약 : 채용공고 데이터, 면접후기 데이터와 각 테이블의 직무, 기업 테이블, 공통화를 위한 마스터 테이블, 면접 진행/결과/리포트 테이블로 구성 ○ 테이블 수 : 14개 ○ 텍스트 정제, 불필요한 컬럼 필터링 완료 ○ 이상치, 결측치 처리 불필요하여 데이터 크기 변화 없음 ● 향후 데이터 사용계획 ○ 사용자 구직 정보(이력서/포트폴리오/자소서) 데이터를 활용해 맞춤 채용공고 추천 제공 ○ 사용자 구직 정보, 추천 공고 데이터를 활용해 면접 예상 질문 생성 ○ 사용자 정보 데이터와 면접 답변 데이터를 면접 평가 생성에 활용 예정