SK네트웍스 Family AI과정 5기

데이터 수집 보고서

□ 개요

- 산출물 단계 : 데이터 수집 및 저장
- 평가 산출물 : 데이터 수집 보고서
- 제출 일자 : 2025-01-03
- 깃허브 경로 : https://github.com/SKNETWORKS-FAMILY-AICAMP/SKN05-FINAL-4TEAM.git
- 작성 팀원 : 김지연, 박보람

데이터 수집 목적	본 데이터 수집의 목적은 대규모 언어 모델(LLM)의 질문과 답변 생성 성능을 높이기 위함이다. 이로써 실제 면접과 유사한 환경을 조성하고, 고품질의 모의 면접 서비스를 제공하고자 한다.

데이터 수집 방법	1차 데이터 수집 1.1 기본 면접 데이터 (웹 크롤링) - 웹사이트(monster.com)에서 제공하는 유형별 면접 질문과 답변을 크롤링한다. 1.2 기술 면접 데이터 (웹 크롤링, kaggle, pdf) - 다양한 기술 분야에 대한 자료를 웹사이트(monster.com), kaggle 공개 데이터셋, PDF 문서로부터 수집한다. 1.3 CS 면접 데이터 (웹 크롤링) - 개발 블로그의 CS 기술면접 포스트에서 수집한다. 1.4 사용 기술 및 저장 포맷 - 웹 크롤링을 위해 Python의 BeautifulSoup, Selenium, Webdriver를 사용한다. - 저장 포맷 : JSON, CSV, PDF 2차 데이터 수집 2.1 기술 면접 추가 데이터 (웹 크롤링, csv,pdf) - 다양한 웹 사이트에서 제공하는 기술 면접 질문과 답변을 크롤링한다. - ML(Machine Learning) 자료는 여러 웹 사이트, PDF, CSV 파일로 부터 수집하고, LLM 자료는 여러 웹 사이트로부터 수집한다. - ML(Machine Learning), LLM에 대한 데이터를 추가 수집하여 AI 데이터로 병합한다. 2.2 사용 기술 및 저장 포맷 - 웹 크롤링을 위해 Python의 BeautifulSoup, Selenium, Webdriver를 사용한다. - 저장 포맷 : 하나의 JSON 파일

수집 데이터(요약)	1차 데이터 수집 1.1 기본 면접 데이터 - 행동, 상황 대처, 조직 문화, 리더쉽 등 다양한 유형에 대한 질문과 답변 형식으로 되어 있다. 1.2 기술 면접 데이터 - Python, SQL, Java, AWS, ML(Machine Learning) 데이터는 각 기술 개념들에 대한 면접 질문과 답변 형식으로 되어 있다. - LLM 데이터는 PDF 형식의 논문 자료로, LLM에 관한 전반적인 내용을 담고 있다. 1.3 CS 면접 데이터 - 자료구조, 알고리즘, 운영체제, 네트워크 등의 다양한 CS 개념들이 정해진 형식 없이 나열되어 있다. 2차 데이터 수집 2.1 기술 면접 추가 데이터 - LLM, ML(Machine Learning) 데이터는 각 기술 개념들에 대한 면접 질문과 답변 형식으로 되어 있다. - LLM, ML(Machine Learning) 즉, AI에 관한 전반적인 내용을 담고 있다.