SK네트웍스 Family AI과정 5기

데이터 수집 보고서

□ 개요

- 산출물 단계 : 데이터 수집 및 저장
- 평가 산출물 : 데이터 수집 보고서
- 제출 일자 : 2024-12-27
- 깃허브 경로 : https://github.com/SKNETWORKS-FAMILY-AICAMP/SKN05-FINAL-4TEAM.git
- 작성 팀원 : 김지연

데이터 수집 목적	본 데이터 수집의 목적은 대규모 언어 모델(LLM)의 질문 생성 및 답변 평가 성능을 고도화 하기 위함이다. 수집된 데이터는 기술 답변과 cs 답변 평가 모델의 파인튜닝에 활용되어, 실제 면접과 유사한 환경을 조성하고, 모범 답안 제공, 피드백 생성의 정확성 및 일관성을 향상시키는 데 기여한다.
데이터 수집 방법	1.1 기술 면접 관련 자료(웹 크롤링, kaggle, pdf) - 다양한 기술 분야에 대한 질문과 답변 자료를 웹사이트(monster.com), kaggle 공개 데이터셋, PDF 자료로부터 수집하였다 1.2 CS 면접 관련 자료 (웹 크롤링) - 컴퓨터 과학(cs) 기본 지식을 평가하는 질문과 답변 수집 - cs 관련 블로그 활용 1.3 사용 기술 및 저장 포맷 - 웹 크롤링: Python의 BeautifulSoup, Selenium, Webdriver 사용 - pdf 데이터 추출 : PyPDF2, PDFplumber 사용 - 수집된 데이터를 CSV 또는 JSON 포맷으로 저장

수집 데이터(요약)	2.1 기술 면접 질문
	- python, java, aws, sql, LLM, machine learning 분야 관련 질문 내용
	- 다양한 기술 스택에 대한 LLM 학습용 데이터로 활용 예정
	- 면접 질문 생성 알고리즘의 기반 데이터로 사용 예정
	- 기술 스택별로 실무 역량을 검증하기 위한 질문 포함
	
	2.2 cs 면접 질문
	- 자료구조, 알고리즘, 운영체제, 네트워크
	- cs 기본기를 검증하는 질문 및 답변 생성 데이터로 활용 예정
	
	2.3 수집 데이터 개요
	- 총 데이터 양 : 10개 파일(json, csv 파일 기준)
	  - 기술 면접: 9개
	  - cs 면접: 1개