SK네트웍스 Family AI과정 3기

데이터 전처리 인공지능 데이터 전처리 결과서

□ 개요

- 산출물 단계 : 데이터 전처리
- 평가 산출물 : 인공지능 데이터 전처리 결과서
- 제출 일자 : 11월 25일
- 깃허브 경로 : https://github.com/SKNETWORKS-FAMILY-AICAMP/SKN03-FINAL-5Team.git
- 작성 팀원 : 김재성

데이터 전처리 개요	• 문서유형 : CSV • 데이터 수집일자 : 11.22 • 데이터 양 : 총 데이터 셋 2,005개

전처리 과정	• 전처리 도구 : Python
	  – Pandas, NumPy, Selenium, bs4, Matplotlib
	• 데이터 추출 방식 : Crawling
	• 불필요한 데이터 제거 기준 :
	  – 중복 데이터
	  – 난이도 기준 외 데이터(프로그래머스 난이도 기준)
	• 정제 방법 : 결측 데이터 행 삭제
	  – 텍스트 기반 문제 내용 정리
	  – 잘못된 형식 수정(HTML 태그 제거)

데이터 전처리 결과	• 결과 : 총 데이터 1,493개
	  – 컬럼 수 : 4개(데이터 ID, 기술스택, 용어, 설명)
	  – 데이터 키워드 분포
	  모듈 : 37%
	  문자열 : 31%
	  함수 : 15%
	  기타(그리디, 알고리즘 등) : 17%
	• 향후 사용계획
	  – 목적 :
	  – AI 기반 면접 질문 추천 시스템에 데이터 활용
	  – 난이도 평가 데이터 사용
	  – 예상 작업:
	  – 데이터 증강(유사 문제 생성)
	  – 각 난이도별 추천 알고리즘 개발