SK네트웍스 Family AI과정 3기

모델링 및 평가 수집된 데이터 및 전처리 문서

□ **개요**

- 산출물 단계 : 모델링 및 평가
- 평가 산출물 : 수집된 데이터 및 전처리 문서
- 제출 일자 : 2024-12-28
- 깃허브 경로 : https://github.com/SKNETWORKS-FAMILY-AICAMP/SKN03-FINAL-5Team
- 작성 팀원 : 김재성

개요	데이터 설명 ● 주요 데이터는 Python 관련 질문과 답변, 프로그래밍 예제, 또는 특정 기술 문서입니다. 데이터 구조 ● 컬럼 개수: 2개 (질문, 설명) ● 데이터 크기: 약 [n] 행 ● 데이터 형식: CSV 데이터 수집목적 ● Python 프로그래밍 관련 질문 데이터베이스 구축. ● 데이터 기반 모델 학습을 통해 사용자 질문에 대한 답변 생성 자동화. ● 라가스(LAGAS) 지표 기반 성능 평가를 위한 데이터 제공.
데이터 자동화 및 검증	데이터 자동화 ● Selenium 및 BeautifulSoup을 활용하여 특정 웹사이트에서 데이터를 수집. ● 수집한 데이터는 pandas로 처리하여 CSV로 저장. ● 주기적인 데이터 업데이트를 위해 추후 스케줄러(Cron 또는 APScheduler)를 사용. 검증기준 ● 데이터 중복 여부, 형식 일치 여부 확인. ● 'PEP (숫자)'로 시작하는 행 제거. ● 수집된 데이터의 결측치 및 이상치 확인. 검증 결과 ● 중복 제거 후 n개의 데이터 유지 확인.

데이터 저장 및 관리

벡터 데이터베이스

- 데이터베이스 기술
  - FAISS 벡터 데이터베이스를 사용.
- 저장 구조
  - 각 벡터는 다음과 같이 저장:
    - 컨텐츠: 벡터와 연결된 주요 컨텐츠.
    - 문장 ID: 해당 문장이 원본 데이터에서 위치하는 ID.
    - 메타데이터: 출처, 날짜 등 부가 정보.
- 저장 예시

컨텐츠	벡터 (예시)	메타데이터
"Python Libraries and Descriptions…"	[0.12, -0.33, 0.54...]	출처: Python library
"Python Glossary and Key Terms…"	[0.45, 0.22, -0.11...]	출처: Python glossary

키워드 추출

- 방법
  - 이력서와 직무 설명서의 핵심 키워드를 추출.
  - 이력서를 영문번역한 후 sllm 모델을 사용하여 기술키워드를 추출

- 저장 방법
  - 키워드는 관계형 데이터베이스에 저장.

키워드 관리

- 추출주기
  - 새로운 이력서나 직무 설명서가 추가되어 면접실행시 추출.

데이터 전처리 과정	전처리 도구 ● 사용 라이브러리 ○ Pandas, Numpy, Selenium, Beautifulsoup, Matplotlib ● 데이터 추출 ○ Crawling 불필요한 데이터 제거 기준 ● 중복 데이터 제거: ○ 동일한 문제 내용이나 기술 스택이 반복된 데이터는 삭제. ○ `Pandas.drop_duplicates()`를 사용하여 중복 행 제거. ● 난이도 기준 외 데이터 삭제: ○ 프로그래머스의 특정 난이도(예: "쉬움", "어려움") 외의 데이터는 필터링. ○ Pandas의 조건문(`DataFrame.query`)을 활용하여 난이도 기준 데이터를 추출. 정제 방법 ● 결측 데이터 행 삭제: ○ 기술 스택, 문제 내용, 설명 등 주요 컬럼에 결측치가 있는 행을 삭제. ○ `Pandas.dropna()`를 사용하여 결측 행 제거. ● 텍스트 기반 문제 내용 정리: ○ 문제 내용에서 불필요한 공백과 특수문자를 제거. ○ 문자열 정규화 및 소문자 변환을 통해 데이터 일관성 유지. ○ Python의 `re` 모듈을 활용하여 정규 표현식으로 정리. ● 잘못된 형식 수정: ○ HTML 태그와 불필요한 메타데이터 제거. ○ BeautifulSoup의 `.text` 속성을 사용하여 순수 텍스트만 추출.

데이터 전처리 결과	결과 데이터 크기 ● 총 데이터 수: 1,493개 ● 컬럼 수: 4개 ○ 데이터 ID: 각 데이터를 고유하게 식별하기 위한 ID. ○ 기술스택: Python 문제에서 요구하는 주요 기술. ○ 용어: 문제의 핵심 키워드 또는 주요 개념. ○ 설명: 문제 내용 및 세부 요구사항. 데이터 키워드 분포 ● 모듈: 37% ● 문자열: 31% ● 함수: 15% ● 기타(그리디, 알고리즘 등): 17% 향후 사용 계획 ● AI 기반 면접 질문 추천 시스템 ○ 전처리된 데이터를 활용하여 사용자의 이력서 및 경력과 연관된 맞춤형 면접 질문을 생성. ● 난이도 평가 데이터 활용: ○ 난이도별로 추천 질문을 제공하여 면접 준비 효율성 향상. 예상 작업 데이터 증강 ○ 유사 문제 생성: ■ 기존 데이터에서 텍스트 데이터를 증강하여 새로운 문제 생성. ■ 예: 문제에 변형된 키워드 삽입 또는 난이도 조정. ○ GPT 기반의 데이터 증강 모델 활용. 난이도별 추천 알고리즘 개발 ○ 사용자의 수준에 따라 질문을 추천하는 알고리즘 개발. ○ 벡터화된 데이터를 기반으로 유사도 검색을 통해 질문 매칭.