SK네트웍스 Family AI과정 5기

인공지능 데이터 전처리 결과서

□ 개요

- 산출물 단계 : 데이터 전처리
- 평가 산출물 : 인공지능 데이터 전처리 결과서
- 제출 일자 : 2025 - 01- 15
- 깃허브 경로 : https://github.com/SKNETWORKS-FAMILY-AICAMP/SKN05-FINAL-4TEAM.git
- 작성 팀원 : 박보람, 김지연, 배윤관

데이터 전처리 개요	**데이터 전처리 목적** 데이터들은 고품질의 면접 질문/답변 생성을 위해 LLM을 파인튜닝 할 목적으로 수집되었다. 데이터들의 형식과 포맷이 모두 제각각이므로, 일관성을 위해서 데이터 전처리를 수행하여 형식과 포맷을 통일시키려 한다. 이외에도 구성 언어, 내용 중복 등도 고려하여 파인튜닝의 효과를 높이려 한다. **데이터 개요** 1) 기본 면접 데이터 - 내용 : 영어 텍스트로 구성된 여러 유형별 면접 질문/답변 - 샘플 수 : 75개의 QA 쌍 - 데이터 타입 : JSON 2) 기술 면접 데이터 (LLM 제외) - 내용 : 영어 텍스트로 구성된 기술 개념 질문/답변 - 샘플 수 : ● Python : 50개의 QA 쌍 ● Java : 29개의 QA 쌍 ● AWS : 14개의 QA 쌍 ● SQL : 12개의 QA 쌍 ● Machine Learning : 50개의 QA 쌍 - 데이터 타입 : JSON 3) LLM 데이터 - 내용 : LLM 관련 논문 (https://arxiv.org/abs/2307.06435) - 데이터 타입 : PDF 4) CS 면접 데이터 - 내용 : 한글 텍스트로 구성된 CS 관련 개념 - 데이터 타입 : CSV 5) AI 면접 데이터 - 내용: LLM 관련 내용과 Machine Learning 의 내용을 담고 있는 QA 데이터 셋 - 샘플 수: 총 553개의 데이터셋 - 데이터 타입: JSON

전처리 과정	**데이터별 전처리**
	1) 기본 면접 데이터
	• 전처리 도구
	  ○ 텍스트 전처리를 위해 re 라이브러리 사용
	  ○ JSON 파일을 불러오고 저장하기 위해 json 라이브러리 사용
	  ○ OpenAI API를 사용하여 원하는 출력 형식으로 변환
	• 전처리 방법
	  ○ '1.', '2.', '3.' 같은 불필요한 요소들 제거
	  ○ QA 데이터 생성을 요청하는 prompt와 QA 데이터가 포함된 completion 템플릿 형식으로 변환
	  ○ 모든 답변을 유연한 형식으로 변환
	  (기술 면접 데이터와의 차이점)
	  ■ 예시 : "스트레스와 압박을 어떻게 다루시나요?"
	  → "저는 스트레스 상황에서 [사용하는 방법, 예: 명확한 우선순위 설정]을 활용하여 [성과]를 이뤘습니다."
	
	2) 기술 면접 데이터 (LLM 제외)
	• 전처리 도구
	  ○ 텍스트 전처리를 위해 re 라이브러리 사용
	  ○ JSON 파일을 불러오고 저장하기 위해 json 라이브러리 사용
	• 전처리 방법
	  ○ 이미 QA로 구성된 데이터셋이므로 전처리 간단
	  ○ 형식을 통일하기 위해 'Question #1:', '1)', 'Q1' 이러한 불필요한 요소들을 제거
	  ○ QA 데이터 생성을 요청하는 prompt와 QA 데이터가 포함된 completion 템플릿 형식으로 변환
	
	3) LLM 데이터
	• 전처리 도구
	  ○ 파싱을 위해 re 라이브러리 사용
	  ○ PDF 파일을 읽어들이기 위해 PDFPlumberLoader 사용
	  ○ PDF 내용을 작은 단위로 쪼개기 위해 RecursiveCharacterTextSplitter 사용
	  ○ OpenAI API를 사용하여 질문 및 답변 생성

- JSON 형식으로 저장하기 위해 json 라이브러리 사용

- 전처리 방법
  - 논문 형식의 PDF에서 References 부분부터는 불필요한 내용들을 담고 있으므로 'References' 문자열을 포함하는 페이지의 이전 페이지까지만 불러옴
    PDF 전체 내용을 여러 청크들로 잘게 쪼갬
  - GPT-4 모델을 사용해서 각 청크에 대해 한 개의 QA를 생성하도록 프롬프트 엔지니어링 수행
  - GPT의 응답을 파싱하여 질문 부분과 답변 부분으로 분리
  - QA 데이터 생성을 요청하는 prompt와 QA 데이터가 포함된 completion 템플릿 형식으로 변환

4) CS 면접 데이터

- 전처리 도구
  - 파싱을 위해 re 라이브러리 사용
  - csv 파일을 읽어들이기 위해 Pandas 사용
  - 텍스트 내용을 작은 단위로 쪼개기 위해 RecursiveCharacterTextSplitter 사용
  - OpenAI API를 사용하여 질문 및 답변 생성
  - JSON 형식으로 저장하기 위해 json 라이브러리 사용
- 전처리 방법
  - csv 파일의 데이터를 불러와서 문자열 타입으로 변환하고 하나의 텍스트로 합침
  - 이후에 여러 청크들로 잘게 쪼개고 LLM 데이터 전처리 방법과 동일하게 처리

5) AI 면접 데이터

- 전처리 도구
  - 텍스트 전처리를 위해 re 라이브러리 사용
  - JSON 파일을 불러오고 저장하기 위해 json 라이브러리 사용
  - 중복 제거를 위해 코사인 유사도 사용
- 전처리 방법
  - 노이즈 필터링(Noisy Data Filtering): 광고와 같이 분석에 불필요하거나 방해가 되는 정보를 제거

<table>
<tr>
<td></td>
<td>
- 이미 QA로 구성된 데이터셋이므로 전처리 간단
- 형식을 통일하기 위해 'Question #1:', '1)', 'Q1' 이러한 불필요한 요소들을 제거
- QA 데이터 생성을 요청하는 prompt와 QA 데이터가 포함된 completion 템플릿 형식으로 변환
- 중복 제거 : 문장간 코사인 유사도를 확인하고, 유사도가 0.98 이상인 질문은 나중 질문을 제거

추가 전처리 (공통)
1) 한국어 번역 : 우리가 제공하고자 하는 AI 모의 면접 웹 애플리케이션 서비스가 한국어를 지원할 예정이므로, 영어로 구성된 데이터들을 전부 한국어로 번역해야 한다. 이를 위해 ChatGPT를 사용하였다.
2) 유사한 질문/답변 제거 : 데이터의 품질을 높이기 위해 중복을 최소화하였다.
3) 저품질의 데이터 제거 : 관련성이 적거나 질문과 답변의 품질이 낮다고 판단되는 데이터들을 제거하였다.
</td>
</tr>
<tr>
<td>전처리 결과</td>
<td>
일관된 공통 템플릿 형식
예시)
“prompt” : “Python에 대한 기술 면접 질문과 답변을 생성하세요.”,
“completion” : “질문: Python에서 네임스페이스란 무엇인가요?\n답변: Python에서 네임스페이스는 모든 이름이 존재하는 공간입니다. 네임스페이스는 변수 이름을 객체와 매핑합니다. 변수 검색 시 이 공간에서 객체를 찾습니다.”

샘플 수
• 기본 면접 데이터 : 총 96개의 QA 쌍 (기본 75개, 상황 판단 46개)
• Python : 50개의 QA 쌍
• Java : 46개의 QA 쌍
• AWS : 15개의 QA 쌍
• SQL : 10개의 QA 쌍
• Machine Learning : 64개의 QA 쌍
• LLM : 99 개의 QA 쌍
• CS : 126개의 QA 쌍
- AI : 530개의 QA 쌍 (LLM + Machine Learning + 추가 수집)
</td>
</tr>
</table>