인공지능 데이터 전처리 결과서

1. 데이터 설명 및 구성

1.1. 데이터 구성

구분	데이터명	데이터 수집 목적	데이터 형태	데이터 출처	데이터 크기	비고
DATA-001	Project md	sLLM 학습용 데이터셋 생성	.md 텍스트 파일, 이미지 파일	https://github.com, https://www.notion.so	텍스트 : 문서별 ~100KB 이미지 : 문서별 상이	
DATA-002	ChatGPT 질의응답 데이터	sLLM 학습용 데이터셋 생성	html, json, 이미지 파일	ChatGPT 데이터 내보내기	문서별 ~50MB	
DATA-003	Gemini 질의응답 데이터	sLLM 학습용 데이터셋 생성	html, json, 이미지 파일	구글 데이터 내보내기	문서별 ~50MB	
DATA-004	AI 개발환경 정보.md	sLLM 학습용 데이터셋 생성	.md 텍스트 파일	ChatGPT 데이터 일부 복사		
DATA-005	bedrock-agent.md	sLLM 학습용 데이터셋 생성	.md 텍스트 파일	https://github.com/kyopark2014/korean-chatbot-using-amazon-bedrock	13KB	
DATA-005	chunking-strategy.md	sLLM 학습용 데이터셋 생성	.md 텍스트 파일	https://github.com/kyopark2014/korean-chatbot-using-amazon-bedrock	5KB	
DATA-006	DPO 설명.md	sLLM 학습용 데이터셋 생성	.md 텍스트 파일	ChatGPT 데이터 일부 복사	4KB	
DATA-007	FFNN 설명.md	sLLM 학습용 데이터셋 생성	.md 텍스트 파일	ChatGPT 데이터 일부 복사	5KB	
DATA-008	LSTM 변형모델.md	sLLM 학습용 데이터셋 생성	.md 텍스트 파일	ChatGPT 데이터 일부 복사	9KB	
DATA-009	nGrinder 트래픽테스트.md	sLLM 학습용 데이터셋 생성	.md 텍스트 파일	ChatGPT 데이터 일부 복사	3KB	
DATA-010	opeanai Whisper 사용법.md	sLLM 학습용 데이터셋 생성	.md 텍스트 파일	ChatGPT 데이터 일부 복사	10KB	
DATA-011	parent-document-retrieval.md	sLLM 학습용 데이터셋 생성	.md 텍스트 파일	https://github.com/kyopark2014/korean-chatbot-using-amazon-bedrock	15KB	
DATA-012	prompt-flow.md	sLLM 학습용 데이터셋 생성	.md 텍스트 파일	https://github.com/kyopark2014/korean-chatbot-using-amazon-bedrock	15KB	
DATA-013	query-transformation.md	sLLM 학습용 데이터셋 생성	.md 텍스트 파일	https://github.com/kyopark2014/korean-chatbot-using-amazon-bedrock	7KB	
DATA-014	Qwen MoE 구조.md	sLLM 학습용 데이터셋 생성	.md 텍스트 파일	ChatGPT 데이터 일부 복사	15KB	
DATA-015	SKN 캠프 정보.md	sLLM 학습용 데이터셋 생성	.md 텍스트 파일	ChatGPT 데이터 일부 복사	7KB	
DATA-016	간단 IT 용어.md	sLLM 학습용 데이터셋 생성	.md 텍스트 파일	ChatGPT 데이터 일부 복사	5KB	
DATA-017	네이버 캠프 정보.md	sLLM 학습용 데이터셋 생성	.md 텍스트 파일	ChatGPT 데이터 일부 복사	14KB	
DATA-018	소켓통신 + 마스킹 훈련법.md	sLLM 학습용 데이터셋 생성	.md 텍스트 파일	ChatGPT 데이터 일부 복사	11KB	
DATA-019	스팀오리 프로그램 원리.md	sLLM 학습용 데이터셋 생성	.md 텍스트 파일	ChatGPT 데이터 일부 복사	4KB	
DATA-020	애자일 방법론 설명.md	sLLM 학습용 데이터셋 생성	.md 텍스트 파일	ChatGPT 데이터 일부 복사	11KB	
DATA-021	어텐션 설명.md	sLLM 학습용 데이터셋 생성	.md 텍스트 파일	ChatGPT 데이터 일부 복사	4KB	
DATA-022	예비군 동원훈련.md	sLLM 학습용 데이터셋 생성	.md 텍스트 파일	ChatGPT 데이터 일부 복사	7KB	
DATA-023	내활동.js	sLLM 학습용 데이터셋 생성	.json	Gemini 대화 내역	959KB	

- 전체 수집 데이터 건수: 30건 123MB

2. 데이터 수집 및 활용의 적법성 검토

- 저작권 준수: 해당 데이터의 이용약관을 검토하여 상업적 이용 및 2차 가공 가능 여부를 확인하였습니다.
- 크롤링 이용약관 준수: Robots.txt 규정에 따라 서버에 부하를 주지 않는 방식으로 크롤링을 수행하였습니다.
- 개인정보 보호: 수집 과정에서 개인 식별 정보(이름, 연락처 등)는 즉시 제외하거나 비식별 처리를 완료하였습니다.

구분	상업 이용 가능	학습 사용 허용	크롤링 이용약관 준수	개인정보 보호	비고
DATA-001	O	O	O	O	
DATA-002	O	O	O	O	
DATA-003	O	O	O	O	
DATA-004	O	O	O	O	
DATA-005	O	O	O	O	
DATA-005	O	O	O	O	
DATA-006	O	O	O	O	
DATA-007	O	O	O	O	
DATA-008	O	O	O	O	
DATA-009	O	O	O	O	
DATA-010	O	O	O	O	
DATA-011	O	O	O	O	
DATA-012	O	O	O	O	
DATA-013	O	O	O	O	
DATA-014	O	O	O	O	
DATA-015	O	O	O	O	
DATA-016	O	O	O	O	
DATA-017	O	O	O	O	
DATA-018	O	O	O	O	
DATA-019	O	O	O	O	
DATA-020	O	O	O	O	
DATA-021	O	O	O	O	
DATA-022	O	O	O	O	
DATA-023	O	O	O	O	

3. 데이터 저장 및 관리

1. .md 형식 데이터

- 저장 형식 : .md → 전처리 된 데이터의 경우 jsonl
- 저장 환경 : 로컬 서버, 구글 드라이브
- 데이터 구조 : 작성된 문서 데이터 텍스트 파일, 이미지 파일 → 전처리 후 jsonl, 이미지 파일 별도 관리
- 데이터 정제 및 전처리 : 문서 토큰화 후 라벨링 → 섹션화 후 카테고리 생성

2. LLM 대화내역 전처리 데이터

- 저장형식 : .html, .json → 전처리 된 데이터의 경우 jsonl
- 저장 환경 : 로컬 서버, 구글 드라이브
- 데이터 구조 : 질의응답 데이터 .html, .json 파일, 이미지 파일 → 전처리 후 jsonl, 이미지 파일 별도 관리
- 데이터 정제 및 전처리 : 문서 질의별 섹션화 후 카테고리 생성 및 내용 요약

4. 전처리 프로세스

- 전체 흐름도:

문서 수집 → 민감 정보 마스킹 → 문서 섹션화

- 섹션화 학습 데이터 생성

  sentence 단위로 segmentation → 섹션 경계를 기준으로 라벨링 → 두 문장 사이에 섹션 경계가 존재하는지 라벨링된 데이터셋 생성

- 유사도 학습 데이터 생성

  문서 섹션화 → 해당 섹션과 내용이 유사한, 형식은 유사하지만 내용이 다른 데이터를 생성 → 두 문장을 쌍으로 유사도 라벨링 (0 또는 1)

- 요약 학습 데이터 생성

  문서 섹션화 → 해당 섹션과 이웃한 데이터를 이용하여, 해당 섹션을 잘 설명하는 요약 데이터 생성 → 원본 텍스트와 함께 데이터셋 구성

- 인덱싱 데이터 생성

  문서 섹션화 → 문서 요약 데이터 생성 → 해당 섹션과 이웃한 데이터와 요약 결과 데이터를 이용하여, 해당 섹션을 잘 설명하는 인덱스 데이터 생성 → 원본 텍스트와 함께 데이터셋 구성

- 데이터 생성 과정
  - 섹션화 학습 데이터 생성
    1. 처리하고자 하는 문서 또는 파일을 로드하여 텍스트 형태로 변환
    2. 데이터에서 정규식을 이용하여 링크를 마스킹 후 별도 저장

       ```
       https://github.com/... → {url_01}
       ```

    3. 마스킹된 문서 원본을 LLM에 전달하여, 섹션 분리를 지시
    4. 나눠진 각 섹션을 규칙 기반으로 sentence 단위로 segmentation 및 metadata 생성

       ```
       ('code_block', r'\n```[\s\S]*?```'),
       ('header', r'^(#{1,6}\s[^\n]*|<h[1-6][^>]*>.*?</h[1-6]>)'),
       ('section_boundary', r'←SectionBoundary→'),
       ('list_item', r'^[ \t]*([-*+]|\d+\.)\s[^\n]*'),
       ('html_br', r'<br\s*/?>'),
       ('newline', r'\n'),
       ('sentence', r'[^ \n].*?(?:[.!?]|다\.)(?=\s|$)')
       ```

       - code block : 코드가 작성된 블록으로 코드 블록이 끝나기 전까지 섹션을 나누지 않음
       - header : 마크다운 문서 헤더 및 html 헤더 태그
       - section_boundary : 수작업으로 섹션을 나눈 문서를 처리하기 위한 라벨
       - list_item : 마크다운 리스트 형태로 된 내용
       - html_br, newline : 개행
       - sentence : 실제 문서 내용
    5. 각 나눠진 sentence 별로, 개행 문자거나 내용이 짧은 부분을 병합하여 실제 의미가 있는 각 sentence로 생성
    6. 각 섹션의 마지막 sentence는 라벨 0, 나머지 sentence는 라벨 1으로 설정하여 어떤 sentence들 사이에서 섹션이 분할되는지 라벨링
    7. 생성된 sentence들의 라벨을 이용하여, 연속된 두 sentence 사이의 섹션 분할 여부 데이터 생성

       ```
       {"text_a" : "앞 문장", "text_b" : "뒤 문장", "label" : 0 or 1 }
       ```

       라벨이 1이면 유사도가 높음 - 라벨이 0이면 유사도가 낮음

  - 유사도 학습 데이터 생성
    1. 섹션화 학습 데이터 3번의 결과로 생성된 섹션별로 분리된 데이터를 각 섹션 별로 LLM에 전달하여 형식은 유사하지만 상세 내용 맥락이 다른 문장과, 형식은 다르지만 상세 내용 맥락이 동일한 문장을 생성
    2. 생성한 데이터를 원본 데이터와의 쌍으로 유사도 라벨링 하여 유사도 학습 데이터 생성

       ```
       {"text_a" : "원본 문장", "text_b" : "생성한 문장", "label" : 0 or 1 }
       ```

       라벨이 1이면 유사도가 높음 - 라벨이 0이면 유사도가 낮음

  - 요약 학습 데이터 생성

1. 섹션화 학습 데이터 3번의 생성 과정에서 앞 뒤 내용을 모두 반영한, 각 섹션에 대한 요약 텍스트 생성
2. 생성한 데이터를 요약 학습용 데이터로 사용

- 인덱싱 데이터 생성
  1. 섹션화 학습 데이터의 3번의 생성 과정에서 앞 뒤 내용을 모두 반영한, 각 섹션에 대한 요약 텍스트를 수집
  2. 섹션화 학습 데이터의 4번의 생성과정에서 생성한 metadata를 통해, metadata=header인 정보들을 통해 각 섹션의 header_path를 수집

     ```
     header_path : 1. 개요 / 팀 소개
     ```

  3. 각 섹션별로 공백, 개행문자, html 태그 등을 제거한 텍스트 위주의 문장을 생성
  4. 정제된 문장을 LLM에 header_path, 요약 텍스트와 함께 전달하여 인덱스(색인) 데이터를 생성

---

5. 학습/검증 데이터 분리

1. **학습/검증 데이터 분리**

- 분리 기준 및 방법:
  - 기준: 무작위 분할 및 신규 생성
  - 데이터 분리 비율(건수): 무작위 추출 100건 및 정제 + 신규 생성 106건