

SK네트웍스 Family AI과정 3기

데이터 전처리 인공지능 데이터 전처리 결과서

□ 개요

- 산출물 단계 : 데이터 전처리
- 평가 산출물 : 인공지능 데이터 전처리 결과서
- 제출 일자 : 2024-12-31
- 깃허브 경로 : <https://github.com/SKNETWORKS-FAMILY-AICAMP/SKN03-FINAL-6Team>
- 작성 팀원 : 최연규

데이터 전처리 개요	<ul style="list-style-type: none">● 문서유형 : pdf● 데이터 수집일자 : 2024-11-30● 데이터 양 : 자동차 매뉴얼 pdf 1개
전처리 과정	<ul style="list-style-type: none">● 전처리 도구: Llama Index, OpenAI● 데이터 추출 방식:<ol style="list-style-type: none">1. Llama Index로 문서 로드2. 로드된 텍스트를 256 청크 사이즈로 분할하여 Parquet로 저장3. OpenAI 모델(gpt-4o-mini)을 통해 질문-답변 생성 <pre>from llama_index.core import SimpleDirectoryReader from llama_index.core.node_parser import TokenTextSplitter documents = SimpleDirectoryReader("../data/genesis/", recursive=True).load_data() from autorag.data.legacy.corpus.llama_index import llama_text_node_to_parquet nodes = TokenTextSplitter().get_nodes_from_documents(documents=documents, chunk_size=256, chunk_overlap=128) corpus_df = llama_text_node_to_parquet(nodes) corpus_df = cast_corpus_dataset(corpus_df) corpus_df.to_parquet('../data/test_dataset/corpus_new.parquet')</pre> <p>prompt = """다음은 참고로 제네시스 차량 매뉴얼에 관한 내용을 만들었습니다. 매뉴얼을 보고 할 만한 질문을 만드세요. 반드시 제네시스 차량과 관련한 질문이어야 합니다. 만약 주어진 매뉴얼 내용이 제네시스 차량과 관련되지 않았다면, '제네시스 차량과 관련 없습니다.'라고 질문을 만드세요.</p> <p>매뉴얼: {text}</p> <p>생성할 질문 개수: {num_questions}}</p> <p>예시: [Q]: 제네시스 차량의 타이어 공기압은 얼마인가요? [A]: 제네시스 차량의 타이어 공기압은 35 PSI입니다.</p> <p>제네시스 차량과 관련이 없는 매뉴얼의 경우 예시: [Q]: 제네시스 차량과 관련 없습니다. [A]: 제네시스 차량과 관련 없습니다.</p> <pre>import pandas as pd from llama_index.llms.openai import OpenAI from autorag.data.legacy.dacreation import make_single_content_qa, generate_qa_llama_index corpus_df = pd.read_parquet('../data/test_dataset/corpus_new.parquet', engine='pyarrow') llm = OpenAI(model='gpt-4o-mini', temperature=0.5) qa_df = make_single_content_qa(corpus_df=corpus_df, content_size=10, qa_creation_func=generate_qa_llama_index, llm=llm, prompt=prompt, question_num_per_content=2)</pre>
데이터 전처리 결과	<ul style="list-style-type: none">● 결과: 문서와 관련된 질문과 답변 샘플 추출 <pre>"query": "고속도로 주행 보조 기능은 어떻게 작동하나요?", "generation_gt": ["고속도로 주행 보조는 직통 가능한 도로 주행 시 주행 보조 버튼을 눌러 커먼 작동합니다. 스마트", "query": "고속도로 주행 보조 작동 중 할 차량이 정차하면 어떻게 되나요?", "generation_gt": ["고속도로 주행 보조 작동 중 할 차량이 정차하면 따라서 정차하며, 정차 후 액", "query": "제네시스 차량의 주행 속도 제한 보조 기능은 어떤 상황에서 작동하나요?", "generation_gt": ["주행 속도 제한 보조 기능은 도로의 제한속도가 70 kmh 이상인 경", "query": "제네시스 차량에서 전방 카메라 렌즈에 이물질이 묻었을 때 어떻게 해야 하나요?", "generation_gt": ["전방 카메라 렌즈에 이물질이 묻으면 인식 성능이 저하되어 지능", "query": "주차 거리 경고 기능이 작동하지 않을 때 어떤 점검을 해야 하나요?", "generation_gt": ["주차 거리 경고 기능이 작동하지 않을 때는 초음파센서가 손상되었는지, 외부", "query": "주차 거리 경고 기능이 정상적으로 작동하지 않는 경우 어떤 상황이 있을까요?", "generation_gt": ["주차 거리 경고 기능이 정상적으로 작동하지 않는 경우는 초음파센", "query": "제네시스 차량의 에어컨은 어떤 상황에서 작동하나요?", "generation_gt": ["제네시스 차량의 에어컨은 정면에서 보통 이상 강도로 증폭하거나, 측면에서 보통 이상의 강", "query": "제네시스 차량에서 이물질이 묻었을 때 청소하기 위한 권장 사항은 무엇인가요?", "generation_gt": ["제네시스 차량에서는 이물질을 등화하여 설치된 이물질 보조 장치에", "query": "제네시스 차량의 후방 뷰 주차 가이드라인은 어떤 거리를 나타내나요?", "generation_gt": ["후방 뷰 주차 가이드라인은 차량으로부터 0.5 m, 1 m, 2.3 m 거리를 나타", "query": "시라운드 뷰 모니터의 자동 커징 기능을 설정하는 방법은 무엇인가요?", "generation_gt": ["시라운드 뷰 모니터의 자동 커징 기능을 설정하려면 시동 'on' 상태에서 인</pre> <ul style="list-style-type: none">● 향후 사용계획: 작성된 매뉴얼 기반 Q&A챗봇의 평가 데이터 셋으로 활용