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

<table>
<tr><td>데이터 전처리 개요</td><td>

- 문서유형 : pdf
- 데이터 수집일자 : 2024-11-30
- 데이터 양 : 자동차 매뉴얼 pdf 1개

</td></tr>
<tr><td>전처리 과정</td><td>

- **전처리 도구**: Llama Index, OpenAI
- **데이터 추출 방식**:
  1. Llama Index로 문서 로드
  2. 로드된 텍스트를 256 청크 사이즈로 분할하여 Parquet로 저장
  3. OpenAI 모델(gpt-4o-mini)을 통해 질문-답변 생성

```
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_over
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')
```

```
prompt = """다음을 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 없었습니다.'라고 질문을 만드세요.

매뉴얼:
{{text}}

생성할 질문 개수: {{num_questions}}

예시:
[Q]: 제네시스 차량의 타이어 공기압은 얼마인가요?
[A]: 제네시스 차량의 타이어 공기압은 35 PSI입니다.

제네시스 차량과 관련이 없는 매뉴얼의 경우 예시:
[Q]: 제네시스 차량과 관련 없었습니다.
[A]: 제네시스 차량과 관련 없었습니다.
"""
```

```
import pandas as pd
from llama_index.llms.openai import OpenAI
from autorag.data.legacy.qacreation 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_
                               llm=llm, prompt=prompt, question_num_per_content=2)
```

</td></tr>
<tr><td>데이터 전처리 결과</td><td>

- 결과:
  문서와 관련된 질문과 답변 샘플 추출

- 향후 사용계획:
  작성된 매뉴얼 기반 Q&A챗봇의 평가 데이터 셋으로 활용

</td></tr>
</table>