데이터 전처리

이 프로젝트에서는 SK매직 서비스 센터가 제공하는 정수기 제품 사용 설명서를 데이터로 활용했습니다. AI 모델이 이 데이터를 학습하여 정확한 답변을 제공할 수 있도록 여러 단계의 전처리 과정을 거쳤습니다. 이 과정은 데이터의 품질을 높이고, AI 모델의 학습에 적합한 형태로 변환하게 진행되었습니다.

데이터 전처리 과정

1. **데이터 수집**
   - **SK매직의 정수기 제품 사용 설명서를 원본 데이터로 활용**했습니다. 이 설명서는 SK매직의 베스트셀러 정수기 모델이 선정되었으며, 사용자가 자주 접하는 주요 정보를 포함하고 있습니다. 더불어, 다른 제품 설명서에도 적용할 수 있도록 여러 설명서를 함께 수집했습니다.
2. **PDF 데이터 로드 및 분할**:
   - `PyPDFLoader`를 사용해 PDF 파일에서 데이터를 로드하고, 이를 페이지 단위로 분할했습니다. 각 PDF 페이지가 개별적으로 처리될 수 있도록 분할했습니다.

     ```
     loader = PyPDFLoader("data/product_guide.pdf")
     pages = loader.load_and_split()
     ```
3. **텍스트 정제 및 청크 분할**:
   - **RecursiveCharacterTextSplitter**를 사용해 PDF의 텍스트 데이터를 일정한 크기의 청크로 분할했습니다. 각 청크는 500자 단위로 설정했으며, 청크 간 중복을 방지했습니다. 이 과정에서 `chunk_overlap` 을 0으로 설정하여 데이터 양을 줄이고 처리 속도를 높였습니다. `chunk_overlap`은 인접한 청크 사이에 중복으로 포함될 문자의 수입니다. 즉 본 코드에서는 각 청크들이 연결부분에서 중복되지 않습니다.

     ```
     text_splitter = RecursiveCharacterTextSplitter(chunk_si
     ze=500, chunk_overlap=0)
     splits = text_splitter.split_documents(pages)
     ```
4. **텍스트 벡터화 및 저장**:

- **OpenAI의 Embedding 모델**을 사용해 분할된 텍스트 청크를 벡터 형태로 변환한 후, **Chroma** 벡터 데이터베이스에 저장했습니다.

```
vectorstore = Chroma.from_documents(documents=splits, e
mbedding=OpenAIEmbeddings(openai_api_key=openai_api_ke
y)
```

  - **벡터화 과정**: OpenAI의 임베딩 모델(text-embedding-ada-002)을 사용해 텍스트 청크를 수치 형태의 벡터로 변환했습니다.
  - **벡터 데이터베이스 저장**: 벡터화된 텍스트 청크를 Chroma 벡터 데이터베이스에 저장하고 인덱싱하여 추후 검색 및 질의응답 작업에 사용할 수 있게 했습니다.

1. **리트리버 설정**:
   - **리트리버 구성**: 벡터 데이터베이스에 저장된 텍스트 청크를 효과적으로 검색하기 위해 리트리버(retriever)를 설정했습니다. 리트리버는 사용자의 질문에 가장 관련성 높은 텍스트 청크를 검색하는 역할을 합니다.

```
retriever = vectorstore.as_retriever(search_kwargs=
{'k':10})
```

     - **검색 파라미터 설정**: `search_kwargs` 파라미터로 `k=10`을 설정했습니다. 이는 사용자 질문에 대해 상위 10개의 관련 텍스트 청크를 검색하여 반환하도록 하는 설정입니다. 이를 통해 사용자는 더 나은 검색성능과 사용자 경험을 최적화 할 수 있습니다.
     - **효율적인 검색**: 리트리버는 벡터 데이터베이스에서 유사성을 기준으로 가장 관련성 있는 청크를 신속하게 검색합니다. 이는 AI 모델이 사용자에게 정확한 정보를 제공하는 데 필수적인 역할을 합니다.

전처리 결과

- 이렇게 전처리된 데이터로 AI 모델은 사용자 질문에 대해 정확하고 신뢰할 수 있는 답변을 제공할 수 있게 되었습니다. 또한, **다양한 제품 설명서에도 적용할 수 있는 범용성을** 고려하여 전처리를 진행했습니다.