SK네트웍스 Family AI과정 3기

데이터 전처리 데이터 전처리 결과서

개요

- 산출물 단계: 데이터 전처리
- 평가 산출물: 데이터 전처리 결과서
- 제출 일자: 2024-12-25
- 깃허브 경로: https://github.com/SKNETWORKS-FAMILY-AICAMP/SKN03-FINAL-6Team

1. 자동차 사용자 매뉴얼

구분	내용
문서유형	자동차 사용자 매뉴얼 PDF
사용 목적	사용자 매뉴얼 기반 Q&A 챗봇 구축
데이터 양	600페이지 가량의 16개의 PDF 파일

- **전처리 과정**

항목	상세 내용
텍스트 추출	- pdfminer.six를 사용하여 PDF 문서 내 텍스트 추출
청킹(Chunking)	- chunk_size=100, overlap=20, max_length=2500 설정- 문맥 보존을 위해 일정 부분 중첩(Overlap) 적용
벡터화	- BAAI/bge-m3 모델로 sparse 및 dense 벡터 생성- 벡터는 Milvus와 같은 벡터 데이터베이스에 저장 가능
중복/노이즈 제거	- 50자 미만의 청크 제거- 중복되는 청크(문단) 제거- 특수문자 및 불필요한 공백 제거

- **데이터 전처리 결과**

결과 항목	내용
전체 문서 수	16개의 문서
총 엔티티 수	26,389개의 엔티티 생성

벡터화 완료	sparse, dense 벡터 각각 생성 (BAAI/bge-m3)
모델 입력 준비	Q&A 챗봇에 활용할 수 있도록 전처리된 텍스트 및 벡터 데이터 확보
품질 개선	- 중복/노이즈 제거- 신뢰도 높은 데이터셋 구성
활용 방안	사용자 매뉴얼 Q&A 챗봇, 검색/추천 시스템, 추가 다국어 확장 가능성

문서2