데이터 전처리 결과서

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

2. 자동차 후기, 차량 데이터

- 차종별 csv 스펙 파일(RDS)

구분	내용
문서유형	자동차 차종별 CSV 파일
사용 목적	사용자 후기 및 자동차 차량별 스펙 데이터 기반 자동차 추천 챗봇 구축
데이터 양	20개의 csv 파일(추가예정)

- 차종별 후기 데이터 파일(Vecter DB → Milvus DB)

구분	내용
문서유형	자동차 차종별 후기 데이터(json 파일)
사용 목적	사용자 후기 및 자동차 차량별 스펙 데이터 기반 자동차 추천 챗봇 구축
데이터 양	5개의 후기 파일(추가예정)

- **전처리 과정(후기 데이터)**

항목	상세 내용
텍스트 임베딩 생성	- `monologg/kobert` 모델과 토크나이저를 사용하여 텍스트 임베딩 생성 - 임베딩 차원: 768 - [CLS] 토큰의 벡터를 추출하여 임베딩으로 사용
Milvus 설정	- parameter(AWS)를 통한 Milvus의 URI와 토큰 가져오기 - Milvus 클라이언트를 통해 데이터베이스 연결
컬렉션 생성	- 컬렉션 스키마 정의 - Primary Key: `id` - 벡터 필드 : `vector` (768차원, 거리 계산 방식: L2) - 정렬 인덱스 생성: `car_id`
데이터 검증 및 삽입	- 임베딩이 768차원인지 확인 -

	`car_id`가 존재하면 업데이트 - `car_id`가 없으면 삽입
메타데이터 처리	- 추가 동적 필드(`metadata`)를 삽입하거나 업데이트 - 기존 데이터 삭제 없이 동적 데이터 확장 가능
데이터 로드 및 관리	- 생성된 컬렉션을 메모리에 로드하여 빠른 데이터 검색 및 삽입 지원

- **데이터 전처리 결과**

결과 항목	내용
전체 문서 수	5개 (추가예정)
총 엔티티 수	총 500개의 데이터 엔트리 생성 (데이터 추가에 따라 데이터 엔트리 추가)
벡터화 완료	- KoBERT 모델을 사용한 768차원 dense 벡터 생성 - 벡터 데이터는 Milvus에 저장 및 관리
모델 입력 준비	- Q&A 검색, 추천 시스템에서 활용할 수 있도록 벡터 임베딩 기반 데이터베이스 구성
품질 개선	- 중복 텍스트 제거 - 노이즈 데이터(짧은 문장, 특수문자 등) 제거 - 신뢰도 높은 데이터셋 구성
활용 방안	사용자 Q&A 검색, 유사 차량 추천, 추가 메타데이터 기반 확장 기능 지원

3. 자동차 / 운전자 보험

구분	내용
문서유형	자동차 / 운전자 보험 약관 PDF
사용 목적	자동차/ 운전자 보험 Q&A 챗봇 구축
데이터 양	150~200페이지 가량의 10개의 PDF 파일

- **전처리 과정**

항목	상세 내용
텍스트 추출	- pdfminer.six를 사용하여 PDF 문서 내 텍스트 추출
청킹(Chunking)	- chunk_size=500, overlap=50, max_length=2000 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	10개의 문서
총 엔티티 수	5142개의 엔티티 생성
벡터화 완료	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	보험 Q&A 챗봇, 검색/추천 시스템, 추가 다국어 확장 가능성