청주 교통사고 위험도 분석 및

LLM-RAG 기반 사고 설명 시스템

(개인프로젝트)

1. 프로젝트 개요

본 프로젝트는 청주 지역 교통사고 데이터를 분석하여
사고 발생 패턴과 위험 지역 및 시간대를 도출하고,
사용자가 교통사고 데이터를 쉽게 이해할 수 있도록
LLM-RAG 기반 사고 설명 시스템을 구현하는 것을 목표로 진행되었습니다.

단순 통계 제공이 아닌,
사고 데이터와 정책 정보를 결합하여 사고 원인과 위험 요인을 자연어로 설명하는 시스템 구현에 초점을 두었습니다.

2. 문제 정의

기존 교통사고 데이터 분석은 다음과 같은 한계를 가지고 있었습니다.

- 사고 건수 및 사망자 수 중심 통계 제공
- 사고가 많이 발생하는 이유 설명 부족
- 정책 결정이나 시민 활용으로 연결되기 어려움

즉,

"데이터는 있지만 왜 위험한지는 알기 어렵다"

는 문제가 존재했습니다.

따라서 본 프로젝트에서는 다음 질문을 중심으로 분석을 진행했습니다.

“교통사고 데이터를 누구나 이해 가능한 설명으로 제공할 수 있을까?”

3. 데이터 설명

청주 지역 교통사고 데이터를 활용했으며, 데이터에는 다음 정보가 포함되었습니다.

- 사고 발생 시간
- 사고 발생 위치
- 사고 유형
- 도로 및 환경 조건
- 사고 발생 건수 및 피해 정보

이를 기반으로 사고 패턴을 분석했습니다.

4. 분석 및 해결 접근 방식

4-1. 사고 발생 패턴 분석

데이터 분석을 통해 다음을 도출했습니다.

- 특정 시간대(특히 야간 시간)에 사고가 집중됨
- 사고 발생이 특정 지역에 집중되는 패턴 확인
- 사고 유형과 환경 조건에 따른 위험도 차이 존재

이를 기반으로 위험 지역 및 시간대 시각화를 수행했습니다.

4-2. 위험 지역 시각화

사고 발생 빈도와 위험 조건을 종합하여

- 위험 지역
- 보통 지역
- 안전 지역

으로 구분하고 지도 기반 시각화를 구현했습니다.

이를 통해 사고 집중 지역을 직관적으로 확인할 수 있도록 했습니다.

4-3. LLM-RAG 기반 사고 설명 시스템 구현

단순 그래프 제공을 넘어
사용자가 질문하면 사고 원인을 설명하는 시스템을 설계했습니다.

구조는 다음과 같습니다.

1. 사고 데이터 및 정책 정보를 텍스트 형태로 변환
2. Vector DB(FAISS)에 저장
3. 사용자 질문 입력
4. 관련 사고 정보 검색
5. LLM이 검색된 정보를 기반으로 설명 생성

이를 통해

- 특정 지역 사고 원인 설명
- 위험 요인 안내
- 정책 정보 연결

이 가능하도록 구현했습니다.

5. 문제 해결 과정

문제 1. 통계 결과만으로는 사고 원인 이해가 어려운 문제

단순 사고 건수 그래프로는 위험 요인을 설명하기 어려웠습니다.

→ 사고 조건 및 환경 정보를 함께 분석하여 패턴 설명 가능하도록 개선

문제 2. 사용자가 데이터를 직접 해석해야 하는 문제

데이터 이해에 전문 지식이 필요했습니다.

→ LLM-RAG 기반 자연어 설명 구조를 도입하여 해결

문제 3. 정책 정보와 사고 데이터를 연결하기 어려운 문제

사고 데이터만으로는 개선 방향 제시가 어려웠습니다.

→ 정책 문서와 사고 데이터를 함께 검색하는 구조 설계

6. 프로젝트 의의

- 교통사고 데이터를 설명 가능한 정보로 변환
- 데이터 분석과 LLM 기반 서비스 구조를 결합한 프로젝트 수행
- 분석 결과를 실제 사용자 활용 가능 구조로 확장

7. 사용 기술 및 개발 환경

데이터 분석

- Python
- Pandas, NumPy
- 데이터 전처리 및 시각화

LLM-RAG 시스템

- FAISS 기반 Vector DB
- RAG 검색 구조
- LLM 기반 응답 생성

서비스 구성

- Streamlit 기반 인터페이스
- End-to-End 데이터 분석 → 응답 생성 구조 구현

GitHub Repository

https://github.com/rjaekawpxm1-netizen/portfolio/tree/main/Projects/cheongju