자연언어처리 기말 대체 과제 중간 보고서

생명과학전공 2303050 박은서

작업의 진척도

완료 초기 data 수집, data pre-processing, semantic search 구현

진행 LLM & RAG 활용 답변 생성, 실시간 검색 및 데이터 수집 기능 추가

예정 multi-hop PAR 방식으로 확장, 초기 데이터 추가 수집

진행 방식

초기 data 수집 Entrez api로 pubmed에 접근, 데이터(pubmed id, article title, abstract) 수집
data pre-processing 문장 분리 → SciSpacy의 en_ner_bc5cdr_md 모델을 활용해 NER 처리 → gene 이름, disease로 normalization

Embedding SentenceTransformer all-mpnet-base-v2로 임베딩 → faiss로 벡터 인덱스 구축

Semantic search 임베딩 벡터 활용, 활용해 top-k 문장 검색 방식의 semantic search 구현

LLM & RAG llama 모델 사용해 LLM & RAG 활용 답변 제공 개발 중

문제점

초기에는 LLM 모델 관련하여 chat-gpt 프롬프트 이용 방식을 생각하였으나, 해당 방법의 경우 api token 사용량에 따른 과금 문제가 있습니다. 우선 오픈소스 LLM인 llama를 직접 돌리는 방식으로 진행해보려고 하며 실패 시 타 오픈소스 LLM을 활용하거나 전부 여의치 않을 경우 약간의 과금은 감수하는 방향으로 실험은 진행하고자 합니다.

제안서 지적사항 피드백

multi-hop 구현 가능성 현재 진행 중인 기초 모델 개발 후 모듈을 붙여 구현 가능(semantic search를 반복 호출하여 각 반환된 문장을 LLM 프롬프트에 전부 집어넣는 방식) / PAR 방식을 이용할 것이므로 Plan 단계에서 사용자 질문을 분석, 중간 질의를 계획하는 LLM 혹은 룰 기반의 모듈을 구현(기본적으로 룰, 시간이 허락하면 LLM으로 확장) → Act 단계에서 1단계 검색 결과를 근거로 해 2단계 검색 쿼리를 생성하고 재검색을 반복하는 multi hop 과정 → Review 단계에서 각 단계별 결과를 통합 및 모순 점검 이후 LLM 기반 RAG로 답변 생성

기존 벡터-db 적용 처리 SciSpacy(NER 모델)로 entity 추출 → 문장별 gene-disease pair 만들어 메타 데이터로 저장 → Sentence Transformer (임베딩 모델)로 벡터화 → FAISS(벡터 데이터베이스)에 인덱싱 / 이후 사용자 쿼리 입력시 해당 쿼리를 임베딩, 벡터 인덱스에서 Top-K 검색