문서 기반 AI 챗봇 시스템 개발

오픈소스 URL : https://github.com/Quinsie/doqmate

2025학년도 2학기
SW 캡스톤디자인 경진대회

팀 명 도큐메이트(DoQ-Mate)

팀 원 표지호(IT정보공학과, 3), 이은재(영어영문학과, 4), 허예림(경영학과, 4), 박제성(컴퓨터공학과, 4)

지도교수 박영진

산업체 메디앙시스템(주)

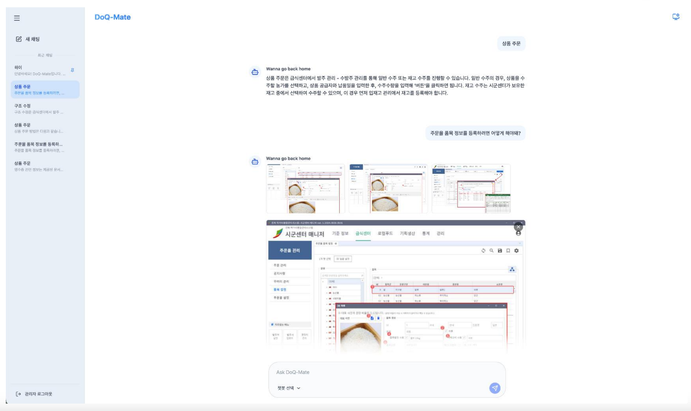
개발 동기 및 목적

■ 개발 동기

기업 내에는 PDF 등 다양한 문서가 방대하게 축적되어 있음에도 최신 정보를 빠르게 찾기 어렵고 검색 효율이 낮아 업무 정확성과 생산성에 제약이 발생하고 있다.
이에 따라 단순 키워드 검색을 넘어 **문서를 정확히 이해하고 출처를 명확히 제시하는 문서 기반 AI 검색이 요구**되며, 이는 **신뢰성 있는 답변 제공과 기존 검색 방식의 비효율성 개선**에 필수적

■ 개발 목적

① 보안이 중요한 환경에서 사용 가능한 로컬 LLM 기반 시스템 구축
외부 API에 의존하지 않는 LLM 환경을 구축하여 기업 정보 유출 위험을 최소화
② 문서 기반 RAG 검색·응답 기능의 고도화
문서의 텍스트·이미지·표를 통합 분석하고, 텍스트 정제·청킹·메타데이터 구조 개선을 통해 후반부 누락·이미지 인식 실패 등 기존 문제를 해결
③ 도메인 특화 지식 반영 및 관리자 중심 운영 지원
실제 업무 매뉴얼과 기록 문서를 학습 데이터로 활용하여 업무 특화된 AI 문서 비서를 제공
④ 비용 효율적이고 지속 가능한 AI 운영 환경 확보
유지비용이 큰 클라우드 기반 LLM 대신 로컬 모델 기반의 비용 효율적 운영 환경을 구축



주요 기술

■ 프론트엔드

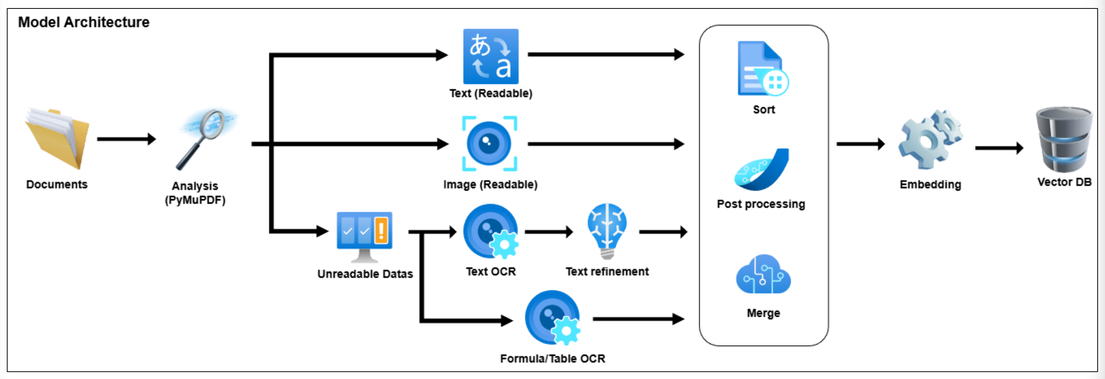
- 개발: Typescript + React + TailwindCSS
- 배포: Vercel

■ 백엔드

① 웹 프레임워크 (Web Framework): Flask
② WSGI 서버 (WSGI Server): Gunicorn
③ 데이터베이스 (Database): PostgreSQL

■ AI

① PDF 파싱 & EasyOCR Fine-tuning
PyMuPDF로 텍스트·이미지를 통합 추출하고, EasyOCR을 파일별 인식률 편차를 보정하도록 파인튜닝하여 비정형·저해상도 PDF에서도 안정적인 텍스트 인식이 가능하도록 설계
② VLM
Qwen 2.5-VL을 사용하여 이미지에서 의미를 추출, 텍스트와 같은 위치에서 임베딩함으로 더욱 정확한 색인이 가능하도록 설계
② 임베딩 & 로컬 벡터DB
bge-m3 임베딩 + Chroma 로컬 벡터DB로 문단 단위 검색 구조 설계
③ 로컬 LLM 기반 RAG 파이프라인
Qwen 2.5-7B-Instruct를 vLLM으로 경량화·최적화해 로컬 환경에서 운용함으로써, 보안성과 응답 속도를 강화한 RAG 시스템 구축



개발 내용

■ 프론트엔드

① 백엔드, 프론트엔드 API 커스텀 Hook 구현
② axiosClient를 사용하여 API 연동
③ Layout 분리를 통해 사용자 권한 별 Routing 구현
④ 채팅 페이지, 관리자 페이지 구현, Local Storage 활용 사용자 채팅 저장

■ 백엔드

1. REST API (Flask)
RESTful 아키텍처를 설계, 프론트엔드가 직관적으로 데이터를 요청하고 제어할 수 있도록 구현.
2. Swagger (OpenAPI)
API 명세서를 보기 좋게 하여 프론트엔드 개발자와의 협업 효율 상승, 실시간 API 테스트 환경을 제공.
3. Gunicorn
Flask 애플리케이션을 구동하여, 다수의 동시 접속 요청을 안정적으로 처리하는 환경을 구축.
4. PostgreSQL
문서의 처리 상태(Pending/Indexing/Ready)와 다양한 데이터들을 저장.

■ AI

① **문단 기반 청킹 및 텍스트 정제:** PyMuPDF·OCR결과를 1·2차 정제로 문장 단절·후반부 누락 문제 해결, 문서 구조 보존과 검색 정확도 향상
② **멀티모달 근거 연동 RAG:** PDF 이미지의 크기·중복을 필터링한 뒤 텍스트와 매핑하여, 검색된 문단과 함께 관련 이미지를 자동 제공하는 근거 기반 응답 시스템 구현
③ **검색 정확도 향상을 위한 쿼리 정제:** 질문 의도 분석·키워드 재구성으로 벡터 검색 재현율 향상 LLM 환각을 줄이기 위한 문서 용어 기반 재작성 전략 적용
④ **문서 품질 관리 및 재색인 체계:** OCR 실패·파싱 오류 감지를 통한 문서 품질 모니터링

결과 및 분석

■ 결과

- 텍스트·이미지·표를 통합 정제하여 문서 전체에서 누락 없는 검색 구조 구현
- 문단 기반 임베딩 + 로컬 벡터DB로 근거 중심 문서 QA 정확도 향상
- 검색된 문단과 함께 관련 이미지·표 자동 연동 → 부분적 환각 및 문맥 탈선 감소
- 생성 LLM이 아니라 검색·정제·요약 중심의 보조 모델로 사용하여 응답의 일관성 확보
- 로컬 LLM·온프레미스 구조로 보안·비용·속도 요구사항 충족

테스트 결과…
OCR/파싱 결과로 나온 조각 대비 문장 재구성률 71.3% → 98.5%, 중복 비율 22.8% → 1.7%
직접 제작·라벨링 진행한 50개 Dataset으로 평가 진행

튜닝 결과…
기존 모델 대비 정확도 65.9% → 85.1%로 19.2% 상승, 추론 시간 2.8초 → 0.9초로 1.9초 단축

■ 분석

- **정확도 개선**: 텍스트 1·2차 정제·문단 청킹으로 후반부 누락·단절 문제 해결
- **신뢰성 강화**: 이미지·표까지 포함된 근거 제공으로 문서 기반 검증 가능성 확보
- **실사용 적합성**: 문서 삭제·재색인·업데이트 대응 구조로 실제 기업 환경에 적용 가능
- **확장성 확보**: 멀티모달 통합 구조로 향후 VLM·표 인식 모델과의 결합 용이