TripPrep

Agent & RAG System

발표자: 서진주

목차

프로젝트 개요

프로젝트명

•TripPrep

한줄 소개

•**해외여행을 떠나는 사람을 위한 여행 준비 서비스**

•**여행 준비 가이드 보고서 생성** -> 노션 연동(보고서, 체크리스트)

•여행 준비 보고서와 기타 필요 문서, 일정을 첨부하여 **RAG 챗봇** 사용 가능

개발 기간

•2025.11.14 ~ 2025.12.05

팀 구성

•서진주

서비스 컨셉 및 핵심 목표

서비스 컨셉
- "정보 분석부터 여행 계획까지, AI 하나로 해결"
- RAG: 복잡한 문서를 읽지 않고 대화로 핵심 정보 파악
- TripPrep: 목적지만 입력하면 수십 개의 웹사이트를 검색해 나만의 여행 가이드 완성

핵심 목표
1. 통합성 (Integration): 문서 분석과 정보 생성을 하나의 플랫폼에서 제공
2. 정확성 (Accuracy):
   - RAG: 문서 기반 답변으로 환각(Hallucination) 최소화
   - TripPrep: 실시간 웹 검색(Tavily)으로 최신 정보 제공
3. 편의성 (Usability):
   - 띄어쓰기 없는 문서도 자동 처리
   - Notion 연동으로 결과물 즉시 활용 가능

시연

보고서 생성 페이지 사용 흐름

1. 입력: 목적지(예: 오사카)와 키워드(예: 맛집, 쇼핑) 입력
2. 생성: 멀티 에이전트(Scout -> Architect -> Writer)가 실시간 정보 수집 및 리포트 작성
3. 활용: 생성된 리포트를 Notion으로 내보내기, 생성된 체크리스트를 Notion으로 내보내기

RAG 챗봇 사용 흐름

1. 업로드: PDF/TXT 파일 다중 업로드 또는 텍스트 붙여넣기
2. 분석: 텍스트 추출 -> 정제(중복 문자 제거) -> 문장 분리 -> 벡터 인덱싱
3. 질문: "이 문서의 핵심 내용은?" 질문 입력
4. 답변: 근거 문서(Source)와 함께 답변 생성

사용 모델

Local LLM
- 모델명: jhgan/ko-sroberta-multitask
- 역할: RAG 답변 생성
- 특징: 온디바이스 구동, 보안성 우수, 비용 0원

Cloud LLM
- 모델명: Claude Haiku/Sonnet
- 역할: 여행 리포트 작성
- 특징: 복잡한 추론, 고품질 작문, 대량 정보 처리

Embedding Model
- 모델명: jhgan/ko-sroberta-multitask
- 역할: 문서 검색 (RAG)
- 특징: 한국어 문장 유사도 측정 특화

기술 스택

Backend
- Framework: FastAPI (통합 서버 구축)
- AI Engine: llama.cpp, Anthropic API, Tavily API
- Vector DB: FAISS (고속 유사도 검색)
- Data Processing: pdfplumber, regex

Frontend
- Core: React 19, TypeScript
- Build: Vite
- UI: 반응형 다크 모드 디자인
- Integration: Notion API

핵심 기술 – 보고서 생성

멀티 에이전트 협업 구조

- **문제:** 단일 프롬프트로는 방대한 여행 정보 수집과 정리가 불가능
- **해결:** 3단계 파이프라인 구축(Class 단위로 3개의 Agent 만들어 협업)
  1. Scout: 정보 수집 전담 – Haiku 호출 (Anthropic API)
  2. Architect: 목차 설계 전담 – Haiku 호출 (Anthropic API)
  3. Writer: 최종 작성 전담 – Sonnet 호출 (Anthropic API)
- **결과:** 각 에이전트가 전문화된 작업을 수행하여 고품질 리포트 생성
- **부가기능 :** 보고서와 체크리스트를 자동적으로 Notion에 업로드 (Notion API)

핵심 기술 – RAG 챗봇

- PDF, TXT 파일 업로드, 텍스트 입력칸으로 텍스트 입력도 가능
  - **Pdfplumber 라이브러리 사용하여** pdf -> txt 변환
- **벡터DB화**
  - chunking (100토큰 정도의 n문장)
  - 임베딩 - jhgan/ko-sroberta-multitask
  - context 무제한 설정 (청크된 것을 SLM에 전달할 때 잘리지 않고 전달)
  - top k = 5, 답변 생성시 일부 제외 가능
- **챗봇 SLM** - A.X-4.0-Light-Q4_K_M.gguf (양자화 모델)

회고 & 향후 계획

성과

- RAG, Agent 개념 활용 앱
- 로컬 SLM과 클라우드 API(LLM)를 적재적소에 배치하여 비용 효율적인 하이브리드 아키텍처 완성

향후 계획

- Pdf 처리 개선
- 작동 시간 단축(현재 기능당 1~2분) - Slm 모델 변경, Agent 연동 코드 개선
- 답변 품질 향상 – 키워드 기반 검색 추가
- 에이전트 개선 – Agent가 능동적으로 tool을 선택하는 구조로 변경
- 여행 준비 체크리스트 알고리즘 강화, 형태 개선(표 형태 고려)
- 데이터베이스 구축 – 이전에 업로드한 문서의 VectorDB 저장되도록(매번 임베딩 필요 x)
- 배포 – 챗봇 모바일 환경 이용 - 여행 시 사용
- BM 구상

감사합니다