SK네트웍스 **Family AI**과정 **3**기

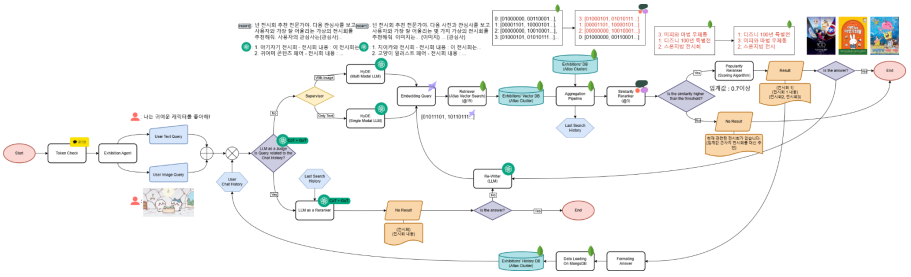
모델링 및 평가 **LLM** 활용 소프트웨어

☐ 개요

- 산출물 단계 : 모델링 및 평가
- 평가 산출물 : LLM 활용 소프트웨어
- 제출 일자 : 2024.12.26
- 깃허브 경로 : https://github.com/SKNETWORKS-FAMILY-AICAMP/SKN03-FINAL-2Team
- 작성 팀원 : 이준석, 박중헌

목적	• **개인화된 전시회 추천 제공**: 사용자의 취향을 반영한 맞춤형 추천으로 전시회 선택 과정을 간소화 • **다중 모달 데이터 활용**: 텍스트와 이미지를 조합한 입력 데이터를 분석하여 정확성을 향상 • **사용자 경험 개선**: 보다 직관적이고 신뢰할 수 있는 추천 시스템을 통해 사용자 만족도를 향상 • **문화 콘텐츠 접근성 향상**: 다양한 전시회 정보를 중앙화하여 사용자가 쉽게 탐색할 수 있도록 지원

데이터 전처리	**데이터 수집 및 저장** ● 전시회 메타데이터 수집: 전시회 이름, 내용, 주제, 이미지 등의 정보를 공식 API 또는 웹 크롤링을 통해 확보 ● 벡터화된 데이터 저장: 전시회 이미지 데이터를 Embedding방식으로 변환하여 Vector Database에 저장 **Retriever** ● HyDE 모델 활용: ○ Single-Modal LLM: 텍스트 입력을 기반으로 전시회와 유사도를 계산 ○ Multi-Modal LLM: 텍스트와 이미지를 결합하여 다중 모달 기반으로 임베딩 생성 ● 유사도 비교: 사용자의 입력과 전시회 임베딩 간 코사인 유사도를 계산하고, 특정 임계값(0.7 이상)을 기준으로 필터링 **Reranking** ● 가중치 기반 재정렬: 유사도 점수 외에도 전시회의 인기, 리뷰 점수 등을 반영하여 결과를 정렬 ● 사용자 과거 검색 이력 반영: 사용자가 이전에 검색한 관심사와 관련된 전시회 정보를 추가로 제공 **입력 데이터** ● 텍스트 입력 ○ 예시: "나는 귀여운 캐릭터를 좋아해!" ○ 처리 과정: 텍스트를 토큰화하여 주요 키워드를 추출하고 임베딩으로 변환 ● 이미지 입력 ○ 예시: 사용자가 업로드한 귀여운 캐릭터 이미지 ○ 처리 과정: 이미지 데이터 OCR Model을 거쳐 텍스트 추출 후 벡터화하여 텍스트 데이터와 결합 **데이터베이스 소스** ● 전시회 정보: MongoDB, Mongo Vector DB ● 기타 사용자 데이터: 사용자의 과거 검색 기록 **출처** ● 데이터 소스: 전시회 공식 API, 문화재청 데이터베이스, 웹 크롤링 ● 임베딩 모델: Upstage Solar Embedding Model

기대효과	• **개인화 추천**: 사용자 취향을 정교하게 분석하여 적합한 전시회 추천 • **전시회 접근성 강화**: 사용자와 전시회를 효과적으로 연결하여 문화 콘텐츠 소비 확대 • **다중 모달 데이터 활용**: 텍스트와 이미지를 통합 분석하여 기존 추천 시스템 대비 정확성 향상 • **사용자 경험 강화**: 직관적 인터페이스와 빠른 응답으로 추천 과정 간소화
도구 및 환경	**개발 언어 및 라이브러리** • **Python**: 데이터 처리 및 모델 구축에 사용 • **RAG (Retrieval-Augmented Generation)**: 검색 기반 생성 모델을 활용하여 사용자 질의와 연관된 전시회 정보 생성 • **RAGAS**: RAG 기반 시스템의 성능 평가와 모니터링 도구로 활용 • **LangChain**: 대화형 에이전트 설계 및 워크플로 관리 • **LangGraph**: RAG 기반 워크플로를 시각적으로 설계하고 최적화 • **EasyOCR, Upstage Embedding** : 이미지 및 텍스트 임베딩 모델 구현 • **OpenAI**: **OpenAI API**: GPT 기반 자연어 처리 모델을 활용하여 고도화된 생성, 프롬프트 엔지니어링 **데이터 저장 및 처리** • **MongoDB**: 전시회 메타데이터, 임베딩 데이터 저장 **운영 및 배포 환경** • **클라우드 환경**: AWS 기반 서버리스 환경 활용 (Lambda, S3 등) • **Kubeflow**: 기계 학습 워크플로의 자동화, 모델 학습 파이프라인 설계 및 배포 관리 • **RAG 워크플로 관리**: LangChain과 LangGraph를 통해 RAG 파이프라인 구현 • **모니터링 및 평가**: RAGAS를 통해 모델 성능 평가 및 사용자 피드백 반영
흐름도	

결론	**주요 성과** ● 텍스트와 이미지 기반 추천 시스템 구현 ● 다중 모달 데이터를 활용한 유사도 계산 및 리랭킹 구조 설계 ● MongoDB를 활용한 확장성 높은 데이터 관리 **향후 계획** 1. **모델 최적화**: HyDE 모델의 성능 개선 및 사용자 피드백을 기반으로 재학습 2. **추천 영역 확대**: 전시회 외에 공연, 영화, 뮤지컬 등 다양한 문화 콘텐츠로 확장 3. **사용자 피드백 수집**: 추천 결과에 대한 사용자 만족도 조사 및 데이터 반영 4. **실시간 추천**: 더 빠른 응답 속도를 위한 캐싱 및 서버 최적화