SK네트웍스 **Family AI**과정 **6**기

모델링 및 평가 LLM 활용 소프트웨어

☐ 개요

- 산출물 단계 : 모델링 및 평가
- 평가 산출물 : LLM 활용 소프트웨어
- 제출 일자 : 2025.03.07
- 깃허브 경로 : https://github.com/SKNETWORKS-FAMILY-AICAMP/SKN06-FINAL-2Team
- 작성 팀원 : 박미현, 안형진

개요	**1. 개요** 1.1 목적 ● 본 프로젝트는 웹툰과 웹소설을 사용자 맞춤형으로 추천해주는 대화형 챗봇을 개발하는 것을 목표로 한다. ● LLM을 활용하여 사용자의 취향을 분석하고, 방대한 콘텐츠 데이터 속에서 적합한 작품을 제안함으로써 사용자 경험을 향상 ● 사용자의 선호도(장르, 키워드, 평점, 선호 플랫폼, 댓글 트렌드)를 분석해 맞춤형 웹툰·웹소설을 추천함으로써 콘텐츠 탐색 시간 단축 ● LLM 과 RAG 기술 결합을 통해 기반 추천 정확도 향상 ● 성격이 다른 5개의 모델을 통해 사용자가 추천을 받거나 챗봇을 이용할 때 만족도 향상.

데이터 전처리	**2. 데이터 전처리** 2.1 입력데이터 ● 사용자 데이터 : 플랫폼에서 제공하는 사용자가 입력한 선호도(최초 취향, 피드백) 데이터, 사용자의 정보 입력. ● 콘텐츠 데이터 : 웹툰 및 웹소설의 제목, 줄거리 요약, 장르, 작가 정보, 평점(좋아요, 조회 수, 별점, 댓글 수를 바탕으로 책정한 점수 데이터) 등 메타데이터. 2.2 전처리 단계 ● 노이즈 제거 : 특수 문자/중복 데이터 필터링 (정규식 활용), 무의미한 데이터 삭제 ● 표준화 · 정규화 : 텍스트 소문자 변환, 숫자, 날짜 표기 통일 (예: "2025.02.10.점" → "2025.02.10") ● 결측값 처리: 플랫폼마다 제공하는 데이터 차이로 인한 데이터 불균형 존재. ○ 카카오페이지: 관심 수 없음 ○ 카카오웹툰: 별점 없음 ○ 네이버 시리즈: 조회수 없음, 랭킹의 기준: 별점, 판매 기준으로 작성되었으며 자세한 기준 알 수 없음 ○ 네이버 웹툰: 조회수 없음 →데이터를 정규화하여 score데이터 생성. LLM은 생성된 score를 기준으로 사용자에게 추천. ● 이상치 처리 : 조회수나 좋아요 수의 경우는 극단적으로 높은 경우가 있어 4분위수를 최고치로 하여 대체하거나 로그 변환을 이용하여 데이터의 치우침을 방지(데이터가 양극화되면 LLM이 높은 점수의 데이터만 추천할 확률이 있기 때문)하여 다양한 추천 지향. ● 임베딩 생성: LangChain을 이용해 텍스트를 벡터로 변환 후, ChromaDB에 저장 ● 임베딩 모델 : "BAAI/bge-m3” ○ 해당 모델의 사용 이유: “text-embedding-3-small”,”text-embedding-ada-002”,”Bespin SRoBERTa” , "BAAI/bge-m3”, "nomic-embed-text”,” bge-m3-korean” 모델을 사용하여 임베딩 하여 연관 있는 데이터를 가져오는 경우의 수를 측정하여 점수화 한 결과 "BAAI/bge-m3”이 300으로 제일 높았고, 그 다음은 “bge-m3-korean”이 270, ”Bespin SRoBERTa”가 165 등으로 측정되어 점수가 제일 높은 "BAAI/bge-m3” 최종 채용. ● 출처 : 공개 웹툰 플랫폼(네이버웹툰, 카카오웹툰, 카카오페이지 웹툰, 네이버 시리즈 웹소설, 카카오페이지 웹소설)의 크롤링 데이터

기대효과	**3. 기대효과** ● 사용자 취향에 맞추어 추천된 웹툰과 웹소설을 쉽게 발견할 수 있어 콘텐츠 소비가 증가할 것으로 예상. ● LLM의 시맨틱 분석을 통해 사용자 쿼리를 더 잘 이해하고, 정확한 검색을 수행하여 “판타지 장르 중 능력 먼치킨 캐릭터의 성장형 스토리"와 같은 세부 선호도에 반영. ● LLM의 자연어 처리 능력을 통해 보다 정확하고 개인화된 추천 제공. 신규 사용자에게 인기 작품과 유사도 기반 추천 ● 플랫폼 내 사용자 체류 시간 증가 및 만족도 향상. ● 추천을 세분화하여(로맨스 모델은 로맨스 추천, 판타지 모델은 판타지 작품 안에서 추천) 사용자가 기대하는 취향과 동떨어진 추천 방지 ● 다양한 성격을 가진 모델을 사용하여 사용자가 여러 캐릭터와 대화하는 것과 같은 효과
도구 및 환경	**4. 도구 및 환경** ● LLM 모델: GPT-4o ● 프로그래밍 언어: Python. ● 프레임워크 및 라이브러리: LangChain (RAG 통합), Hugging Face Transformers, ChromaDB, ChatOpenAI, DJango, MySQL, AWS ● 데이터베이스: SQL을 Amazon RDS에 연동하여 웹툰 작품 정보 및 사용자 데이터를 효율적으로 저장 및 관리. ● 개발 환경: AWS 클라우드 기반 컴퓨팅 자원 활용.

결론	**5. 결론** 5.1 주요 성과 ● SQL과 RDS에 데이터 저장 ● 콘텐츠 데이터에 따른 추천 시스템(score) 구현 ● agent를 이용하여 사용자의 의도 파악 후 그에 알맞은 답변 생성 ● LLM이 추천 시스템을 인식하여 추천에 반영 5.2 향후 계획 ● 챗봇의 개별적 성격을 더욱 뚜렷하게 개발할 예정 ● 데이터 확장: 더 많은 웹소설·웹툰 신규 데이터셋 추가 ● 강화학습 도입: 사용자의 피드백을 반영한 추천 알고리즘 개선 ● 한번 추천한 작품은 DB에 저장하여 재추천하지 않게 할 예정