

SK네트웍스 Family AI과정 3기  
 데이터 수집 및 저장 데이터 수집 보고서



□ 개요

* 산출물 단계 : 데이터 수집 및 저장
* 평가 산출물 : 데이터 수집 보고서
* 제출 일자 : 2024.11.14
* 깃허브 경로

: [https://github.com/SKNETWORKS-FAMILY-AICAMP/SKN03-FINAL-2Team](https://github.com/SKNETWORKS-FAMILY-AICAMP/SKN03-FINAL-3Team)

* 작성 팀원 : 이준석

| 데이터 수집 목적 | **본 프로젝트는 2030을 대상으로 뮤지컬과 전시회를 LLM을 이용해 추천하는 목표로 한다.**   * 뮤지컬의 경우: DeepFM 모델을 사용한 배우, 장르 기반의 사용자 맞춤 뮤지컬을 추천한다. (y축 - 예매율) * 또한, RAG를 활용하여 배우 유사도와 내용 유사도를 분석하고, LLM(Large Language Model)을 통해 사용자에게 유사한 배우나 작품을 추천한다. * 전시회의 경우: OCR모델을 사용해 상세정보의 이미지 를 텍스트로 변환하여 이미지에서 추출된 텍스트 또는 사용자 텍스트 간의 유사도 검색을 수행하여 사용자에게 유사한 전시회를 추천한다.   **이를 통해 사용자에게 맞춤형 추천 서비스를 제공하고, 문화 예술 분야에서의 사용자 경험을 향상시키는 것이 데이터 수집의 주요 목적이다**. |
| --- | --- |
| 데이터 수집 방법 | **뮤지컬 데이터 수집**   * 웹 크롤링:   뮤지컬의 캐스팅 보드 데이터를 수집   * API 활용:   뮤지컬의 포스터, 제목, 장소, 배우, 제작진, 일시, 런타임 등의 메타데이터 수집   * LLM 활용:   제목을 LLM에 넣어서 줄거리와 장르를 수집  **전시회 데이터 수집**   * 웹 크롤링:   전시회의 포스터, 제목, 장소, 일시, 가격 등의 메타데이터 수집  OCR할 전시회의 상세정보 이미지파일 수집   * API 활용:   플랫폼에서 API를 제공하는 경우, API 키를 발급받아 데이터를 수집 |
| 수집 데이터(요약) | **뮤지컬 데이터**   * 이미지 데이터:   공연 사진: 뮤지컬을 대표하는 이미지 파일.  활용 목적: 공연의 분위기와 스타일을 시각적으로 전달 텍스트 데이터:  1. 공연 정보: 제목, 일시, 위치, 내용, 배우, 가격, 링크 등 2. LLM을 이용한 줄거리 및 장르 정보 3. 캐스팅 보드 데이터 (2차 에자일)   활용 목적: DeepFM Model을 이용한 예매율 예측 및 사용자에게 정보 제공   * 임베딩 데이터:   위치, 런타임, 가격, 라이센스, 장르, 줄거리, 배우, 상영기간, 제작진, 공연일 & 시간  **전시회 데이터**   * 이미지 데이터:   공연 사진: 전시회를 대표하는 이미지 파일.  활용 목적: 전시회의 분위기와 스타일을 시각적으로 전달  상세 정보 이미지: 전시회의 상세 정보를 OCR Model을 이용해 텍스트 추출 텍스트 데이터:  1. 전시회 정보: 제목, 일시, 위치, 내용, 가격, 링크 등 2. OCR Model을 이용해 추출된 전시회 상세 정보 텍스트 3. 화가 및 작품 데이터: 작품의 제작 배경과 의도에 대한 설명 및 작가의 경력 및 주요 작업 배경 등 (2차 에자일)   활용 목적: OCR Model과 RAG Pipeline을 이용한 전시회 추천 및 사용자에게 정보 제공   * 임베딩 데이터:   OCR Model로 추출한 전시회 상세정보 텍스트 |
| 데이터 특성, 품질관리 및 기대기능 | 데이터 특성 및 품질 관리  * 다양성 확보 : 뮤지컬과 전시회 모두 다양한 장르와 시대의 작품을 포함하여, 데이터의 폭을 넓힘으로써 사용자에게 보다 풍부한 선택지를 제공. * 전처리 과정:  1. 데이터 정제: 노이즈 제거, 불필요한 HTML 태그나 특수 문자 제거. 2. 이미지 처리: 해상도 조정 및 포맷 변환. 3. OCR Model을 사용한 텍스트 LLM을 이용해 요약  * 품질 검사  1. 중복 제거: 동일한 데이터가 중복되지 않도록 관리. 2. 정확성 검토: 수집된 정보의 정확성 검증. 3. 최신성 유지: 최신 데이터로 업데이트하여 정보의 유효성을 지속적으로 유지.  기대 기능뮤지컬추천 정확도 증가: DeepFM이 예측한 예매율을 사용하여 추천 정확도 상승개인화 추천: RAG와 LLM을 활용하여 선호 배우와 유사한 작품을 추천  1. 대화형 인터페이스로 누구나 친근하게 사용할 수 있는 추천 시스템  전시회이미지, 텍스트에 관계 없이 사용자의 입력쿼리와 유사한 전시회 추천개인화 추천: RAG와 LLM을 활용하여 사용자 입력쿼리에 알맞는 전시회 추천  1. 대화형 인터페이스로 누구나 친근하게 사용할 수 있는 추천 시스템 |