FILM: FInd with LLM

LLM(Large Language Model)과 RAG(Retrieval-Augmented Generation)를 활용한 영화 추천 서비스 개발

프로젝트 데모 비디오: https://youtu.be/cLDNgXLfgQU

Github 코드: https://github.com/krx0896/LLM-and-RAG-based-Movie-Recommendation-System_Upstage-LLM-Innovators-Challenge

ID SQUARE LAB

고려대학교 KOREA UNIVERSITY

고려대학교 산업경영공학과 김윤성 (yunseongkim@korea.ac.kr)
고려대학교 산업경영공학과 이창호 (lch0783@korea.ac.kr)
고려대학교 산업경영공학과 이준혁 (leejh9787@korea.ac.kr)
고려대학교 산업경영공학과 김지환 (kjh4295233@korea.ac.kr)

프로젝트 개요 및 목표

프로젝트 개요 (Project Overview)

- RAG (Retrieval-Augmented Generation) 및 LLM (Large Language Model) 기술을 활용한 개인화된 영화 추천 시스템을 개발하는 것을 목표로 함.
- 해당 시스템은 사용자의 자연어 쿼리와 요구사항을 바탕으로, 벡터 스토어에서 영화 데이터 임베딩을 검색하고 LLM을 거쳐 **가장 적합한 영화 3편 이상**을 추천함.
- 사용자는 단순한 장르나 태그 기반 추천이 아닌, 캐릭터 성장, 미래적 배경, 특정 영화 요소 등과 같은 **고급 요구사항에 맞는 영화를 추천** 받을 수 있음.

프로젝트 목표 (Project Objectives)

- **개인화된 영화 추천 시스템 개발**: LLM을 통해 사용자의 쿼리와 요구사항을 정확하게 이해하고 처리하여 복잡한 조건에도 맞춤형 영화를 추천하는 시스템 구축.
- **설명 기반의 추천 제공**: 해당 영화가 추천되었는지 구체적으로 설명하여 사용자 신뢰도와 만족도를 높이는 투명한 추천 시스템 구축.
- **확장 가능하고 실시간으로 동작하는 시스템 설계**: 최신 영화나 새로운 영화 데이터가 추가될 때마다 벡터 스토어에 효율적으로 임베딩을 추가하여 LLM 학습 과정 없이 데이터 업데이트 및 검색 성능을 유지.
- **RAG 기반 검색 및 추천**: 사용자 요구와 가장 유사한 영화를 찾기 위한 효율적인 정보 검색 프로세스 구축.
- **상업적 가능성 검토**: 시스템을 OTT 플랫폼이나 영화 큐레이션 서비스 등에 적용할 수 있도록 상업화 가능성 탐구.

문제 정의 및 해결 방안

문제 정의 (Problem Definition)

- **기존 추천 시스템의 문제점**
  - 기존 노래 추천 시스템은 *collaborative filtering (CF) 와 **content-based models (CBM)을 사용하여 사용자의 **맥락적 정보 반영이 어려움**(Mayank, M., 2023).
  - 전통적인 추천 시스템에서는 아티스트 이름, 노래 제목, 앨범 제목(태그 수준)과 같은 미리 정의된 태그를 검색에 활용함. 태그 수준 입력을 사용하여 음악을 검색할 때 **데이터베이스에서 지정된 태그가 없는 음악을 검색하는 것이 어려움**(Doh, S., Won, M., Choi, K., & Nam, J., 2023).
- **LLM의 문제점**
  - LLM은 실제 데이터에 근거하지 않은 허구의 정보나 잘못된 정보를 생성하는 **Hallucination 문제로 영화 추천 과정에서 신뢰할 수 없는 정보를 제공**할 가능성이 있음.
  - LLM은 주기적인 학습이 필요하며, 새로운 영화나 콘텐츠가 추가될 때마다 **실시간으로 학습시키는 것이 어렵고**, 이는 **상당한 시간과 비용이 소모**됨.

해결 방안 (Solution Approach)

- **LLM 활용을 통한 기존 추천 시스템 문제점 보완**
  - LLM을 추천 시스템에 통합함으로써, 제한된 데이터 가용성 시나리오에서도 **맥락적으로 관련성 있는 추천으로 개인화된 제안을 제공**할 수 있음(Di Palma, D., 2023).
  - LLM은 **자연어를 이해하는 탁월한 능력을 보유**하고 있어 사용자 선호도, 항목 설명 및 맥락 정보를 이해하고 보다 정확하고 관련성 있는 추천을 생성할 수 있음. 이는 **사용자 만족도와 참여도를 향상시키는 결과**를 가져옴(Hua, W., Li, L., Xu, S., Chen, L., & Zhang, Y., 2023).
- **RAG 활용을 통한 LLM의 문제점 보완**
  - **정확한 정보 기반 추천**: RAG 기반 추천 시스템은 기존의 정보 검색 시스템보다 다양한 정보를 제공할 수 있으며, 이를 통해 LLM의 성능을 효과적으로 향상 가능(Contal, E., & McGoldrick, G., 2024). 또한 실제 영화 데이터를 검색한 후 이를 바탕으로 LLM이 결과를 생성하는 방식으로 **Hallucination 문제를 줄일 수 있음**(Béchard, P., & Ayala, O. M., 2024).
  - **효율적인 리소스 관리 및 최신 데이터 반영**: RAG는 **대규모 학습 과정을 줄이고**, 검색된 실제 데이터를 활용하여 LLM이 필요한 정보만을 생성하기 때문에 **재학습 없이도 최신 추천을 유지 가능**(Gao, Yunfan, et al., 2023).

*collaborative filtering (CF): 협업 필터링은 유사한 사용자의 청취 행동 및 평가를 기반으로 사용자의 음악 선호도를 예측하는 방식으로 작동
**content-based models (CBM): 콘텐츠 기반 모델은 음악 자체의 속성을 중심으로 노래의 리듬, 피치, 장르 등 음향적 특성을 분석하여 음악을 추천

기술적 특징

추천 시스템 기술적 설명

LLM Method

Foundation LLM

- 사용자 프롬프트에 대한 맥락 정보를 잘 반영할 수 있는 Upstage solar-pro LLM을 API로 불러와 Foundation LLM으로 사용하고자 함.

Prompt Engineering

- LLM의 명확한 역할을 영화 추천 시스템으로 설정하고 관련 영화 정보에 대한 context를 제공하여 프롬프트의 구조, 내용, 형식을 조정하여 원하는 응답하도록 유도함.

RAG Method

Knowledge Source

- 사용자 query에 만족하는 영화 retrieval을 하기 위해 영화 'Description', 'Review' 변수를 가지는 IMDb 영화 데이터셋을 사용하여 Pinecone vector store에 저장하고 API를 통해 사용함.

Embedding Model

- 사용자 query에 대해 최적의 영화를 retrieval 하기 위해 영화 변수인 'Title', 'Year', 'Genre', 'Description', 'Director', 'Cast', 'Review' 의 텍스트 데이터를 Upstage solar-embedding-1-large-passage, 사용자 query를 Upstage solar-embedding-1-large-query 모델을 활용하여 임베딩함.

RAG + LLM

Lang-Chain

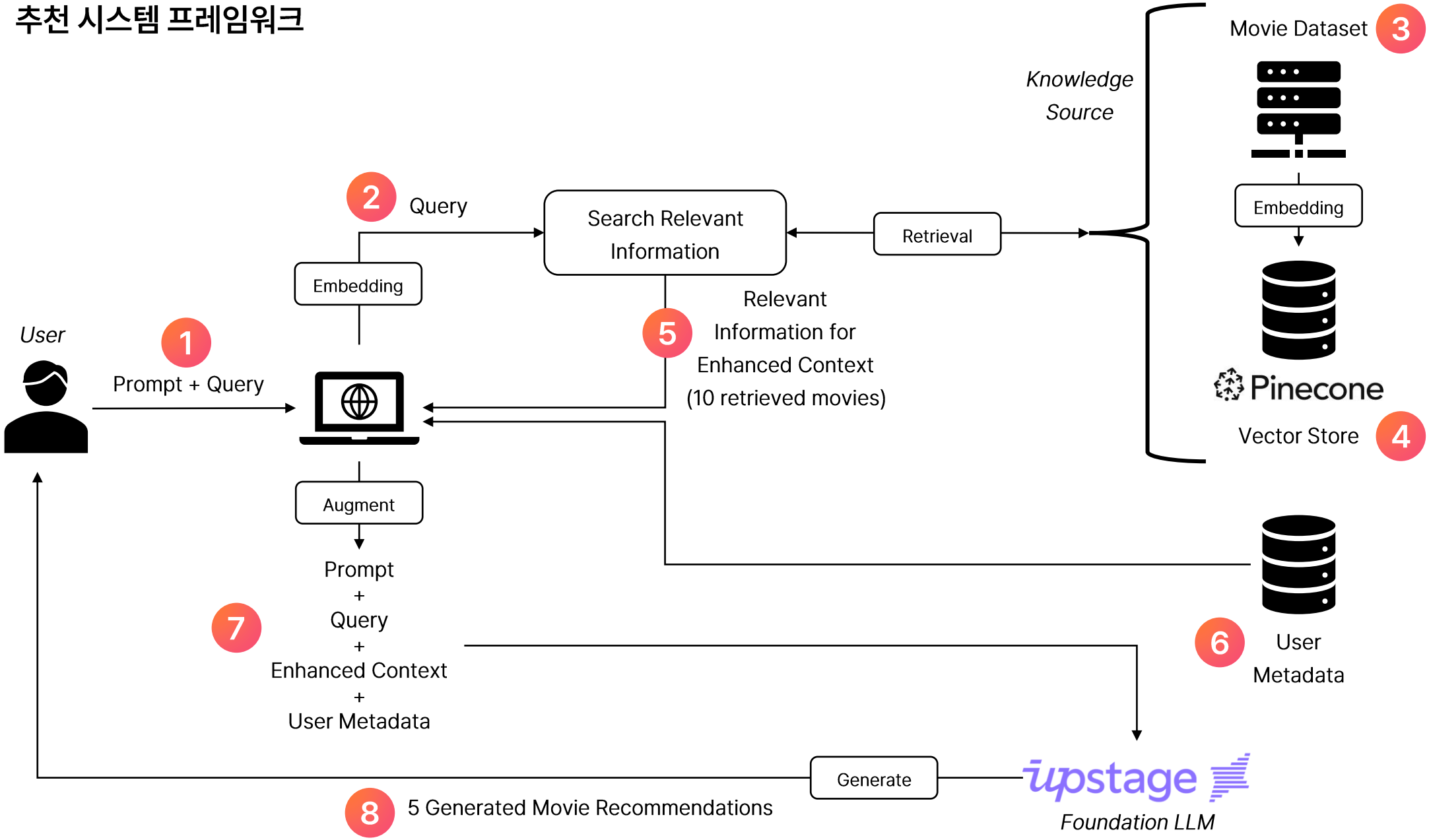
- RAG의 vector store와 사용자 query에 대한 임베딩 모델, 추천 결과 생성을 위한 LLM을 통합하는 언어 모델 파이프라인을 구축하고, 효율적으로 운영하고자 함.

Process

- 사용자가 영화 추천에 대한 요구 사항을 query로 입력하면, 먼저 벡터 스토어에 저장된 영화 설명 임베딩과 비교하여 코사인 유사도를 기반으로 가장 유사한 상위 10개의 영화가 1차적으로 추려짐. 그 후, 이 영화들의 'Title', 'Genre', 'Description', 'Review' 정보와, 사용자 query와 LLM의 역할을 포함한 프롬프트가 결합되어, 최종적으로 가장 적합한 3편 이상의 영화를 선택하고 간단한 설명과 함께 추천함.

기술적 특징

추천 시스템 프레임워크



기술적 특징

추천 시스템 결과 확인

미래적 배경을 가진 SF 영화 중에서 캐릭터의 성장이 돋보이는 작품을 추천해 주세요.

제목: 아카이브 (Archive) - 2023년에 개봉한 SF 영화로 미스터리 요소를 담고 있으며, 진정한 인간에 가까운 AI를 연구하는 조지 알마의 여정을 그립니다.

제목: 블레이드 러너 (Blade Runner) - 2023년에 개봉한 SF 영화로 드라마 요소를 담고 있습니다. 지구로 도망친 네 명의 복제인을 추적하고 제거해야 하는 블레이드 러너의 이야기를 다룹니다.

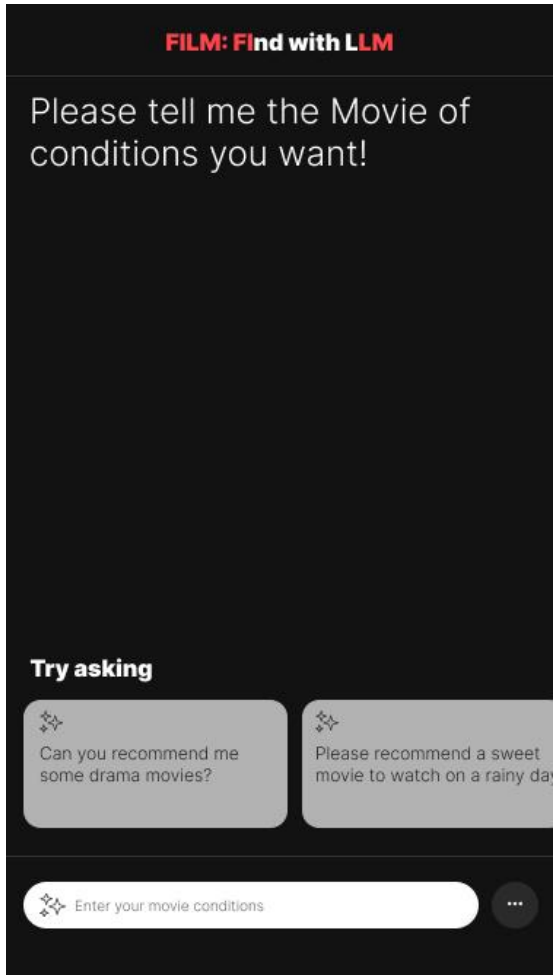
제목: 사이언스 픽션 볼륨: 오시리스 시그너스 (Science Fiction Volume: The Osiris Cygnus) - 1995년에 개봉한 SF 영화로 액션과 모험 요소를 포함하고 있습니다. 이 영화는 좋은 연기와 특수효과, 그리고 탄탄한 스토리와 캐릭터 성장을 보여줍니다.

제목: 유랑지구 (Liu lang di qu) - 1983년에 개봉한 SF 영화로 모험 요소를 담고 있으며, 이 목록에서 유일하게 강한 여성 주인공이 등장하는 영화입니다. 인간을 위한 새로운 행성을 찾는 이야기를 다룹니다.

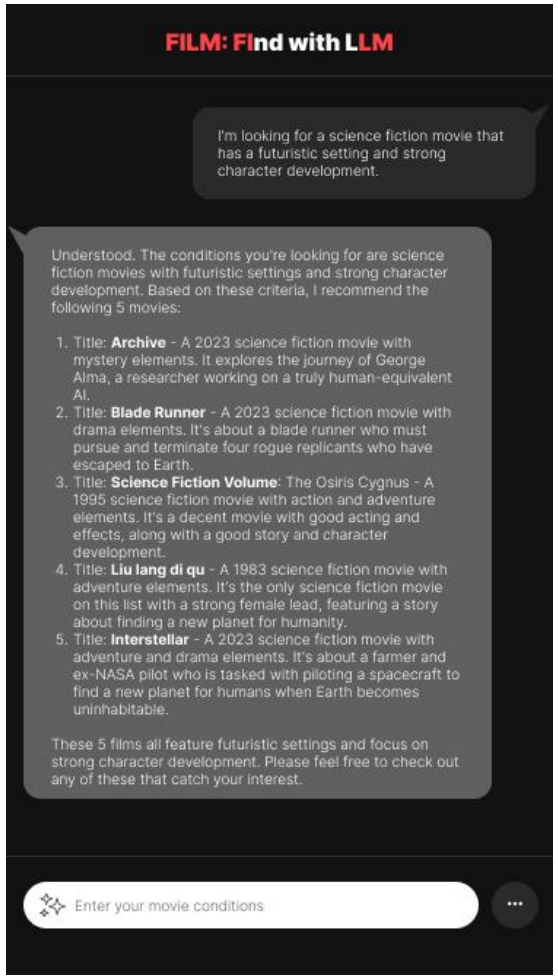
제목: 인터스텔라 (Interstellar) - 2023년에 개봉한 SF 영화로 모험과 드라마 요소를 담고 있습니다. 지구가 살 수 없게 되었을 때, 새로운 행성을 찾기 위해 우주선을 조종하는 임무를 맡은 전직 NASA 파일럿의 이야기를 다룹니다.

주요 기능

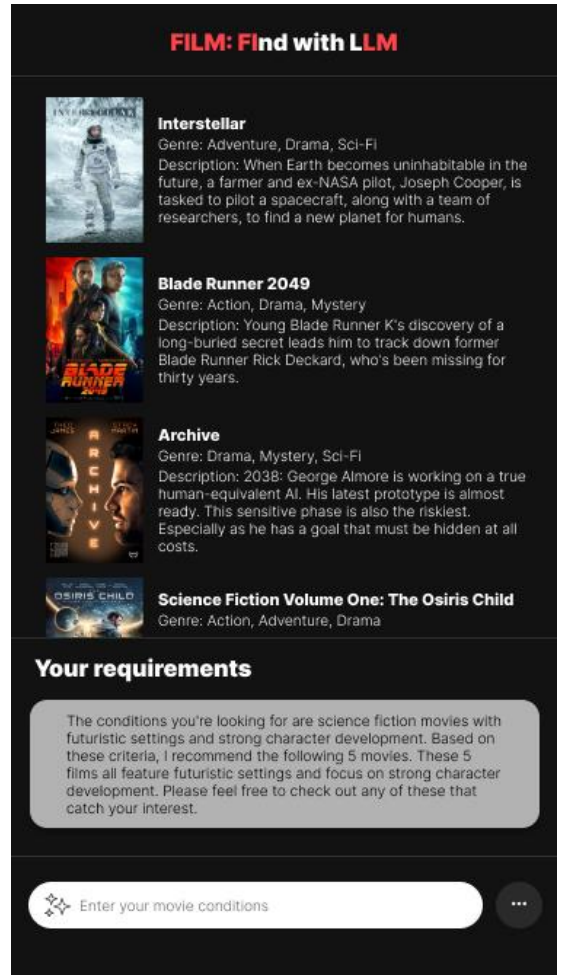
서비스 인터페이스



초기 화면



추천 받을 영화 정보 입력



추천 결과 및 설명

서비스 미래 가치

실용성 (Practicality)

- **개인화된 추천 및 유연한 질의 처리**: LLM을 활용하여 사용자 질의의 맥락을 이해하고, 단순한 태그 기반 추천보다 훨씬 더 유연한 응답하여 개인화된 추천 제공.
- **실시간 정보 반영**: RAG 구조를 통해 벡터 스토어에서 영화 데이터를 실시간으로 검색하여 사용자 질의와 가장 유사한 영화 정보를 제공 가능함.
- **설명 기반 추천**: 추천된 영화에 대해 설명을 제공하여 왜 특정 영화가 추천되었는지, 영화의 특징이 사용자 요청과 어떻게 부합하는지를 명확하게 설명 가능함.

상업성 (Commercial Potential)

- **OTT 플랫폼과의 통합**: 넷플릭스, 디즈니 플러스, 아마존 프라임 같은 대형 OTT 서비스에서 개인 맞춤형 영화 추천을 제공함으로써 사용자의 체류 시간을 늘리고 구독을 유도 가능함.
- **콘텐츠 큐레이션 서비스**: 사용자가 요구하는 특정 영화 주제나 스타일에 맞는 맞춤형 콘텐츠를 제공하는 유료 큐레이션 전문 서비스로 확장 가능함.
- **인터랙티브 추천 엔진**: 사용자와 대화형으로 상호작용하며 추천하는 시스템은 엔터테인먼트 웹사이트, 영화 포털, 혹은 심지어 모바일 앱에서도 상업적 가치를 발휘 가능함.

향후 발전 계획 (Future Development Plans)

- **다양한 도메인 적용 가능**: 영화 이외에도 해당 추천 시스템을 사용하여 다른 콘텐츠(책, 음악 등)로도 확장이 가능함.
- **다양한 데이터 소스 통합**: 현재는 영화 데이터베이스와 같은 한정된 소스만 사용하고 있지만, 향후에는 SNS, 리뷰 사이트, 트렌드 데이터를 실시간으로 반영하여 더욱 정확하고 동적인 추천이 가능할 것임.
- **멀티모달 추천**: 단순히 텍스트 기반 정보뿐만 아니라, 영화 예고편 영상, 시각적 콘텐츠, 포스터 등 다양한 형식의 데이터를 학습하여 시청자의 시각적 선호도까지 반영할 수 있는 멀티모달 추천 시스템으로 발전할 수 있음.
- **인터랙티브 콘텐츠 추천**: 사용자가 추천받은 영화를 본 후, 추가 질문이나 피드백을 통해 더 나은 맞춤형 추천을 제공하는 방식으로 발전할 수 있음.

References

- Mayank, M. (2023, December 9). Mood based music recommendation system. Medium. https://medium.com/@UTMSBA24/mood-based-music-recommendation-system-5afb8bb90082
- Doh, S., Won, M., Choi, K., & Nam, J. (2023, June). Toward universal text-to-music retrieval. In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1-5). IEEE.
- Di Palma, D. (2023, September). Retrieval-augmented recommender system: Enhancing recommender systems with large language models. In *Proceedings of the 17th ACM Conference on Recommender Systems* (pp. 1369-1373).
- Contal, E., & McGoldrick, G. (2024). RAGSys: Item-Cold-Start Recommender as RAG System. arXiv preprint arXiv:2405.17587.
- Hua, W., Li, L., Xu, S., Chen, L., & Zhang, Y. (2023, September). Tutorial on large language models for recommendation. In *Proceedings of the 17th ACM Conference on Recommender Systems* (pp. 1281-1283).
- Béchard, P., & Ayala, O. M. (2024). Reducing hallucination in structured outputs via Retrieval-Augmented Generation. arXiv preprint arXiv:2404.08189.
- Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., ... & Wang, H. (2023). Retrieval-augmented generation for large language models: A survey. arXiv preprint arXiv:2312.10997.

ID SQUARE LAB

고려대학교 KOREA UNIVERSITY

Q&A

ID SQUARE LAB

고려대학교 KOREA UNIVERSITY