부산대학교 정보컴퓨터공학부

[2025전기] 졸업과제 착수보고서

웹크롤링 기반 블루레이 추천 서비스

담당교수 : 조준수

Lapis blue

202055539 박덕형

202055549 박태준

202055564 안형찬

1.과제 개요

1) 과제 배경:

블루레이는 HD 비디오를 저장하기 위한 디지털 광 기록 방식의 저장매체로, 전 세계적으로 CD나 DVD를 대체할 고화질 매체로 주목받고 있다.

-블루레이 플레이어와 블루레이-

일부에서는 블루레이를 수집하는 것이 하나의 취미로 자리 잡았지만, 한국에서는 여전히 인지도가 낮아 관련 정보를 얻기가 어렵고, 적절한 콘텐츠를 접하기도 쉽지 않다.

이러한 문제를 해결하기 위해 웹 크롤링과 LLM 기술을 활용해 블루레이 관련 데이터를 수집하고 전처리하여, 보다 풍부하고 정확한 정보를 제공하는 추천 서비스를 구축하고자 한다.

2) 과제 목표:

사용자 취향에 맞는 블루레이 추천과 블루레이 구매 사이트 정보를 제공함으로써 한국에서의 블루레이의 접근성을 높이는것을 목표로 한다.

3) 프로젝트 개요:

부족한 블루레이 정보를 보완하기 위해 웹 크롤링을 활용하고, 다양한 사이트에서 수집한 데이터를 기반으로 LLM을 사용해 주요 feature를 자동으로 추출한다. 이렇게 얻은 정보는 기존의 MovieLens 영화 데이터와 결합하여 추천 시스템을 구성하며, 각 블루레이 항목에는 웹 크롤링을 통해 확보한 구매 링크도 함께 제공한다.

2.대상 문제 및 요구조건 분석

1) 기존 문제점 - 특화된 추천시스템의 부재

기존의 블루레이 판매 사이트에서는 단순히 상품 정보를 나열하거나 판매량에 기반한 추천을 제공할뿐, 이용자의 개인 취향이나 선호도에 따른 추천을 해주는 서비스가 부족하다.

따라서 사용자가 상품을 탐색하는데 다양한 사이트에서 정보를 수집해야하는 번거로움이 있고, 원하는 정보를 찾기 까지 많은 시간과 노력이 필요하다.

2) 요구 사항 분석

1. 추천 시스템 구현

- 사용자 활동 기반 추천 리스트 갱신.
- 사용자의 최신 활동(좋아요, 조회, 구매 등)을 기반으로 추천 리스트를 실시간 또는 주기적으로 갱신하여 제공.
- 다양한 요소를 고려한 맞춤형 추천
- 영화의 장르, 배우, 감독, 사용자 평가 내역 등 콘텐츠 정보뿐 아니라 블루레이 상품의 가격, 화질 등의 요소를 함께 고려하여 개인 맞춤형 추천 시스템을 구현.

2. 블루레이 상품 크롤링 및 데이터 처리

- 정보 수집
- 데이터베이스 최신화
  추천 시스템의 정확도를 높이기 위해 데이터마이닝 기반 분석을 주기적으로 수행
- 최소 주 1회 크롤링 및 분석을 통해 최신 블루레이 정보와 IMDB의 평론가/일반 사용자 평점을 반영하여 데이터베이스를 업데이트하고, 추천 알고리즘을 보정.
- 검색 시스템 구현
  수집한 블루레이 정보를 바탕으로 키워드 기반 검색 기능을 제공.
- 정렬 기능 구현
  사용자가 원하는 기준(판매량, 장르, 평가 등)에 따라 블루레이 목록을 정렬할 수 있는 기능 제공.

3. 사용자 관리 시스템

- 회원가입 및 로그인 기능
  사용자 인증을 위한 회원가입 및 로그인 시스템을 구현하여 개인별 서비스 제공.
- 사용자 행동 데이터 수집
  사용자별 조회, 구매, 좋아요, 평가 등의 활동 데이터를 수집하여 추천 시스템에 반영.

3. 현실적 제약 사항

- 현재 서비스에 사용할 유저 기반 데이터가 없다.

  원래 추천 시스템을 개발하기 위해서는 유저들이 취향을 분석할 수 있도록 유저 기반 데이터들을 사용해야 한다. 즉, 유저들의 취향을 분석할 수 있는 데이터들이 있어야 효과적으로 적합한 영화 및 블루레이를 선택할 수 있는 것이다. 그러나 현재 서비스의 시작 단계임에 따라 이러한 데이터들이 축척되지 않았다는 제약사항이 있었다.

  이를 해결하기 위해 MovieLens 영화 데이터셋에 들어있는 유저 기반 평가 데이터셋을 사용하여 처음에 부족한 유저 기반 데이터를 어느 정도 마련하기로 하였다. 이 데이터셋에는 각 유저별로 특정 영화를 평가한 별점 평가가 들어있기 때문에 협업 필터링과 같은 추천 연산 과정에 적합할 것이라는 판단을 하였다.

- 데이터셋이 영화에만 한정되어 있고 블루레이 관련 데이터가 없다.

  결국 우리 서비스는 블루레이를 추천하는 것인데, 영화에 대한 데이터는 많지만 정작 블루레이 관련 데이터, 예를 들면 한화 가격이나 화질(일반, 4K), 한정판 버전 등등에 대한 데이터는 없다. 이러한 정보들은 추천 연산 과정을 거칠 때 필수적으로 들어가는 속성들이다. 이는 실제 추천 연산에 사용되는 값이므로 유저에게 보여지는 값이랑은 다르다. 유저에게 보여주는 값은 실시간 웹스크래핑을 통해 실제 현재 값을 보여줄 예정이므로 이와는 상관이 없다. 이러한 가격이나 화질 등등의 속성들을 가지고 연산을 해야 되므로 데이터셋 구축이 필요한 것이다.

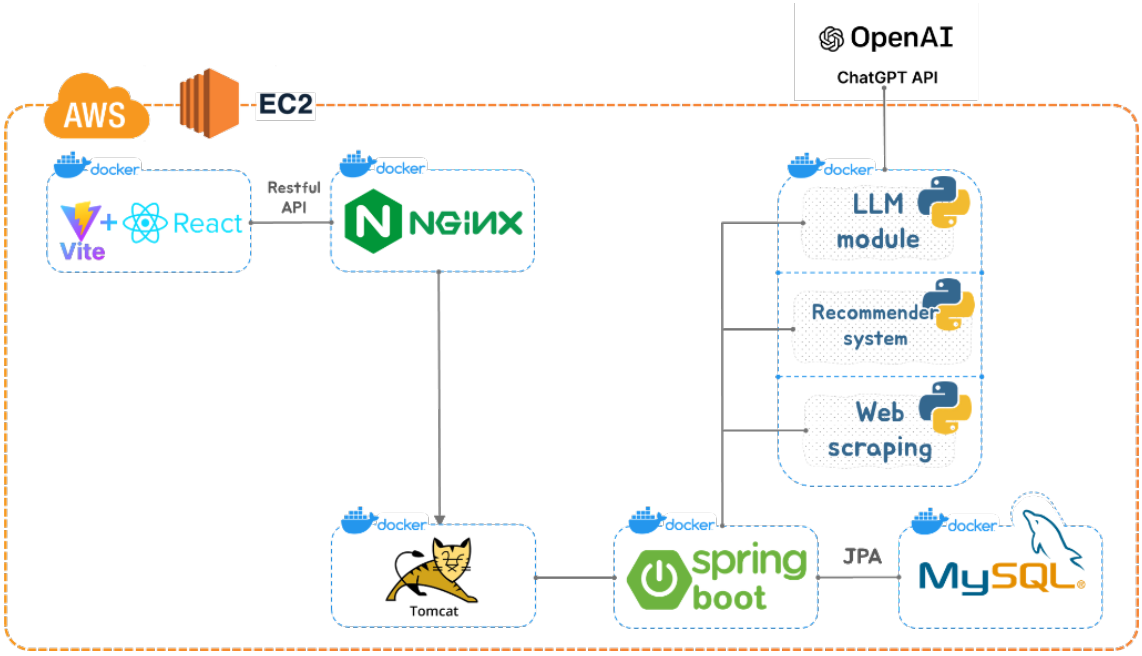
  이를 해결하기 위해서 웹스크래핑을 활용하여 데이터를 축척시키로 하였다. 이를 통해 블루레이 관련 데이터셋을 어느 정도 축척시킬 수 있을 뿐만 아니라 실제 거래되는 가격을 구할 수 있는 장점도 있다. 그러나 축척해야 할 데이터 양이 많은 만큼 서비스 시작 전에 틈틈히 데이터를 계속 추가해 나갈 예정이다.

- 웹스크래핑을 통해 블루레이 정보를 축척해도 가격에 대한 데이터가 모호하다.

  웹스크래핑으로 가격 데이터들을 가져온다 하더라도 여러 상점 사이트들을 기반으로 하는 데이터의 가격이 많이 상이할 수도 있다. 그러면 데이터셋에 들어갈 가격이 모호해진다는 문제점이 있다. 앞에서 말했듯이 이 가격은 추천 연산 용이므로 유저에게 보여지는 가격은 웹스크래핑 기반 실시간 가격이다. 그러나 추천 시스템의 연산 요소로 계속 변동되는 값을 넣을 수 없으므로 가격 속성이 모호하다는 문제가 있는 것이다.
  이는 가격의 평균 등 특정 대푯값을 연산하여 데이터셋에 기입하는 방법이 있다. 이렇게 되면 추천 연산에 모호한 값이 아닌 정형화 된 값을 사용할 수 있다는 면에서 효과적일 것이라고 생각이 된다.
  또한 가격에 대한 변동이나 할인 이벤트와 같은 요소들에 둔감할 수도 있는 문제점이 있다. 이는 틈틈히 웹스크래핑을 통해 데이터셋 자체를 최신화 할 수 밖에 없다. 물론 앞에서 작성했듯이 유저에게는 실제 할인값 등이 보여져서 문제가 없을 수도 있으나, 추천 연산 과정에서는 이러한 값이 사용되지 않는다는 문제가 여전히 남아 있다.

4. 설계 문서

1) 설계 구조도



1.프론트엔드:

기술 스택: React.js, vite

- 사용자 인터페이스 제공 (로그인, 추천 목록, 상품 정보 등)
- 추천 요청 및 결과 표시
- 사용자 클릭/피드백 데이터 수집
- API 서버와 통신하여 백엔드 연동

2.백엔드:

- 기술 스택: NGINX
- React.js에서 요청한 RESTful API를 분석
- HTML, CSS, JavaScript와 같은 정적 요청이라면 직접 서빙
- 데이터베이스에 접근하는 동적 요청이라면 Spring Boot 내장 Tomcat에 리버스 프록시

- 기술 스택: Spring Boot
- RESTful API 구현: JSON 기반 데이터 교환 방식 채택
- JPA를 통한 데이터 액세스 레이어 구축
- 내장 Tomcat 서버 활용

3.인프라 구조:

기술 스택: Docker

- 프론트엔드, 백엔드, 데이터베이스를 독립적인 컨테이너로 관리
- 각 서비스의 독립적 개발 및 테스트 환경 구축 가능
- 개발자 간 동일한 환경에서 작업 가능

4.데이터베이스

기술 스택: MySQL

- 사용자 ID/PW 및 평가 정보 저장
- 상품 정보, 카테고리, 추천 로그 저장

2) 사용 기술

1.LLM 모듈

기술 스택: Python, OpenAI API

- 웹스크래핑에서 가져온 순수 HTML 코드를 LLM이 해석해서 정리
- 상품 설명 기반 자연어 이해
- LLM에게 질문 → 응답 결과를 프론트로 전달

2.웹 스크래핑

기술 스택: Python

-학습을 위한 웹 스크래핑

- MovieLens에 없는 가격 정보를 수집
- 최신영화에 대한 정보를 수집

-상품 정보를 위한 웹 스크래핑

- 추천 결과에 따른 블루레이 상품 정보를 수집
- 상품명, 가격, 이미지, 상세 설명 등 추출
- 웹스크래핑된 데이터는 DB에 저장

3.추천 시스템

기술 스택: Python

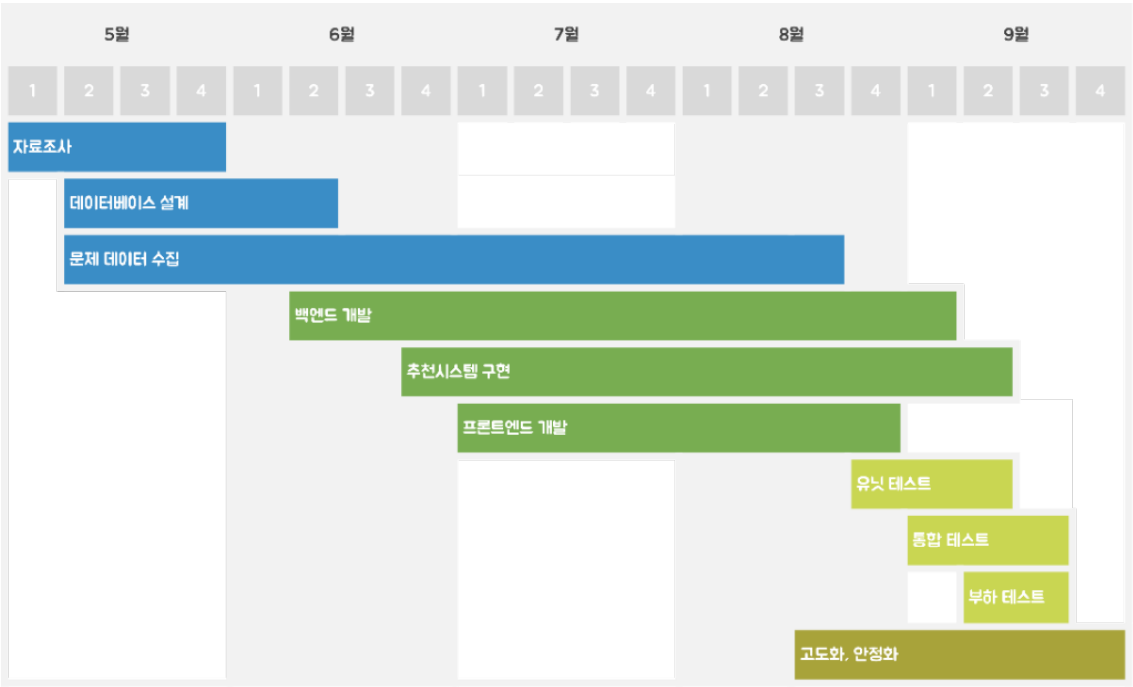
- 상품 데이터, 사용자 데이터 등을 사용하여 적합한 블루레이 제품을 추천
- MovieLens 데이터를 활용
- 웹스크래핑으로 MovieLens 데이터셋에 대응하는 블루레이 정보를 수집
- 협업 필터링 기반 영화 유사도 계산 (MovieLens 기반)
  → 사용자 간 혹은 영화 간 유사도를 DB 기반으로 계산
  → 유사도가 높은 영화 리스트 확보

- 해당 유사 영화 리스트를 기반으로 블루레이 후보 셋 생성
  → 각 영화에 해당하는 블루레이들(에디션 포함)을 매핑
- 블루레이 특성을 활용한 콘텐츠 기반 필터링 실행
  → 가격, 화질, 한정판 여부, 재고, 출시일, 패키징 정보 등 고려
  → 사용자 선호 정보와 매칭하여 최종 추천

3) 시스템 흐름

- 사용자는 웹 브라우저를 통해 ReactJS로 구현된 블루레이 추천 서비스에 접속
- 사용자의 요청이 정적 요청이라면 nginx가 처리하고, 동적 요청이라면 nginx에서 Spring Boot 내장 tomcat 서버로 전달
- 사용자 요청 처리에 필요한 데이터를 mysql 데이터베이스에서 가져옴.
- 서버에서 LLM, 웹 스크래핑, 추천 알고리즘 등의 Python 스크립트들을 활용해, 불러온 데이터를 기반으로 상황에 적합한 연산을 수행함
- 연산 결과는 다시 사용자에게 전달되어 웹 페이지에 표시됨

5. 추진 체계 및 일정 - 간트 차트



6. 구성원 역할분담

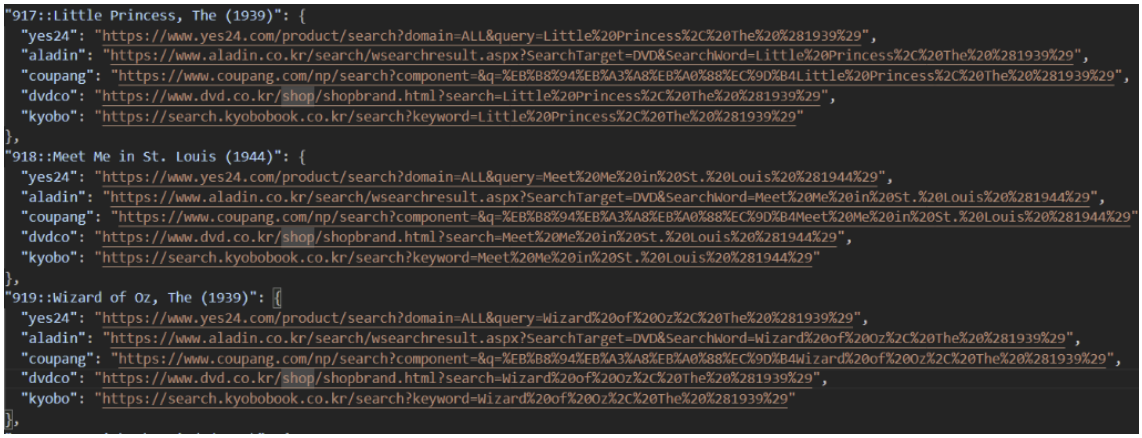
이름	역할
박덕형	스프링 부트 개발, 웹스크래핑, 서버 인프라 구성
안형찬	프론트 엔드, LLM, RESTful API 관리
박태준	추천시스템 연구 및 설계, DB설계 및 구축

7. 구성원별 진척상황

박덕형	**MovieLens Dataset 전처리** - MovieLens Dataset 구조 분석 - 영화별 ID, 제목, 장르 정보를 체계적으로 추출 및 가공 *웹크롤링 및 JSON화 툴은 Python으로 수행함 **블루레이 정보 시험적 웹크롤링** - MovieLens dataset과 블루레이 매핑을 위하여 시험적인 블루레이 상품 정보 웹크롤링 실행 - yes24, aladin, coupang, dvdco, kyobo 등 국내 주요 블루레이 판매 사이트들의 블루레이 정보 탐색 - 상품 페이지에 있는 정보들을 바탕으로 웹크롤링 방식을 수정 및 관리 - 효과적인 웹크롤링 방식을 설계 및 탐색 **블루레이 상품 페이지 JSON화** - yes24, aladin, coupang, dvdco, kyobo 등 국내 주요 블루레이 판매 사이트의 검색 URL들을 분석 - 이러한 URL들과 앞서 전처리한 MovieLens Dataset을 활용하여 영화 Id, 영화 이름, 영화 출시년도 정보를 사용해 웹스크래핑에 필요한 URL들을 생성 - 이러한 URL들을 관리하기 편리하도록 JSON 형식으로 정리 - 사이트 접속에 문제가 없도록 URL 인코딩 작업 - 이를 통해 MovieLens의 영화 데이터와 실제 블루레이 상품 정보를 효과적으로 매핑할 수 있는 기반을 마련
안형찬	**프론트엔드 프로토타입 구축** - React + Vite 환경을 기반으로 블루레이(Blu-ray) 상품 정보를 시각적으로 보여주는 프로토타입 페이지를 구현 - 초기 단계에서는 정적 데이터를 활용했으며, 추후 백엔드 또는 API

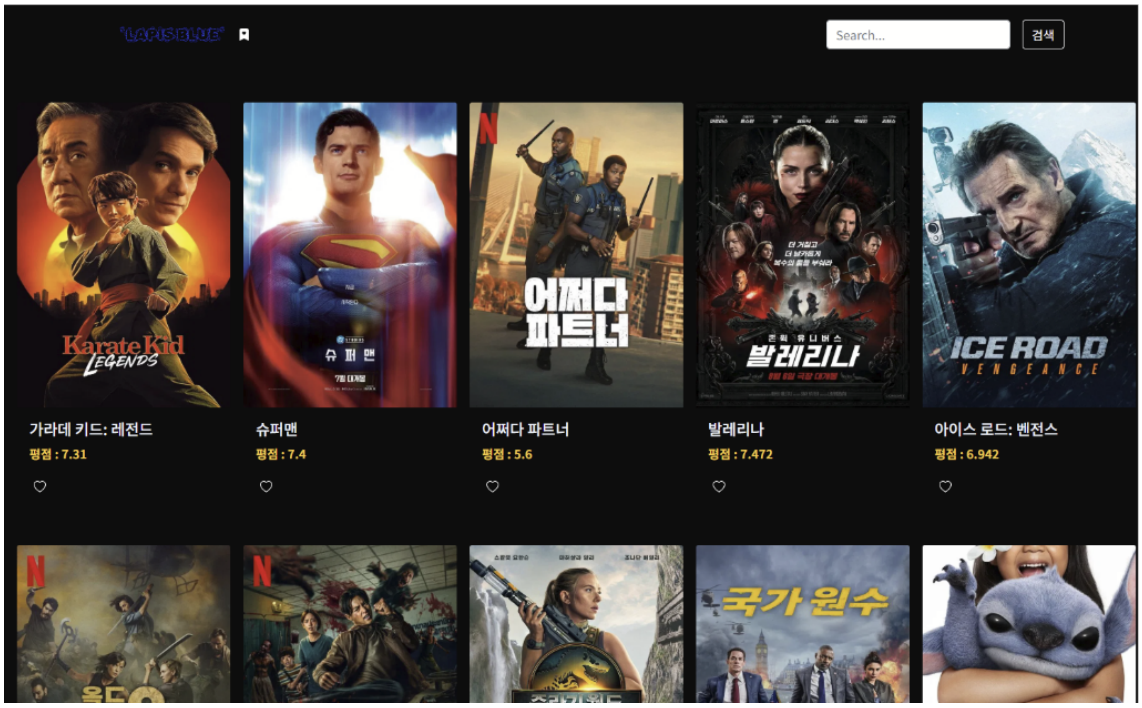
	연동을 통해 동적 데이터로 확장 필요 **LLM API 설계** - 전달받은 상품 상세 URL에서 필요한 상품 관련 정보만을 자동으로 추출 하는 코드 작성 - 입력 데이터의 복잡성, 구조의 비일관성을 고려하여 GPT-4.0 mini 모델을 사용 - LLM API 를 활용하여 상품명, 가격, 설명, 사양 등 핵심 정보만을 추출하여 JSON 객체 형태로 정제 - LLM 결과에서 형식에 맞지 않는 부분을 정리하여 하나의 JSON 파일로 병합 - 정보 추출 코드와 ,API는 다양한 도메인의 상품 페이지에서도 유사한 방식으로 활용 가능
박태준	**백엔드 인프라 구축** - 백엔드 구축을 위해 AWS에서 EC2 server를 free tier로 대여 - 추후 용량이 더 필요하다면 tier를 올릴 계획임 **데이터베이스 구축** - 블루레이 메타데이터 및 사용자 활동 기록을 효율적으로 관리하기 위한 데이터베이스 스키마 설계 **CSV 파일 기반으로 데이터베이스에 삽입** - Python을 활용하여 CSV 형식의 데이터를 MySQL 데이터베이스에 삽입하기 위한 SQL 변환 모듈 구현 - 장르와 같이 다중 값을 갖는 속성을 CSV 파일로 처리하는 데 한계 발생 - 이를 해결하기 위해 블루레이 정보를 JSON 형식으로 표현하기 표현하기로 하였고 이를 위해 코드를 수정 및 개선 중

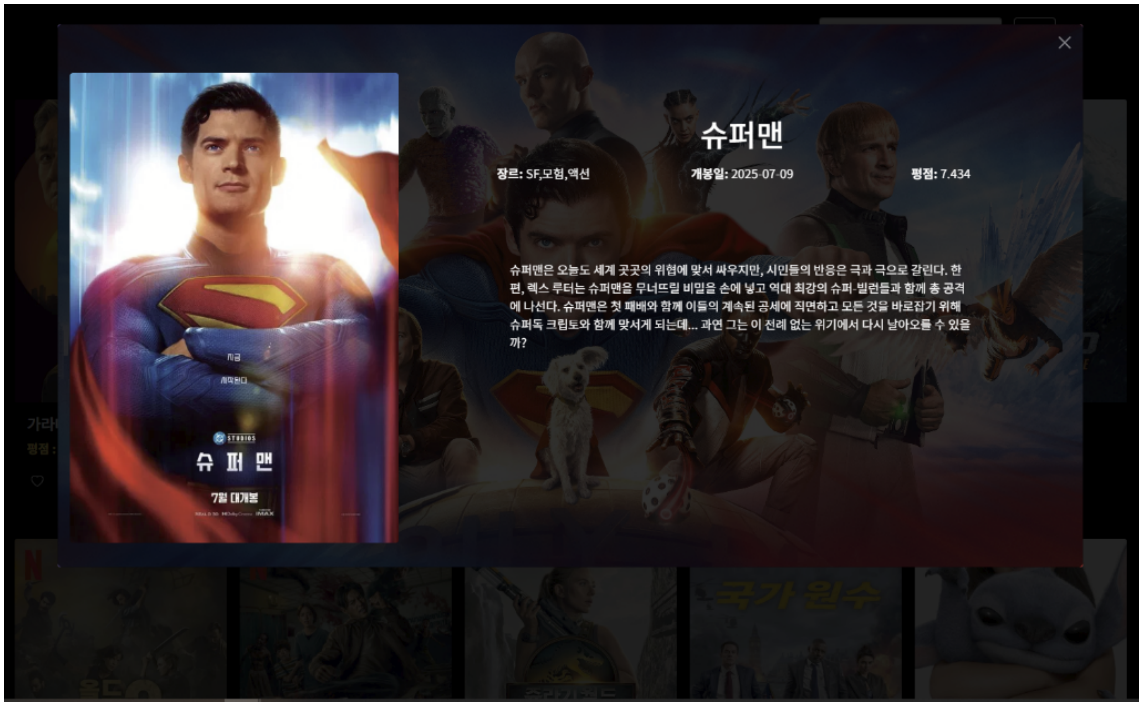
8. 과제 수행 내역 및 중간결과



<URL JSON파일>

<GPT 결과 JSON파일>


LAPIS BLUE
Search...
검색
가라데 키드: 레전드
평점 : 7.31
슈퍼맨
평점 : 7.4
어쩌다 파트너
평점 : 5.6
발레리나
평점 : 7.472
아이스 로드: 벤전스
평점 : 6.942
국가 원수


슈퍼맨
장르: SF,모험,액션
개봉일: 2025-07-09
평점: 7.434
슈퍼맨은 오늘도 세계 곳곳의 위협에 맞서 싸우지만, 시민들의 반응은 극과 극으로 갈린다. 한편, 렉스 루터는 슈퍼맨을 무너뜨릴 비밀을 손에 넣고 역대 최강의 슈퍼-빌런들과 함께 총 공격에 나선다. 슈퍼맨은 첫 패배와 함께 이들의 계속된 공세에 직면하고 모든 것을 바로잡기 위해 슈퍼독 크립토와 함께 맞서게 되는데... 과연 그는 이 전례 없는 위기에서 다시 날아오를 수 있을까?

<프론트 프로토타입>

9. 추진 계획

1) 추천 시스템 개발

- 협업 필터링(Collaborative Filtering) 기반의 추천 알고리즘을 구현하여 사용자 간의 선호도를 분석하고 유사한 취향의 콘텐츠를 추천하도록 설계

- 블루레이 상품의 가격, 화질, 장르 등의 특성을 반영한 콘텐츠 기반 필터링 (Content-based Filtering) 로직도 함께 개발하여, 사용자 취향에 맞는 정교한 추천이 가능하도록 구현

- 실시간 사용자 행동 데이터를 기반으로 추천 결과를 지속적으로 업데이트하며, 추천 정확도와 반응성 테스트를 통해 알고리즘의 성능을 검증

2) 백엔드 프로토타입 및 시스템 구축

- Spring Boot 기반의 RESTful API 서버를 구축하여 안정적인 백엔드 환경 마련

- 서비스 로직, 데이터베이스 연동, 권한 관리 등을 포함한 기본 백엔드 기능 구현

- 환경 간 일관성을 유지하고, 배포 및 유지보수 효율성을 높이기 위한 도커 기반 백엔드 환경 구성

3) 프론트엔드/백엔드 통합 개발

- React.js를 활용한 사용자 인터페이스(UI) 개발을 통해 직관적이고 반응성 높은 화면 구성

- NGINX 및 Docker를 활용해 프론트엔드와 백엔드, API 서버 등의 요소들을 통합하고, 전체 시스템에 대한 연동 테스트 및 배포 자동화 수행

4) 서비스 고도화 및 안정화

- 초기 사용자 피드백을 수집하고 이를 기반으로 UX/UI 개선 및 기능 보완

- 추천 알고리즘의 정밀도 향상과 데이터 품질 개선을 위한 지속적인 튜닝 작업 진행

- 트래픽 증가, 다양한 사용자 환경에 대응할 수 있도록 성능 최적화 및 서버 안정성 강화

- 보안 취약점 대응, 개인정보 보호 강화를 위한 기술적·관리적 조치 마련

참고문헌:

가자마마사히로, 이즈카고지로, 마쓰무라유야. 『추천 시스템 입문』. 한빛미디어 2023.
이호수. 『넷플릭스 인사이트』. 21세기북스 2020.
김영우. "블루레이 디스크(Blu-ray Disc)". 스포츠 동아. 2011-02-14 10:50.