**주간프로젝트 기획안**

기획안 작성일자 : 2024년 5월 10일

|  |  |
| --- | --- |
| 조 명 | 2조: 파천내천 (파묘는 천만관객 내 연봉은 천만달러) |
| 조 원 | 조장: 김선규 부팀장: 정사라 조원: 김현우, 박소현, 이의준 |
| 프로젝트 주제 및 개요 | **주제** : <파묘> 흥행 요인 분석 – 온라인 사용자 반응을 중심으로  **개요** :   * **배경**: 코로나 이후 침체된 한국영화 시장에서 4번째로 천만관객을 동원한 <파묘>는, 천만 영화 중 ‘공포’, ‘오컬트’라는 가장 독특한 장르적 요소를 가지고 흥행에 성공함. 이 독특한 흥행 요인이 무엇이며, 어떤 함의를 갖고 있는지 온라인 사용자 반응을 중심으로 분석하여 인사이트를 얻고자 함 * **목적**:  1) 파묘의 흥행 요인 중 온라인 사용자 반응을 분석하여 사용자 반응과 실제 관객수 증가에 상관관계가 있는지 확인한다.  2) 사용자 리뷰의견 모델링 및 분석을 통해 인사이트를 도출하고 비즈니스 적용 가능성을 발견한다. * **타겟**: 영화 관계자, 콘텐츠 비즈니스 관계자, 영화에 관심있는 시민   **기대효과**: 파묘의 흥행 요인 분석을 통해 영화산업 및 관련 콘텐츠 산업을 활성화할 인사이트를 얻을 수 있다. |
| 프로젝트 수행 방향 및 내용  (수행 방법/도구) | ● 해결하고자하는 문제, 최종 산출물의 청사진  1. 데이터 수집:  - API 이용 (네이버, 구글, 영화진흥위원회)  - 크롤링 (유튜브, 메가박스, 왓차피디아, 키노라이츠, CGV)  2. 전처리:  - 연속형 데이터와 자연어 데이터 별 데이터 merge  - 결측치 확인 및 제거  - 텍스트 정제 및 정규화  - 토큰화 및 벡터화  3. 분석(EDA)  - 기초 통계 확인, Scatter plot, Histogram  - 텍스트 전처리 (정제, 토큰화, 사용자사전 구축, 불용어처리)  - 감성분석 모델링, LDA 분석  - 모델 문제점 분석 및 최적화, 모델 간의 성능비교 및 고도화  →최적 모델 선정  4. 시각화    5. 대시보드 구현    6. 웹 애플리케이션 프로토타입 구현      ● 데이터소개  a. 연속형 데이터  - 일일 박스오피스 (영화진흥위원회 API)  - 네이버, 구글 검색량 데이터 (네이버, 구글API(pytrend) )  b. 자연어 데이터  - 평점이 있는 리뷰 데이터 (메가박스, 왓챠피디아, 키노라이츠, 쿠팡플레이)  - 평점이 없는 리뷰 데이터 (CGV , 롯데시네마)  - 파묘 영상에 관한 유투브 댓글 데이터  ● 수행도구  - 프로그래밍 : Vs Code / Google Colab /  - 협업 도구 : Google Drive / Slack (허들) / Jira  - 배포 및 시각화 도구 : Streamlit |
| 프로젝트 조직  (구성원 및 역할) | ● 역할분담   * 조장: 김선규 – 네이버 API를 이용한 데이터 수집, 전처리 및 모델링, 웹 구현 * 조원1: 정사라(부팀장) – 구글 API를 이용한 데이터 수집, 전처리 및 모델링, 웹 구현 * 조원2: 김현우 – 영화관 리뷰 데이터 수집, 전처리 및 모델링, 웹 구현 * 조원3: 박소현 – 영화진흥위원회 데이터 수집, 전처리 및 모델링, 웹 구현 * 조원4: 이의준 – 영화관 리뷰 데이터 수집, 전처리 및 모델링, 웹 구현 |
| 프로젝트 추진 일정 | ● 일정  5/10 ~17 : 주제 선정, 기획, 일정 수립  5/18 ~ 5/24: 데이터 수집  5/25 ~ 6/2 : 데이터 전처리 & 텍스트 전처리  5/28 ~ 6/2 : 데이터 탐색을 위한 기초 EDA  6/2 ~ 6/7 : 감성 분석 및 모델링  6/8 ~6/10 : LDA 분석, 데이터 분석을 위한 EDA  6/10 ~ 6/11: 웹 애플리케이션 구현, 결과 자료 시각화 및 보고서 작성  6/12 : 프레젠테이션 |