데이터 분석 프로젝트 : 영화산업 데이터 분석

지능형 소프트웨어
기말고사 대체 과제

협성대학교 컴퓨터공학과
20170677 오융택

목차

- 선정 동기
- 데이터 수집 방법
- 데이터 분석 목표
- 데이터 분석 방법

선정 동기

- • 최근에 영화를 한 편 관람하였는데 어떤 영화는 기대받지 못했지만 실제로 개봉했을 때 예상외의 성적을 거두어 관람객의 평가가 좋았고, 어떤 영화는 좋은 성적을 기대했지만 예상했던 것 보다 좋지 못한 결과를 보여주고 있습니다.

- 이에 영화의 예매율과 관객 리뷰가 실제 영화 매출에 미치는 영향을 알아보고 싶어 주제를 선정하게 되었습니다,

데이터 수집 방법

- • 영화 상영회사 **3사(CJ CGV, 롯데시네마, 메가박스)의 예매율, 평점/리뷰, 성별과 연령 분포** 데이터를 수집한다.

- • 파이썬을 기반으로 **Selenium과 BeautifulSoup** 라이브러리를 통해 **동적웹 크롤링**을 통해 데이터를 수집한다.

분석 목표

- • 극장 위치, 상영 날짜에 따라 영화 예매율과 실제 관람객 수의 차이에 영향이 있는지를 분석한다.
- • 관람객의 성별, 나이 등에 따라 선호하는 영화 장르를 비교하고 이 것이 영화의 예매에 미치는 영향을 분석한다.
- • 관람객이 영화를 관람한 뒤 작성한 리뷰가 추후 예매율에 미치는 영향을 분석한다.

분석 방법

- • 리뷰데이터는 감정 분석과 빈도 분석을 통해 자연어 처리를 먼저 수행한 뒤 긍정 리뷰와 부정 리뷰로 다시 분류한다.
- • 리뷰를 제외한 다른 데이터들은 우선 군집화를 통해 특징을 추출해낸다.
- 추출된 데이터들이 서로에게 미치는 영향을 파악하기 위해 회귀분석을 수행한다.

결론 도출

- • 분석을 수행한 뒤 matplotlib을 사용하여 분석한 데이터들의 상관계수를 시각화 하여 영화 예매율에 영향을 미치는지 확인한다.
- 만약 회귀분석 결과로 높은 상관계수(-1 ~ -0.5 or 0.5 ~ 1)가 도출된다면 실제로 리뷰가 영화 예매율에 유의미한 영향을 미친 것이니 유의미한 데이터 분석임을 알 수 있다.
- 반대의 경우(-0.5 ~ 0.5) 영화 리뷰가 예매율에 영향을 미치지 않은 것이니 가설을 폐기한다.