1 과제수행 동기 및 배경

(1) 과제 수행 동기

- 취업 포트폴리오 제작
- 공모전 입상
- 적절한 기술력을 통해 사회에 문제를 해결

(2) 과제 선정 배경

현재 영화 산업의 규모는 16억 달러 (1조 7천억원[1])로 2010년부터 5% 안팎의 성장률을 보이는 시장입니다. 크기와 성장률에 불구하고, 현재 영화 산업은 제작단계에서 투자자들이 투자금 회수에 대한 확신 없이 주관적인 판단에 의지하여 제작 및 투자가 결정되고 그 결과로 상당한 투자에도 불구하고 흥행하지 못한 채 손익분기점조차도 넘기지 못한 영화가 다수입니다.[2] 이 때 영화진흥원이 제공하는 데이터를 통하여 과제를 접근해, 투자의사결정에 정보를 지원 할 수 있는 솔루션을 제공하는 것이 과제의 목표입니다.

일반 영화 시청자들과의 접촉을 위한 B2C와 투자자, 기업체에 접근하기 위한 B2C시장으로 구분하였습니다. B2C시장의 경우 기초 통계를 이용하여 배우에 대한 정보, 감독에 대한 정보 등을 제공해 소비자에게 흥미를 제공하는 용도로 이용될 것이며 이를 통해 솔루션의 홍보가 가능합니다. B2B시장의 경우 각종 통계기법, 데이터 마이닝 기법을 통해 도출해낸 예상 결과를 배우 별, 감독 별 등 다양한 항목을 입력 시 대응되는 예상 값을 제공함으로써 영화의 제작, 투자의사결정시 필요한 근거를 제공합니다.

---

1) THEME-Report-2017

2) 중앙일보 기사 https://news.joins.com/article/21411450

2 과제수행 내용

(1) 과제내용

과제의 수행 방법은 전형적인 데이터 분석 프로젝트의 진행방향을 따르되, 신뢰성 있는 데이터 분석 결과를 제공하기 위하여, 각 분석 단계에서 이전 단계 완료시 다음 단계로 진행되는 폭포수(Water Fall) 진행 방식과 같은 계층적 프로세스 모델이 아닌 각 단계 별로 피드백의 적응이 가능한 반복적 정련 방식을 이용합니다. 데이터 분석의 수행 도중 발견되는 문제를 즉각적으로 해결하고 결과의 평가를 반복함으로써 신뢰성 있는 분석결과를 도출할 수 있습니다.

세부 진행 방향은 데이터 준비, 데이터 분석, 시스템 구현, 평가 및 반복 정련, 프로젝트 평가 순으로 진행됩니다. 데이터 준비는 영화진흥청에서 제공하는 API를 통한 데이터 수집 및 네이버영화에서 제공하는 영화 정보 크롤링을 통해 1차적으로 데이터베이스를 구축하고 분석에 알맞게 이상값 제거, 표준화, 정규화를 진행합니다. EDA(탐색적 데이터 분석)의 단계를 통해 데이터 분석 시 필요할 주요 변수들을 파악하고 각 변수들에 대한 가중치 조정 및 통계 기법 적용을 통해 분석 결과 및 예측치를 생성합니다. 이를 바탕으로 B2C, B2B 시장에 필요로 되는 시스템을 웹을 통해 구현하고 실제로 이용함으로써 평가를 진행합니다. 도출된 부정적, 긍정적 요인을 이용해 반복 정련을 실시하고 최종 결과물을 산출, 평가합니다.

(2) 작품의 기술성

1. 데이터 수집 과정
   pytho의 웹 크롤러 라이브러리 BeautifulSoup, Selenium을 이용하여서 수집
2. 데이터 전처리 과정
   R과 python의 pandas 그리고 sql을 이용한 데이터 전처리 기법적용
3. 데이터 분석 과정
   R과 python을 이용해 각종 통계모델 기법을 적용
4. 데이터 시각화 과정
   R과 python을 이용한 분석결과 시각화하고 군집을 나눠서 insight를 얻고 다시 진행
5. 솔루션접근 과정
   python의 django를 통해서 웹페이지를 제작
   분석하여 만들어진 수치를 MySQL에 저장
   분석결과를 토대로 시각화.

(3) 제작과정

수행내용	일정															
	9 월				10 월				11 월				12 월			
	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
데이터 수집, 정제, EDA	■	■	■	■												
데이터 1차 분석				■	■											
데이터 2차 분석						■	■									
솔루션 적용							■	■	■							
웹 운영							■	■	■	■	■					
평가 및 피드백											■	■	■			
최종 평가 및 보고서 정리												■	■			

3 과제해결 방안 및 고찰

(1) 과제 수행과정에서의 문제점 및 시행착오

1. 데이터 무결성을 해치는 동명이인 문제
   배우 황정민의 경우 무명의 배우까지 포함하여 3명의 사람이 존재.
   각 배우의 영향력을 평가하기 위해서는 동명이인을 반드시 분류해야할 필요성을 인지.
2. API 처리량 제한
   KOFIC API는 인당 일 3000회만 수집이 가능
3. Django의 버전차이
   기존 진행했던 reference의 django가 1.x인데 진행한 프로젝트는 2.x인 것을 확인

(2) 문제점 해결과정

1. 데이터 무결성을 해치는 동명이인 문제
   1) KOFIC(코픽)이라는 곳에서 제공하는 OPEN API를 사용
   2) 배우별 고유 코드를 부여
2. API 처리량 제한
   1) CRON을 이용해서 매일 오후12시에 python 코드 실행
   2) 자동화하여 일주일간 수집

=> 총 5382개의 동명이인 데이터를 분류할 수 있었습니다.

3. Django의 버전차이
   1) 버전차이로 인해 발생하는 setting파일의 변화로 참고서적을 통하여 지식 업데이트

(3) 추후 개선방향

미국 할리우드 메이저 영화사 중 하나인 20th Century Fox사에서는 영화 관람객을 예측하는데 영화 팬에 대한 이해도를 높이기 위해 세분화된 고객 데이터와 영화 시나리오를 바탕으로 훈련된 사내 딥 러닝 모델을 개발했다는 것을 확인 했습니다. 이것으로 다양한 유형의 영화에 대한 선호도 패턴을 파악하기 위한 것이었습니다.

현재 캡스톤 프로젝트에서 진행한 프로젝트는 걸음마 수준에 불과하며 할리우드 영화사의 딥러닝 모델은 영화의 톤, 핵심 관람객 및 확장 관람객과의 어피니티, 영화의 잠재적 재무성과를 평가하기 위한 가장 객관적이고 데이터 기반이며 효과적인 지표 중 하나를 제공한다는 것을 확인 할 수 있었습니다.

또한, 시나리오 분석이 가장 중요하다고 생각하였던 우리 팀의 예측에 반해 미국은 텍스트 분석이 영화 관람객을 유인할 수 있는 추가적인 역동적 기제 없이 스토리의 주요 골격만 제시하였기 때문에 많은 정보를 얻기 어려웠다는 말을 내놓았습니다. 미국은 팀에서는 영화의 전체적 마케팅 캠페인에 있어 가장 중심적이고 단 하나의 요소로 남아 있는 영화 예고편을 연구하였고, 회사로서는 예고편이 영화 팬의 호기심과 흥미를 제대로 돋움으로써 흥행에 연관이 있다고 판단하였습니다.

그래서 향후 개선방향으로써는 현재 머신러닝을 통한 컴퓨터 비전을 학습 중에 있으며 좀 더 나아가 컴퓨터 비전의 머신러닝을 저희 프로젝트에 적용시키도록 하겠습니다.

※ 인용사이트 :
https://cloud.google.com/blog/products/ai-machine-learning/how-20th-century-fox-uses-ml-to-predict-a-movie-audience

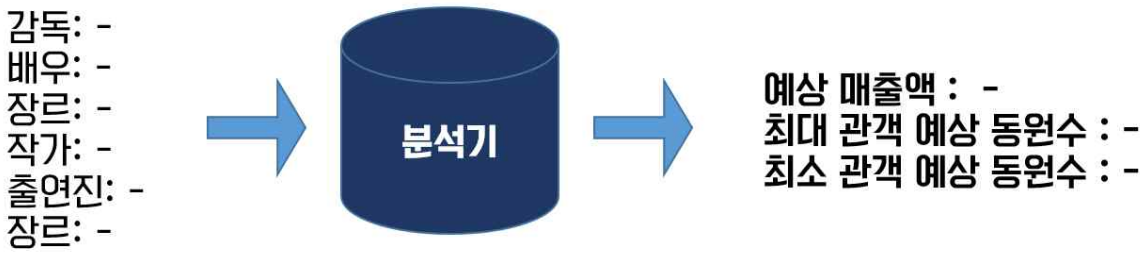
※ 관련 논문 : https://arxiv.org/abs/1810.08189

4 과제 기대효과 및 활용계획

◆ 과제 수행 후 얻을 수 있는 결과 및 효과

-영화제작 의사결정 지원 솔루션
-각종 항목별 기초 통계를 통한 영화 정보 열람
-데이터 분석 프로세스에 대한 이해도 증대

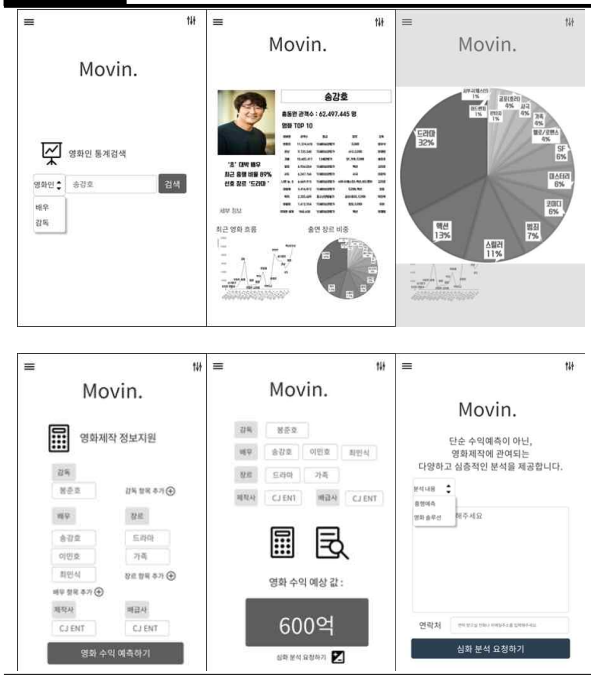
결과물의 형태 및 수익 모델



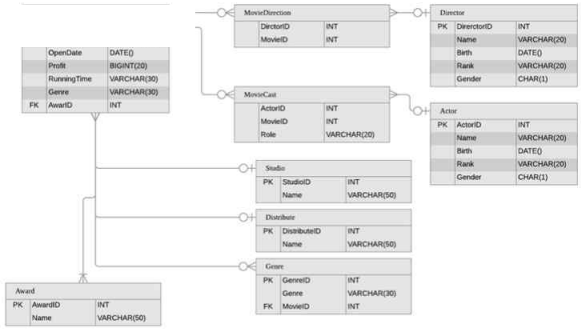
영화 시장은 크게 영화인, 제작사, 투자자 3요소로 구성이 되어있다. B2B시장의 경우 투자자가 투자하고자 하는 영화기획의 정보를 입력하면, 정보를 감독, 배우, 작가, 장르, 제작사에 따른 흥행 가능성을 제시합니다. 변수로 입력된 사항들을 통해 흥행과의 상관성을 예측해내고, 예상되는 최소 수익과 최대 수익의 범위로 투자자에게 정보를 제공합니다. 기초 통계만을 바탕으로 하지않고 각종 분석기법을 통해 도출된 요인을 통해 의미있는 자료를 제공하고, 정보의 열람을 통해서 투자자는 기획된 영화의 투자 유무를 결정합니다. 이 과정에서 기본적인 자료 제공의 경우 월간, 년간 정액제를 통해서 제공하되 투자를 결정하면 투자 수익의 0.5~1.5%의 수익을 대가로 심화적인 분석 자료를 제공합니다. 그 분석자료의 예로는 감독 변수에 따른 배우 추천, 작가 추천, 나아가 인공지능 학습을 통한 시나리오의 분석 및 수정 방안 제시 등이 있습니다. 즉 기본적인 정액제의 형태로 제공되는 정보는 자동화를 통해 추가적인 작업의 필요성을 최소화하고, 영화 제작 및 기획 단계에서 수익성을 높일 수 있는 솔루션을 제공해 수익을 창출하는 방식을 고객의 요청에 맞춰 제공할 수 있습니다. 일 예로 캐스팅 과정에서 최선의 배우를 놓친 상황에서 차선책이 될 수 있는 배우를 데이터적 접근을 통해 제공할 수 있습니다.

B2C시장에서 제공될 결과물은 배우에 대한 검색 시 배우의 상위 10개 영화와 배우가 출연한 영화의 흥행 비율에 대한 정보를 보여준 후 배우를 스타 배우, 초대박 배우, 대박 배우, 일반 배우 등과 같은 분류를 통해 제공하고 출연 장르에 따른 흥행 비율, 협업한 감독과의 흥행 비율 등을 세부항목으로 제공합니다. 영화와 감독 또한 같은 방식으로 정보를 제공하고, 영화에 흥미를 가지고 있는 인터넷 사용자를 공략합니다. 예상되는 수익모델은 광고 배너 정도이지만, 이용자들이 만들어내는 검색자료를 통한 데이터 수집과 B2B 시장에 대한 홍보가 주요 목표입니다. 사용자가 생성하는 데이터를 통해 영화시장에 대한 트렌드를 도출해내, 분석 자료 및 솔루션에 적용이 가능 합니다. 그리고 B2C 사이트를 통해 유입된 영화시장 관련 종사자들을 통해 광고효과를 얻을 수 있습니다.

5 작품 사진



프로그램 실행화면



프로그램 구성 ERD

6 과제역할분담

(1) 학생

순번	성 명	전공	학년	학번	연락처	담당업무
1	정광진	컴퓨터정보공학부	4			데이터전처리 / 분석 / 프로그래밍
2	유준호	경영학과	3			데이터 시각화 / 분석 / 시장조사

7 실습비 사용내역

구분	품목	세부내용	용도	금액
실습비	회의비	중앙도서관에서 진행	과제수행을 위한 회의 진행	60,800
	문헌구입비	데이터분석의 이해를 위함	과제수행을 위한 책 구매	45,000
총계				105,800

8 참고문헌

손민규 (2018). 데잍터 분석을 떠받치는 수학. 위키북스

브레트 란츠 지음 ; 윤성진 옮김 (2017). R을 활용한 머신 러닝 2/e : R로 머신 러닝 알고리즘 작성, 데이터 준비, 데이터 예측 기법 깊이 파기. 에이콘

김석훈 (2018). 파이썬 웹프로그래밍 : Django(장고)로 배우는 쉽고 빠른 웹개발. 서울: 한빛미디어

Vanderplas, Jake (2017). Python data science handbook : [essential tools for working with data]