Data analysis on Box office

Group member: Shuo Liu, Qian Wang, TianRun Sun

1.Data Science Problem

Imagine that you are going to see a movie tonight, you may think What’s in theaters right now? What’s that new movie my friend was just talking about a couple days ago? That trailer I saw for another film a few weeks ago looked interesting. Another movie review I read sounded promising...what should I see?

Actually, these elements have influenced moviegoers’ film options and people’s choices of movies are directly related to box office performance. It is very important to understand the moviegoers’ pattern so that movie marketers are in a better position to align and adjust their strategies to capture the interest of potential moviegoers.

In 2013, google published a research paper on quantifying movie magic with google search(<https://ssl.gstatic.com/think/docs/quantifying-movie-magic_research-studies.pdf)>. Basically, it dived into data analysis on google search volume, amount of paid clicks, the movie trailer engagement and come to the conclusion that the timing of trailer searches on both Google and YouTube are leading indicators of box office success.

2.Potential Analysis

In our project, we will try to take three aspects to predict the box office. For movies itself, the rating, genre,3d imax or 2d. From moviegoers’ perspective, we believe many movie lovers will choose movies based on the heat of the movies on social media since most people will see popular movies with someone familiar. Also, people often decide if they see the movie on whether the director is good or the actor/actress is good. From the movie industry’s perspective, the release date of the movie and the opening theatres leads to different competition (studio always wants to avoid another blockbuster title)

Therefore, we want to collect such data through 1.Tmdb api and 2.twitter api

1. Motion Picture Rating, Genres, film format, running time, director, actor/actress, budget, prequel
2. Twitter heat(amount of related twitter for further analysis), twitter review
3. Release date, opening theatres, competition

Possible hypotheses:

1. Films’ genres are important elements to box office, superhero movies are usually huge success like guardians of galaxy and some drama not
2. If a film is highly rated by movie critic or moviegoers, box office will last and do better for a long time. Else, they will drop sharply.
3. Great director and actors/actresses attract people to see their films
4. Films rated PG-13 have more chances to have better box office records than films rated R
5. If release date of two big titles is close, they may have less box office number than they should have.

……….

3.Data Issue

as we collect data from tmdb and twitter , we have thousands of raw data

for the movie data base, we will try to clean several important attributes like budget, genres, homepage, production companies, production countries, revenue, runtime, spoken languages, title, cast, crew; for the twitter part, we focus on the movies in 2017, and try to collect movie related tweets to show their popularity and evaluate their effect on box office, and we will try to delete some informal type data or non-related data. Further details are included in README.txt