Big Data Analytics Symposium - Fall 2018

Analytics Project: Moviepedia

Team:

Xinyi Liu Xinsen Lu Yiming Li

Abstract: This analytic provides insights in movies for script writers and movie investors. By analyzing gross box office incomes, user ratings, directors, actors and actress, and commonly mentioned keywords from reviews and their sentiments regarding to genres, it provides evidence of decision in an early stage of movie production.

Motivation

Who are the users of this analytic?

Script writers and movie producers.

Who will benefit from this analytic?

Script writers and movie producers.

Why is this analytic important?

Movie producers can determine whether an outline is a cliché and get tips on how to get desired cast and crews at an early stage, so that they can avoid wasteful investment.

Goodness

What steps were taken to assess the "goodness" of the analytic?

1. Compare the results, including something unexpected, with IMDb
2. Compare general sentiment and commonly mentioned words from IMDb reviews and Twitter

Data Sources

Name: IMDb 5000 Dataset

Description: This dataset provides metadata of about 5,000 movies in IMDb. There are over 4,600 records after removing duplicates.

Size of data: ~20MB

Name: IMDb Reviews Dataset

Description: This dataset contains IMDb movie IDs, user reviews and ratings of movies listed in the IMDb 5000 Dataset. It has about 220,000 reviews.

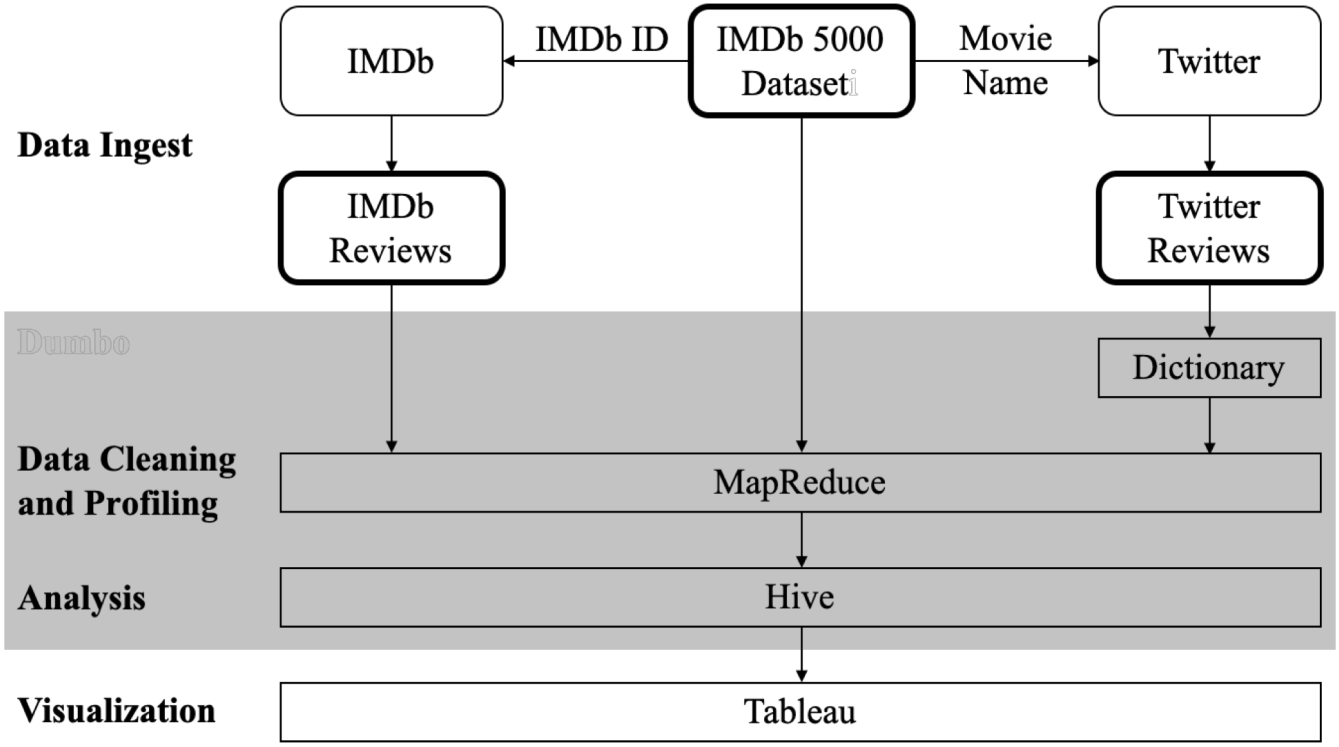
Size of data: ~200MB

Name: Twitter Reviews Dataset

Description: This dataset has tweets sharing reviews of movies listed in the IMDb 5000 Dataset. It has about 180,000 reviews.

Size of data: ~60MB

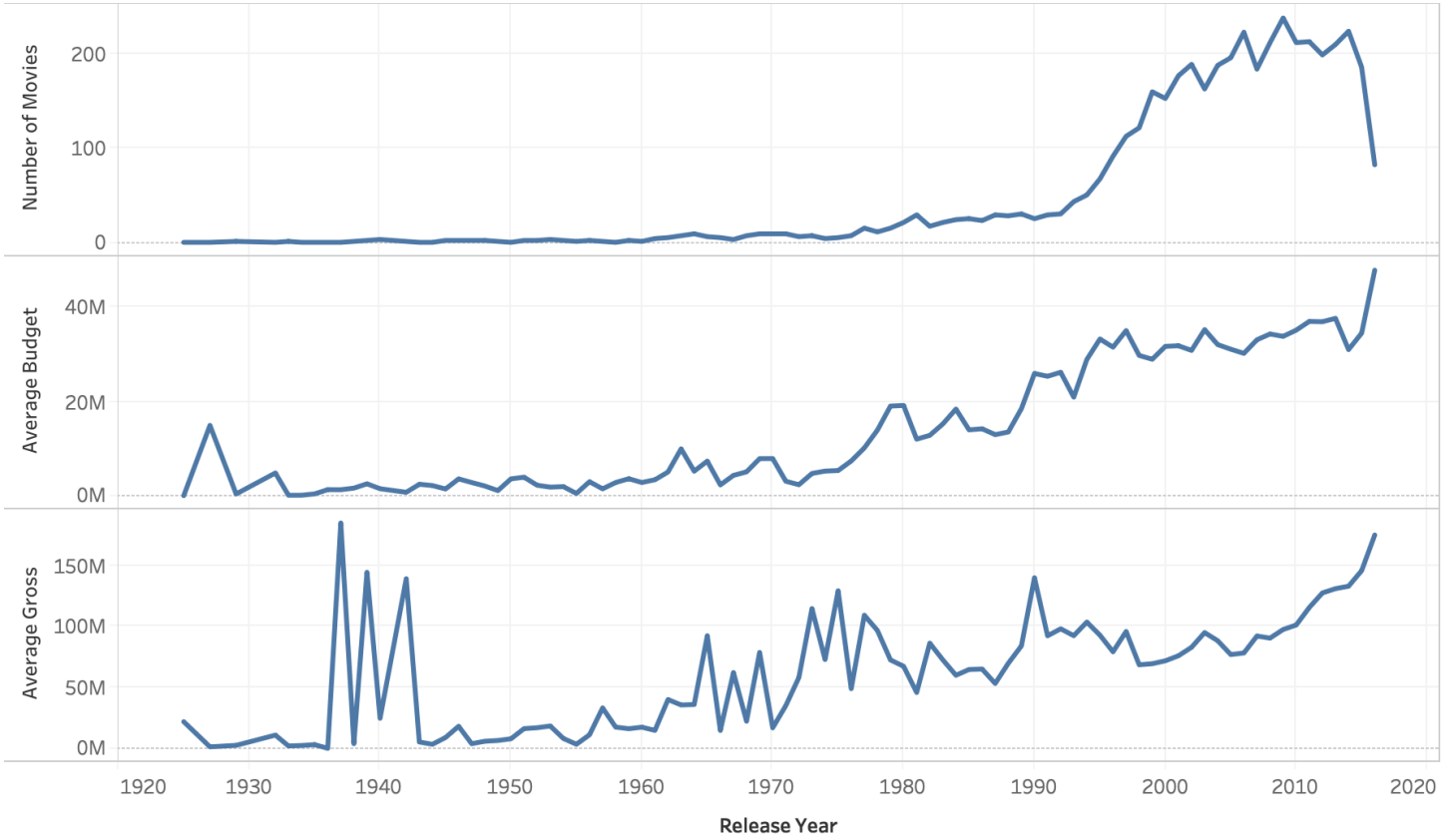
Design Diagram



Platform on which the analytic ran: NYU Dumbo

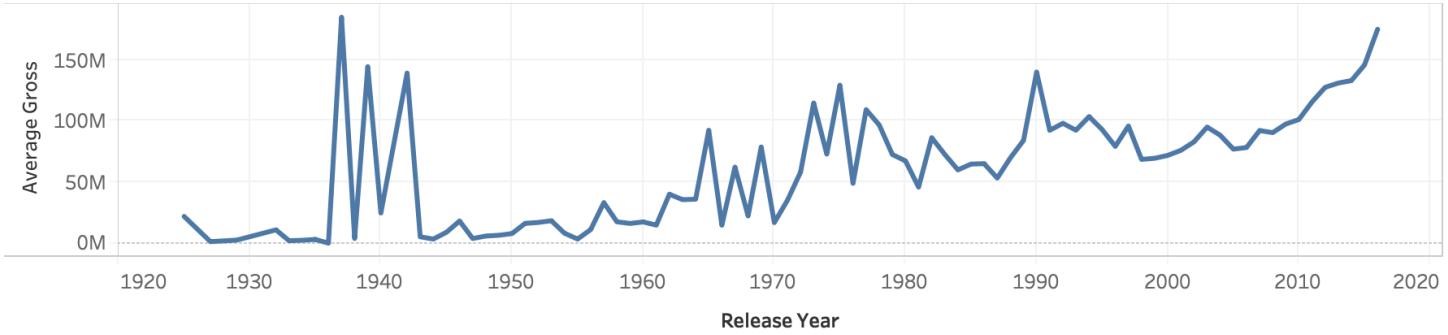
Results

1. Development trend: Producers are shooting more, investing more, and earning more on movies.



Results

1. Development trend



Movie	Worldwide Gross
Snow White and the Seven Dwarfs (1937)	184,925,486
The Wizard of Oz (1939)	33,953,637
Mr. Smith Goes to Washington (1939)	9,000,000
Gone with the Wind (1939)	390,525,192
Casablanca (1942)	10,462,500
Bambi (1942)	268,000,000

Results

2. Correlation matrix: A movie with a higher budget tends to get more box office income.

	actor_1..	actor_2..	actor_3..	cast_to..	director..	duration	imdb_s..	movie_..	num_cr..	num_us..	num_vo..	product..	release..	worldw..
actor_1_fa..	1.000	0.433	0.339	0.887	0.096	0.089	0.063	0.138	0.192	0.148	0.184	0.186	0.080	0.144
actor_2_fa..	0.433	1.000	0.717	0.724	0.177	0.179	0.110	0.304	0.394	0.294	0.344	0.381	0.161	0.327
actor_3_fa..	0.339	0.717	1.000	0.692	0.215	0.204	0.112	0.366	0.463	0.359	0.412	0.475	0.204	0.402
cast_total_..	0.887	0.724	0.692	1.000	0.166	0.166	0.096	0.292	0.375	0.287	0.341	0.391	0.162	0.326
director_fa..	0.096	0.177	0.215	0.166	1.000	0.238	0.228	0.229	0.266	0.333	0.386	0.149	-0.084	0.181
duration	0.089	0.179	0.204	0.166	0.238	1.000	0.286	0.214	0.263	0.325	0.317	0.265	-0.150	0.245
imdb_score	0.063	0.110	0.112	0.096	0.228	0.286	1.000	0.288	0.325	0.316	0.428	0.055	-0.218	0.195
movie_face..	0.138	0.304	0.366	0.292	0.229	0.214	0.288	1.000	0.719	0.458	0.580	0.353	0.213	0.431
num_critic..	0.192	0.394	0.463	0.375	0.266	0.263	0.325	0.719	1.000	0.608	0.630	0.507	0.270	0.527
num_user_..	0.148	0.294	0.359	0.287	0.333	0.325	0.316	0.458	0.608	1.000	0.816	0.452	-0.003	0.566
num_voted..	0.184	0.344	0.412	0.341	0.386	0.317	0.428	0.580	0.630	0.816	1.000	0.446	0.010	0.624
production..	0.186	0.381	0.475	0.391	0.149	0.265	0.055	0.353	0.507	0.452	0.446	1.000	0.176	0.741
release_year	0.080	0.161	0.204	0.162	-0.084	-0.150	-0.218	0.213	0.270	-0.003	0.010	0.176	1.000	0.090
worldwide_..	0.144	0.327	0.402	0.326	0.181	0.245	0.195	0.431	0.527	0.566	0.624	0.741	0.090	1.000

Results

3. Word cloud: Action films

IMDb

Twitter

Obstacles

1. Non-English characters

Some special characters crawled from web may bring error to Hadoop Streaming jobs based on python 2.6.6. We solve it by setting

```
UTF8Writer = codecs.getwriter('utf-8')
sys.stdout = UTF8Writer(sys.stdout)
```

in mapper.

2. Cache file in MapReduce

We have to read in a file apart from the input into the MapReduce job. We try to scan the file in mapper class but failed. The solution is adding the cache file in driver and read in by buffer reader.

Summary

This analytic provides insight for movie producers and script writers based on statistics and reviews from audience. When evaluating an idea or a script, this analytic can be an evidence of decision.

Acknowledgements

- We would like to appreciate Prof. McIntosh for her help and guidance on this project.
- We would like to appreciate Chuan Sun for providing the IMDb 5000 Dataset.

References

1. T. Ashwitha, A. Rodrigues and N. Chiplunkar. Movie Dataset Analysis using Hadoop-Hive. In 2nd IEEE International Conference on Computational Systems and Information Technology for Sustainable Solutions, 2017.
2. G. Mishne and N. Glance. Predicting Movie Sales from Blogger Sentiment. Accessed at https://www.aaai.org/Papers/Symposia/Spring/2006/SS-06-03/SS06-03-030.pdf.
3. M. Mestyán, T. Yasseri and J. Kertész. Early Prediction of Movie Box Office Success Based on Wikipedia Activity Big Data. Accessed at https://doi.org/10.1371/journal.pone.0071226.
4. R. Paul. Big Data Analysis of Indian Premier League using Hadoop and MapReduce. In 2017 International Conference on Computational Intelligence in Data Science (ICCIDS), 2017.
5. M. Kumar and A. Bala. Analyzing Twitter Sentiments Through Big Data. In 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), 2016.
6. B. Liu, E. Blasch, Y. Chen, D. Shen and G. Chen. Scalable Sentiment Classification for Big Data Analysis Using Naïve Bayes Classifier. In 2013 IEEE International Conference on Big Data, 2013.
7. C. Sun. Predict IMDB Movie Rating. Accessed at https://github.com/sundeepblue/movie_rating_prediction.
8. IMDb. IMDb Website. Accessed at https://www.imdb.com.
9. Twitter. Twitter Developer. Accessed at https://developer.twitter.com.
10. The Numbers. The Numbers - Where Data and the Movie Business Meet. Accessed at https://www.the-numbers.com.
11. K. Bougé. Stop Words. Accessed at https://sites.google.com/site/kevinbouge/stopwords-lists.
12. HLT - Natural Language Processing. SentiWords. https://hlt-nlp.fbk.eu/technologies/sentiwords.

Thank you!