Script Evaluation & Production Company Prediction

By

Arshiya Sayadda – as12855@nyu.edu

Badhri Narayanan Suresh – bns332@nyu.edu

Bhaskar Venkatraman – bv640@nyu.edu

Table of Contents

1. Abstract
2. Problem Statement
3. Target Audience
4. Dataset
5. Architecture Diagram
6. Approach / Methodology
7. Data Cleaning
8. Exploratory Data Analysis
9. Content Based Recommendation System
10. Evaluation of Model
11. Conclusion
12. Further Improvements
13. References

Abstract

We attempt to create a description based recommendation system using Pyspark ML that can take in a few lines about the movie as input and output a possible list of production companies that can likely accept your script based on your input about the movie. This model is supplemented by exploratory data analysis which is used to delve deeper into the data and get an understanding of the dataset. We then create a new metric ‘return’ from this data analysis which we will use in our recommendation system.

Problem Statement

The film industry puts in billions of dollars into making movies every year but very few turn up a profit. Which scripts get selected to be made into movies is a very subjective process. It has more to do with luck than any logical thought process. They are operating on limited data.

Even if a script is great and it gets picked to be made, the wrong production company could squander its potential for profit.

We believe that for an industry with such a huge investment, there should be a more concrete and objective way of decision making.

Target Audience

Script writers with a bound script can narrow down the list of production companies for selling their script and eventually sell it to the company that has historically made great box office hits with similar scripts.

Production companies receive hundreds of scripts and therefore, take time to evaluate each of them. Prediction of expected rating and return by a system would expedite the process and eventually yield high efficiency.

Dataset

The files used contain metadata for all 45,000 movies listed in the Full MovieLens Dataset. The dataset consists of movies released on or before July 2017. Data points include cast, crew, plot keywords, budget, revenue, posters, release dates, languages, production companies, countries, TMDB vote counts and vote averages.

This dataset also has files containing 26 million ratings from 270,000 users for all 45,000 movies. Ratings are on a scale of 1-5 and have been obtained from the official GroupLens website.

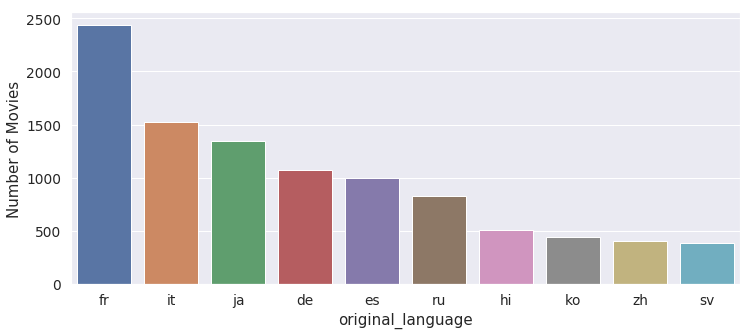
* movies\_metadata.csv

The main Movies Metadata file. Contains information on 45,000 movies featured in the Full MovieLens dataset. Features include posters, backdrops, budget, revenue, release dates, languages, production countries and companies.

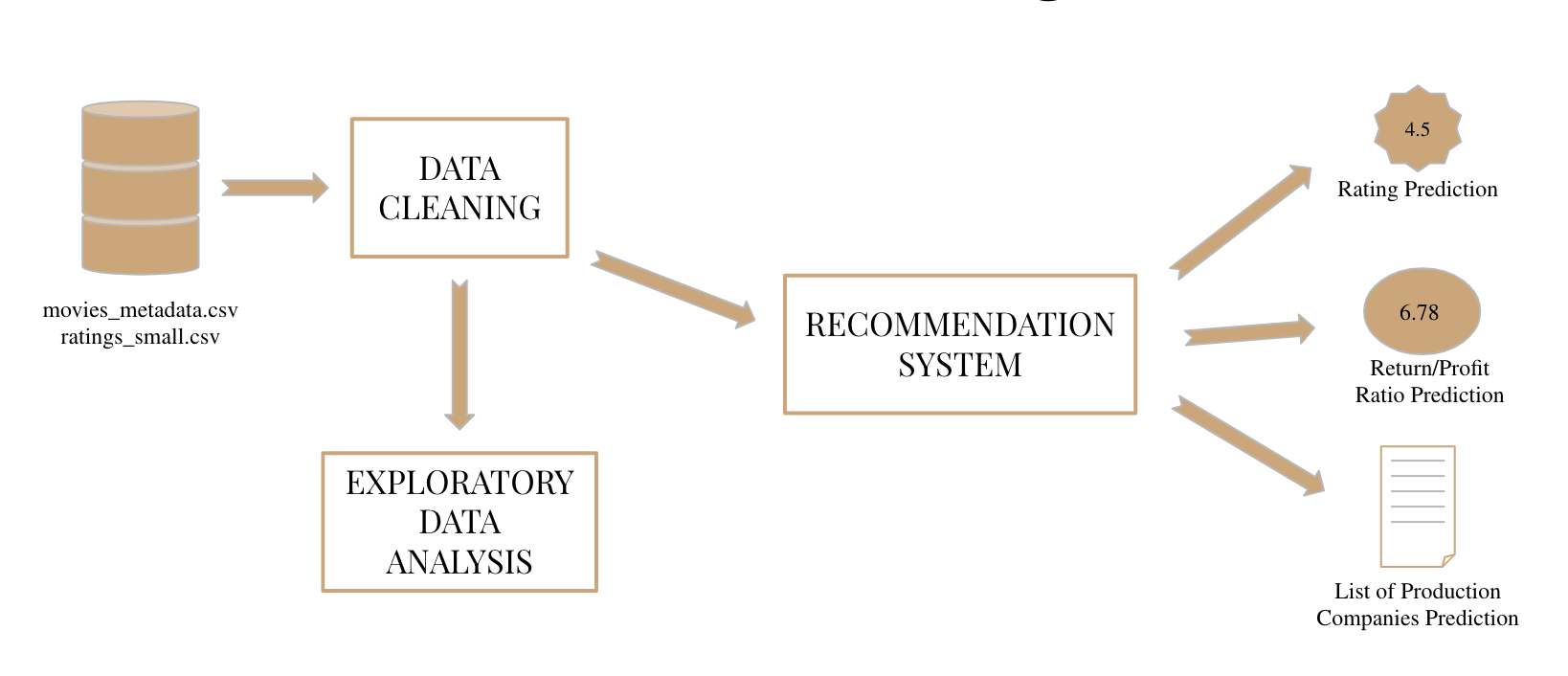
* ratings\_small.csv

The subset of 100,000 ratings from 700 users on 45000 movies

* Number of Unique Languages : 91
* Number of Movies in the Dataset : 45000
* Number of English Movies : 32185
* Majority of the movies released before 2017 have English as their original language.
* About 71.5% of the movies are in English followed by French and so on.



Architecture Diagram



Approach / Methodology

Build a Recommender System solely based on plot line.

Provide list of Production companies that could most likely accept a movie script

Predict the revenue-budget ratio of movies taking as input its one-line plot.

Predict the Rating (IMDB) for unreleased movies

Data Cleaning

Drop duplicates and redundant columns : title vs original title, adult movies

Remove rows with incomplete data on revenue and budget

Remove entries with inconsistent datatypes: titles with ascii

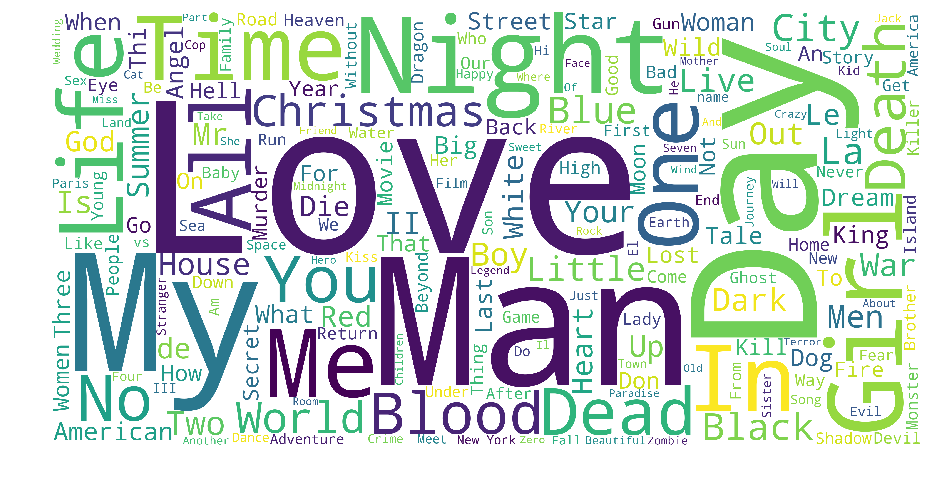
Extract relevant information as new columns: return = revenue/budget, year from release date

Exploratory Data Analysis

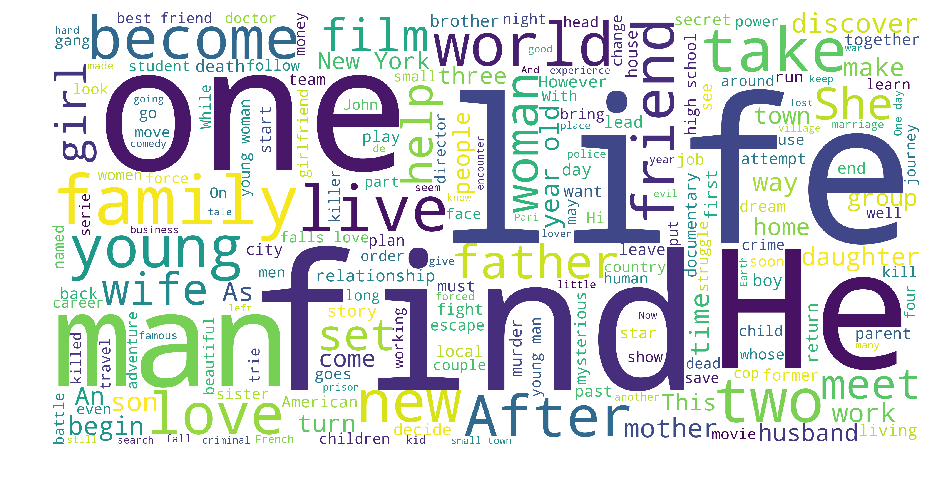
# 

# Word Occurrence Analysis :Word Cloud Representation

# Movie Titles

The word **Love** is the most commonly used word in movie titles. Words like Man, Day ,My are also commonly used in movie titles.

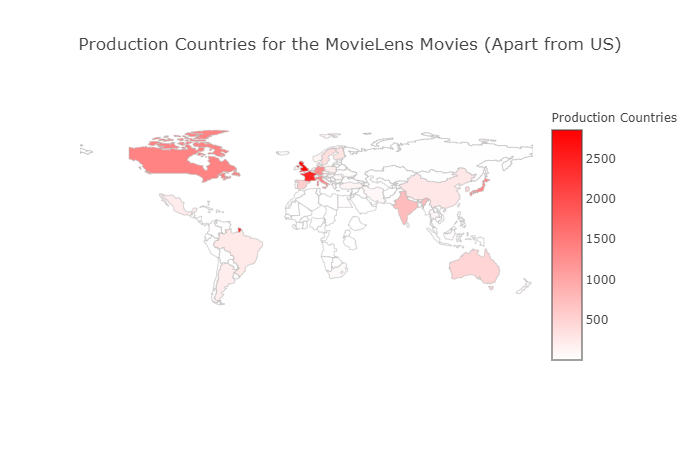
# Movie Overview



**Life** is the most commonly used word in Movie Blurbs. **One** and **Find** are also popular in Movie Blurbs.

Together with **Love**, **Man** , these wordclouds give us a pretty good idea of the most popular themes present in movies.

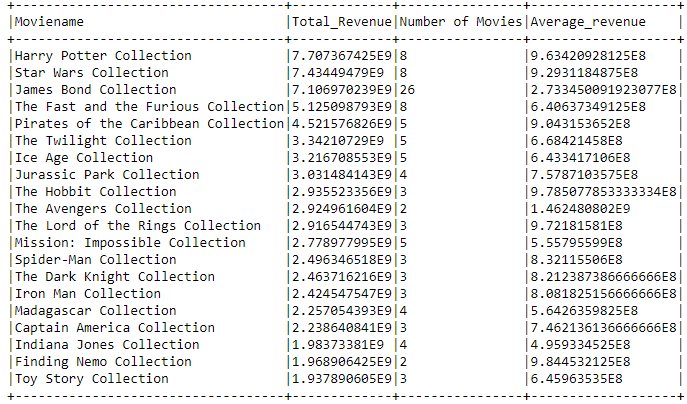
# Filming Locations Analysis : World Map Representation



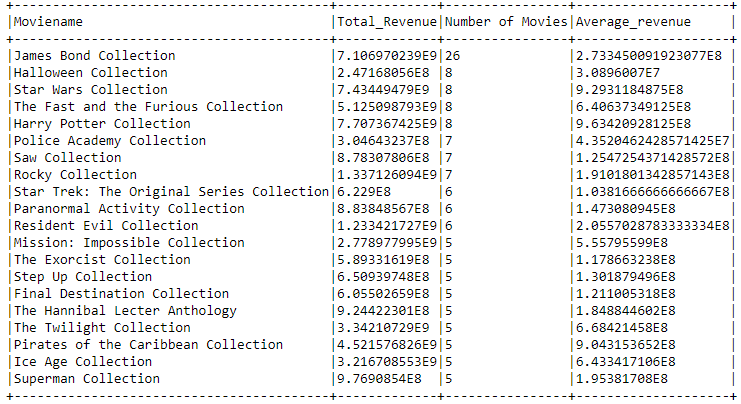
Europe is also an extremely popular location with the UK, France, Germany and Italy in the top 5. Japan and India are the most popular Asian countries when it comes to movie production.

# Franchise Movies Analysis

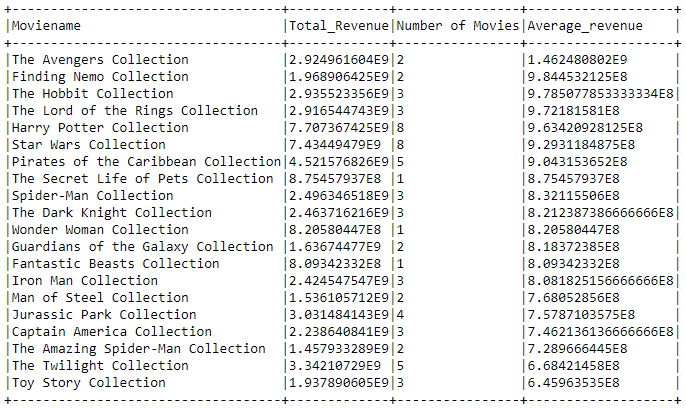
# Highest Grossing Most Successful (by Avg Gross)

Harry Potter Franchise is the highest grossing movie franchise to date. Star Wars is a close second.

# Longest Running

James Bond is the longest running movie franchise with a staggering 26 movies. Although ,the average revenue of this series is not that high it still made a good amount of money in the movies.

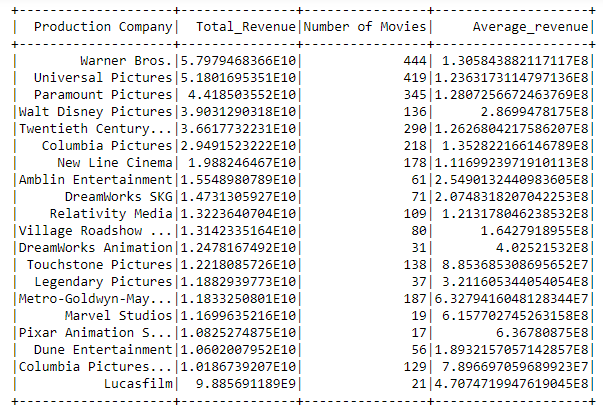
# Most Successful



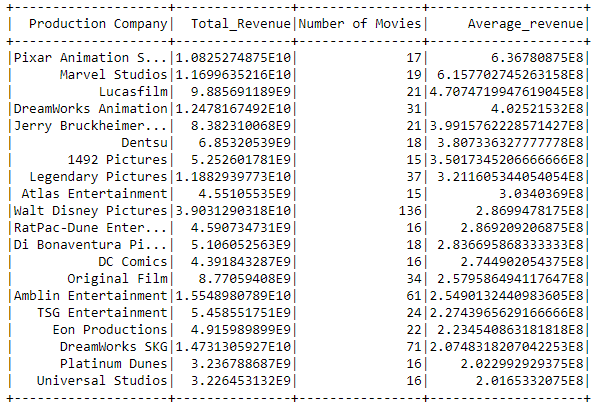
The Avengers Collection is the most successful movie franchise in this collection .

# Production Companies Analysis

# Highest Earning

Looking at Production Houses , Warner Bros boasts the highest total revenue followed by Universal Pictures. It is easy to see that there is a lot of money involved in producing movies and the revenue is staggering.

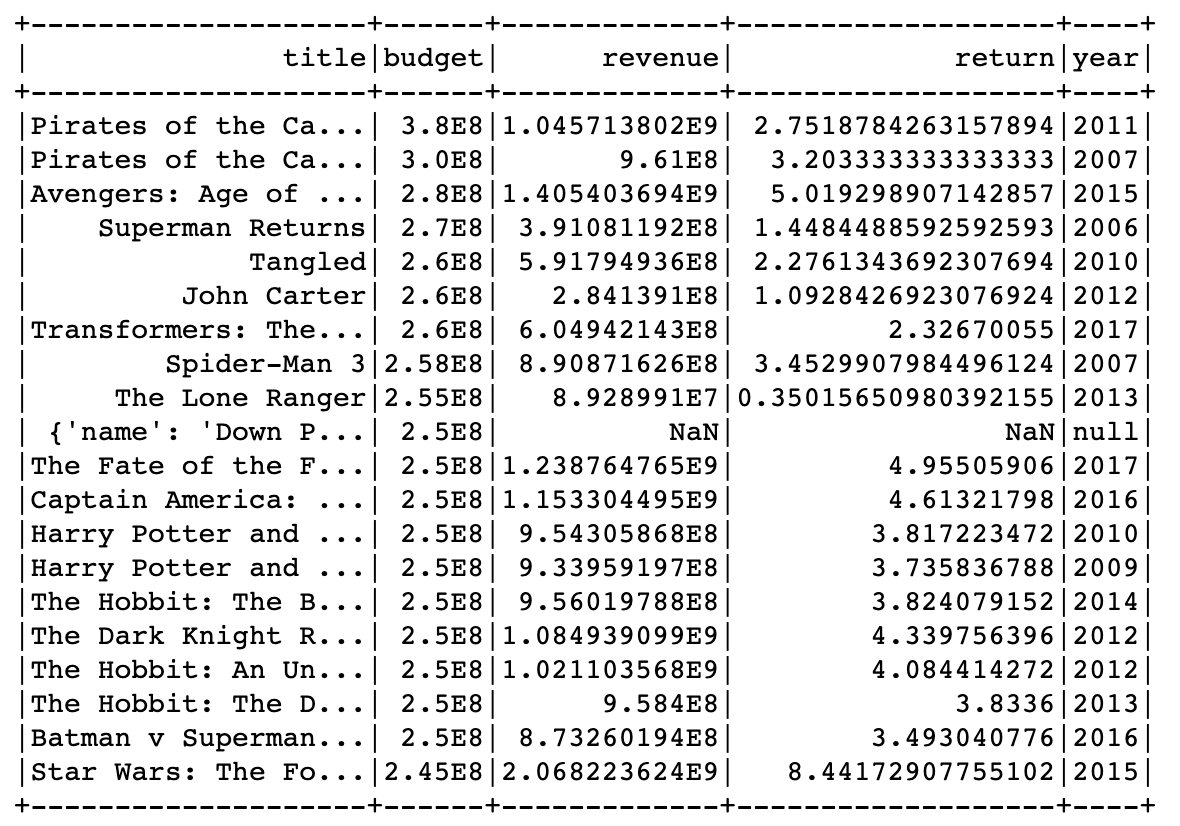
# Most Successful



Pixar Animation has the highest average revenue in the list , even though it has not made nearly the same number of movies as some of the other companies in this list. This shows how important every movie is to the revenue of a production company.

# 

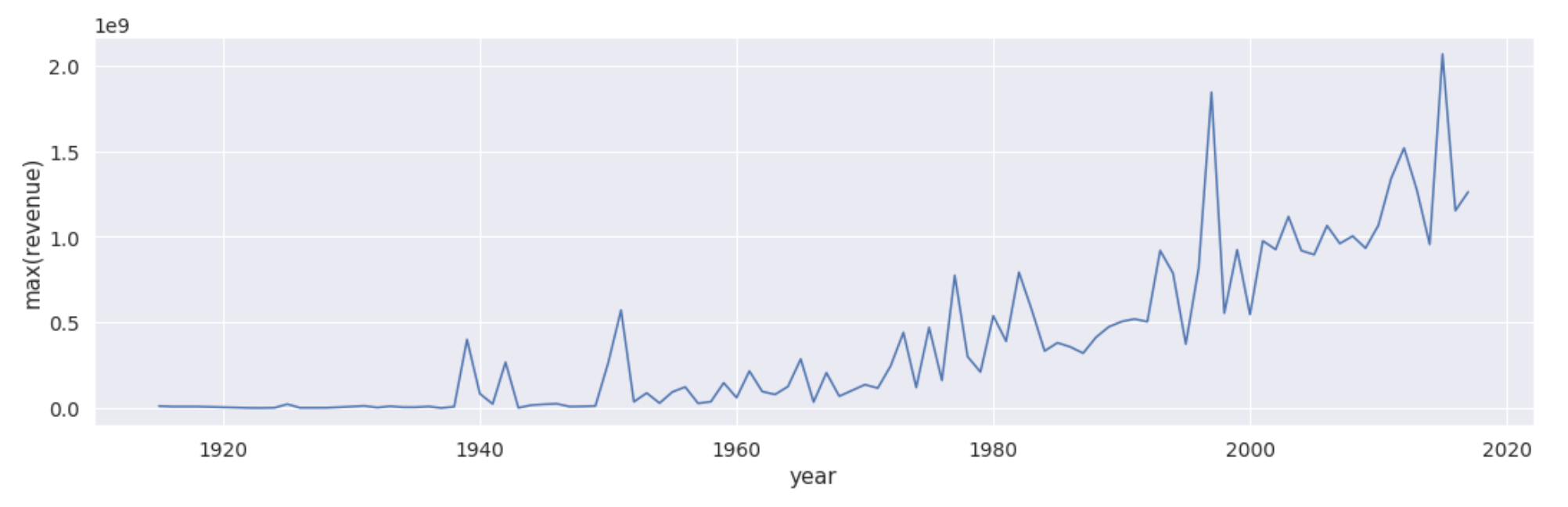
# Most Expensive Movies of all Time



Pirates of the Carribean with a staggering budget of nearly 380 million is the most expensive film made to date. The second most expensive film is another Pirates of the Carribean movie with a nearly similar budget.

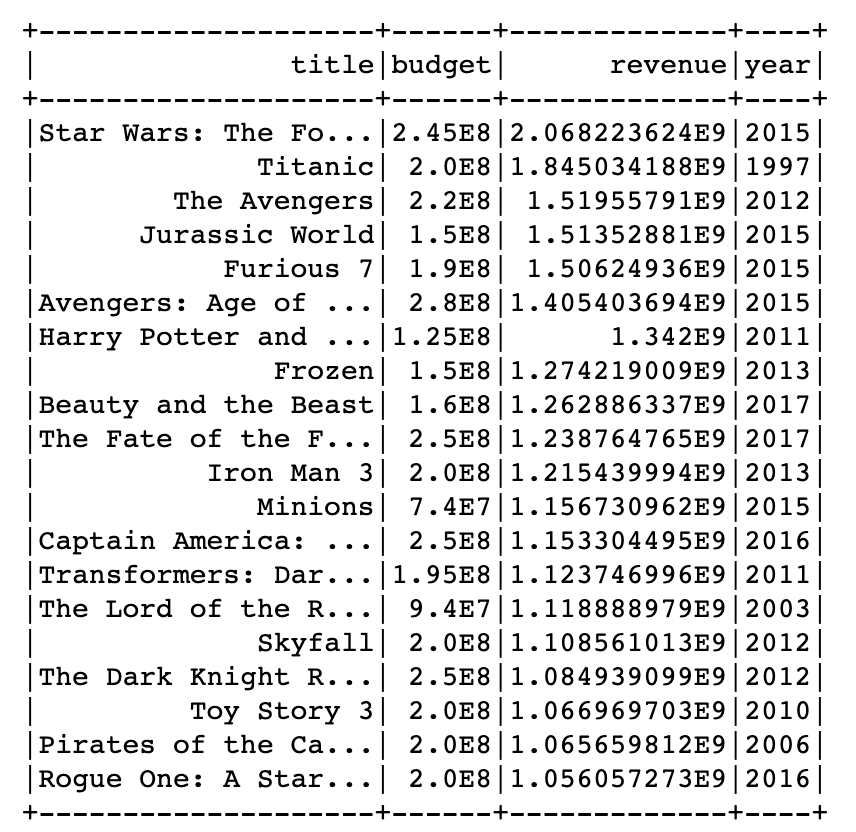
# Revenue Analysis

# Insights from Summary Statistics



It is easy to see that year on year the maximum revenue made from movies is increasing , and it is forecasted to increase even more.

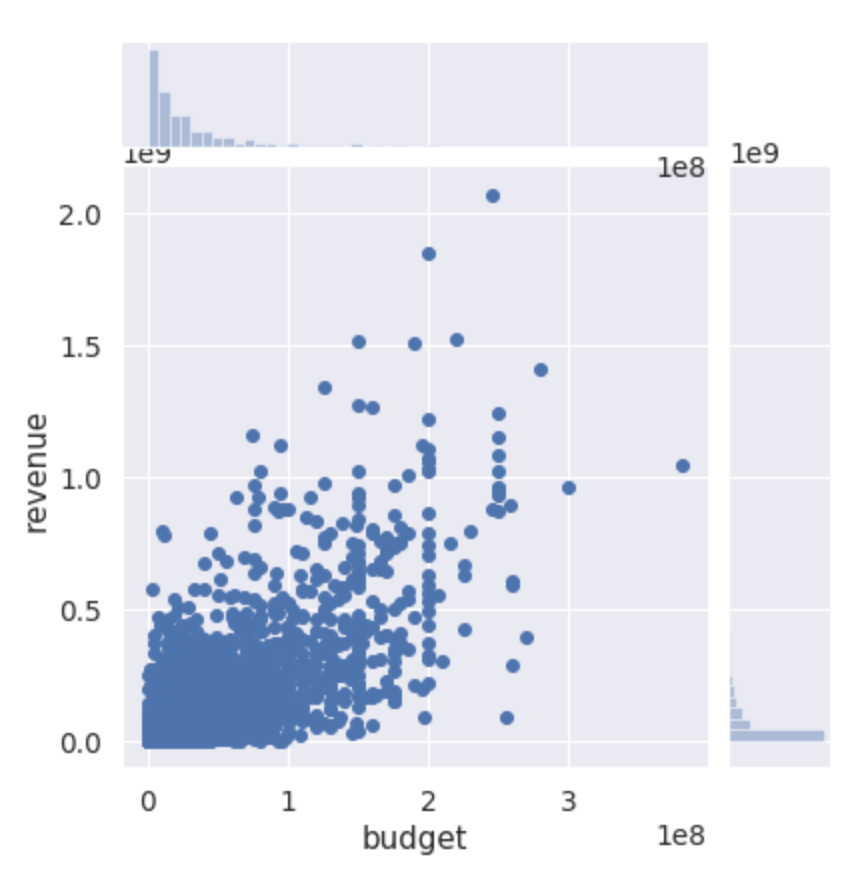
# Highest Grossing Films of All Time



Star Wars is the highest grossing individual film of all time with revenue of almost 2 billion dollars. The classic movie Titanic still figures high in the list in second place despite its release time and inflation.

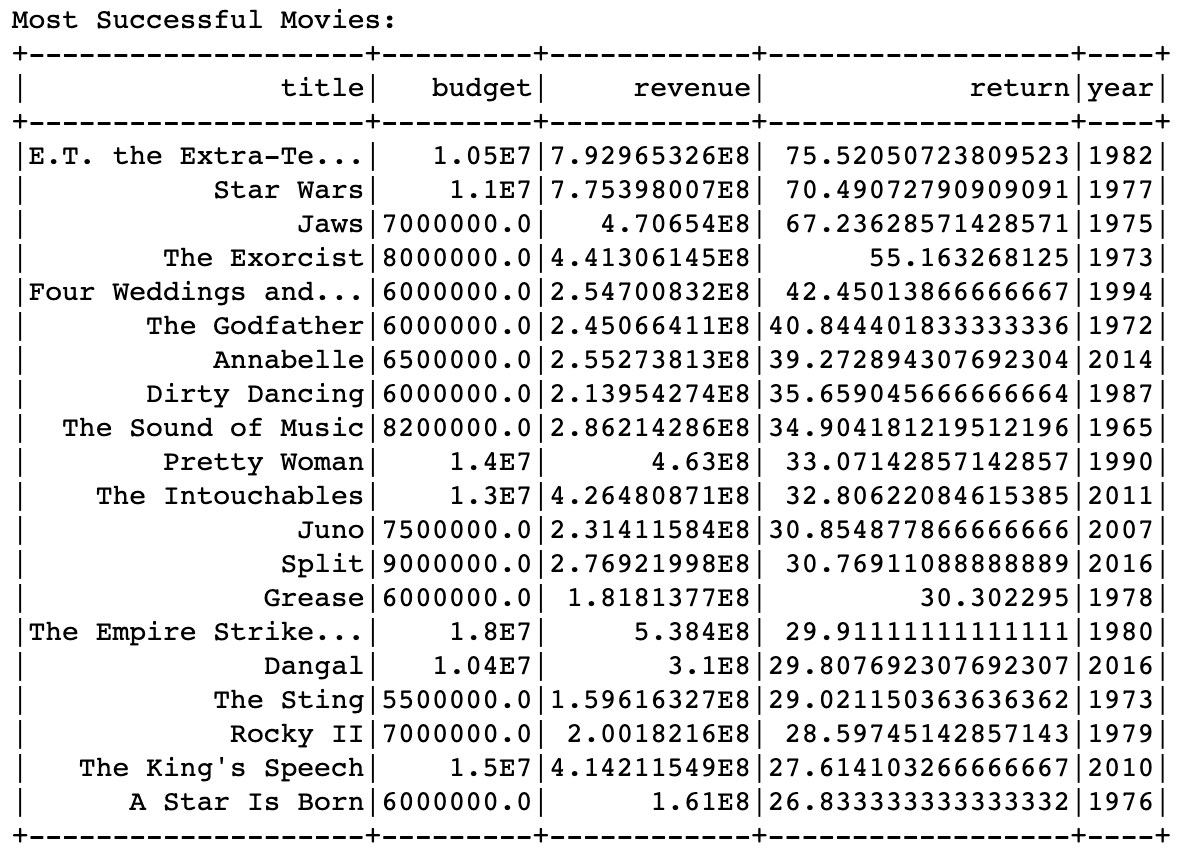
# 

# Returns or Movie Success Analysis



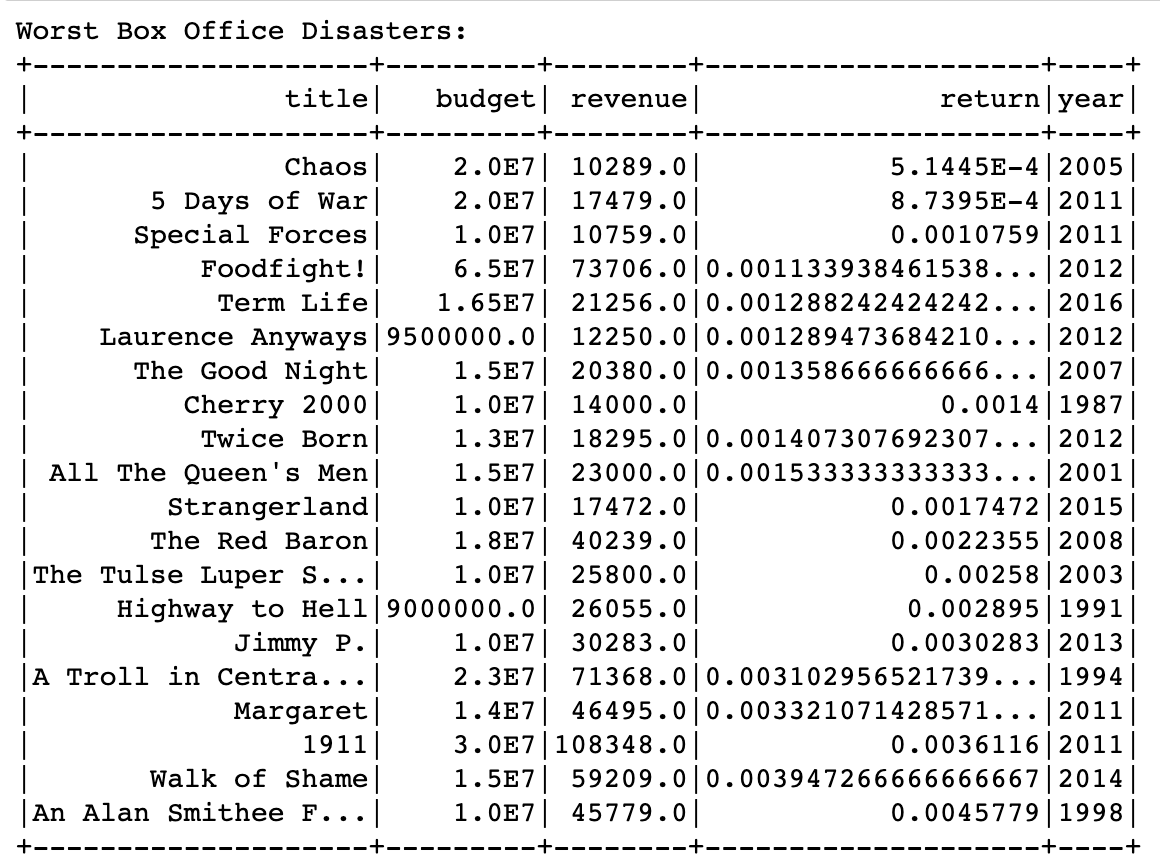
There’s a strong correlation between revenue and budget and this is evident from the fact that majority of the highest grossing movies are produced by the biggest production companies with high budgets. We take this metric return to evaluate the success of a movie and will also use it in recommendation.

# Most Successful Movies



Based on the return metric defined above , E.T. is the most successful individual movie with a revenue of nearly 75 times the outlay. Star Wars , which has featured highly in the other lists is unsurprisingly second in this list.

# Worst Box Office Disasters



Using the return metric , we found the Worst Box Office Disasters . Unironically , Chaos was the movie that featured highest in the list , giving a negative return on it’s outlay.

# 

# 

# Number of movies released on a particular day

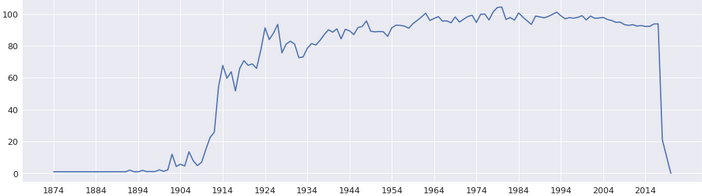
# 

# 

Unsurprisingly , most of the movies are released on Friday in preparation for the weekend . The rest of the days are closely separated but Friday is clearly when most of the movies are released .

# 

# Average Runtime through the years

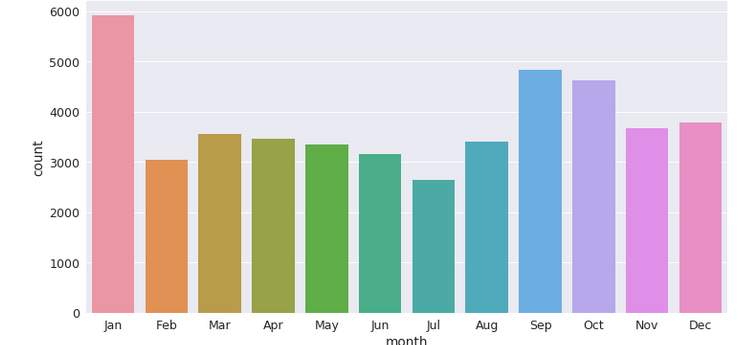


The average runtime of movies steadily increased with a peak of around 100 minutes. This may have to do with the better quality movies that have been slowly increasing.

**Number of movies released per month.**

# 

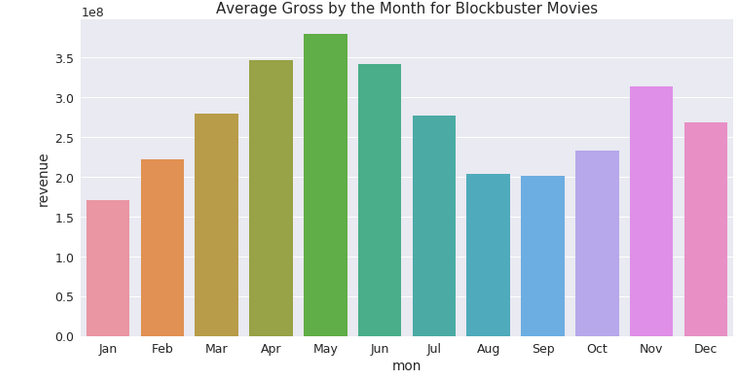
# 



January is the most popular month for releasing movies with nearly 6000 movies released in that month. As with the data on the number of days , the rest of the months are closely separated.

# Average Gross per month

# 

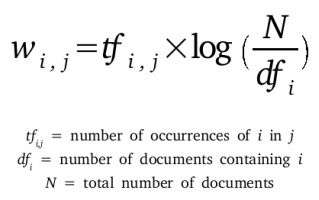


The data is evenly spread and it is very close , but May has the highest average gross for the movies with nearly 350million , April is a close second followed by June.

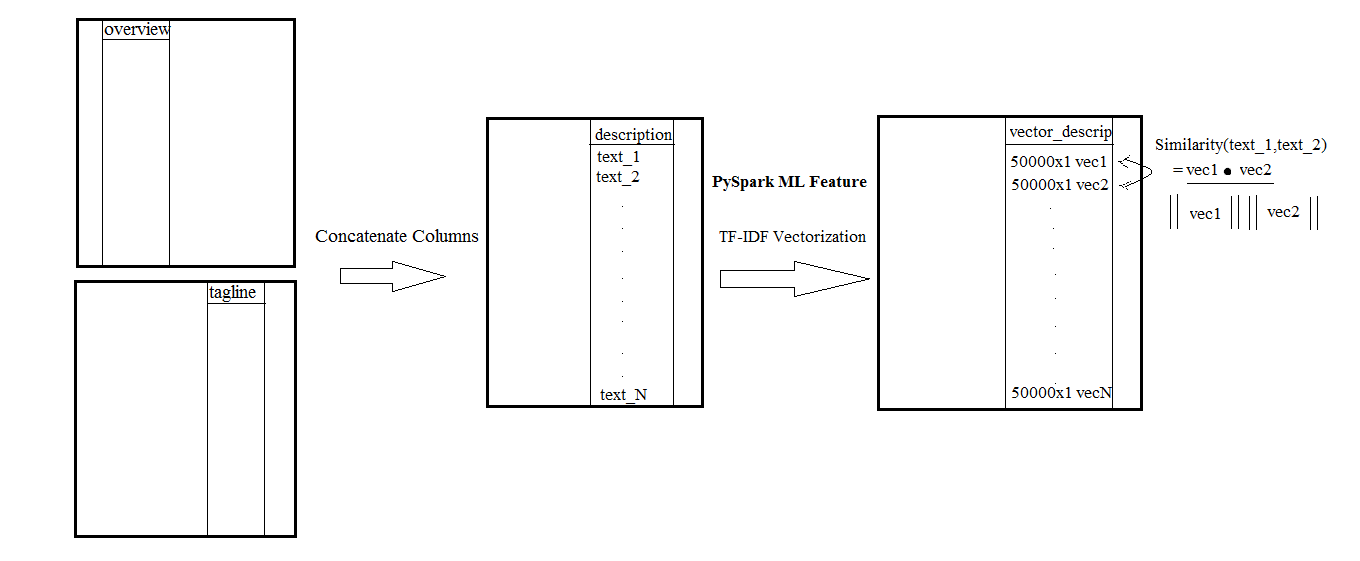
Description Based Recommendation System

Traditional Recommendation systems take into account the user preferences , attributes of a movie such as actors, directors. But we are focussing solely on the movie description for accomplishing our task. We need the recommendations to have movies with similar plot as the input query and therefore, we use a similarity metric and provide the movies with highest similarity with the input. The dataset has two columns overview and tagline of each movie and we concatenate these columns and make it into a single column as name it as description.

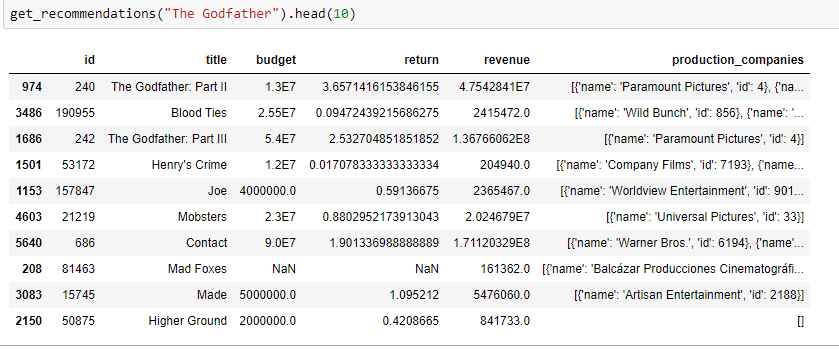
The transformation procedure we use for converting the text to a fixed size vector is known as TF-IDF Vectorizer ( Pyspark ML ). The algorithm is a statistical way to score words in a document. TF-IDF score of each word in all the documents are computed as following:



The first term is called as the term frequency and it is the number of occurences of the word I in document j. The second term is called as inverse document frequency and it indicates how common the word i is among all the documents. The score essentially rewards words with high frequency in the document and penalizes them for their prevalence among the documents. For instance, the word ‘the’ has no significant meaning as it occurs in all the documents and therefore the IDF penalizes it by giving a score of zero. We use this procedure for transforming our text column of description to a fixed size real values vector for carrying out further computations. The vectorizer first creates a corpus from all the rows of text and tokenizes them. For each word in each row, it computes the score as described above and finally, creates a vector for each entry. The entire process is depicted below in the diagram.



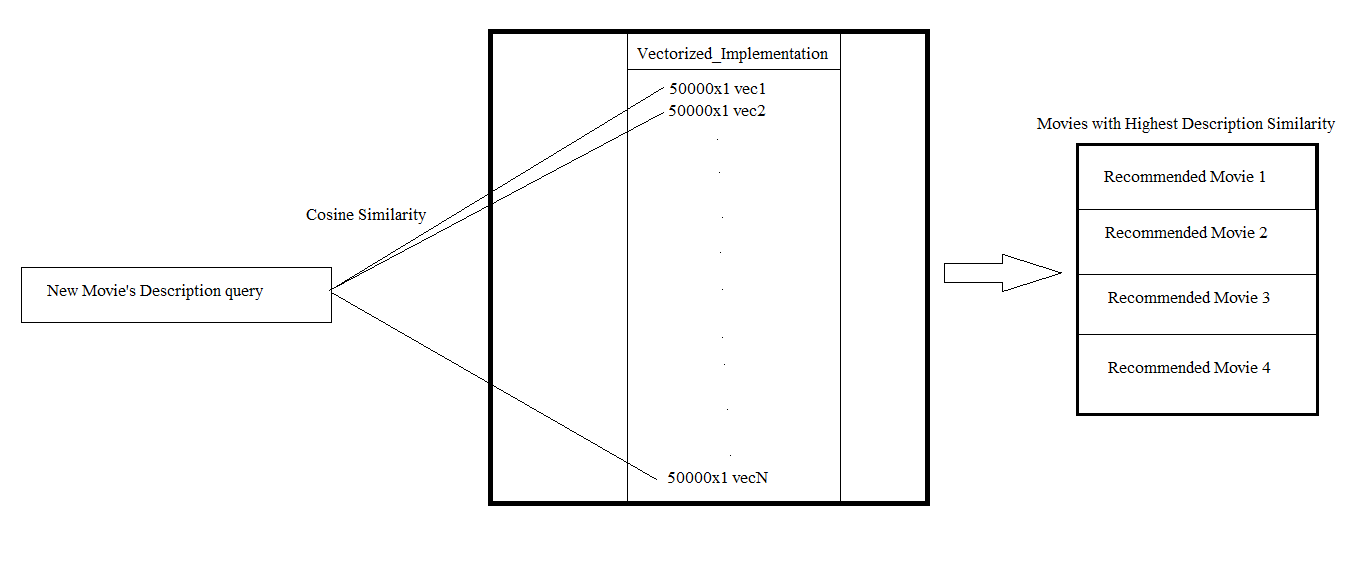
The similarity metric we use for assessing how similar two vectors are is cosine similarity metric. The metric results in 1 if the description is identical and results in 0 if the vectors are orthogonal. The Recommendation results are impressive in the sense that movies with similar plot lines are given as output with just the description given as input. The following two results indicate the type of recommendation the system makes.



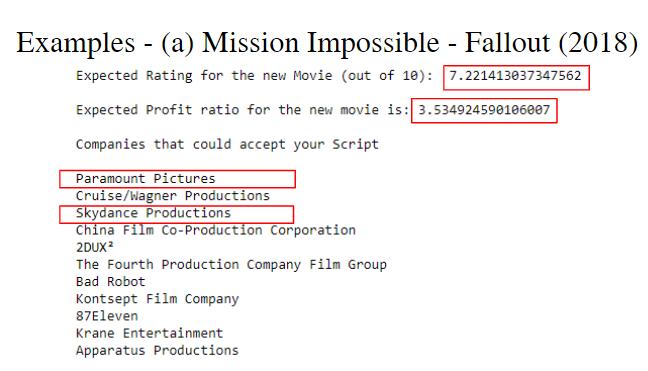
The first example is using the movie “The Godfather” as input. Widely regarded as one of the greatest films of all time, this is a classic example move for Crime fiction and Mafia. The results above are obviously close to the movie and it is apparent with just the titles. In particular, we see that the other Godfather sequels appear even without explicitly taking into account the movie name. Other movies also related to crime and gangsters. If we take for instance the movie “Mad Foxes” at the bottom, the plot of the movie is about a man taking revenge on a gang after they murder his family. We could easily see that there a lot of similar factors between the plots and this is the output by just taking a two-line description of the movie.

**Movie Plot Evaluation**

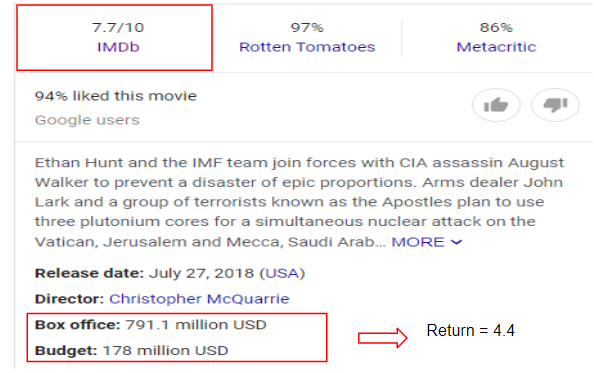
We take the recommender built above one step further and attempt to predict how good a movie plot is and what production companies are most likely to accept this script. This would be a good tool for script writers and production companies for objective decision making. There are about 6100 production companies in the US and it’s a laborious task for script writers to narrow down their search for selling their script. Also, for script writers to take appropriate inspiration and reference while working on a script, could use this tool to get the list of movies that have similar plots. The process is similar to recommendation but instead of giving as input an already existing movie, we create a dummy entry with our new movie’s description and then find the top recommendations for our new movie. For an input movie description query, we compute the cosine similarity with all the movies in the dataframe and pick the top 10 movies with highest similarity. The entire process in depicted below in the diagram

 For the recommended movies, we compute the average rating, average return (revenue/budget) and the set of production companies that produced these recommended films. The reason why the two metrics would work is that movie with similar plot elements would require similar type of manual cost and therefore, budget and revenue would have some correlation. For instance, if the script a scriptwriter is working on is a space movie, it would definitely require certain amount of high end computer graphics and high set costs. This system being driven by similar plot lines would give space movies as recommendation and most of those movies would also have used graphics and costly sets. This gives a minimum guarantee that the expected or predicted rating and return would act as baseline or a rough estimate. The movie descriptions we use for validating the system belong to movies that got released after 2017 since the database only has movies before 2017. The results of the plot evaluation are as follows.

The first movie description we gave as input is of Mission Impossible : Fallout which was released in 2018. This movie is part of the franchise Mission Impossible starring Tom cruise.



We see that the expected IMDB rating is about 7.2 and the expected profit ratio is about 3.53. The predicted list of companies that could accept this movie script are also present in the result. Now, we can look at the actual values and production company.

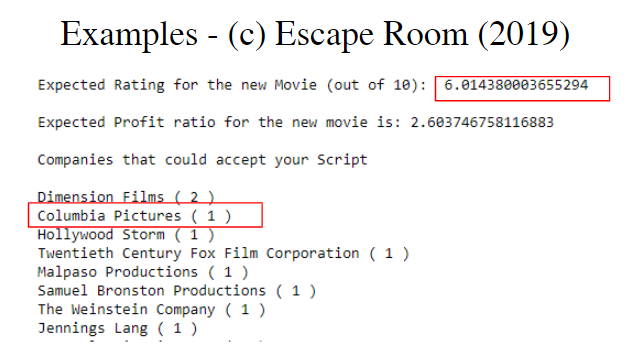


We can see that the actual IMDB rating the movie is about 7.7 and the revenue to budget ratio is about 4.4. These two values match with the predicted values with just reasonable error. We will also look at the actual production company that produced this movie to see if it exists in our predicted list.

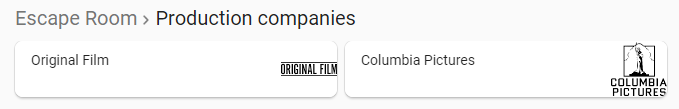


We see that the two companies which produced the movie are right at top of the predicted companies list which are marked using a red box.

Another Example that is more interesting is of a movie called Escape Room (2019). We will first see the result outputted by our system and then move on to the actual values.



We see that the IMDB rating is around 6 and also a list of predicted companies that could accept this kind of script. Now, we will take a look at the actual ground-truth and compare.



One of the production companies that produced this movie is Columbia Pictures and interestingly, the same company has made it into the predicted list right at the top. This is a fascinating result because there are about 6100 production companies in the US and for a system to narrow it down to the most optimal value would be extremely useful for script writers searching for companies for selling their script. We will also look at the actual IMDB rating to see if the expected rating is close enough.



We see below that the rating is 6.3 and therefore, close to our prediction.

**Conclusion**

Therefore, our system works reasonably well with just a two line description or plot of the movie. This system just as it is or expanded further would be of immense help to script writers for narrowing down their search and finding the most profitable production company with respect to their particular script, and would be of great help to production companies to analyze and evaluate hundreds of plots received everyday and make adjustments to the script to increase the expected rating and return.

Further Improvements could be done by expanding the description to a brief summary of the movie thereby including some specifics of the plot and as result, improving the recommendation. The second improvement could be extending our prediction of production companies to predicting the cast and crew that would most likely accept a particular script.

References

<https://www.kaggle.com/rounakbanik/the-movies-dataset>

<https://spark.apache.org/docs/latest/api/python/pyspark.ml.html>

<https://seaborn.pydata.org/>