Movie Profit Trend Visualizer and Predictor

Authors : Abikamet Anbunathan, Hyerin Seok, Scott Scheraga
CSCI 5622 - Spring 2020, Professor Quigley

Problem Space

The film industry involves costly and risky investments and movie studios want to maximize profit and reduce risk especially for large-budget productions. We believe that this risk can be reduced through the use of unsupervised machine learning techniques to cluster together movie genres of movies made in the last 20 years, note the profit the movies of that genre over time, through the use of genre and descriptive terms, forecast the profit of a future movies

This could potentially reduce the need for production studio's resources to be allocated to research tasks.

Manually labeling broad film categories would be time consuming for studios, and unsupervised clustering methods may be able to determine categories that a human categorizer would not notice.

Lastly, the sub-problem of automatically clustering movies together based on characteristics could be useful for digital distribution companies such as Hulu or Netflix, in order to display similar movies together in their interface, easing the browsing of their content.

Our solution is novel because of our use of a two-stage k-means classifier, optimized with elbow analysis in order to better analyze films through subcategorization.

Approach

Our approach involved processing movie genre and overview text columns in the TMDB dataset (discussed below) using NLTK. To process this data, we ran each text column through filtering steps that made the text lowercase, removed punctuation and most stop-words, lemmatized, stemmed and then tokenized the words. We then used Tf-idf (term frequency-inverse document frequency) vectorization in order to find the frequency of words used in each text block. We then performed Kmeans clustering on the movie genres, with distance being set as 1-cosine_similarity distance between the films. We utilized elbow analysis to find that 7 clusters was the optimal k for genre clusters.

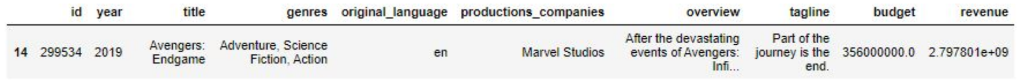
Next, we performed the same TF-idf process on movie overviews, and performed 7 separate k-means clustering operations. The top 3 (lowest) TF-idf word scores for each genre, and top 5 word score for each overview are displayed in this poster's main chart.

Once we were able to label each film with a genre cluster and overview cluster, we were able to compare similar films with linear regression. While we had initially set out to predict only a single year's film profits (revenue-cost) for a given overview cluster, we decided to train our regression models on the years of 2000 to 2015, and predict profits from 2015 to 2019, in order to get a better sense of the accuracy of our model.We also recorded the accuracy of our model with Mean Absolute error, Mean squared error and Root mean squared error.

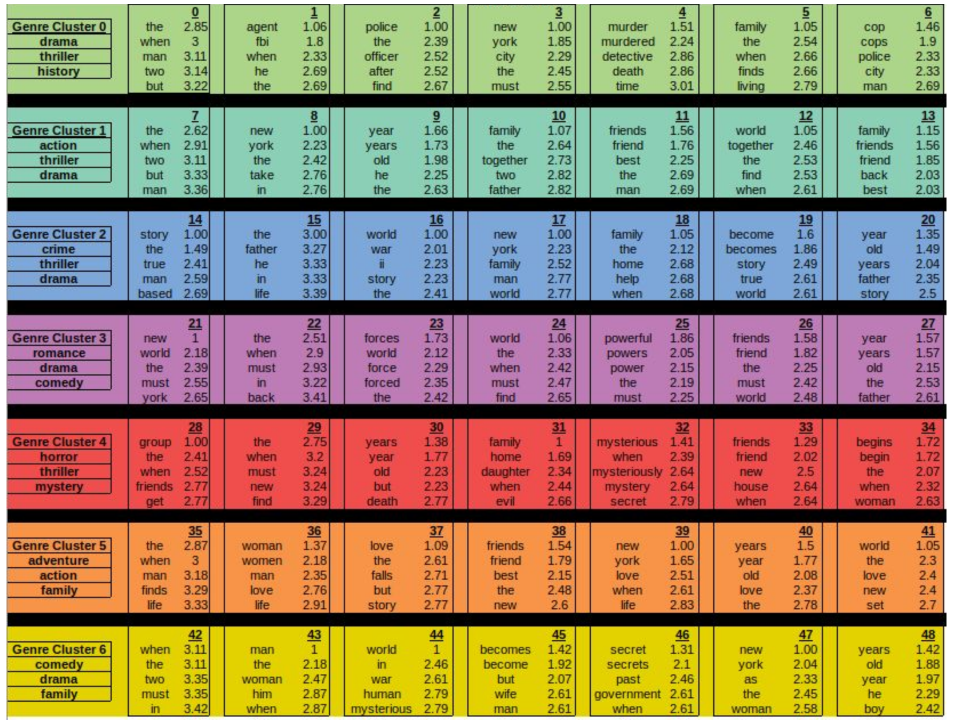
Data

TMDB API was our choice of database for collecting movie data. We compiled the 3685 entries of the released year, movie title, genres, original language, movie description(synopsis, tagline), budget and revenue. We ended up with 10 features(id, year, title, genres, language, prod company, overview, tagline, budget, revenue).

▼ Example entry after filtered and compiled of the movie Avenger: Endgame



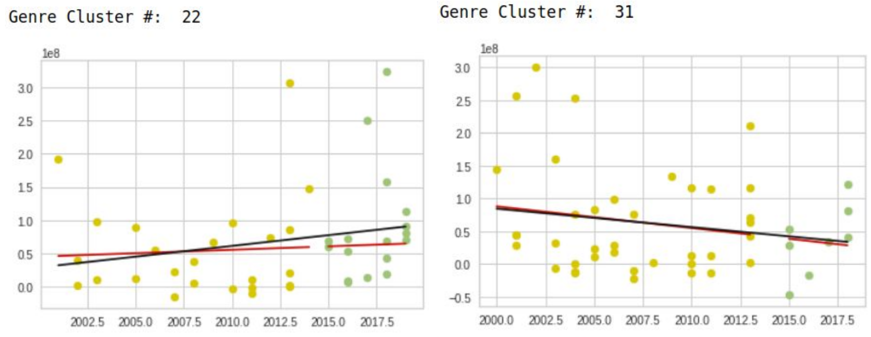
	id	year	title	genres	original_language	productions_companies	overview	tagline	budget	revenue
14	299534	2019	Avengers: Endgame	Adventure, Science Fiction, Action	en	Marvel Studios	After the devastating events of Avengers: Infi...	Part of the journey is the end.	356000000.0	2.797801e+09



		0		1		2		3		4		5		6
Genre Cluster 0	the	2.85	agent	1.06	police	1.00	new	1.00	murder	1.51	family	1.05	cop	1.46
drama	when	3	fbi	1.8	the	2.39	york	1.85	murdered	2.24	the	2.54	cops	1.9
thriller	man	3.11	when	2.33	officer	2.52	city	2.29	detective	2.86	when	2.66	police	2.33
history	two	3.14	he	2.69	after	2.52	the	2.45	death	2.86	finds	2.66	city	2.33
	but	3.22	the	2.69	find	2.67	must	2.55	time	3.01	living	2.79	man	2.69
		7		8		9		10		11		12		13
Genre Cluster 1	the	2.62	new	1.00	year	1.66	family	1.07	friends	1.56	world	1.05	family	1.15
action	when	2.91	york	2.23	years	1.73	the	2.64	friend	1.76	together	2.46	friends	1.56
thriller	two	3.11	the	2.42	old	1.98	together	2.73	best	2.25	the	2.53	friend	1.85
drama	but	3.33	take	2.76	he	2.25	two	2.82	the	2.69	find	2.53	back	2.03
	man	3.36	in	2.76	the	2.63	father	2.82	man	2.69	when	2.61	best	2.03
		14		15		16		17		18		19		20
Genre Cluster 2	story	1.00	the	3.00	world	1.00	new	1.00	family	1.05	become	1.6	year	1.35
crime	the	1.49	father	3.27	war	2.01	york	2.23	the	2.12	becomes	1.86	old	1.49
thriller	true	2.41	he	3.33	ii	2.23	family	2.52	home	2.68	story	2.49	years	2.04
drama	man	2.59	in	3.33	story	2.23	man	2.77	help	2.68	true	2.61	father	2.35
	based	2.69	life	3.39	the	2.41	world	2.77	when	2.68	world	2.61	story	2.5
		21		22		23		24		25		26		27
Genre Cluster 3	new	1	the	2.51	forces	1.73	world	1.06	powerful	1.86	friends	1.58	year	1.57
romance	world	2.18	when	2.9	world	2.12	the	2.33	powers	2.05	friend	1.82	years	1.57
drama	the	2.39	must	2.93	force	2.29	when	2.42	power	2.15	the	2.25	old	2.15
comedy	must	2.55	in	3.22	forced	2.35	must	2.47	the	2.19	must	2.42	the	2.53
	york	2.65	back	3.41	the	2.42	find	2.65	must	2.25	world	2.48	father	2.61
		28		29		30		31		32		33		34
Genre Cluster 4	group	1.00	the	2.75	years	1.38	family	1	mysterious	1.41	friends	1.29	begins	1.72
horror	the	2.41	when	3.2	year	1.77	home	1.69	when	2.39	friend	2.02	begin	1.72
thriller	when	2.52	must	3.24	old	2.23	daughter	2.34	mysteriously	2.64	new	2.5	the	2.07
mystery	friends	2.77	new	3.24	but	2.23	when	2.44	mystery	2.64	house	2.64	when	2.32
	get	2.77	find	3.29	death	2.77	evil	2.66	secret	2.79	when	2.64	woman	2.63
		35		36		37		38		39		40		41
Genre Cluster 5	the	2.87	woman	1.37	love	1.09	friends	1.54	new	1.00	years	1.5	world	1.05
adventure	when	3	women	2.18	the	2.61	friend	1.79	york	1.65	year	1.77	the	2.3
action	man	3.18	man	2.35	falls	2.71	best	2.15	love	2.51	old	2.08	love	2.4
family	finds	3.29	love	2.76	but	2.77	the	2.48	when	2.61	love	2.37	new	2.4
	life	3.33	life	2.91	story	2.77	new	2.6	life	2.83	the	2.78	set	2.7
		42		43		44		45		46		47		48
Genre Cluster 6	when	3.11	man	1	world	1	becomes	1.42	secret	1.31	new	1.00	years	1.42
comedy	the	3.11	the	2.18	in	2.46	become	1.92	secrets	2.1	york	2.04	old	1.88
drama	two	3.35	woman	2.47	war	2.61	but	2.07	past	2.46	as	2.33	year	1.97
family	must	3.35	him	2.87	human	2.79	wife	2.61	government	2.61	the	2.45	he	2.29
	in	3.42	when	2.87	mysterious	2.79	man	2.61	when	2.61	woman	2.58	boy	2.42

▲ Final clusters with their top 3 genres, in their genre cluster

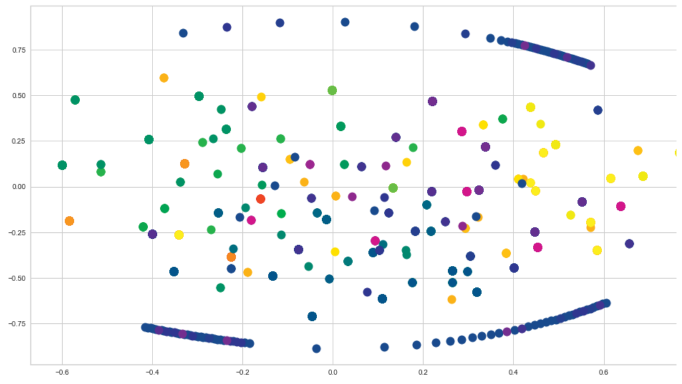
▼ Overview Cluster 22 & 31 Linear Regression Plot



```
Mean Absolute Error: 52477290.298334315
Mean Squared Error: 7404443370197355.0
Root Mean Squared Error: 86049075.3593399
```

```
Mean Absolute Error: 40215282.968290925
Mean Squared Error: 2721900994373635.0
Root Mean Squared Error: 52171841.01000879
```

(Overview Clusters 22 and 31. Yellow Data: Pre-2015 Profits, Green: Post-2015 profits, Red line: linear regression curve of pre-2015 data, Black line: linear regression curve of all data)



▲ Clustering of Multidimensional scale tf idf data

Results

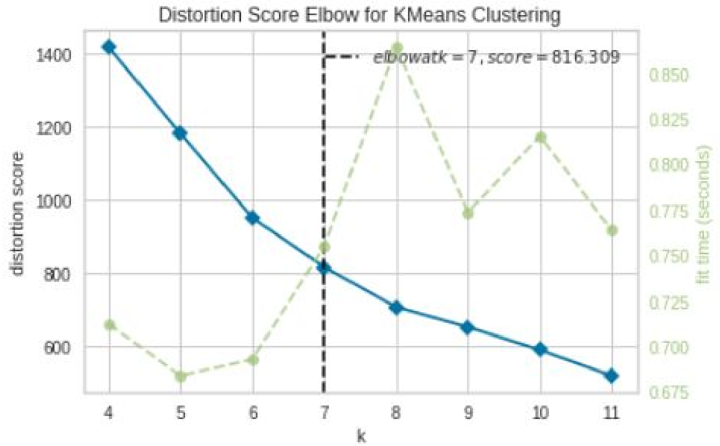
While we are unable to display all of our linear regression plots here, we show overview clusters 22 and 31 as examples of upward and downward trending plots. In both cases, our regressor (shown in red) was fairly accurate to the overall data's trendline, shown in (black). The movie profit trend data often has a large number of outliers, and even distinct trends often have wide scatterplots.

A potential source of error in determining movie genres was stop-words. We were unable to completely remove them, despite a large amount of undertaken effort. We noticed that the word "the" is in a wide variety of overview clusters. While the English stopword set and a custom stopword set both included the word "the", further investigation will need to occur in order to determine if the word "the" reappears due to other words turning into "the" due to our sequence of lemmatization and stemming.

Discussion

Our results show some promise in predicting movie trends. The movie overview k-means clustering results showed surprisingly good results, such as picking out WW2 films, and horror films. While further regression models may need to be investigated, in order to reduce error, we believe that even standard linear regression can distinguish upward from downward trends. As this dataset has a very large number of outliers, a higher-order regressor (ie. a 3+ order polynomial regressor) may likely overfit the data.

Other potential future work may involve lasso or ridge regression in order to create a better-fitting model, with a smaller risk of overfitting.



(Elbow Analysis results for determining optimal k for Genre-cluster K-means clustering)