

Objective is to make a lightweight app that can be hosted on GitHub pages like the Belly Button Biodiversity that provides a list of movies that might be of interest for viewing.

Dataset

(ml-25m) describes 5-star rating and free-text tagging activity from [MovieLens](<http://movielens.org>)

Files to leverage: movies.csv, genome-score.csv, ratings.csv, tags.csv

ML Strategy

The genome-scoring file contains tags like action packed, alternate endings, android, animals which are then scored with a relevance for each movie.

Using this relevance scoring on the movie tags it is desired to apply an unsupervised machine learning model to group movies that might be also of similar interest.

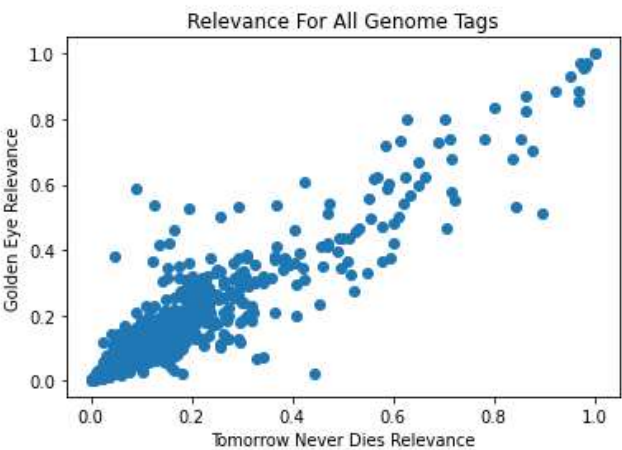
ML Strategy

Table of tagId 1 '007' which with a relevance filter of .95 shows bond movies

```
1 # Lets see how well this genome scoring seems to find relevant movies by examining tagId 1 which represent texts '007'
2 check1_df=df.loc[(df['tagId']==1) & (df['relevance']>.95)]
3 check1_df.head(20)
```

movieId		title	genres	tagId	relevance
10152	10	GoldenEye (1995)	Action Adventure Thriller	1.0	0.99950
1725966	1722	Tomorrow Never Dies (1997)	Action Adventure Thriller	1.0	0.99975
2385890	2376	View to a Kill, A (1985)	Action Adventure Thriller	1.0	0.99975
2991659	2947	Goldfinger (1964)	Action Adventure Thriller	1.0	0.99950
2992787	2948	From Russia with Love (1963)	Action Adventure Thriller	1.0	0.99975
2993915	2949	Dr. No (1962)	Action Adventure Thriller	1.0	0.99975
3026632	2983	Ipcress File, The (1965)	Thriller	1.0	0.97025
3033400	2989	For Your Eyes Only (1981)	Action Adventure Thriller	1.0	0.99975
3034528	2990	Licence to Kill (1989)	Action Adventure Thriller	1.0	0.99950
3035656	2991	Live and Let Die (1973)	Action Adventure Thriller	1.0	0.99975
3037912	2993	Thunderball (1965)	Action Adventure Thriller	1.0	0.99975
3131541	3082	World Is Not Enough, The (1999)	Action Adventure Thriller	1.0	0.99975
3668538	3633	On Her Majesty's Secret Service (1969)	Action Adventure Romance Thriller	1.0	0.99950
3670794	3635	Spy Who Loved Me, The (1977)	Action Adventure Thriller	1.0	0.99975
3674178	3638	Moonraker (1979)	Action Adventure Sci-Fi Thriller	1.0	0.99950
3675306	3639	Man with the Golden Gun, The (1974)	Action Adventure Thriller	1.0	0.99975
4034032	3984	Diamonds Are Forever (1971)	Action Adventure Thriller	1.0	0.99975
4055466	4005	Living Daylights, The (1987)	Action Adventure Thriller	1.0	0.99975
5576150	5497	Our Man Flint (1965)	Adventure Comedy Sci-Fi	1.0	0.99050
5843544	5796	Casino Royale (1967)	Action Adventure Comedy	1.0	0.99725

Check of all Tags for Relevance
Correlation between Two Similar
Bond Movies



ML Strategy

Perform Unsupervised Clustering with X like below

Genome Tags

Movie Ids

tagId	1.0	2.0	3.0	4.0	5.0	6.0	7.0	8.0	9.0	10.0	...	1119.0	1120.0	1121.0	1122.0	1123.0	1124.0
movielid																	
1	0.02875	0.02375	0.06250	0.07575	0.14075	0.14675	0.06350	0.20375	0.2020	0.03075	...	0.04050	0.01425	0.03050	0.03500	0.14125	0.05775
2	0.04125	0.04050	0.06275	0.08275	0.09100	0.06125	0.06925	0.09600	0.0765	0.05250	...	0.05250	0.01575	0.01250	0.02000	0.12225	0.03275
3	0.04675	0.05550	0.02925	0.08700	0.04750	0.04775	0.04600	0.14275	0.0285	0.03875	...	0.06275	0.01950	0.02225	0.02300	0.12200	0.03475
4	0.03425	0.03800	0.04050	0.03100	0.06500	0.03575	0.02900	0.08650	0.0320	0.03150	...	0.05325	0.02800	0.01675	0.03875	0.18200	0.07050
5	0.04300	0.05325	0.03800	0.04100	0.05400	0.06725	0.02775	0.07650	0.0215	0.02975	...	0.05350	0.02050	0.01425	0.02550	0.19225	0.02675

Relevance Values relating Movies to the Tags

As output Get Clustering of movies to utilize

Cluster	title	genre
107	Toy Story (1995)	Adventure Animation Children Comedy Fantasy
50	Jumanji (1995)	Adventure Children Fantasy
41	Grumpier Old Men (1995)	Comedy Romance
41	Waiting to Exhale (1995)	Comedy Drama Romance
136	Father of the Bride Part II (1995)	Comedy
82	Heat (1995)	Action Crime Thriller
2	Sabrina (1995)	Comedy Romance
16	Tom and Huck (1995)	Adventure Children
78	Sudden Death (1995)	Action
48	GoldenEye (1995)	Action Adventure Thriller
125	American President, The (1995)	Comedy Drama Romance
170	Dracula: Dead and Loving It (1995)	Comedy Horror
48	Balto (1995)	Adventure Animation Children
175	Nixon (1995)	Drama
12	Cutthroat Island (1995)	Action Adventure Romance
155	Casino (1995)	Crime Drama
60	Sense and Sensibility (1995)	Drama Romance
119	Four Rooms (1995)	Comedy
16	Ace Ventura: When Nature Calls (1995)	Comedy
12	Money Train (1995)	Action Comedy Crime Drama Thriller
52	Get Shorty (1995)	Comedy Crime Thriller
16	Copycat (1995)	Crime Drama Horror Mystery Thriller
4	Assassins (1995)	Action Crime Thriller
160	Powder (1995)	Drama Sci-Fi
181	Leaving Las Vegas (1995)	Drama Romance
128	Othello (1995)	Drama

Database Strategy

SQLite file contains ML output of clustering groups



Cluster	title	genre
107	Toy Story (1995)	Adventure Animation Children Comedy Fantasy
50	Jumanji (1995)	Adventure Children Fantasy
41	Grumpier Old Men (1995)	Comedy Romance
41	Waiting to Exhale (1995)	Comedy Drama Romance
136	Father of the Bride Part II (1995)	Comedy
82	Heat (1995)	Action Crime Thriller
2	Sabrina (1995)	Comedy Romance
16	Tom and Huck (1995)	Adventure Children
78	Sudden Death (1995)	Action
48	GoldenEye (1995)	Action Adventure Thriller
125	American President, The (1995)	Comedy Drama Romance
170	Dracula: Dead and Loving It (1995)	Comedy Horror
48	Balto (1995)	Adventure Animation Children
175	Nixon (1995)	Drama
12	Cutthroat Island (1995)	Action Adventure Romance
155	Casino (1995)	Crime Drama
60	Sense and Sensibility (1995)	Drama Romance
119	Four Rooms (1995)	Comedy
16	Ace Ventura: When Nature Calls (1995)	Comedy
12	Money Train (1995)	Action Comedy Crime Drama Thriller
52	Get Shorty (1995)	Comedy Crime Thriller
16	Copycat (1995)	Crime Drama Horror Mystery Thriller
4	Assassins (1995)	Action Crime Thriller
160	Powder (1995)	Drama Sci-Fi
181	Leaving Las Vegas (1995)	Drama Romance
128	Othello (1995)	Drama

SQLite file also contains a summary of ratings like average rating, count of ratings

SQLite file also contains a using NLP of user tags from the tags.csv by movieId for the word cloud



Movie Finder

Use the interactive charts below to explore the dataset



Movie Finder

Use the interactive charts below to explore the dataset

D3 word
movieid :
Updates
each mov

948

948

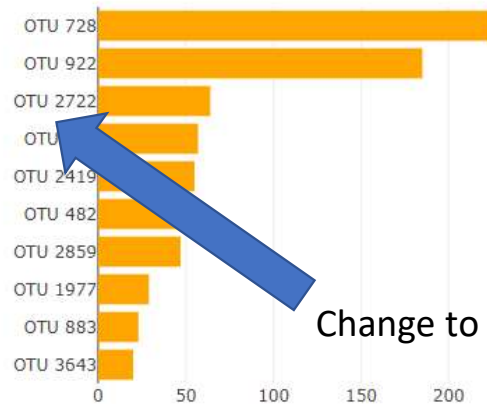
ID: 948

ID: 948

Rating Counts 93

Average Rating: 4.5

Genre:



OTU	Count (approx.)
OTU 728	210
OTU 922	185
OTU 2722	60
OTU 2419	55
OTU 482	50
OTU 2859	45
OTU 1977	30
OTU 883	25
OTU 3643	20

OTU	Count (approx.)
OTU 728	210
OTU 922	185
OTU 2722	60
OTU 2419	55
OTU 482	50
OTU 2859	45
OTU 1977	30
OTU 883	25
OTU 3643	20

