

Clustering of top 250 movies from IMDB

Mustafa Panbiharwala

Siddharth Ajit

Broad goals

- Find out if there's any intrinsic pattern or clusters among top rated movies (Based on IMDb Ratings).
- Identify similarities among the movies of the same cluster.
- Use Dimensionality Reduction to improve clustering output
- Repeat the above steps for different clustering algorithms.

Data acquisition

- Movie features were obtained using OMDB's API.
- Plot summary were scrapped from IMDB

<https://www.imdb.com/chart/top>.

Top Rated Movies


Top 250 as rated by IMDb Users



SHARE

Showing 250 Titles

Sort by: Ranking

Rank & Title	IMDb Rating	Your Rating	
 1. The Shawshank Redemption (1994)	★ 9.2	☆	+ 
 2. The Godfather (1972)	★ 9.2	☆	+ 
 3. The Godfather: Part II (1974)	★ 9.0	☆	+ 
 4. The Dark Knight (2008)	★ 9.0	☆	+ 
 5. 12 Angry Men (1957)	★ 8.9	☆	+ 
 6. Schindler's List (1993)	★ 8.9	☆	+ 

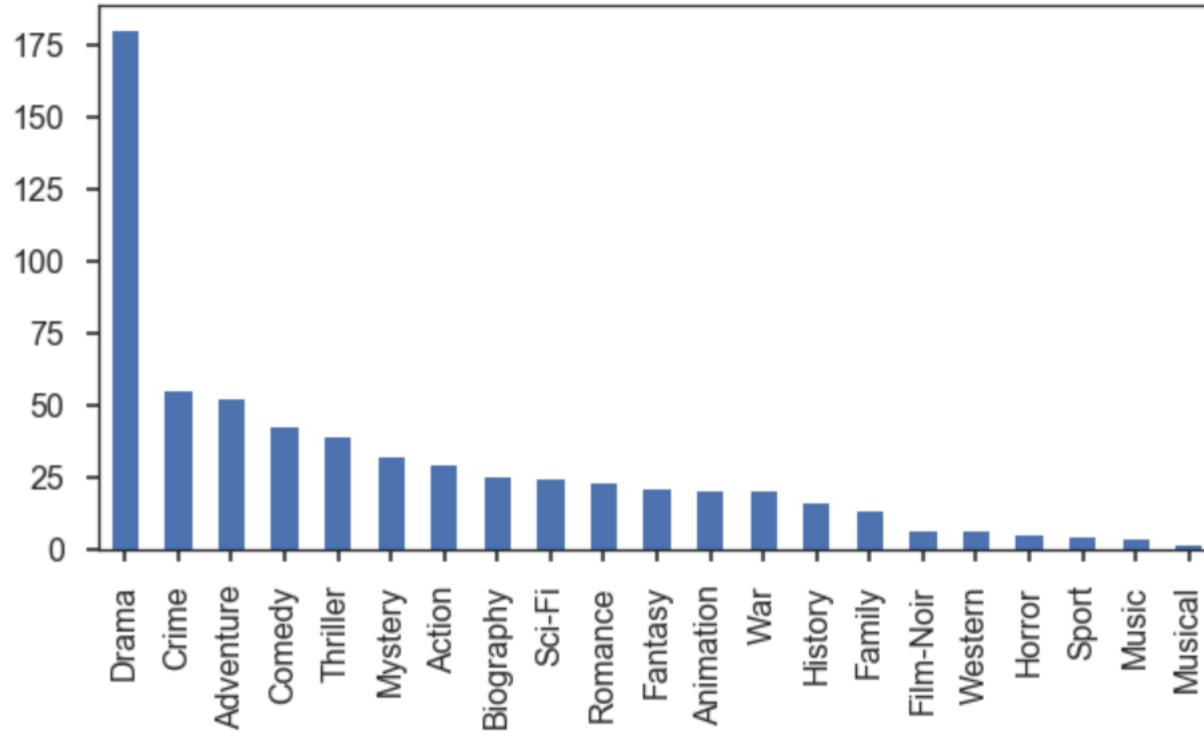
- Data was directly acquired from OMDB's Database through it's public API.
- The API had movie's IMDB id as one of the outputs.
- Used movie's IMDB to scrap the data from IMDB's website.
- OMDB's API had the following Output-

```
Index([u'Plot', u'Rated', u'Response', u'Language', u'Title', u'Country',  
      u'Writer', u'Metascore', u'imdbRating', u'Director', u'Released',  
      u'Actors', u'Year', u'Genre', u'Awards', u'Runtime', u'Type', u'Poster',  
      u'imdbVotes', u'imdbID'],  
      dtype='object')
```

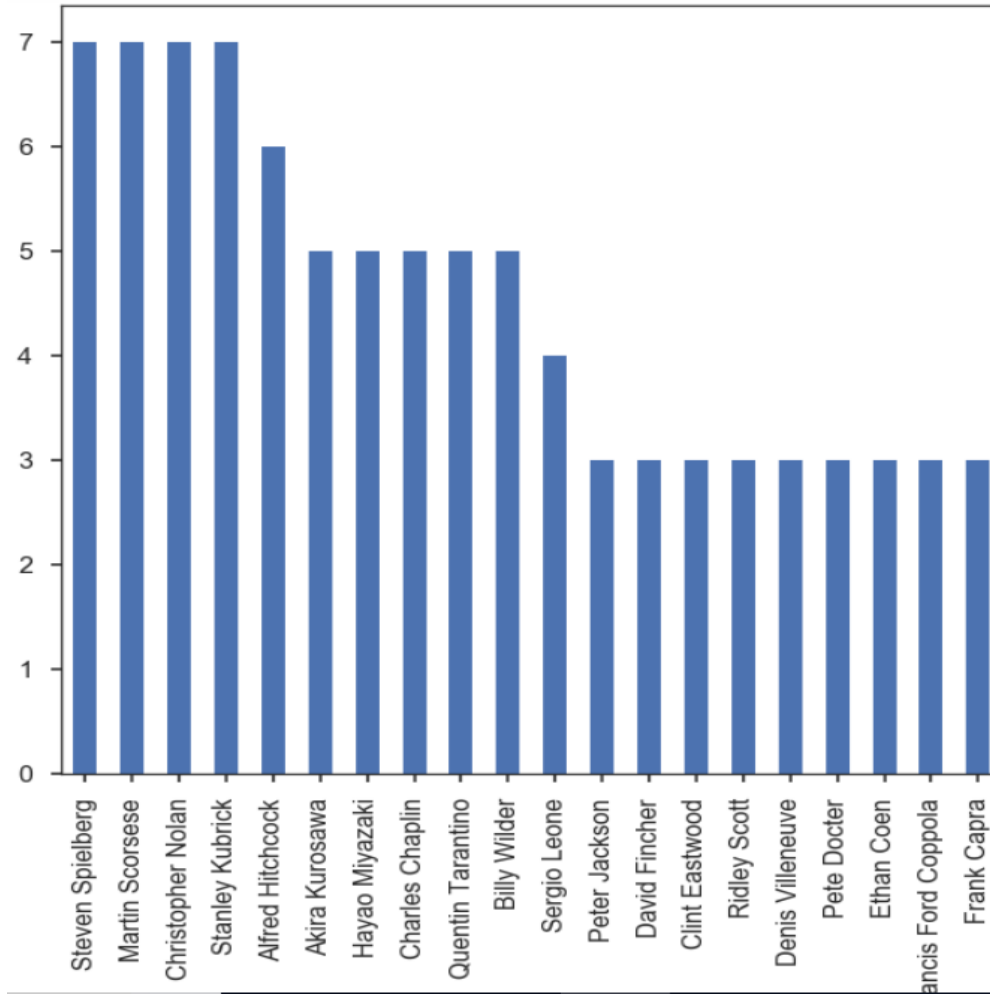
Feature selection

- Features used for clustering : year, runtime ,genre director, actors, plot, language, country.
- Certain features like year of release was discretized to categorical variables by choosing a suitable cutoff value.
- For text heavy columns, important words were pulled after cleaning and sorting according to TF-IDF scores.

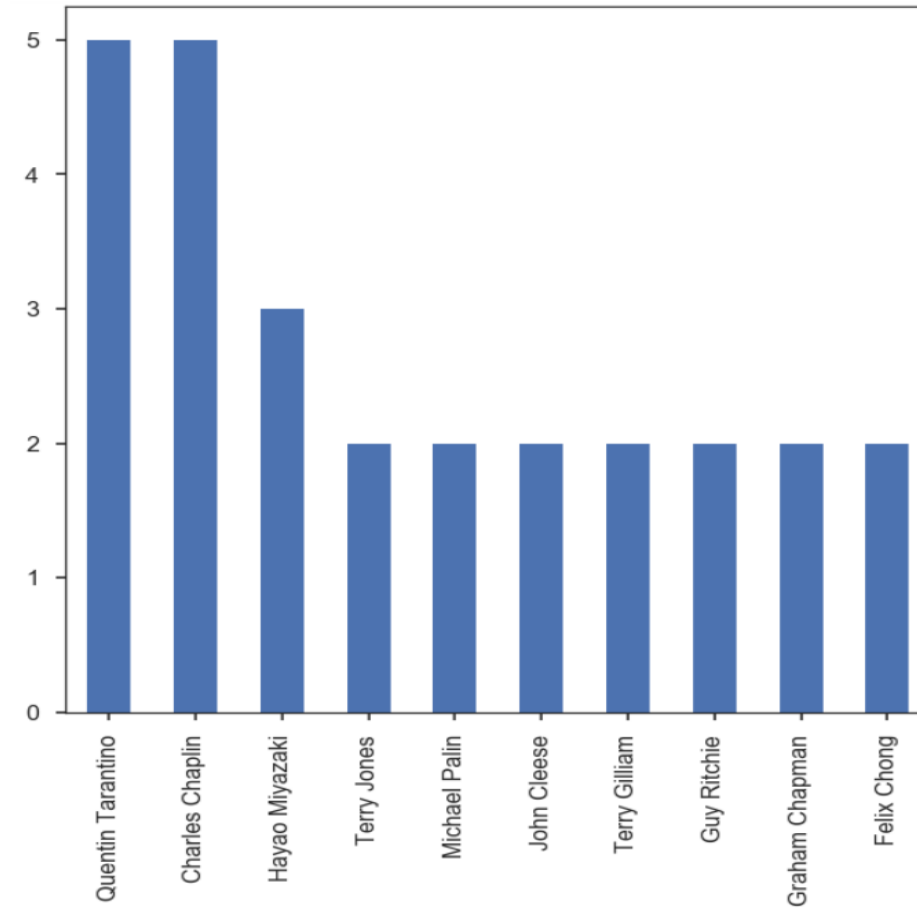
Genre



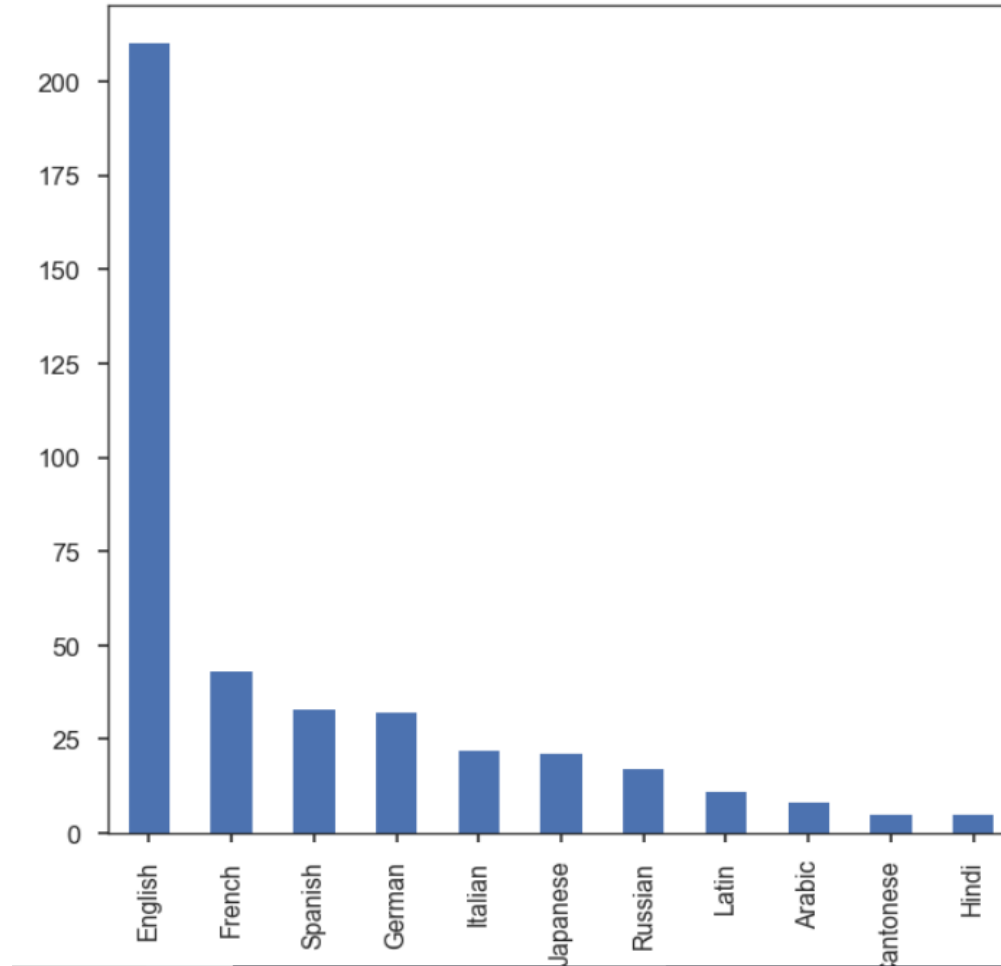
Actors



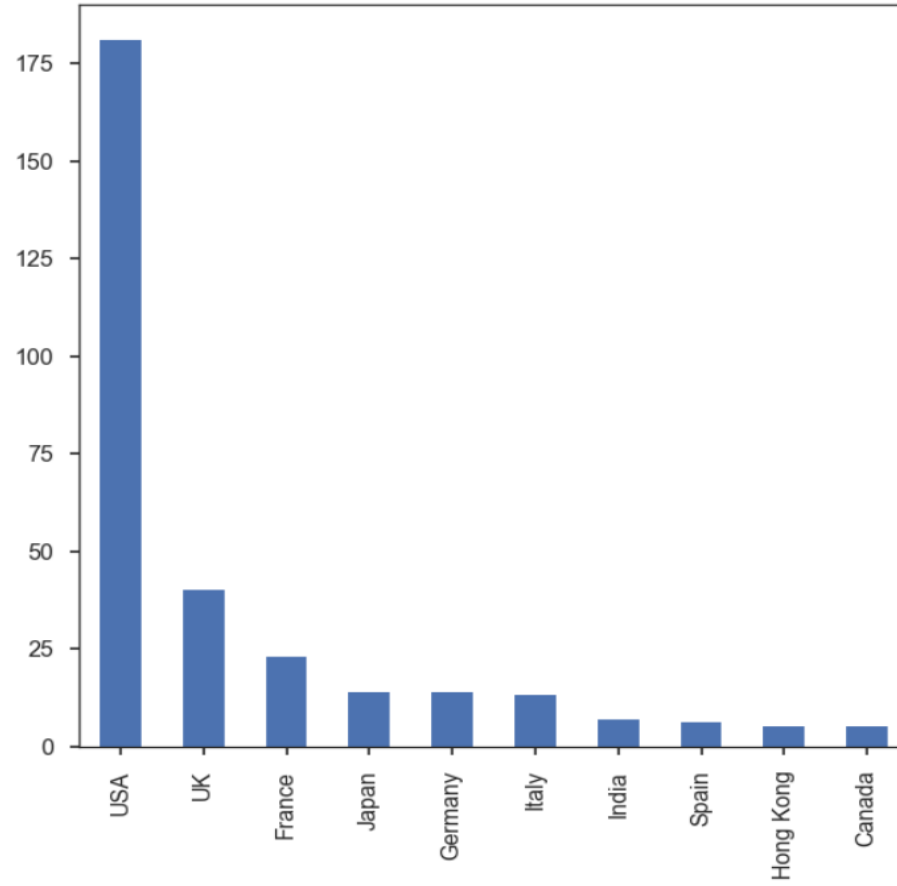
Directors



Language

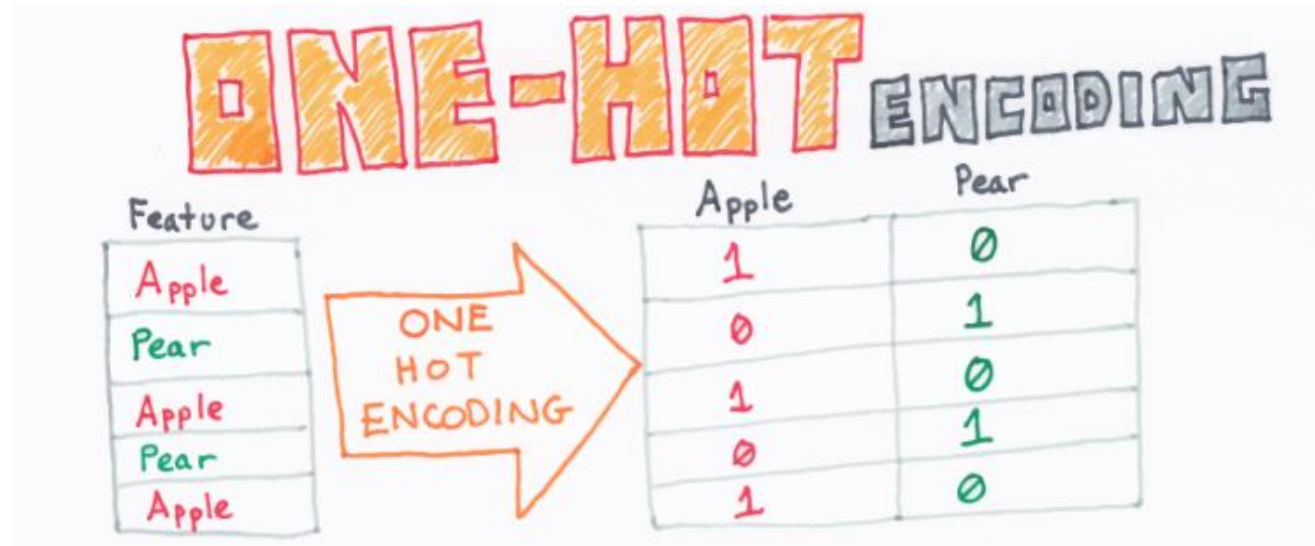


Countries



Feature representation

- Features – Genre, Actors, Directors, Language and Countries were one hot encoded.



Text processing of plot summary

These are standard procedures for NLP problems.

- Stripping of all punctuations, stop words etc.
- Tokenizing – breaking text into sentences and words.
- Stemming – reducing the inflected word to the root form.

TF-IDF

- Stands for term frequency–inverse document frequency
- Reveals the importance of a word to a document in a corpus.

$TF(t) = (\text{Number of times term } t \text{ appears in a document}) / (\text{Total number of terms in the document})$

$IDF(t) = \log(\text{Total number of documents} / \text{Number of documents with term } t)$.

$TF\text{-}IDF = TF(t) * IDF(t)$

Terms with TF-IDF scores

- The list was combed through for relevant words
- Top 20 words were selected and one hot encoded

Terms with Highest TF-IDF Scores:

kill	20.38
father	18.81
police	16.19
wife	15.07
men	14.81
woman	13.62
love	13.57
friend	13.54
family	13.29
war	13.25
child	13.08
car	12.99
son	12.99
home	12.81
money	12.72
old	12.63
room	12.41
film	12.00
live	11.97
mother	11.84
city	11.57
people	11.38
night	11.37
boy	11.34
fight	11.34
story	11.02
world	10.97

words selected:

father, police, family , men , young , war , child , home , young, son , love, money , friend, mother , escape , boy , girl, murder , brother, escape, german , gang

Final Dataframe

	Title	Year	Runtime	genre:Action	genre:Adventure	genre:Animation	genre:Biography	genre:Comedy	genre:Crime	genre:Drama	genre:Family	ge
0	The Shawshank Redemption	1.0	142	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	
1	The Godfather	0.0	175	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	
2	The Godfather: Part II	0.0	202	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	
3	The Dark Knight	1.0	152	1.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	
4	12 Angry Men	0.0	96	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	

5 rows × 113 columns

Curse of dimensionality

- Final data frame is sparse
- As predictor space explodes, clustering becomes difficult due to various reasons – Effect of noise, high correlation among predictors.
- Approach – Dimension reduction (PCA)

Principal component vectors

- Orthogonal transformation to convert set of correlated variables to a linearly uncorrelated variables called principal vectors.
- The first principal component has the highest variance among all the principal components
- 24 top principal components were used for clustering
- They explained 99.8% of the variance in our dataset.

PCA solution

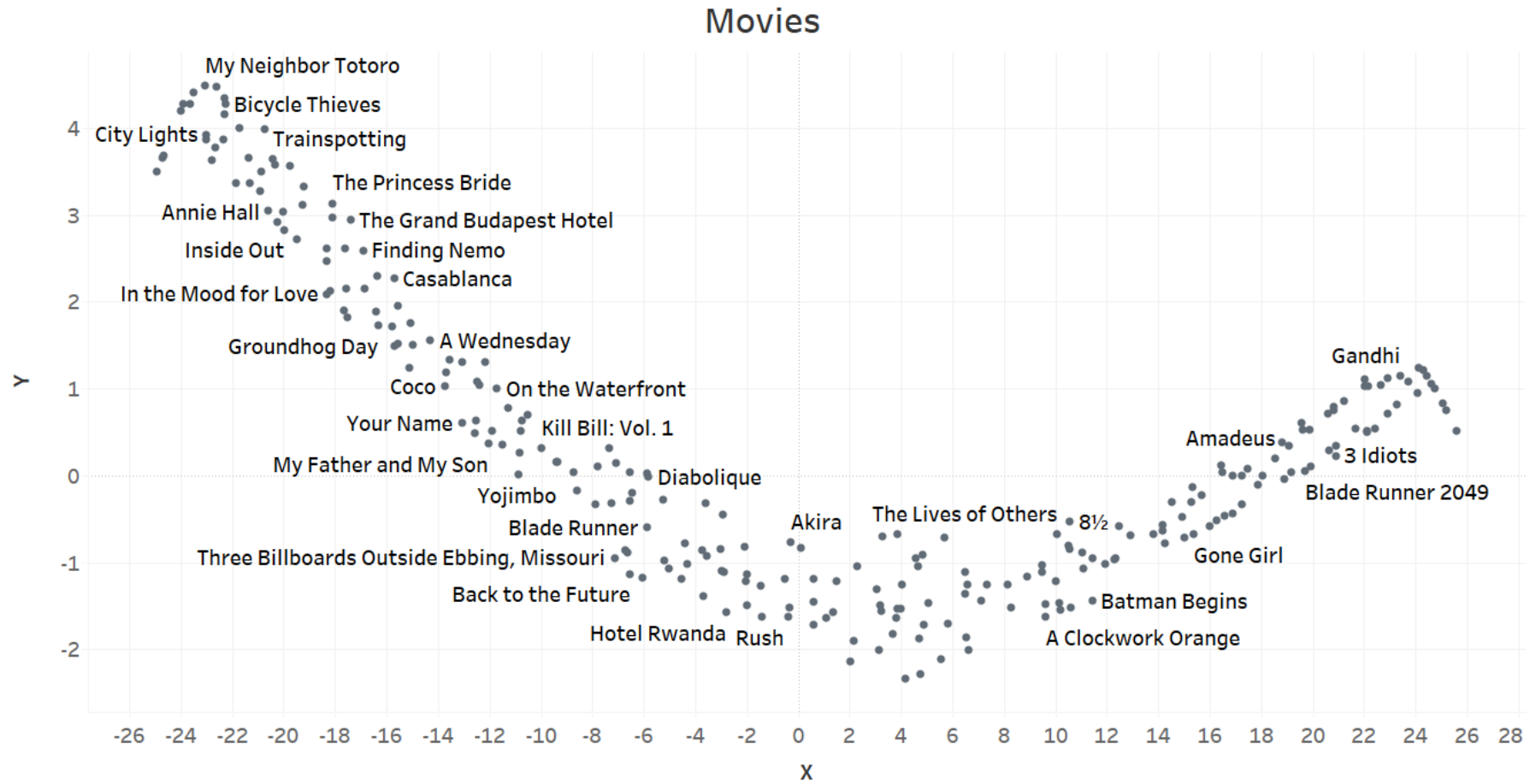
$$S_X \equiv \frac{1}{n-1} \mathbf{X}\mathbf{X}^T$$

- Solve using singular value decomposition $S = A'\Lambda A$
- A matrix contains eigen vectors of S
- Λ is a diagonal matrix corresponding to each eigen vector
- Retain top q eigen values from Λ
- Now, the predictor subspace has been mapped from p dim to q dim

Predictors sorted by loading vector weights

	weights	features	abs_weights
1	0.999947	Runtime	0.999947
103	0.003886	son	0.003886
8	0.003475	genre:Drama	0.003475
6	-0.003337	genre:Comedy	0.003337
100	0.003135	family	0.003135
98	0.002838	war	0.002838
12	0.002246	genre:History	0.002246
4	-0.002012	genre:Animation	0.002012
5	0.001918	genre:Biography	0.001918
77	0.001453	language:Italian	0.001453
90	0.001446	Country:India	0.001446
110	0.001443	murder	0.001443
102	0.001401	jewish	0.001401
95	0.001397	man	0.001397
17	-0.001286	genre:Romance	0.001286
23	0.001261	Actor:Robert De Niro	0.001261
97	0.001251	life	0.001251
99	-0.001238	police	0.001238
9	-0.001221	genre:Family	0.001221
94	0.001167	young	0.001167
82	0.001147	language:Hindi	0.001147
16	-0.001000	genre:Mystery	0.001000
105	0.001000	world	0.001000

Tsne 2-D



Clustering

- K means
- EM clustering
- DBSCAN
- Hierarchical clustering

K means clustering

- Randomly assign data points to (1...K) clusters
- Compute centroids of clusters
- Calculate distance measure from data points to corresponding centroids and reassign the points to centroids by distance
- Repeat the above steps until there is no reassignment of data points

```
K-MEANS( $\{\vec{x}_1, \dots, \vec{x}_N\}, K$ )
1  ( $\vec{s}_1, \vec{s}_2, \dots, \vec{s}_K$ )  $\leftarrow$  SELECTRANDOMSEEDS( $\{\vec{x}_1, \dots, \vec{x}_N\}, K$ )
2  for  $k \leftarrow 1$  to  $K$ 
3  do  $\vec{\mu}_k \leftarrow \vec{s}_k$ 
4  while stopping criterion has not been met
5  do for  $k \leftarrow 1$  to  $K$ 
6    do  $\omega_k \leftarrow \{\}$ 
7    for  $n \leftarrow 1$  to  $N$ 
8    do  $j \leftarrow \arg \min_j |\vec{\mu}_j - \vec{x}_n|$ 
9        $\omega_j \leftarrow \omega_j \cup \{\vec{x}_n\}$  (reassignment of vectors)
10   for  $k \leftarrow 1$  to  $K$ 
11   do  $\vec{\mu}_k \leftarrow \frac{1}{|\omega_k|} \sum_{\vec{x} \in \omega_k} \vec{x}$  (recomputation of centroids)
12  return  $\{\vec{\mu}_1, \dots, \vec{\mu}_K\}$ 
```

EM clustering GMM

EM Algorithm for GMM

Initialize $\boldsymbol{\mu}$ and Σ

E-step

Fix $\boldsymbol{\mu}$ and Σ and Update z_k^i

$$z_k^i = \frac{g_k(\mathbf{x}_i | \boldsymbol{\mu}_k, \Sigma_k)}{\sum_{k=1}^K g_k(\mathbf{x}_i | \boldsymbol{\mu}_k, \Sigma_k)}$$

M-step

Fix z_k^i and Update $\boldsymbol{\mu}$ and Σ

$$\hat{\boldsymbol{\mu}}_k = \frac{1}{z_k} \sum_{i=1}^N z_k^i \mathbf{x}_i$$

$$\hat{\Sigma}_k = \frac{1}{z_k} \sum_{i=1}^N z_k^i (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k)^T$$

Stop if converged

$\{\hat{\boldsymbol{\mu}}_k\} \{\hat{\Sigma}_k\}$

(K means)

- $K = 5$ (From DBSCAN)
- Cluster formations were difficult to interpret at higher K values
- Minimum SSE was obtained at $k = 8$

K means results – Cluster 0

- No of clusters – 5
- Theme : family, biography
- Genre : Fantasy, comedy,
biography

Movies

Star wars : Episode V

Star wars : Episode IV

The great dictator

Mad Max : Fury road

Whiplash

Jaws

Passion of Joan of Arc

K means results – Cluster 1

- Theme : Son, family, Murder
- Runtime is very high in comparison to the median value
- Genre : Dark, History, biography

Movies

The godfather part 1

The godfather part 2

Schindler's list

Once upon a time in America

Gandhi

Gone with the wind

Ben-Hur

K means results – Cluster 2

- Theme :War, kill
- Runtime is very high in comparison to the median value
- Genre : Action, Crime, thriller

Movies

Dark knight

Dark knight rises

Pulp fiction

Inglorious basterds

Django Unchained

Scarface

There will be blood

Once upon a time in west

K means results – Cluster 3

- Theme : young
- Genre : Comedy, Animation, Romance

Movies

Spirited away

Finding Nemo

Modern times

City lights

Casablanca

Toy story

Hachi : A dog's tale

Monty Python

Truman Show

K means results – Cluster 4

- Theme : home, family
- Actor : Al Pacino
- Genre : Thriller, crime

Movies

Shawshank Redemption

Matrix

The shining

Terminator 2

Beautiful mind

To kill a mockingbird

Se7en

Prestige

Good will hunting

V for Vendetta

La la land

Catch me if you can

DBSCAN

- DBSCAN is a density based clustering algorithm.
- Robust towards Outlier Detection.
- The most common distance metric used is Euclidean distance especially for high-dimensional data, this metric can be rendered almost useless due to the so-called “curse of dimensionality”.

DBSCAN Results – Cluster 1

- No of clusters – 5
- Director : Hiyao Miyazaki
- Theme : Young, world
- Genre : Animation, Adventure

Movies
Spirited Away
Finding Nemo
Nausicaa of the valley of wind
Wall -E
The Lion king

Cluster 2

- Director : Christopher Nolan
- Theme : Help, dark
- Actors : Christian Bale, Di Caprio
- Genre : Drama, thriller

Movies
The dark night
Inception
Prestige
Batman begins

Cluster 3

- Actor : Charlie Chaplin
- Theme : love, family
- Genre : Comedy, drama

Movies
City lights
Modern times
The great dictator
The gold rush

Cluster 4

- Director : Stanley Kubrick
- Theme : war, men, German
- Genre :Drama, autobiography
- All the movies in these cluster have extremely high run time

Movies
Barry Lyndon
Gandhi
Judgement at Nuremberg
Lawrence of Arabia
Schindler's list
Seven Samurai

Cluster 5

- Director : Clint Eastwood
- Theme : war, gang
- Genre :Action, fantasy
- Movies in these cluster are also longer in comparison to median run time

Movies
Lord of the rings series
The good, the bad and the ugly
Mad Max : Fury road
Gangs of Wasseypur
Roshomon
Yojimbo

Outlier clusters

- DBSCAN assigned many data points as outliers
- Why ?? – DBSCAN looks for spatial density for clustering. In higher dimensions, it is difficult to find such density clusters.
- The clusters formed by DBSCAN are easier to interpret than K means

Future work

- Scaling of predictors
- In K means clustering, finding the best value of k that is optimum according to gap statistic, silhouette score and has good interpretability
- Different distance measures such as jaccard , Gower distance
- Heirarchical clustering