IMDb

Top 250 Movies Clustering

Starring …

- Introduction
- Data Description
- Pre-Processing
- Natural Language Processing
- PCA : Eliminating Curse of Dimensionality
- K Means Clustering
- Hierarchical Clustering
- Interpretations and Results
- Conclusions

Have you ever wondered why certain movies get nominated for Oscars? What's so special about them?

Introduction

- IMDB (Internet Movie Database) : World’s most popular and authoritative source for movie ratings and user reviews.
- OMDB : Opensource IMDB API used for collecting Top 250 movies
- Goals :
  - Use Clustering to recognize patterns among top rated movies
  - Identify what attributes such as Genre, Plot, Runtime etc contribute towards their high rating
  - Identify similarities of movies within the same cluster based on these attributes

Data Description

Possible Clustering Attributes:

- Year
- Rated
- Runtime
- Director
- Writer
- Actor
- Language
- Genre

Number of Rows : 250

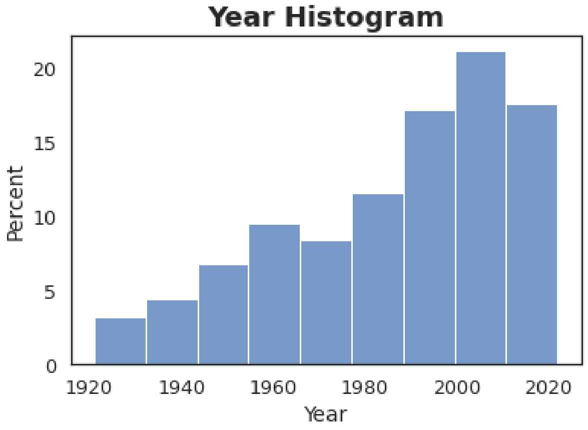
Number of Columns : 26

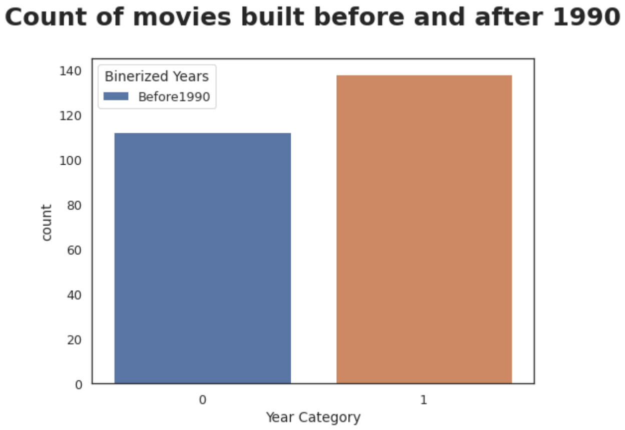
Number of Duplicated Rows : 0

Pre-Processing

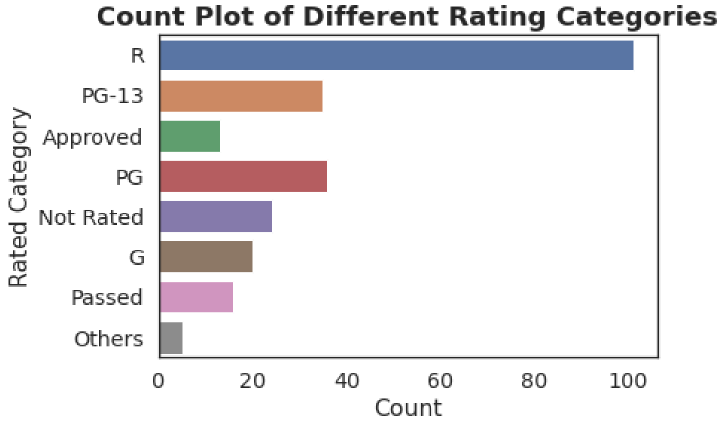
- Year – Continuous to Binary
- Rated – Combined less frequent categories(Unrated, TV-PG, X, GP, TV – MA) as ‘Others’
- Runtime – Binned in 3 categories:
  - Short : <90 minutes
  - Normal : 90 – 150 minutes
  - Long : >150 minutes
- Directors/Actors - Filtered top directors/actors with 4 or more movies
- Language- 5 Most frequent
- Country – 10 Most frequent
- Plot – Processed using NLP
- All the categorical variables were converted into dummy variables

Year

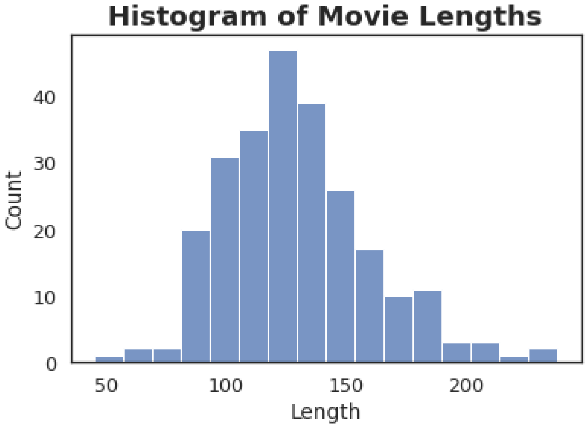


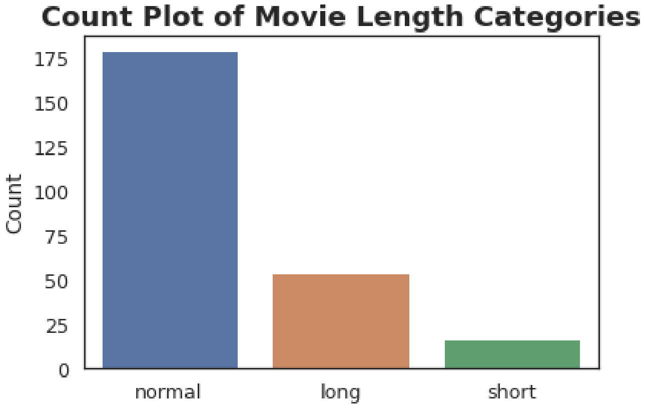


Rating

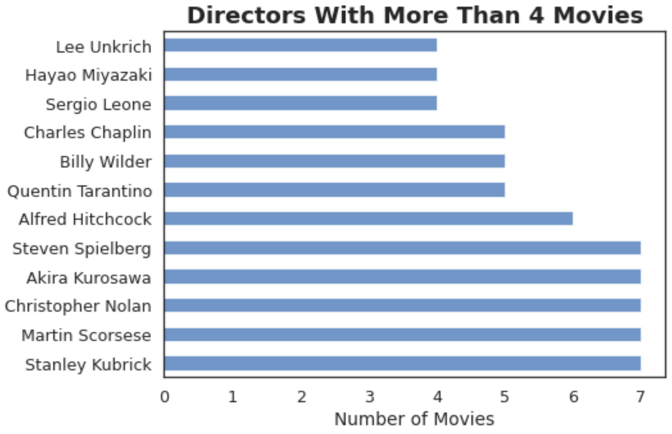


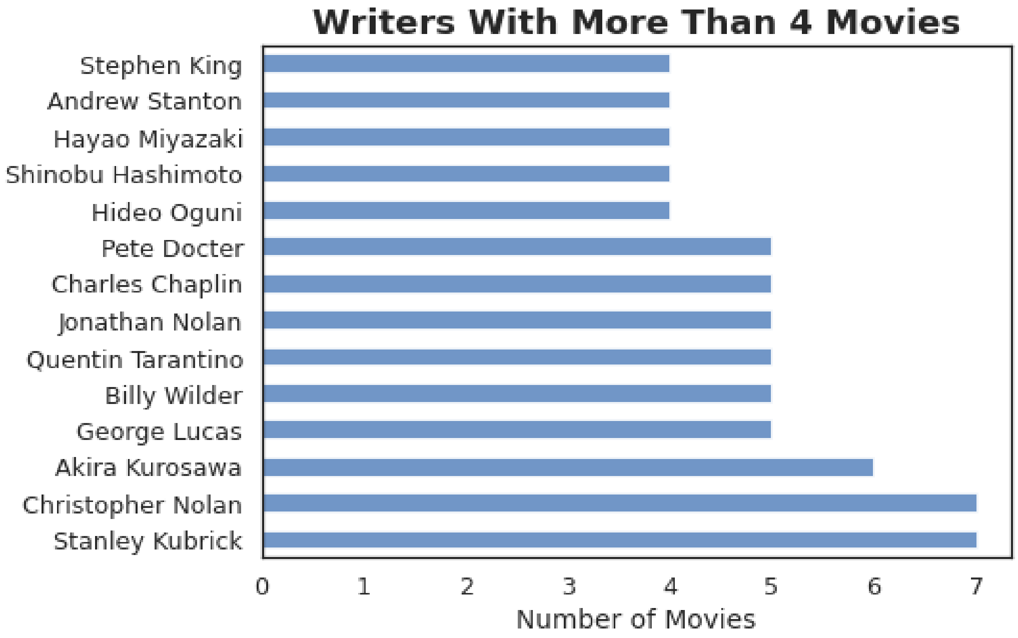
Runtime



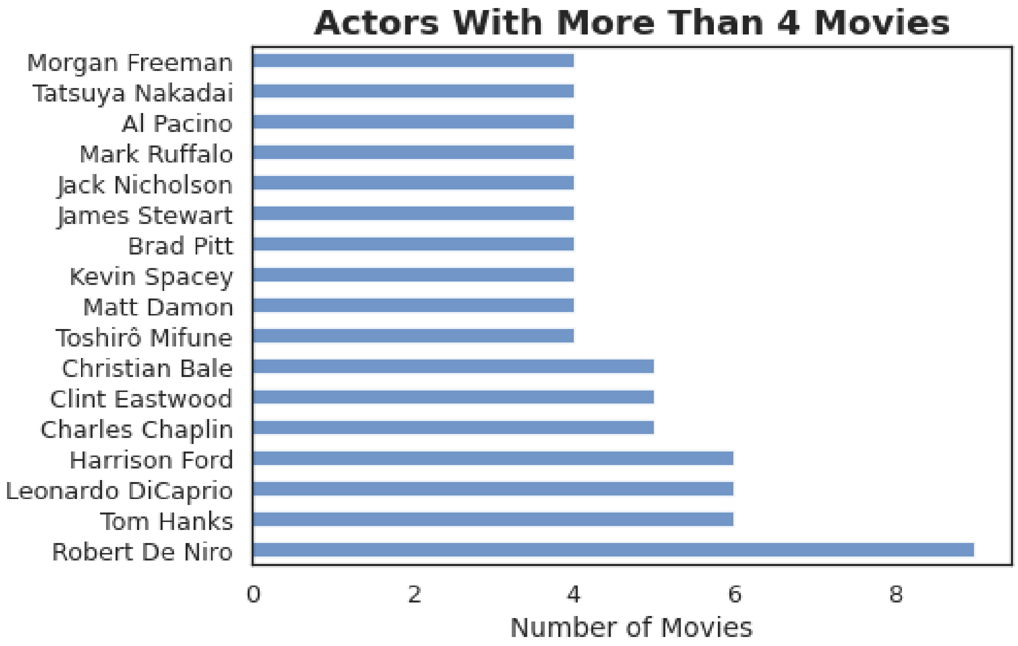


Director

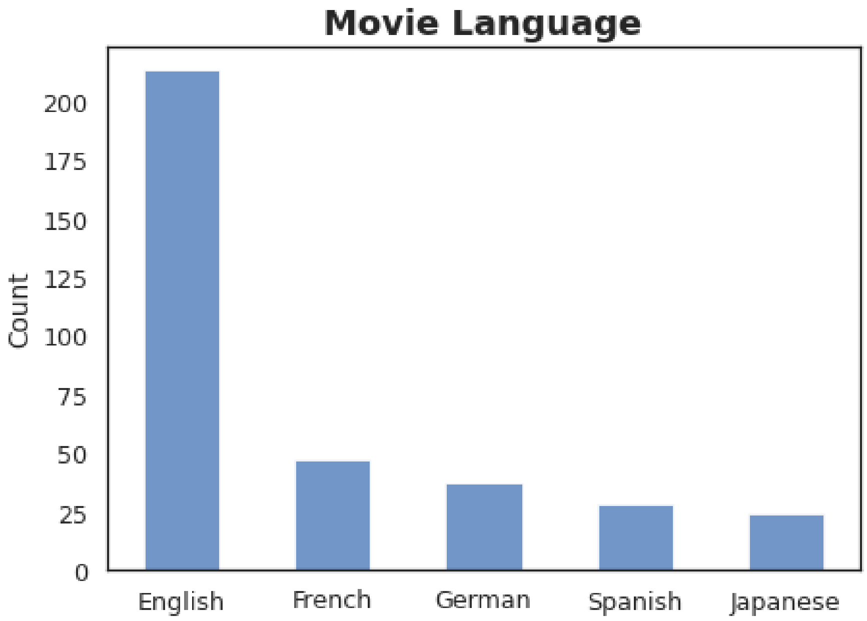




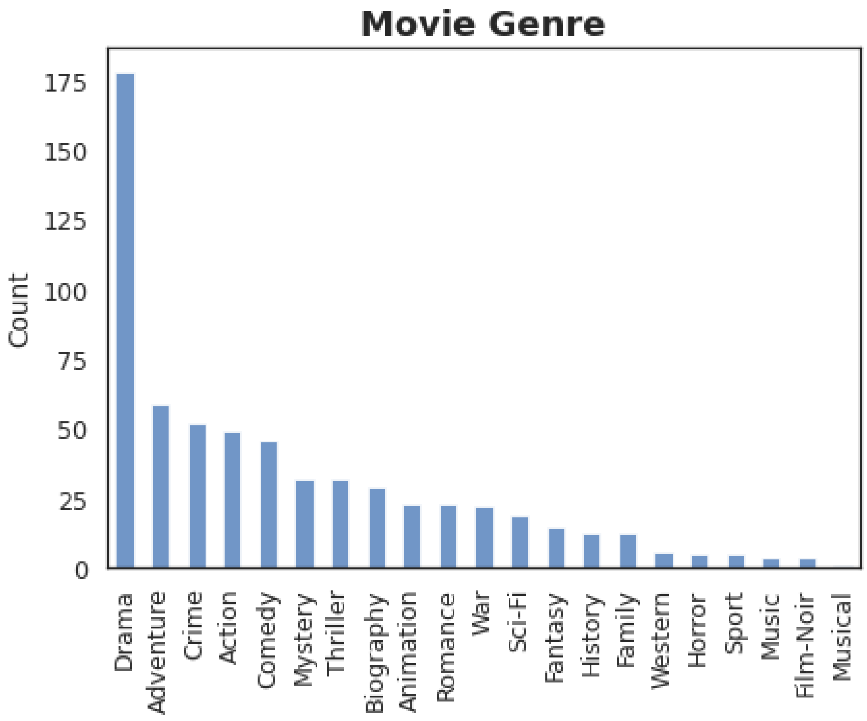
Writers With More Than 4 Movies



Actors With More Than 4 Movies



Movie Language



Movie Genre

NLP

Step 1

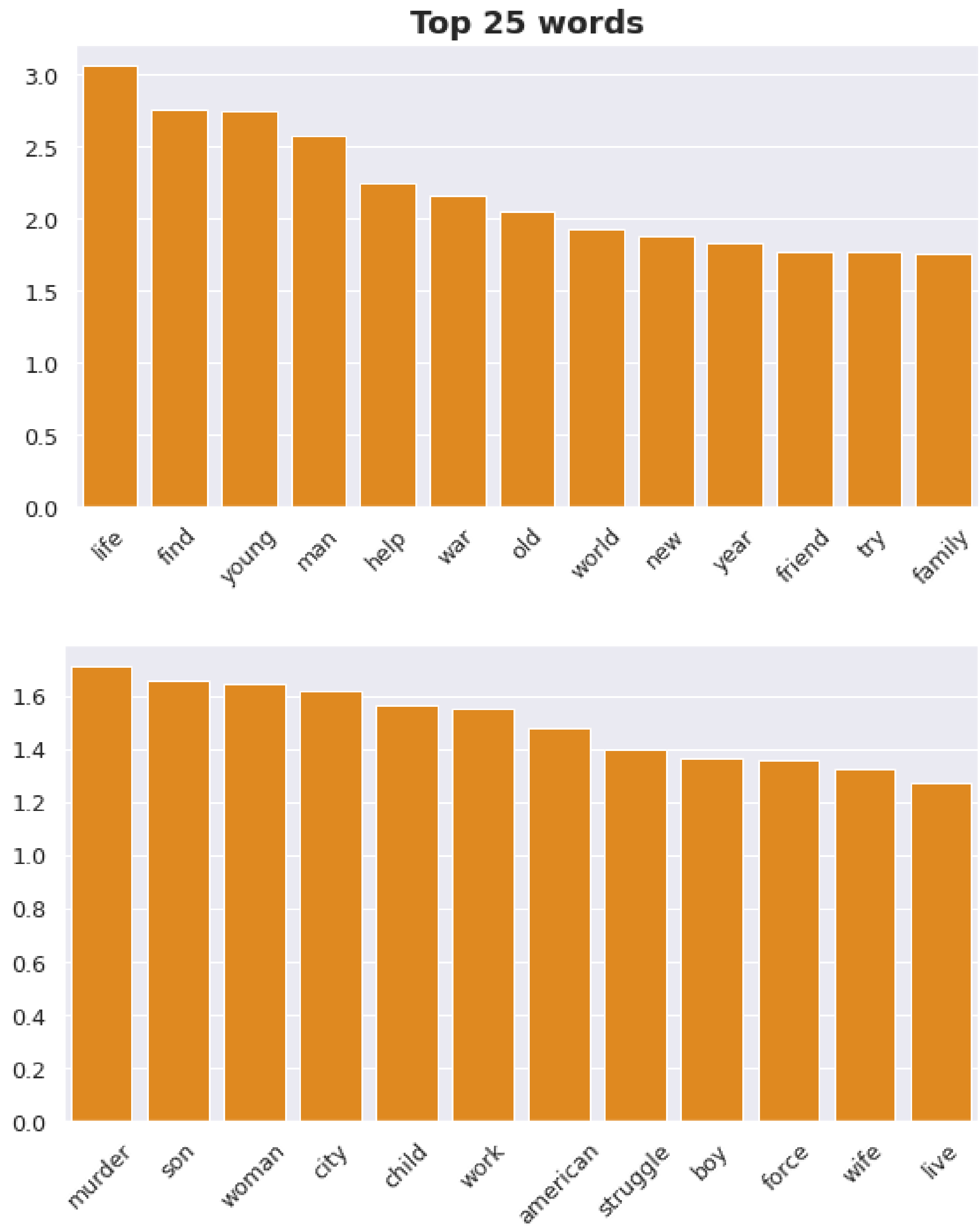
- Movie Description Preprocessing:
  - Tokenization
  - Lemmatization
  - Removing Stop words
  - Removing Punctuations

Step 2

- TF-IDF Vectorization

Step 3

- Choose 25 top words



Final Dataset Attributes for PCA:

- Year (Binary)
- Runtime
- Country
- Genre
- Actors
- Directors
- Top 25 plot words
- Rated

Final Dataset Dimension:

- 250 rows and 95 columns

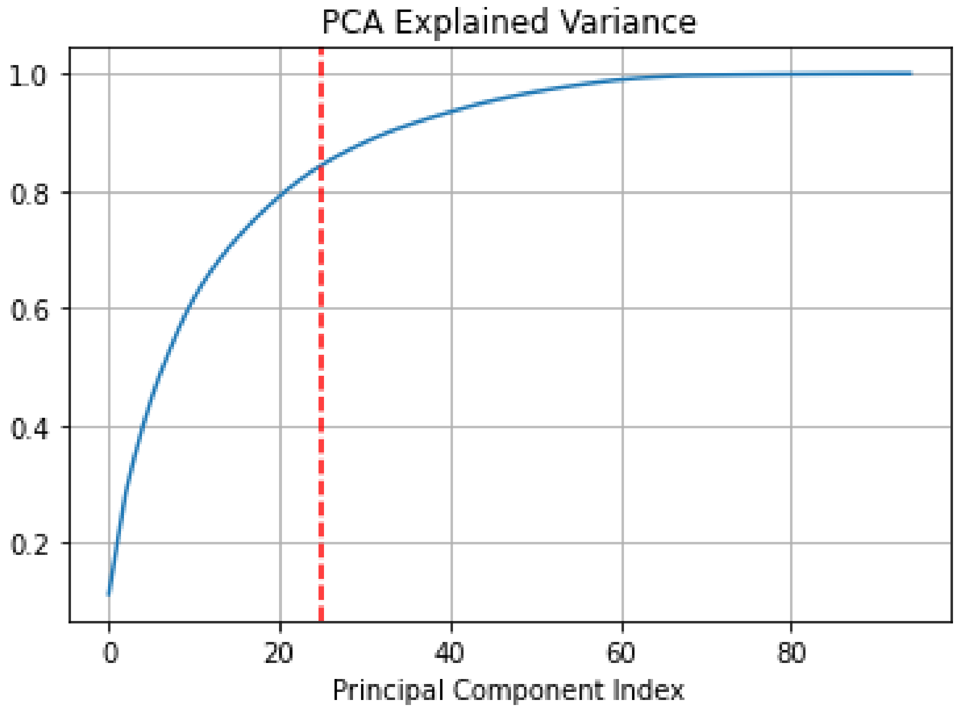
PCA

Purpose:

- Reducing dimension of the data

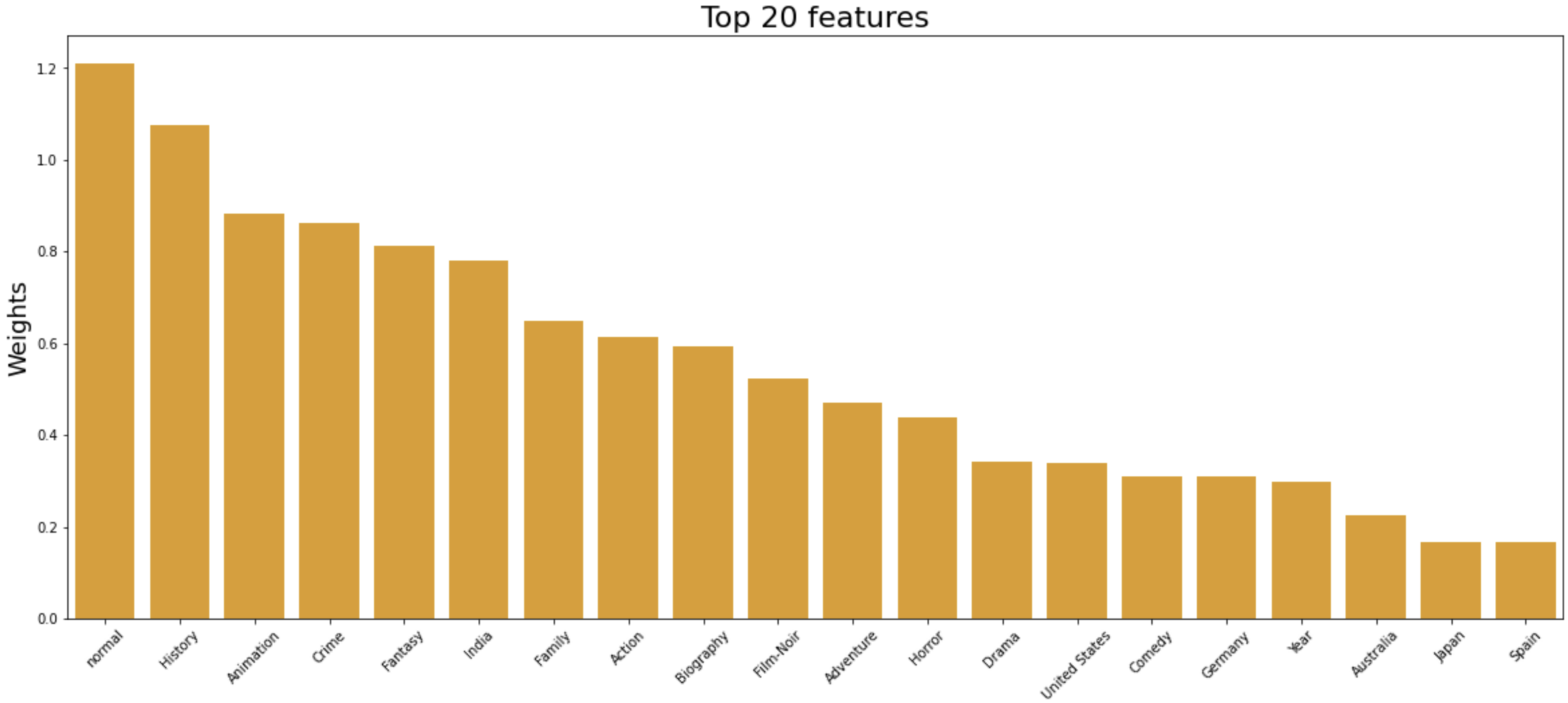
Cutoff Point:

- 80% of explained variance – 22 PCs



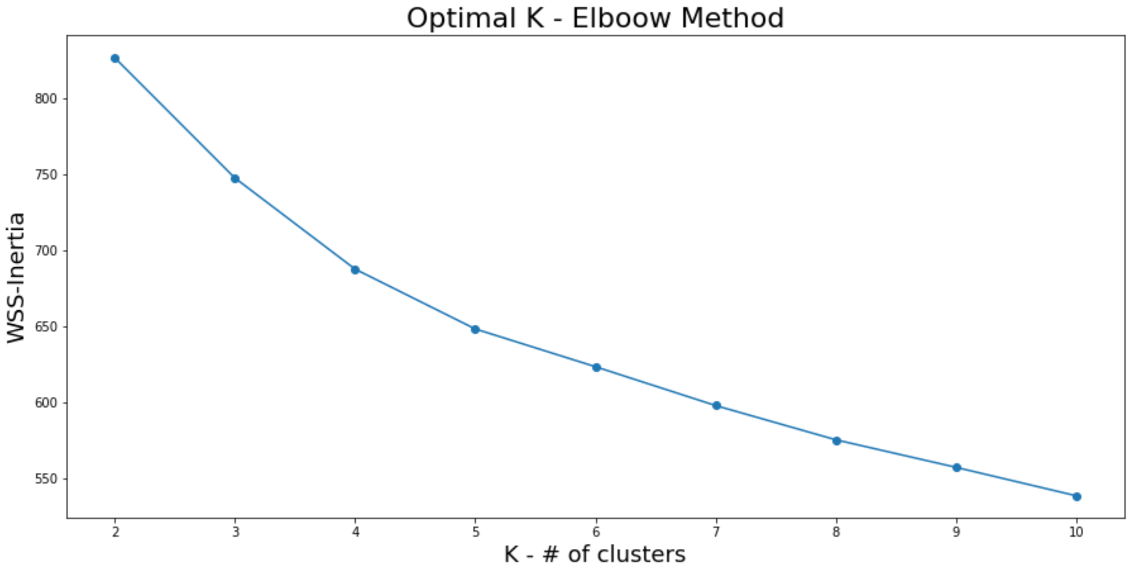
PC	16	17	18	19	20	21	22	23	24	25
EVR	0.72	0.73	0.75	0.76	0.78	0.79	0.8	0.81	0.82	0.83

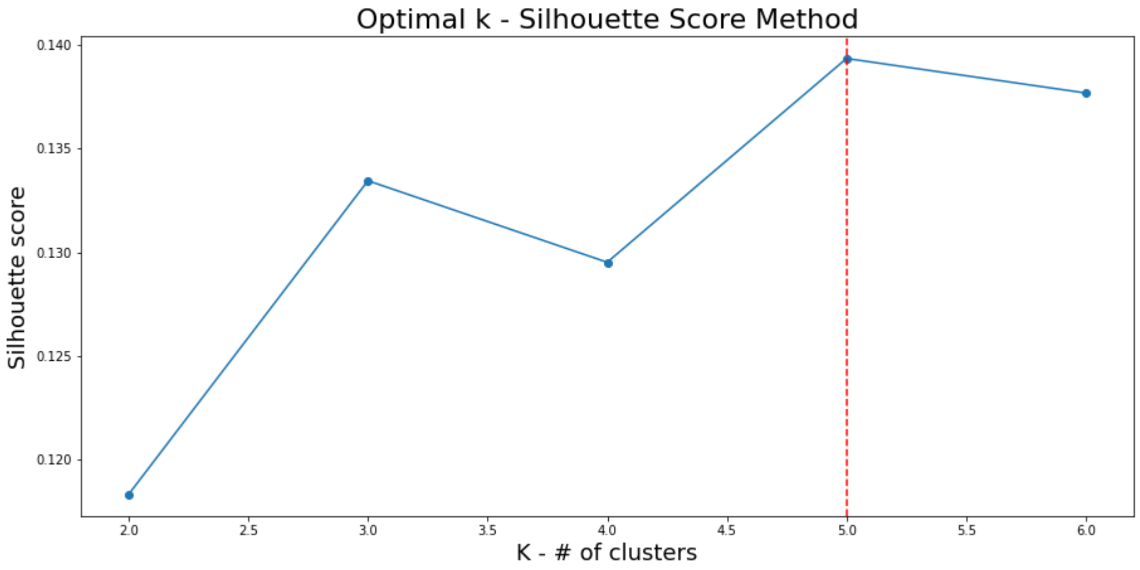
First 20 Principal Component (PC1-20):



K - Means

- Find Optimal K
  - Elbow Method – 4 or 5
  - Silhouette Score – 5


Optimal K - Elboow Method

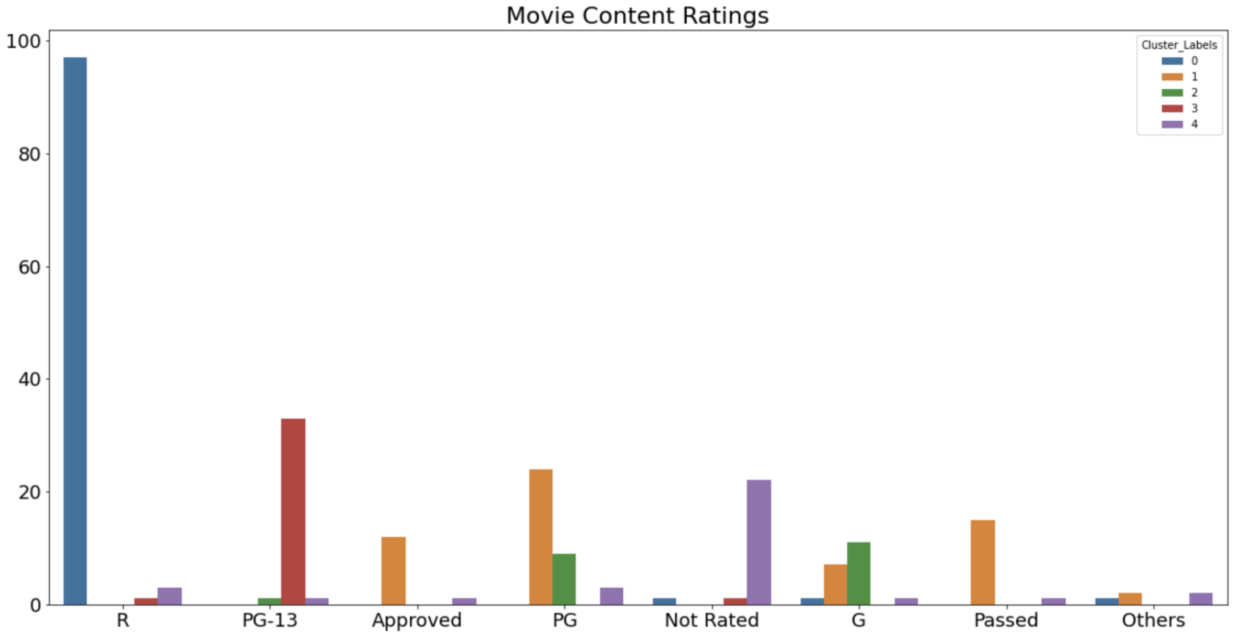

Optimal k - Silhouette Score Method

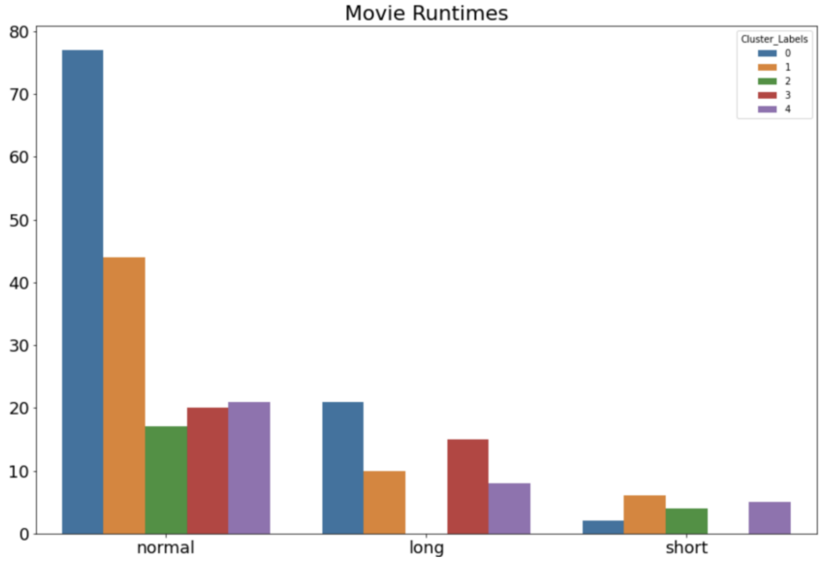
K-Means

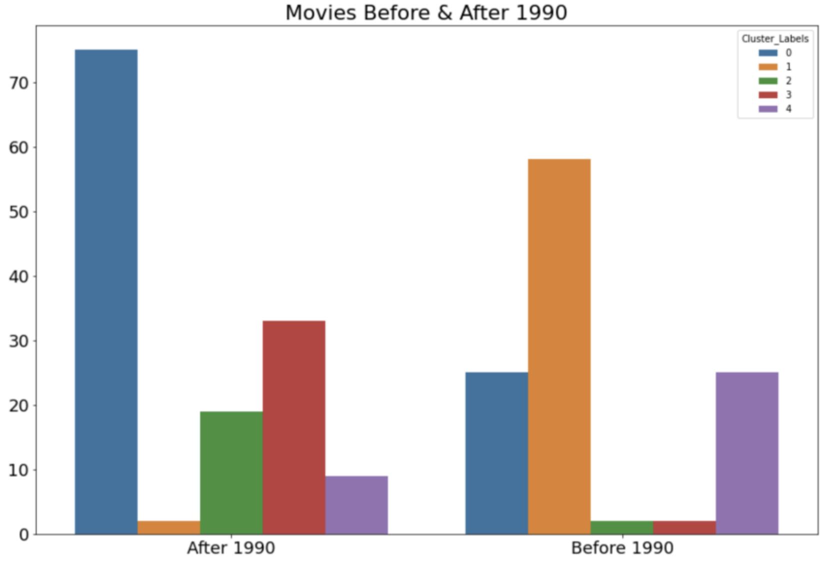
- Mini Batch K-Means
  - K= 5 clusters

- Cluster Sizes :

0	1	2	3	4
100	60	21	35	34


Movie Content Ratings

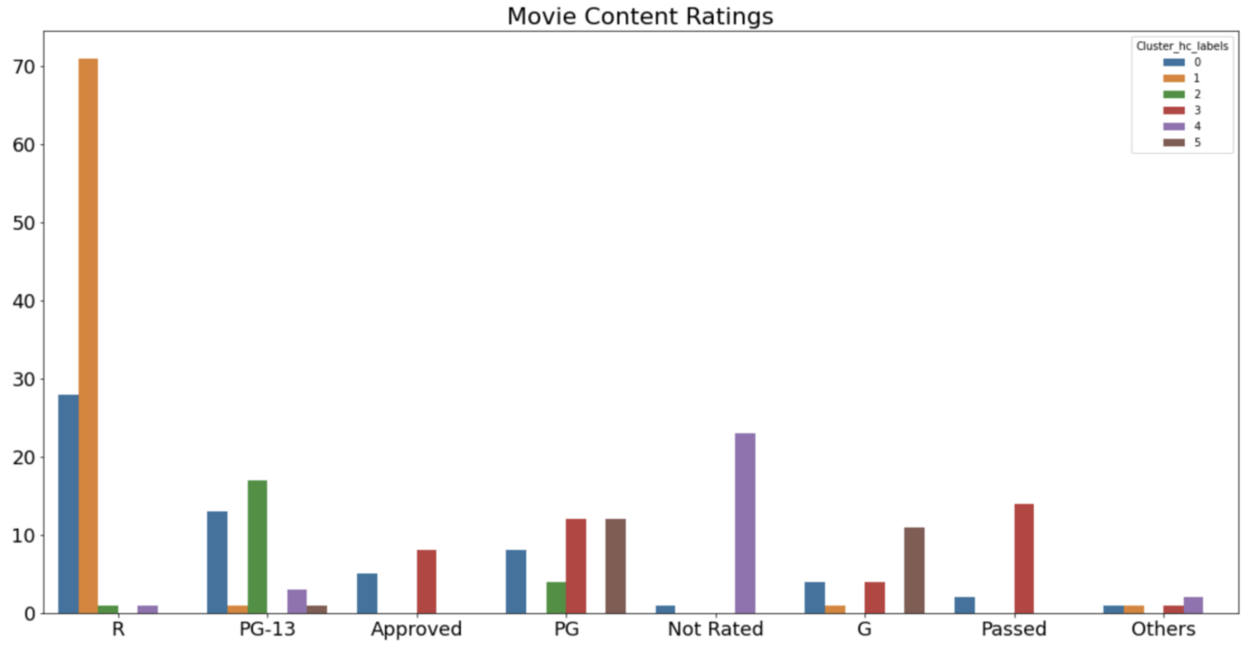

Movie Runtimes

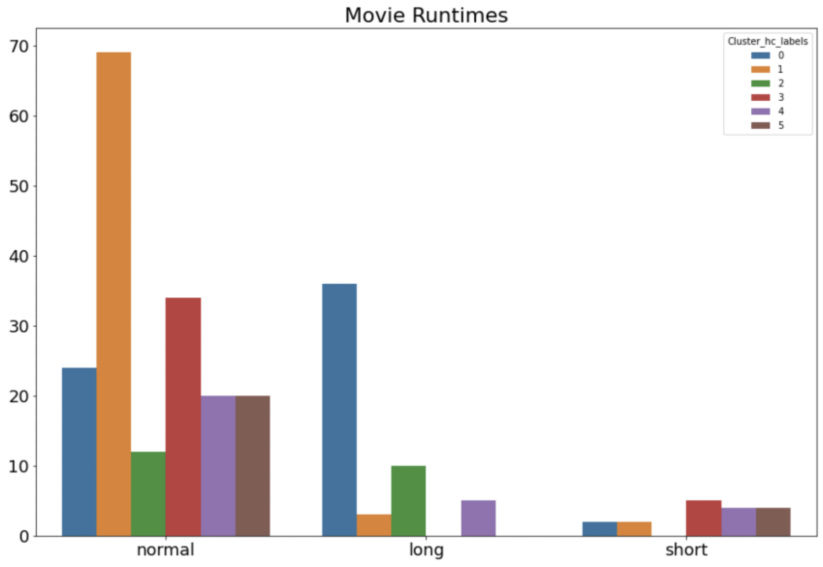

Movies Before & After 1990

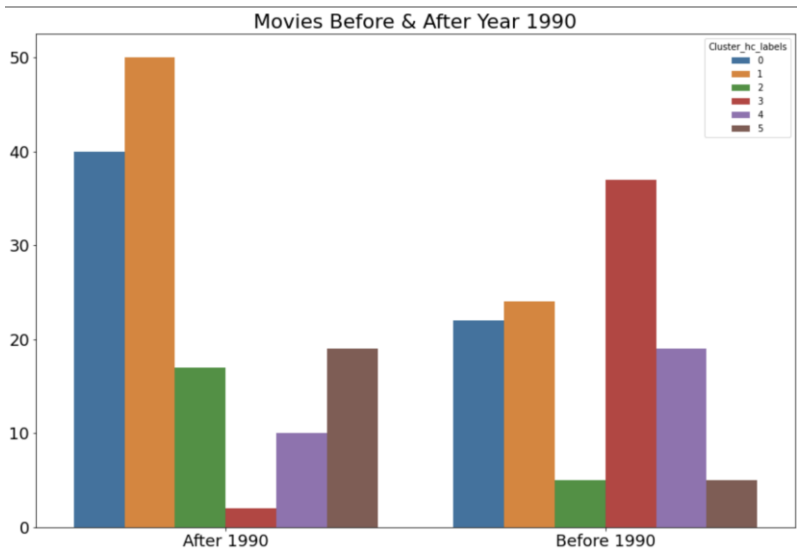
Hierarchical - Agglomerative

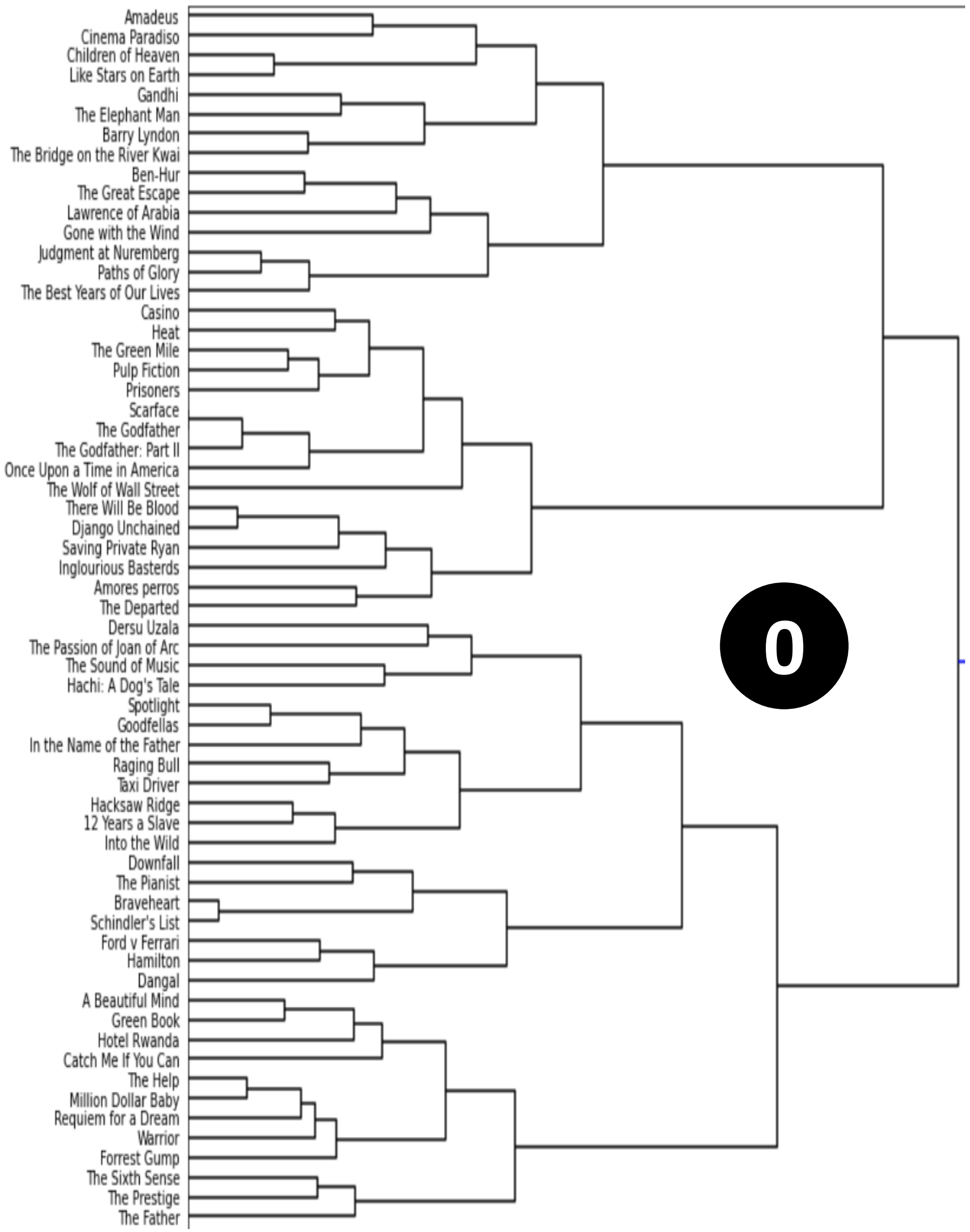
- Dendrogram
  - 6 clusters
  - Linkage : Ward
  - Affinity : Euclidean

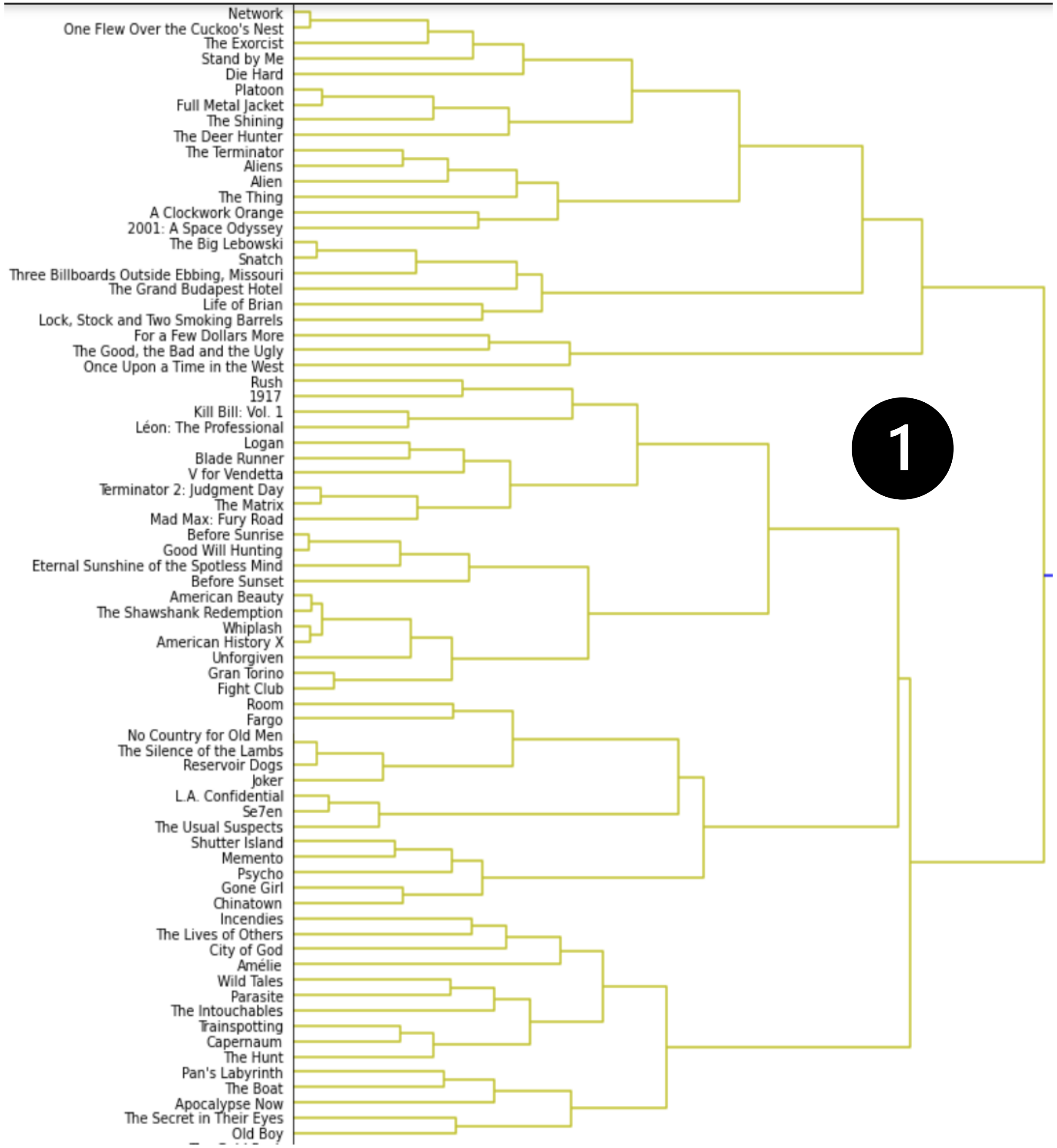
0	1	2	3	4	5
62	74	22	39	29	24

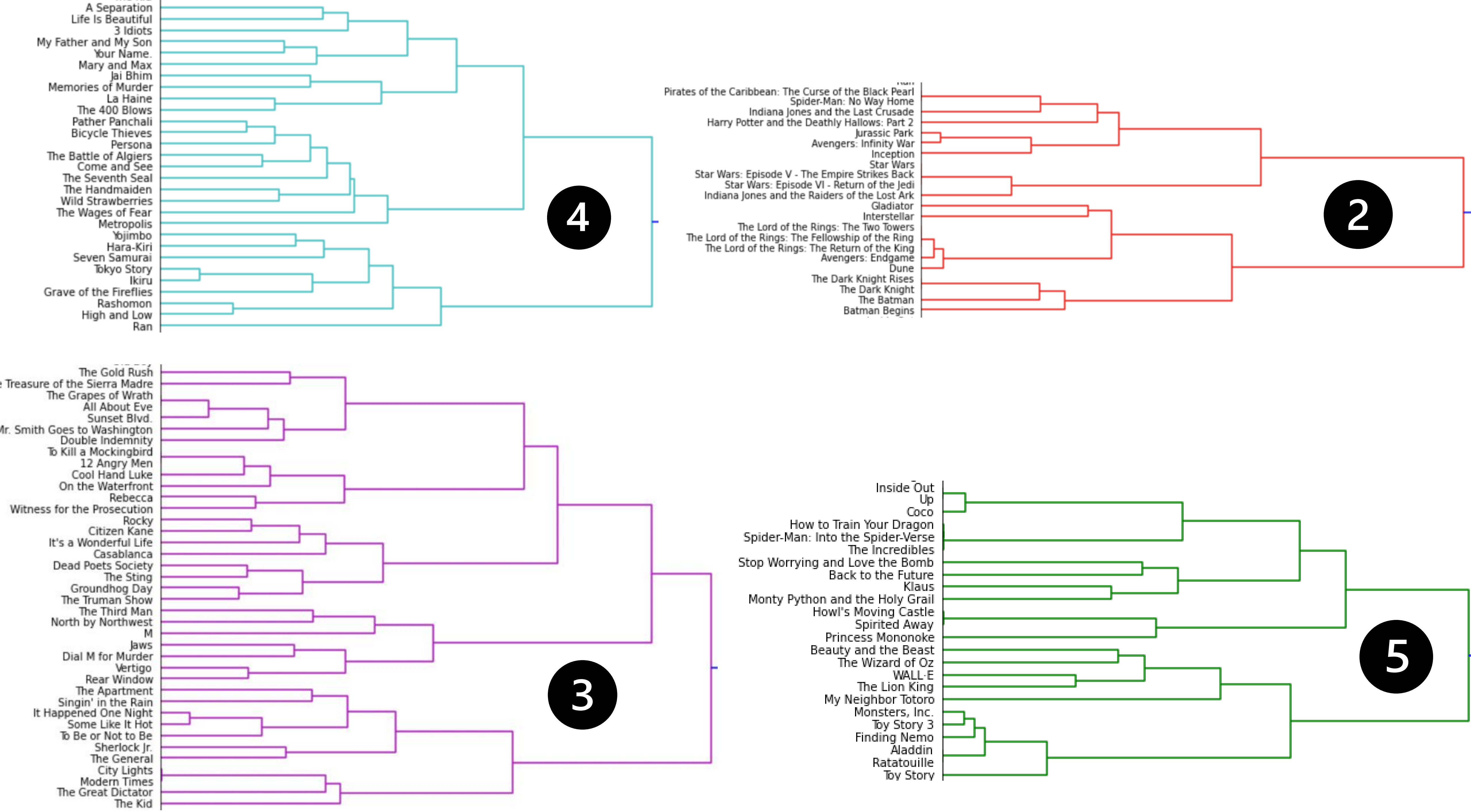

Movie Content Ratings


Movie Runtimes

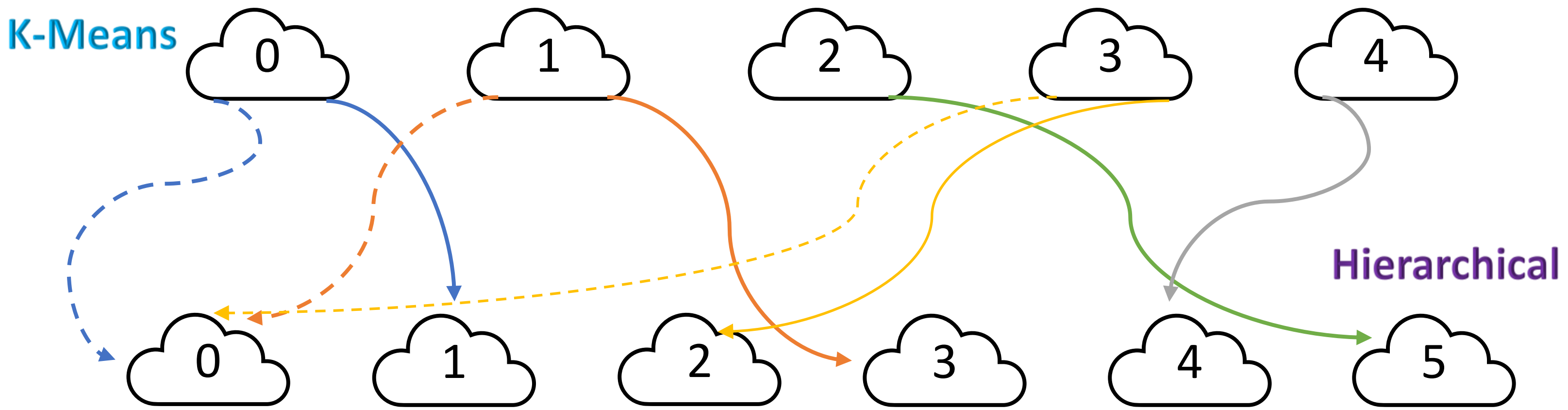

Movies Before & After Year 1990


Amadeus
Cinema Paradiso
Children of Heaven
Like Stars on Earth
Gandhi
The Elephant Man
Barry Lyndon
The Bridge on the River Kwai
Ben-Hur
The Great Escape
Lawrence of Arabia
Gone with the Wind
Judgment at Nuremberg
Paths of Glory
The Best Years of Our Lives
Casino
Heat
The Green Mile
Pulp Fiction
Prisoners
Scarface
The Godfather
The Godfather: Part II
Once Upon a Time in America
The Wolf of Wall Street
There Will Be Blood
Django Unchained
Saving Private Ryan
Inglourious Basterds
Amores perros
The Departed
Dersu Uzala
The Passion of Joan of Arc
The Sound of Music
Hachi: A Dog's Tale
Spotlight
Goodfellas
In the Name of the Father
Raging Bull
Taxi Driver
Hacksaw Ridge
12 Years a Slave
Into the Wild
Downfall
The Pianist
Braveheart
Schindler's List
Ford v Ferrari
Hamilton
Dangal
A Beautiful Mind
Green Book
Hotel Rwanda
Catch Me If You Can
The Help
Million Dollar Baby
Requiem for a Dream
Warrior
Forrest Gump
The Sixth Sense
The Prestige
The Father
0


Network
One Flew Over the Cuckoo's Nest
The Exorcist
Stand by Me
Die Hard
Platoon
Full Metal Jacket
The Shining
The Deer Hunter
The Terminator
Aliens
Alien
The Thing
A Clockwork Orange
2001: A Space Odyssey
The Big Lebowski
Snatch
Three Billboards Outside Ebbing, Missouri
The Grand Budapest Hotel
Life of Brian
Lock, Stock and Two Smoking Barrels
For a Few Dollars More
The Good, the Bad and the Ugly
Once Upon a Time in the West
Rush
1917
Kill Bill: Vol. 1
Léon: The Professional
Logan
Blade Runner
V for Vendetta
Terminator 2: Judgment Day
The Matrix
Mad Max: Fury Road
Before Sunrise
Good Will Hunting
Eternal Sunshine of the Spotless Mind
Before Sunset
American Beauty
The Shawshank Redemption
Whiplash
American History X
Unforgiven
Gran Torino
Fight Club
Room
Fargo
No Country for Old Men
The Silence of the Lambs
Reservoir Dogs
Joker
L.A. Confidential
Se7en
The Usual Suspects
Shutter Island
Memento
Psycho
Gone Girl
Chinatown
Incendies
The Lives of Others
City of God
Amélie
Wild Tales
Parasite
The Intouchables
Trainspotting
Capernaum
The Hunt
Pan's Labyrinth
The Boat
Apocalypse Now
The Secret in Their Eyes
Old Boy
1


A Separation
Life Is Beautiful
3 Idiots
My Father and My Son
Your Name.
Mary and Max
Jai Bhim
Memories of Murder
La Haine
The 400 Blows
Pather Panchali
Bicycle Thieves
Persona
The Battle of Algiers
Come and See
The Seventh Seal
The Handmaiden
Wild Strawberries
The Wages of Fear
Metropolis
Yojimbo
Hara-Kiri
Seven Samurai
Tokyo Story
Ikiru
Grave of the Fireflies
Rashomon
High and Low
Ran
4
Pirates of the Caribbean: The Curse of the Black Pearl
Spider-Man: No Way Home
Indiana Jones and the Last Crusade
Harry Potter and the Deathly Hallows: Part 2
Jurassic Park
Avengers: Infinity War
Inception
Star Wars
Star Wars: Episode V - The Empire Strikes Back
Star Wars: Episode VI - Return of the Jedi
Indiana Jones and the Raiders of the Lost Ark
Gladiator
Interstellar
The Lord of the Rings: The Two Towers
The Lord of the Rings: The Fellowship of the Ring
The Lord of the Rings: The Return of the King
Avengers: Endgame
Dune
The Dark Knight Rises
The Dark Knight
The Batman
Batman Begins
2
The Gold Rush
e Treasure of the Sierra Madre
The Grapes of Wrath
All About Eve
Sunset Blvd.
Ir. Smith Goes to Washington
Double Indemnity
To Kill a Mockingbird
12 Angry Men
Cool Hand Luke
On the Waterfront
Rebecca
Witness for the Prosecution
Rocky
Citizen Kane
It's a Wonderful Life
Casablanca
Dead Poets Society
The Sting
Groundhog Day
The Truman Show
The Third Man
North by Northwest
M
Jaws
Dial M for Murder
Vertigo
Rear Window
The Apartment
Singin' in the Rain
It Happened One Night
Some Like It Hot
To Be or Not to Be
Sherlock Jr.
The General
City Lights
Modern Times
The Great Dictator
The Kid
3
Inside Out
Up
Coco
How to Train Your Dragon
Spider-Man: Into the Spider-Verse
The Incredibles
Stop Worrying and Love the Bomb
Back to the Future
Klaus
Monty Python and the Holy Grail
Howl's Moving Castle
Spirited Away
Princess Mononoke
Beauty and the Beast
The Wizard of Oz
WALL·E
The Lion King
My Neighbor Totoro
Monsters, Inc.
Toy Story 3
Finding Nemo
Aladdin
Ratatouille
Toy Story
5

Cluster Comparison



Hierarchical

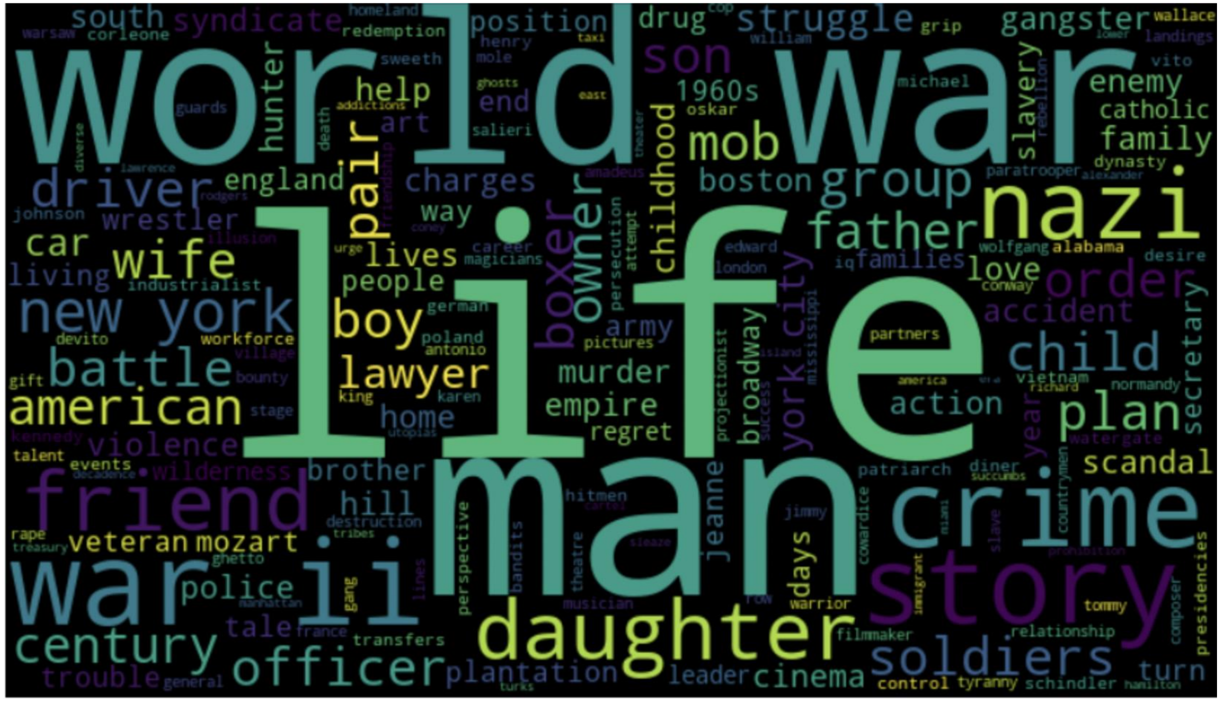
K-Means

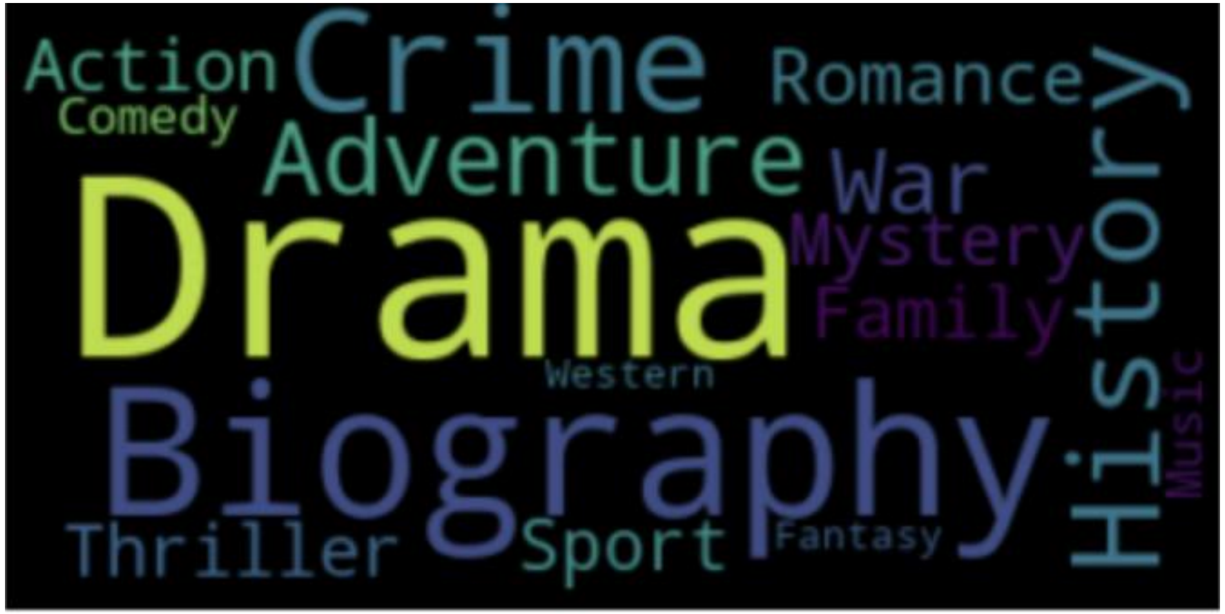
	0	1	2	3	4	5	All
0	30	69	0	0	1	0	**100**
1	12	2	4	39	0	3	**60**
2	0	0	0	0	0	21	**21**
3	14	1	18	0	2	0	**35**
4	6	2	0	0	26	0	**34**
All	**62**	**74**	**22**	**39**	**29**	**24**	**250**

Cluster 0

>> Biographies & Historical Movies

- **Rated** – R, PG-13
- **Runtime** – Long (More than 150 minutes)
- **Directors** – Martin Scorsese, Quentin Tarantino, Steven Spielberg
- **Actors** – Robert De Niro, Al Pacino, Tom Hanks, Leonardo Di Caprio
- **Movies** – The Godfather: I &II, Casino, The Wolf of Wall Street, Catch me If you Can, Gandhi, Dangal

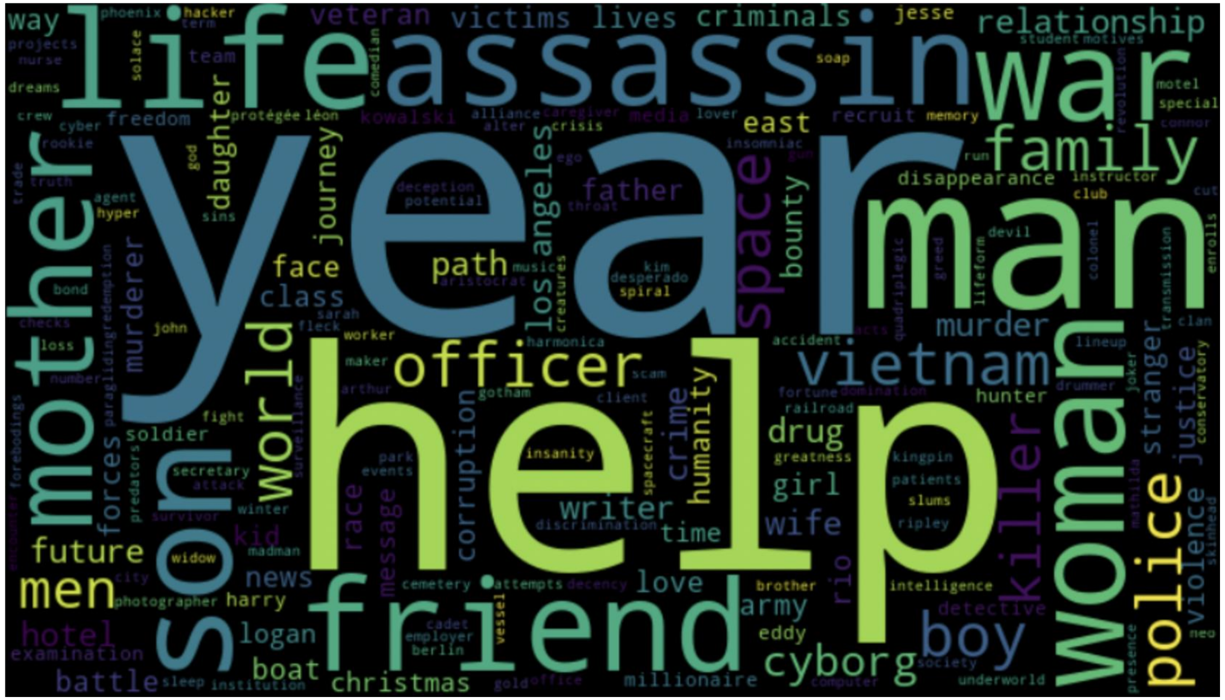


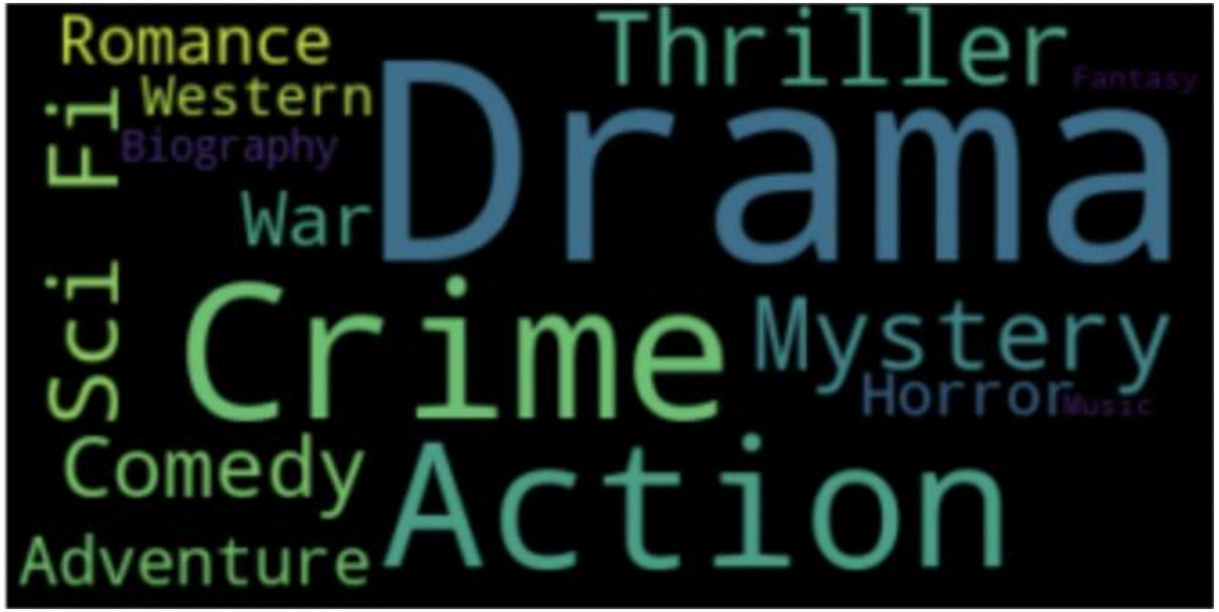


Cluster 1

>> Rated R, Crime Movies

- **Rated** – R
- **Runtime** – Normal (90-150 minutes)
- **Directors** – Stanley Kubric, Sergio Leone
- **Actors** – Kevin Spacey, Clint Eastwood, Morgan Freeman
- **Movies** – Se7en, The Terminator, A Clockwork Orange, The Matrix

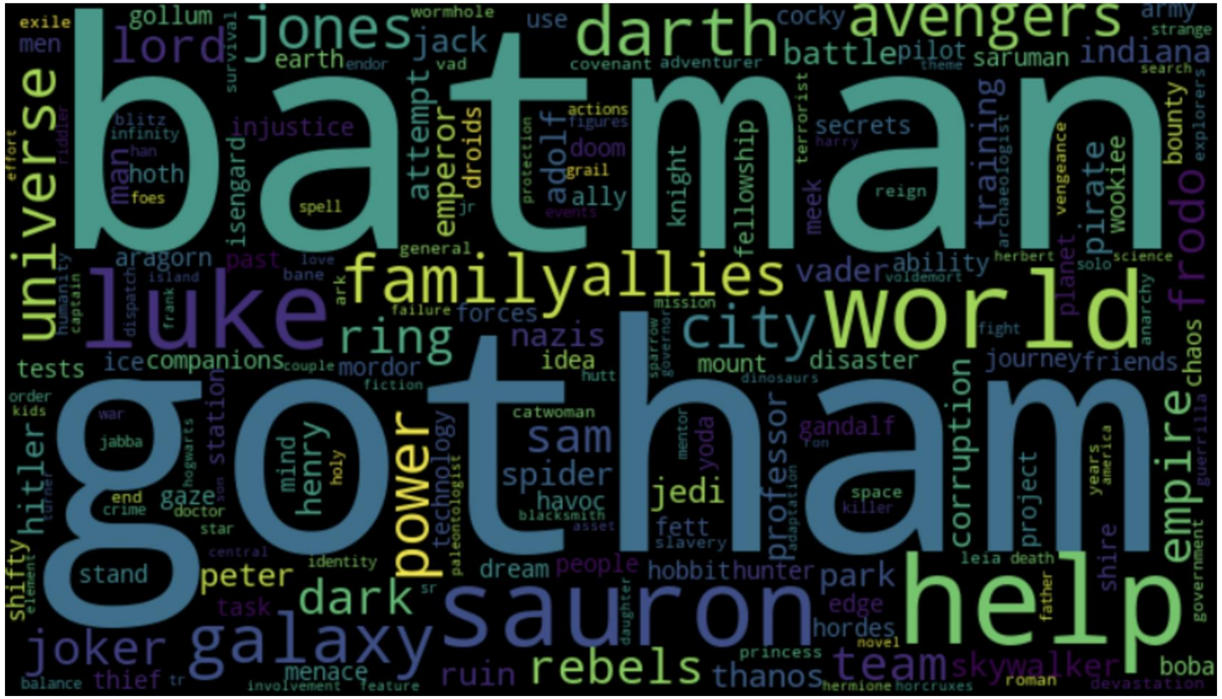


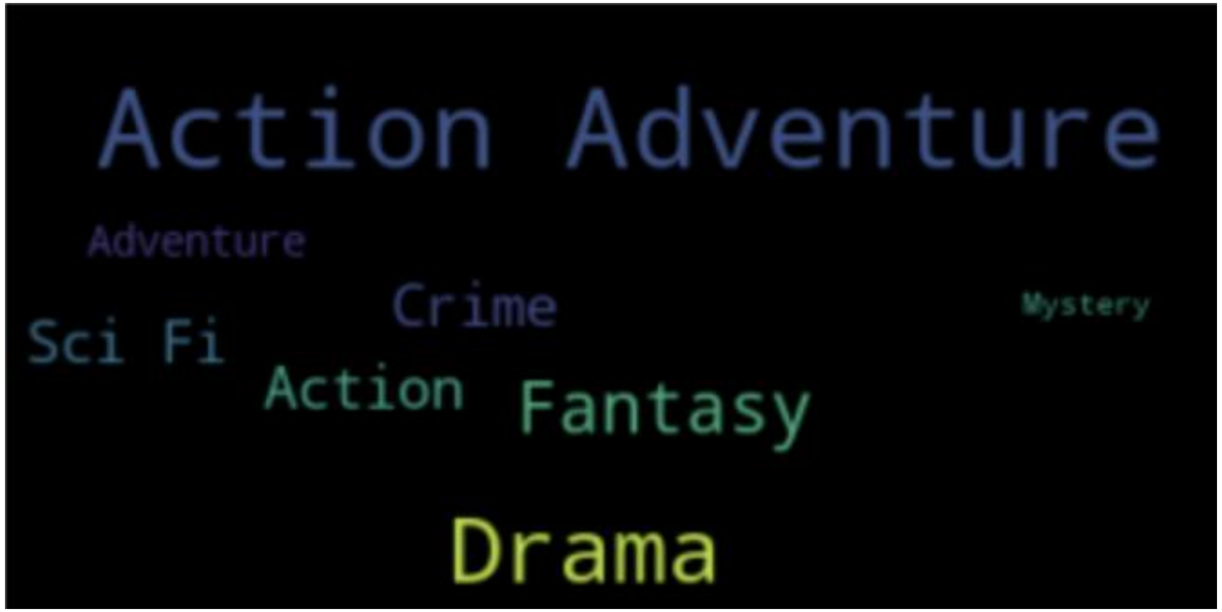


Cluster 2

>> Superhero/Action Movies

- **Year –** After 1990
- **Rated** – PG -13, PG
- **Directors** – Christopher Nolan, Steven Spielberg
- **Actors** – Harrison Ford, Christian Bale
- **Movies** – Batman Trilogy, Interstellar, Avengers-Endgame & Infinity War, Star Wars, Indiana Jones

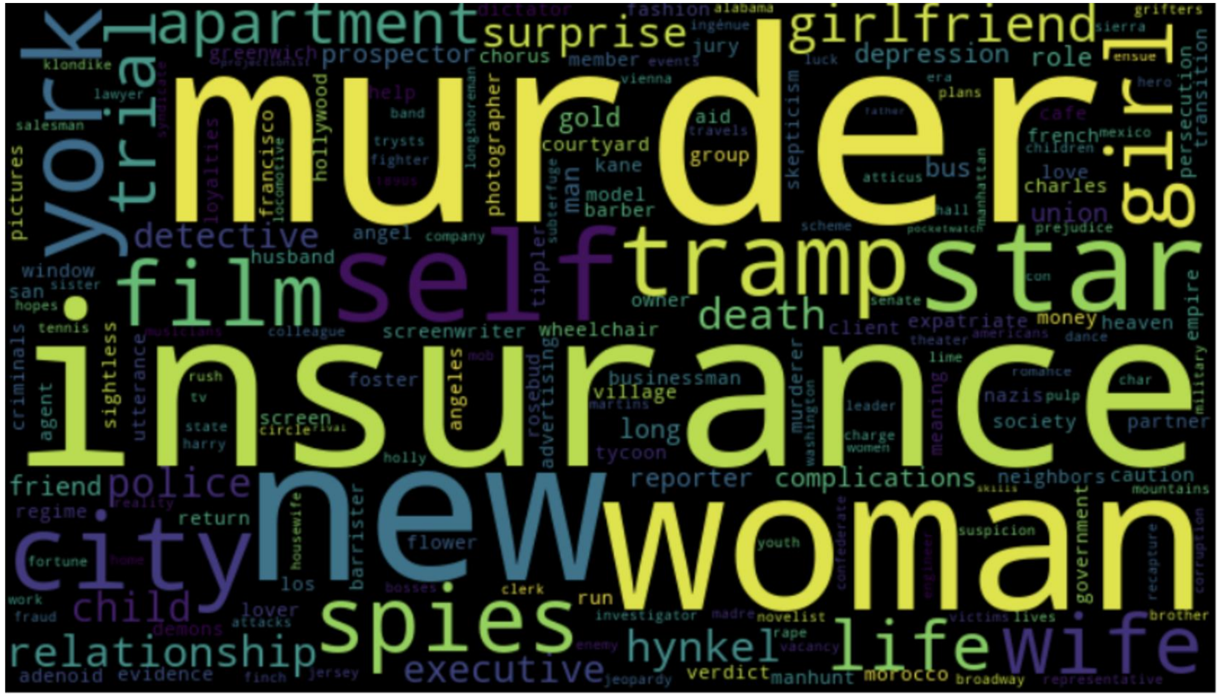


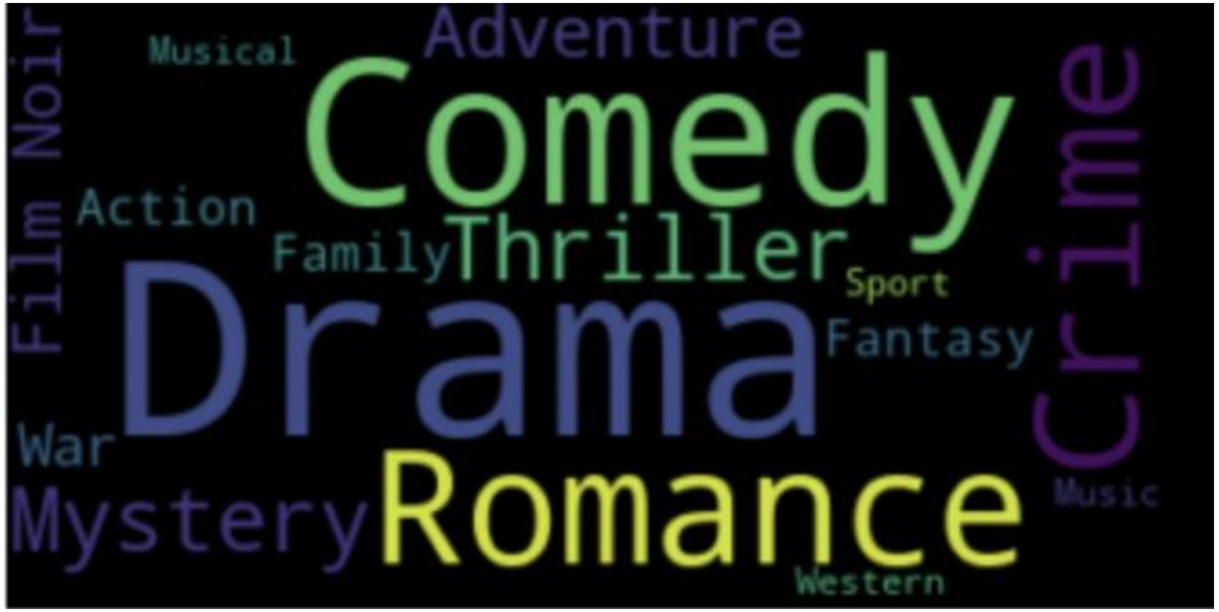


Cluster 3

>> Retro Movies

- **Year –** Before 1990
- **Rated** –PG, Passed, Approved
- **Directors** –Alfred Hitchcock, Billy Wilder, Charles Chaplin
- **Actors** – Charles Chaplin, James Stewart
- **Movies** – Vertigo, The Apartment, Modern Times, Rear Window

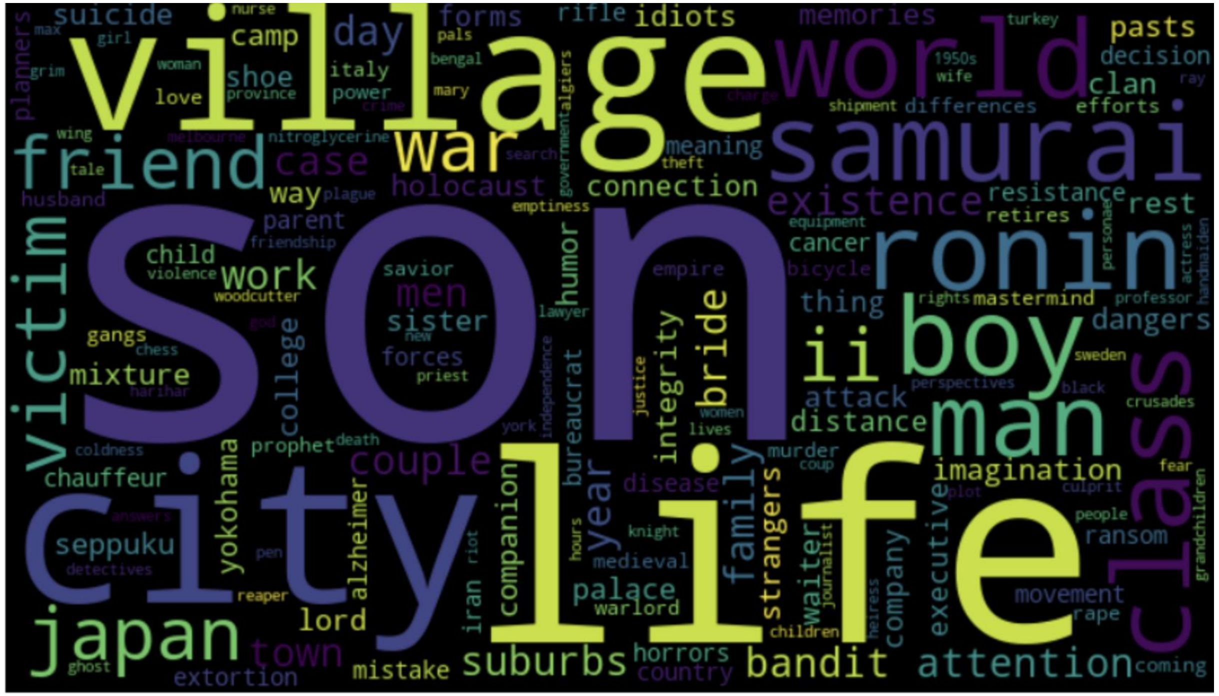


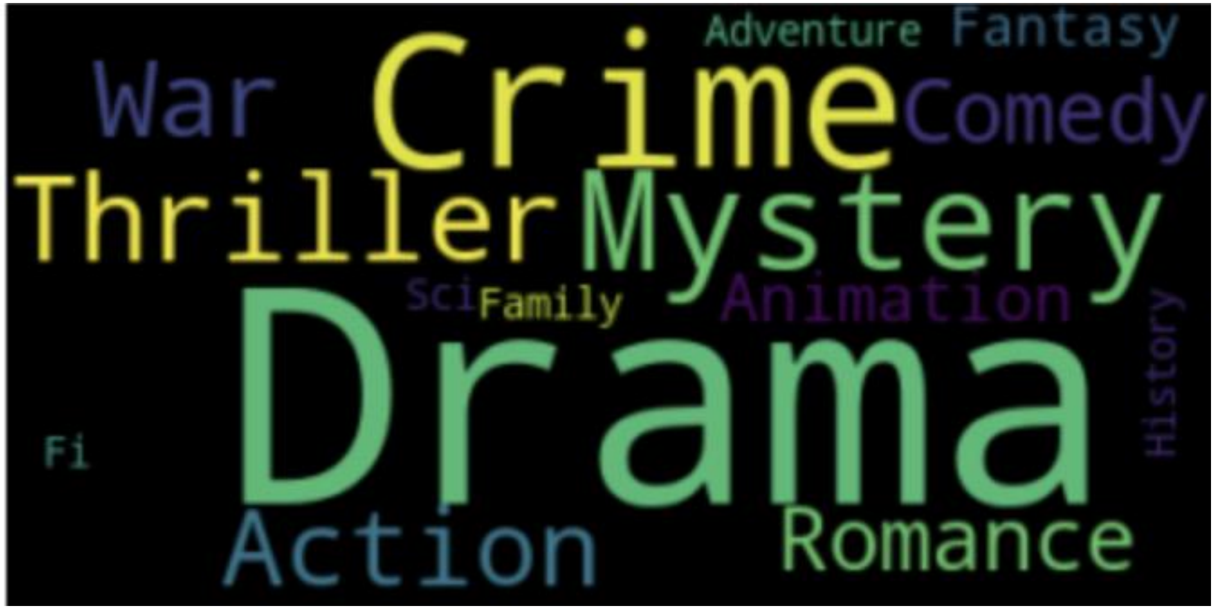


Cluster 4

>> Movies from around the world

- **Rated** –Not Rated
- **Country** – Japan, France, India, Australia, Italy
- **Directors** –Akira Kurosawa
- **Actors** – Toshiro Mifune, Tatsuya Nakadai
- **Movies** – Seven Samurai, 3 Idiots, Yojimbo, Life is Beautiful, A Separation, Ran

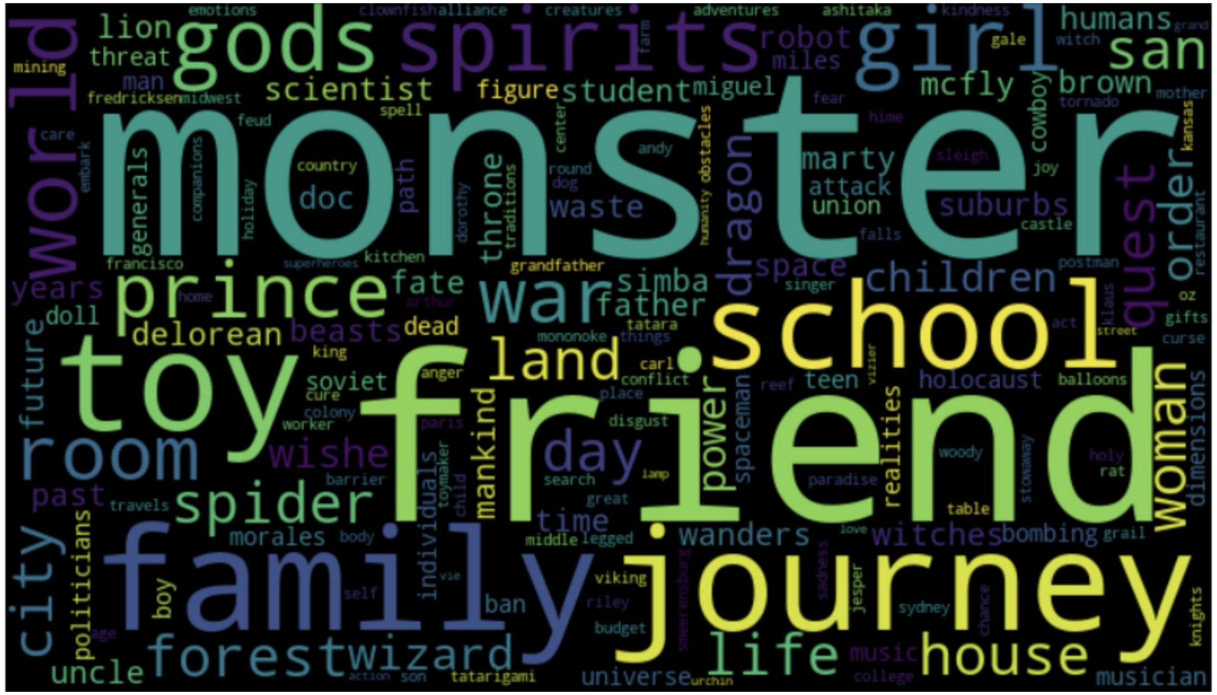


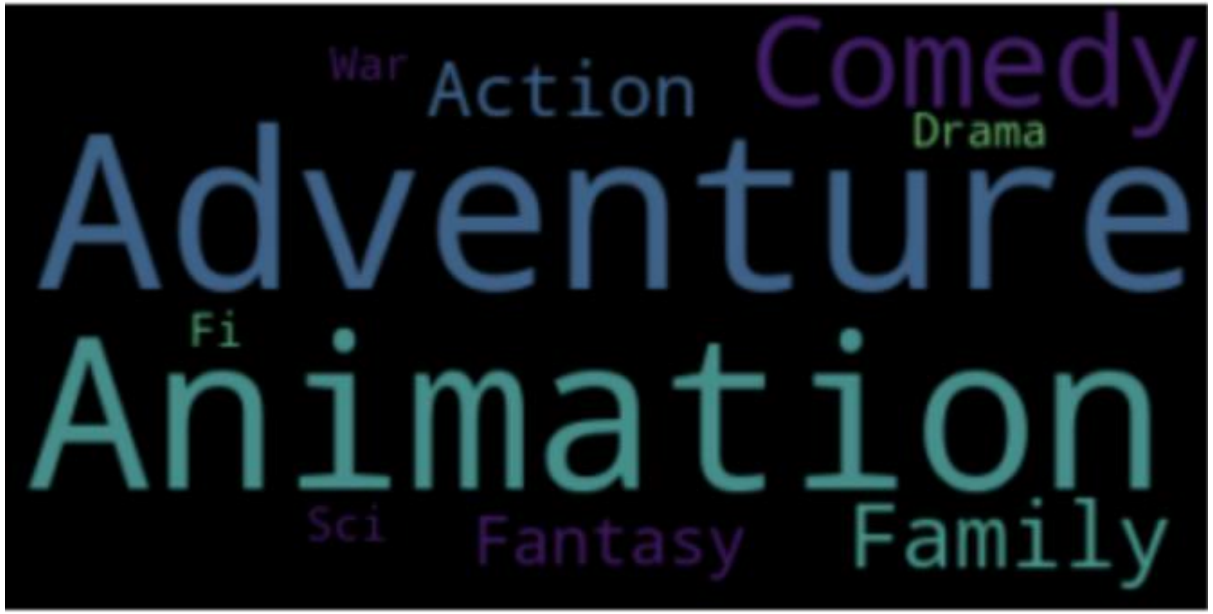


Cluster 5

>> Animated/Kids/Family Movies

- **Year –** Mostly after 1990
- **Rated** –PG, G
- **Directors** –Hayao Miyazaki, Lee Unkrich
- **Actors** – Tom Hanks
- **Movies** – Toy Story 1 &3, Spirited Away, Finding Nemo, Aladdin, Wall-E, The Incredibles, Beauty and the Beast





Conclusions

- Hierarchical Clustering : More Defined Clusters
- Dominating Genre: Crime, Drama, Action & Biography Movies
- Superhero Action & Sci-Fi Movies: More popular after 1990.
  Thanks to better VFX, DC & Marvel.
- Retro Movies : Popular for Comedy & Romance Dramas.
- International Movies: Dominantly from Japan
- Kids & Family Movies: Mostly Anime with Adventurous Plots