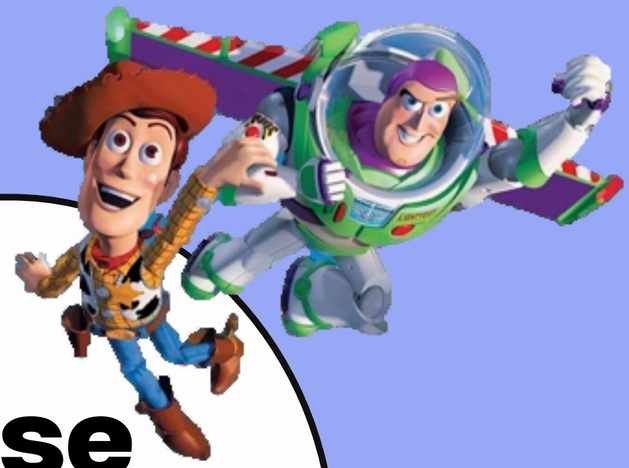


TMDb: The Movie Database

Exploring Genre Similarities and Enhancing Movie Recommendations through Clustering and Topic Modeling

Del Giudice Francesca (912367) - Nava Sara (870885) - Saresini Giulia (864967)

Master's Degree in Data Science - Text Mining & Search Course
Department of Computer Science, Systems and Communication
Academic Year 2024/25 - University of Milano Bicocca

Dataset Presentation

- The dataset contains information on a **wide range of movies (1.4M)**, including attributes such as *titles*, *genres*, *plot overviews*, *keywords*, and other metadata.
- The dataset is updated daily and is available on its **Kaggle** page.
- The **main attributes** used in this project are listed below.

OVERVIEW

Brief description or summary of the movie.

GENRES

List of genres the movie belongs to.

KEYWORDS

Keywords associated with the movie.

Project Goals

Genre Similarities Analysis with Clustering Algorithms
based on movie overviews, that can be classified in more than one genre.

Movie Recommendation System with Topic Modeling
using movie keywords to extract dominant themes and analyzing movie overviews with cosine similarity.

CLUSTERING

Data Cleaning

1 **Remove Missing Values**
Main columns, *overview* and *genres*, must be complete.

2 **Overview Cleaning**
Movie plots must have a length between 250 and 600 characters.

3 **Remove Duplicate Rows**
There are some duplicate movies, and also duplicate plots.

4 **Genre Cleaning**
Tv Movie genre genre was removed due to its limited interpretability.

Data Cleaning

The total **number of genres** in this dataset, after cleaning, is **18**.

DRAMA
COMEDY
ACTION
CRIME
HORROR
THRILLER
FANTASY
SCIENCE FICTION
ANIMATION
FAMILY
MUSIC
WAR
WESTER
HISTORY
ROMANCE
DOCUMENTARY
ADVENTURE
MISTERY

Data Cleaning

The **visualization of genres** has been restructured. Instead of a genre list as in the original dataset, binary columns were created for each movie genre.

['Drama', 'Action', 'Crime']

DRAMA	COMEDY	...	ACTION	CRIME
1	0	...	1	1

Text Representation

SentenceTransformer

all-MiniLM-L6-v2 model was chosen for text representation due to its ability to generate high-quality **sentence embeddings**, capturing semantic similarities effectively, fundamental to capture overviews relationship.

Input Tokenization	Sentence is split into tokens (words or subwords)
Transformer Encoding	Tokens pass through a transformer model to produce context-aware embeddings for each token
Pooling	The token embeddings are aggregated to create a single embedding vector for the entire sentence
Output	The resulting vector is a fixed-size representation of the input sentence, regardless of its length

Text Visualization

t-SNE Dimensionality Reduction

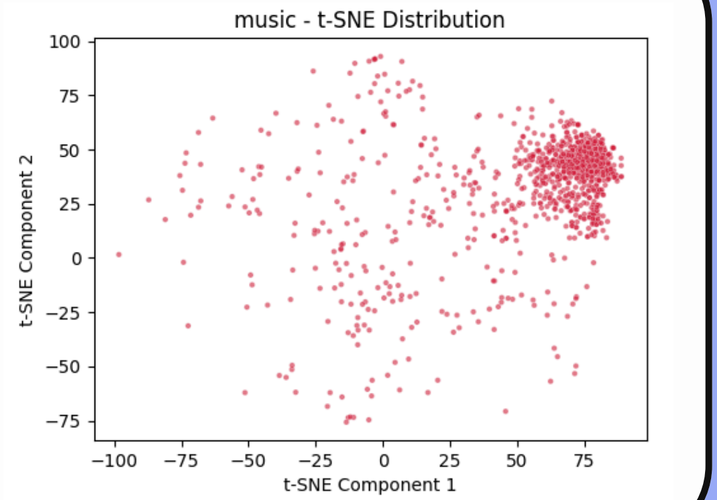
t-SNE is used to project **high-dimensional** movie overview **embeddings into** a **2D space**, enabling a visual exploration of their distribution and revealing clusters or patterns based on semantic similarity.

Text Visualization

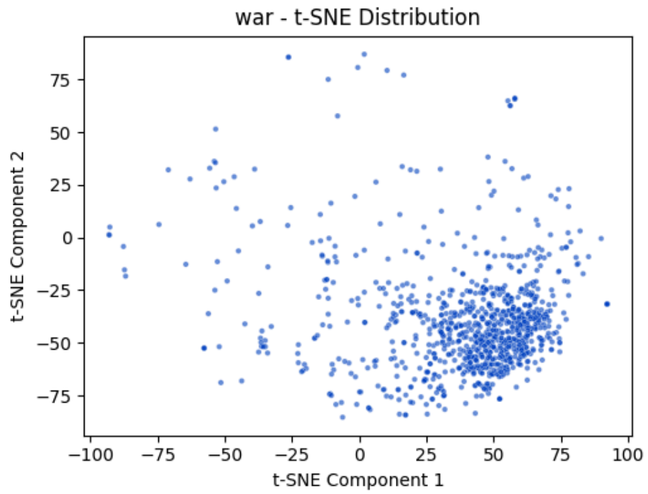
COMPACT GENRES

have points that are tightly clustered, indicating strong semantic or thematic similarities.

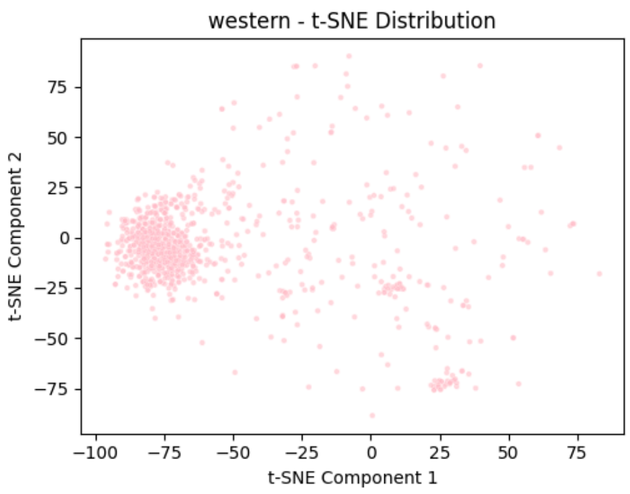
MUSIC



WAR



WESTERN

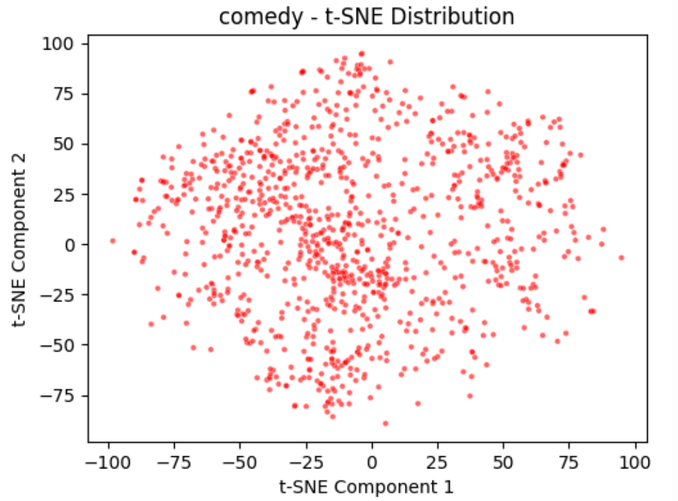


Text Visualization

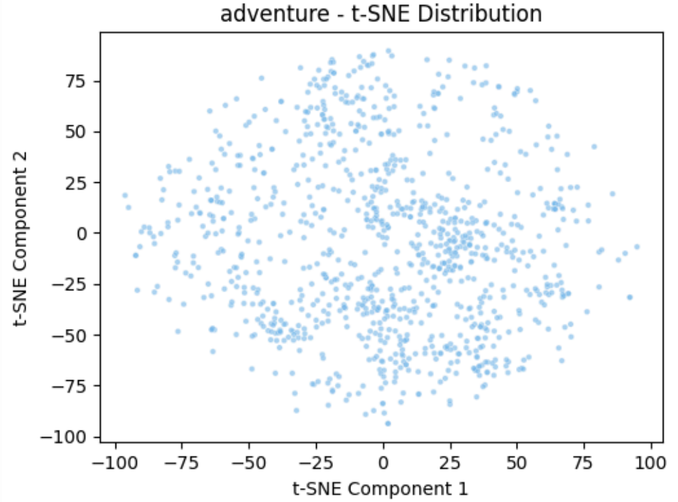
SPREAD-OUT GENRES

have points that are widely dispersed, reflecting greater variability in their semantic representation.

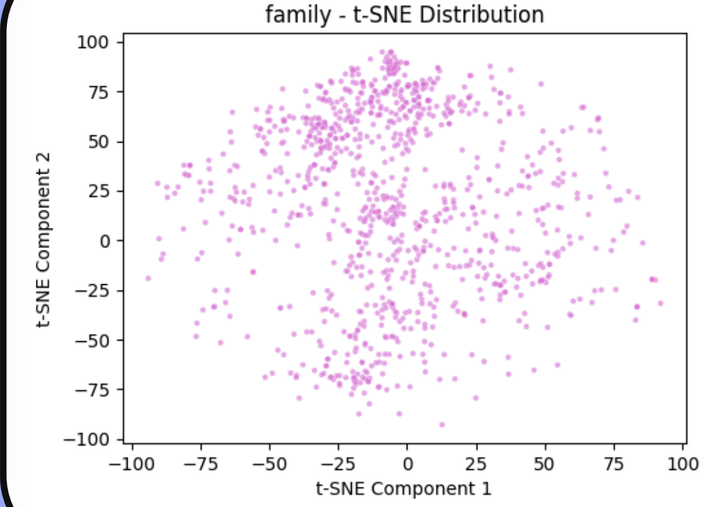
COMEDY



ADVENTURE



FAMILY



So, what is the main problem of our data?

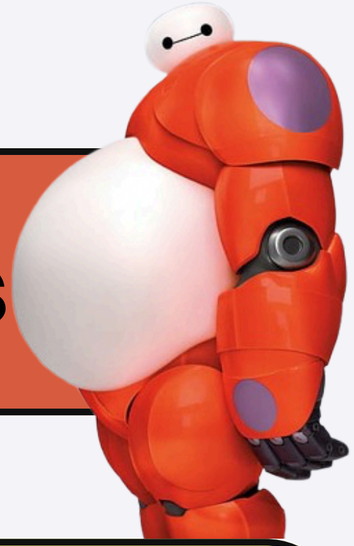
GENRES OVERLAPPING

- Movie can have more than one genre
- Some genres frequently complement many others, almost forming a distinct hybrid genre

Given this problem, how clustering can be useful?

1. Use clustering only to **group movies** based on overview similarity
2. Then analyze **genres distribution** in each cluster
3. Assess genres similarity based on **correlations**, computed on relative frequencies of each genre in each cluster

Select Clustering Algorithms



K-MEANS CLUSTERING

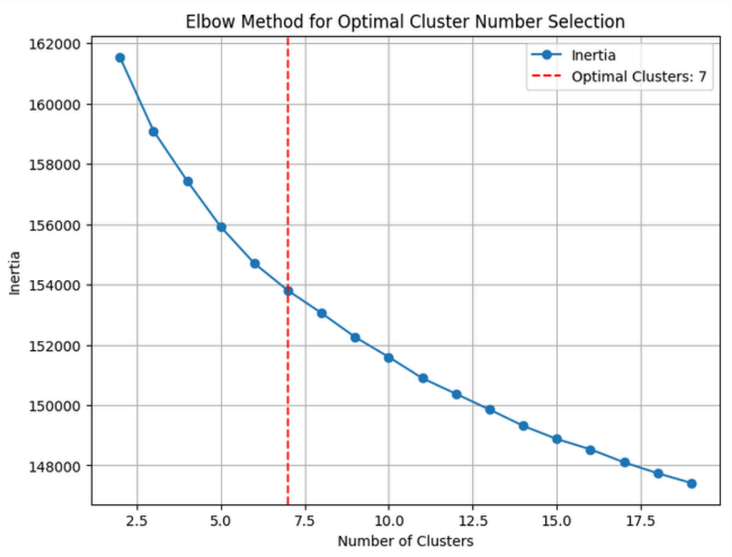
HIERARCHICAL CLUSTERING (4 CLUSTERS)

HIERARCHICAL CLUSTERING (8 CLUSTERS)

FUZZY C-MEANS CLUSTERING

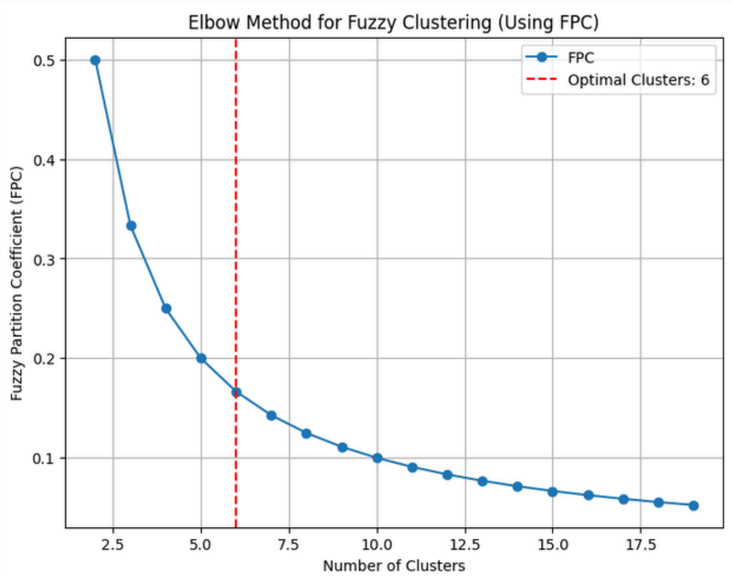
K-MEANS CLUSTERING

Find the optimal number of clusters



FUZZY C-MEANS CLUSTERING

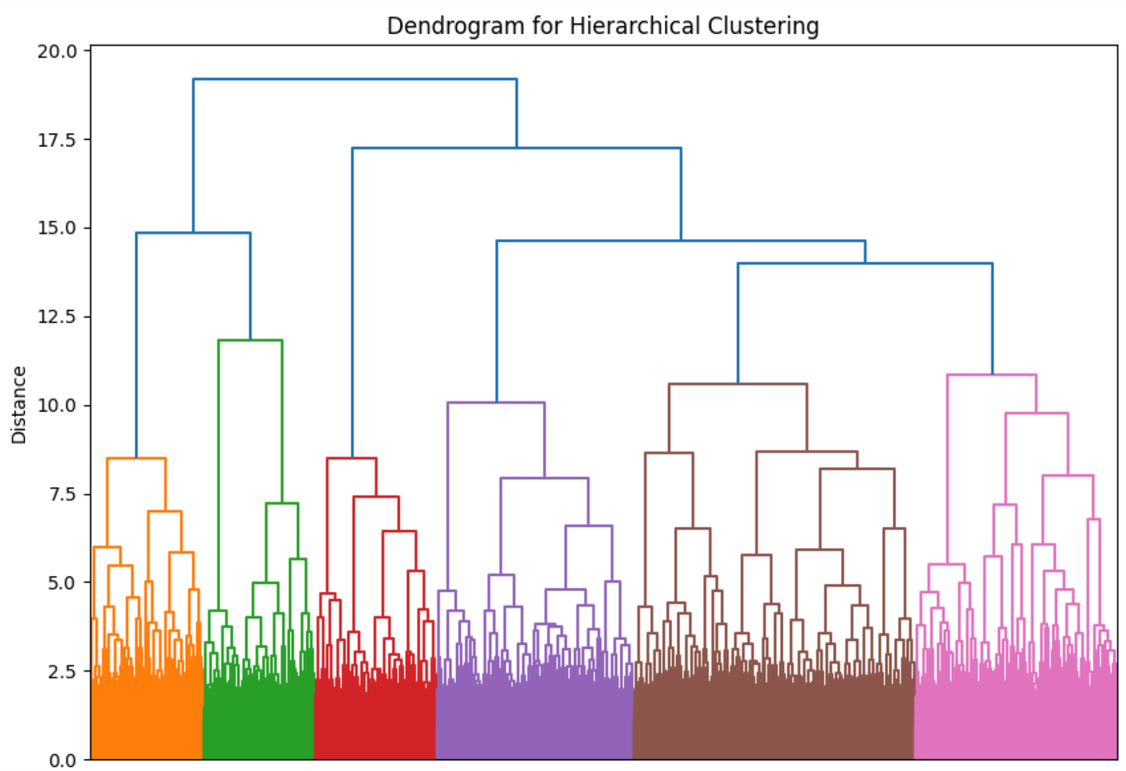
Find the optimal number of clusters



HIERARCHICAL CLUSTERING (4)

HIERARCHICAL CLUSTERING (8)

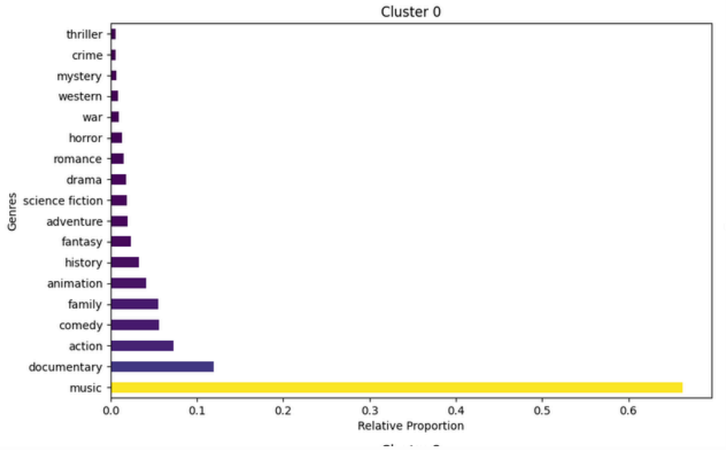
Find the optimal number of clusters



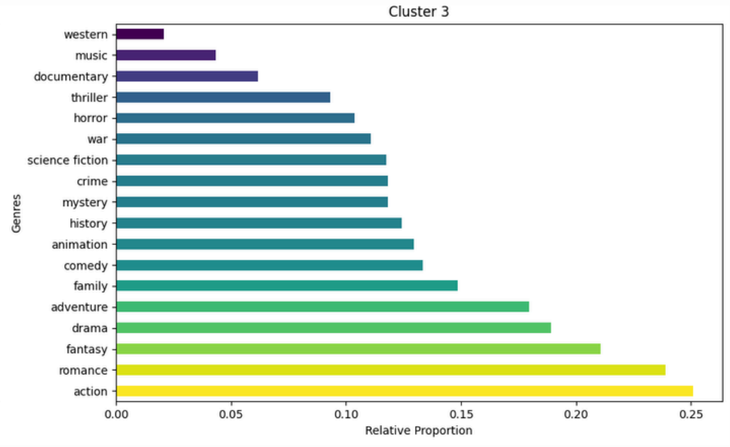
K-MEANS CLUSTERING

Analyze **relative frequencies** distribution of genres in each cluster

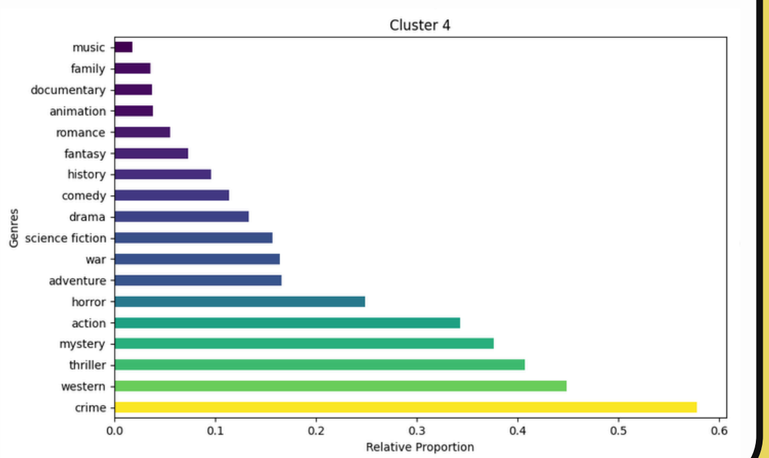
GENRE PREDOMINANCE



RANDOM ASSIGNEMENTS



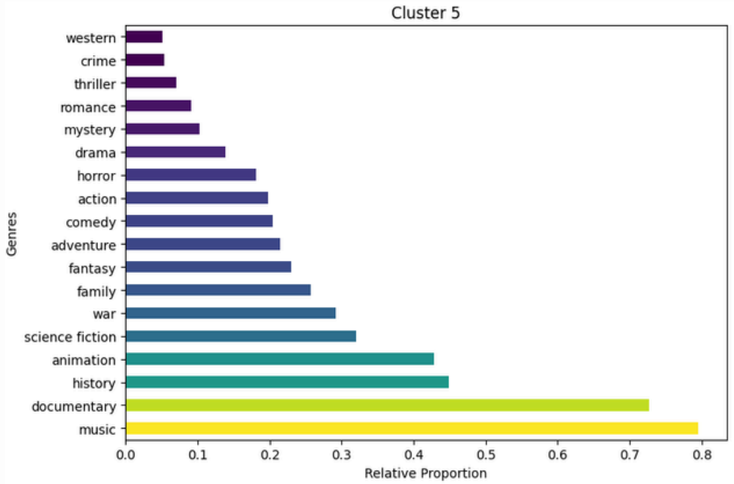
GENRES SIMILARITY



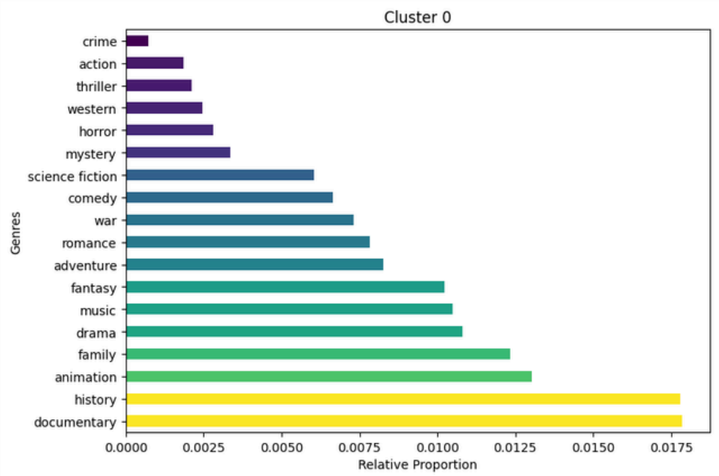
FUZZY C-MEANS CLUSTERING

Analyze **relative frequencies** distribution of genres in each cluster

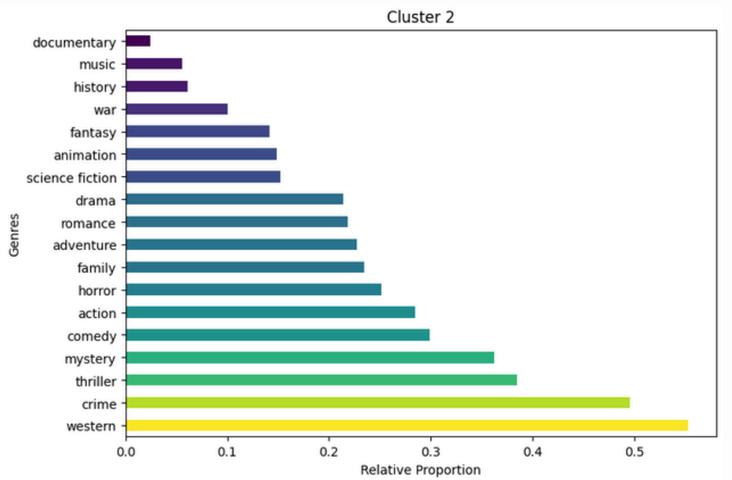
GENRE PREDOMINANCE



RANDOM ASSIGNEMENTS



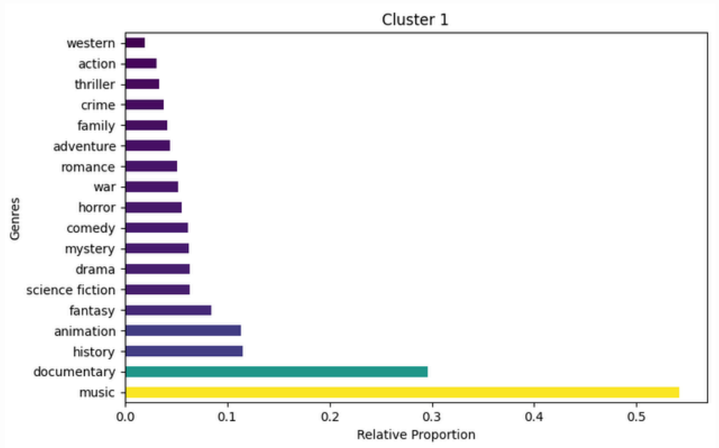
GENRES SIMILARITY



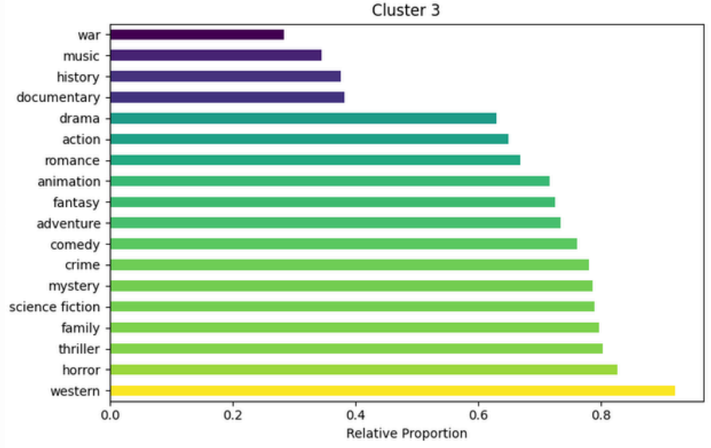
HIERARCHICAL CLUSTERING (4)

Analyze **relative frequencies** distribution of genres in each cluster

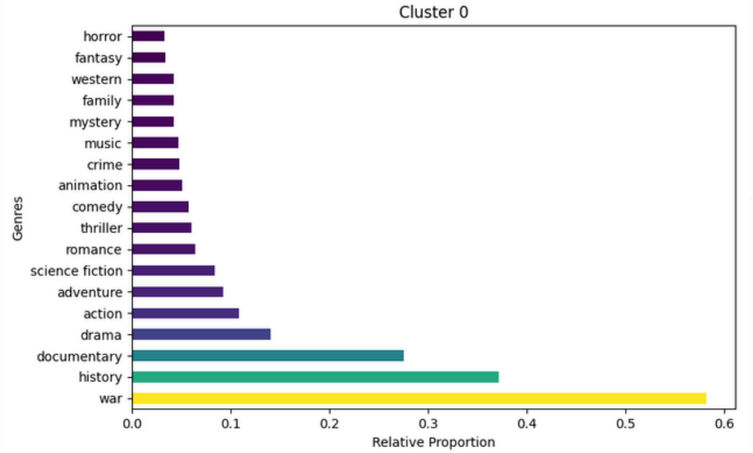
GENRE PREDOMINANCE



RANDOM ASSIGNEMENTS



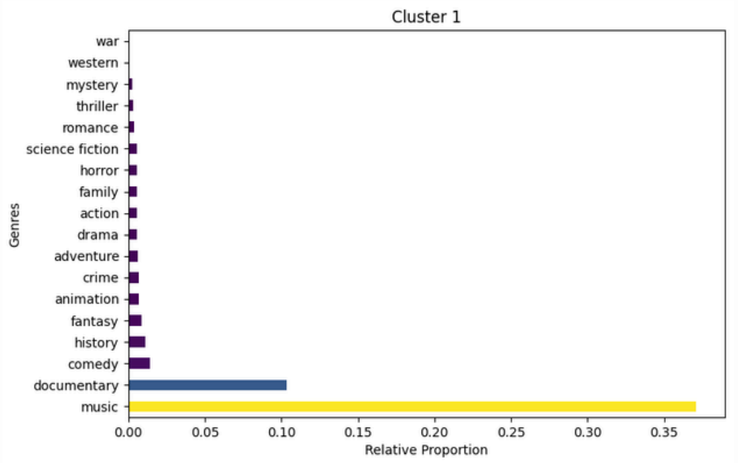
GENRES SIMILARITY



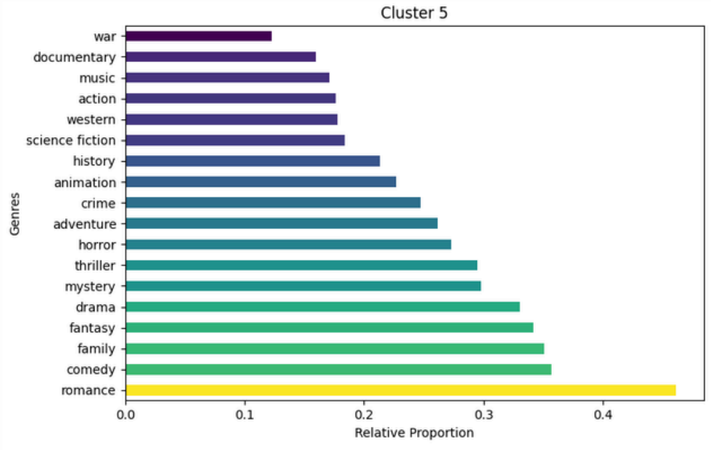
HIERARCHICAL CLUSTERING (8)

Analyze **relative frequencies** distribution of genres in each cluster

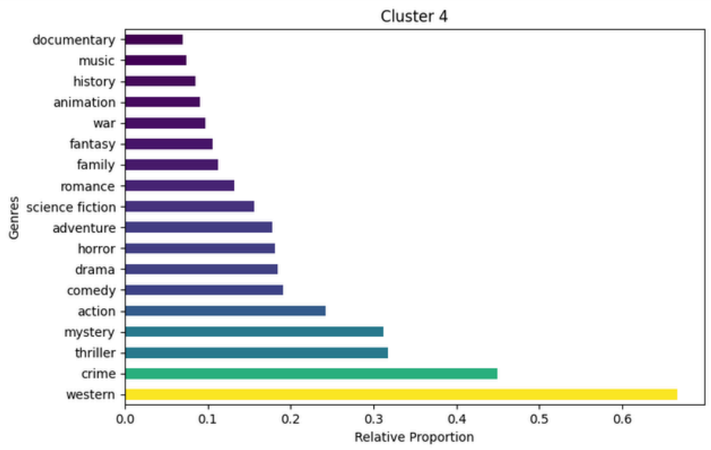
GENRE PREDOMINANCE



RANDOM ASSIGNEMENTS



GENRES SIMILARITY



Correlation Analysis

Count the number of **movies** for **each genre in each cluster**, then divide by the **size of** the respective **cluster**

Compute the **correlation matrix** from the **relative frequencies table** obtained previously, in order to find correlations between genres

Find the **top three most correlated genres** for each genre

Assessment Methods

NETWORK GRAPH

Show **relationships between movie genres** as a graph, where **genres** are represented **as nodes**, and their **similarities** (correlations) **as edges**. The layout organizes genres based on their connections, with node size reflecting the number of links and edge thickness showing the strength of similarity. The visualization highlights how genres cluster together.

MODULARITY

It measures how well genres form clusters in the graph, in particular the **density of links inside communities compared to links between communities**. If the modularity score is **high**, the graph has **clear communities** (clusters of genres). If the score is **low**, then the graph is **more evenly distributed**, and there are no clear clusters.

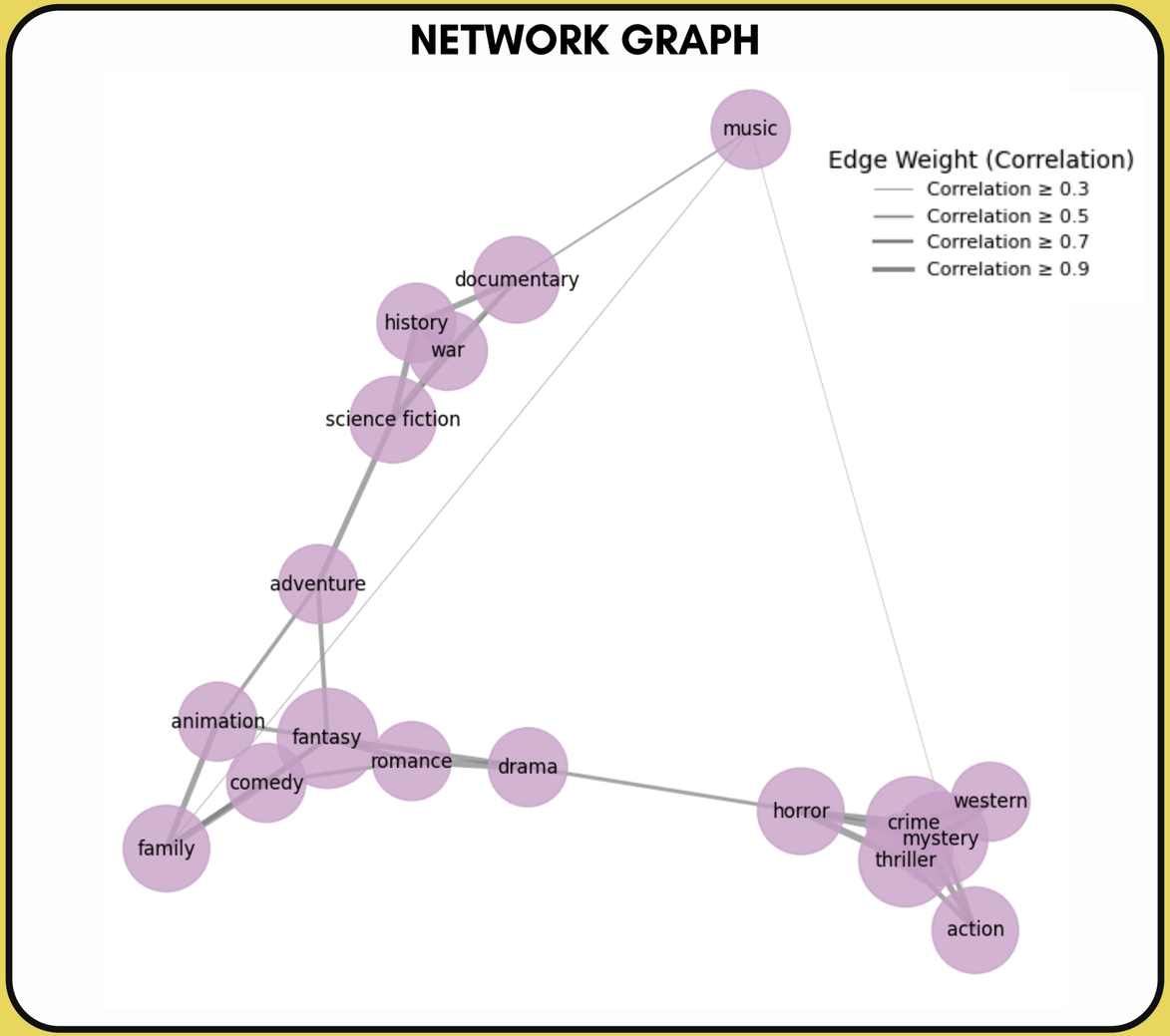
CLUSTERING COEFFICIENT

It quantifies **how close the genres are to forming clusters** by measuring the **tendency** of genres **to form triangles** (where three genres are all connected). A **higher** clustering coefficient means **genres tend to cluster together**.

GRAPH DENSITY

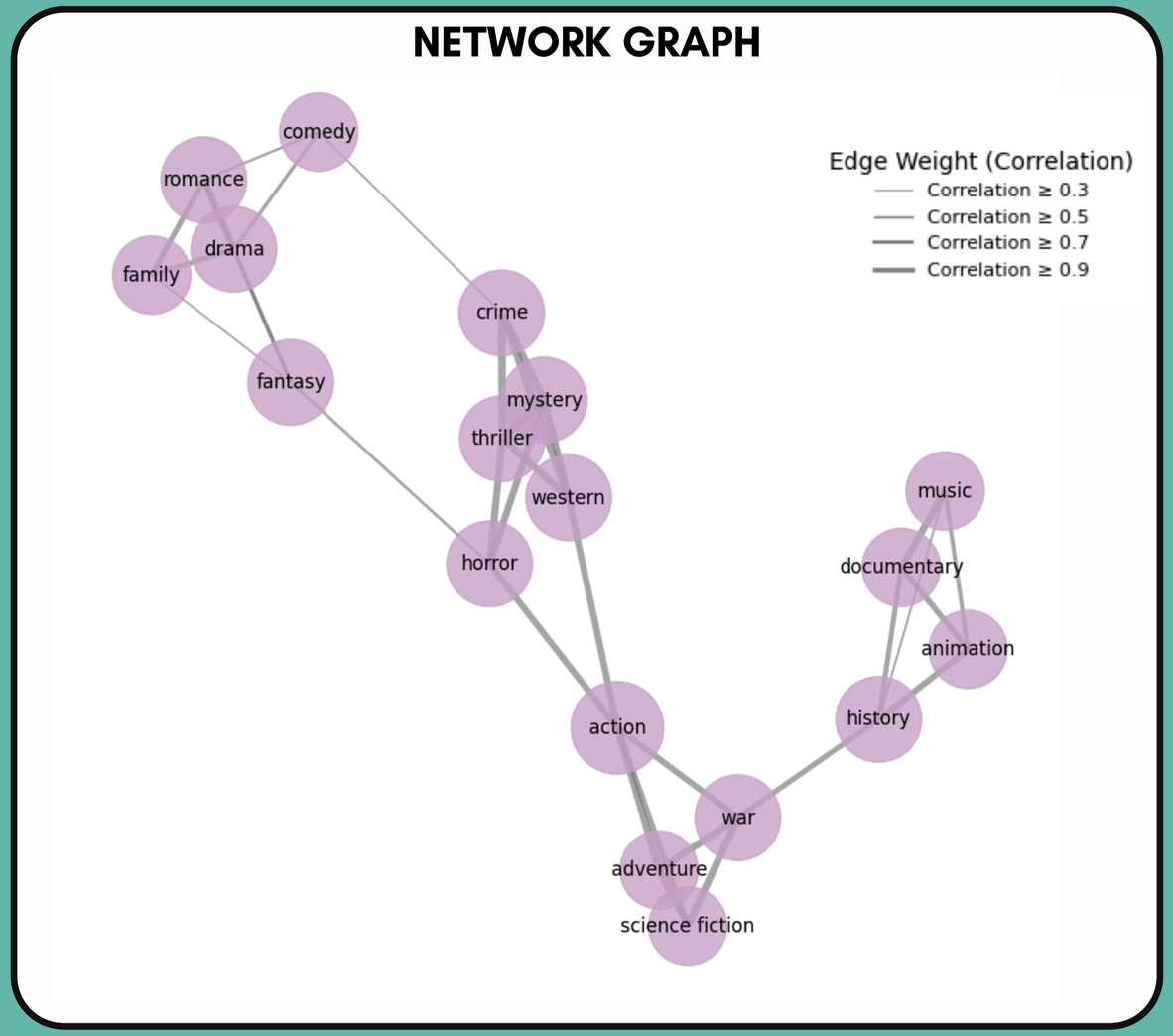
The density of the graph can give a rough idea of **how well-connected the genres are**. If the graph has **high** density (i.e., most genres are connected to each other with high correlation), it suggests there's **no clear cluster structure**.

K-MEANS CLUSTERING



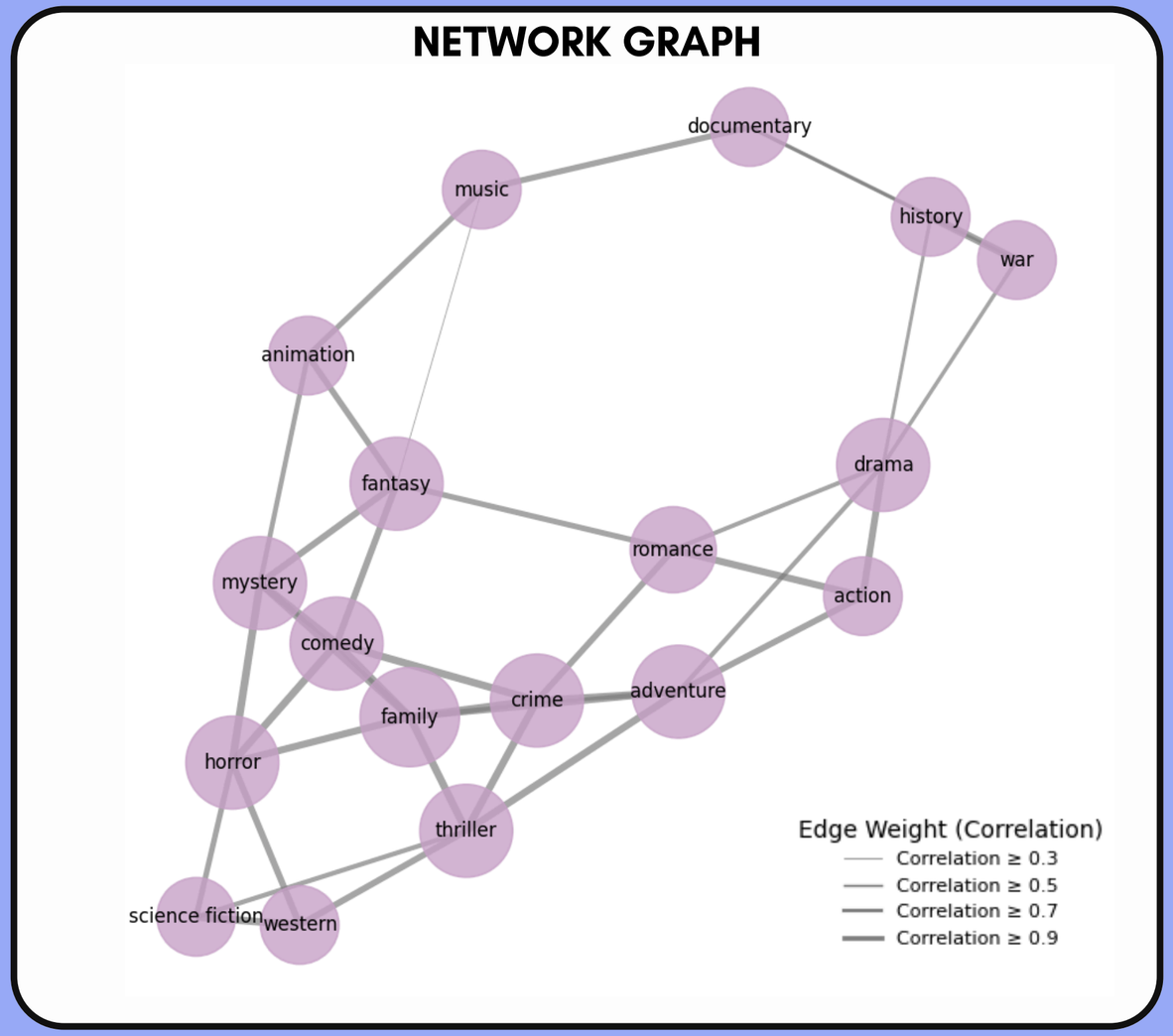
MODULARITY	0.5792
CLUSTERING COEFFICIENT	0.5796
GRAPH DENSITY	0.2222

FUZZY C-MEANS CLUSTERING



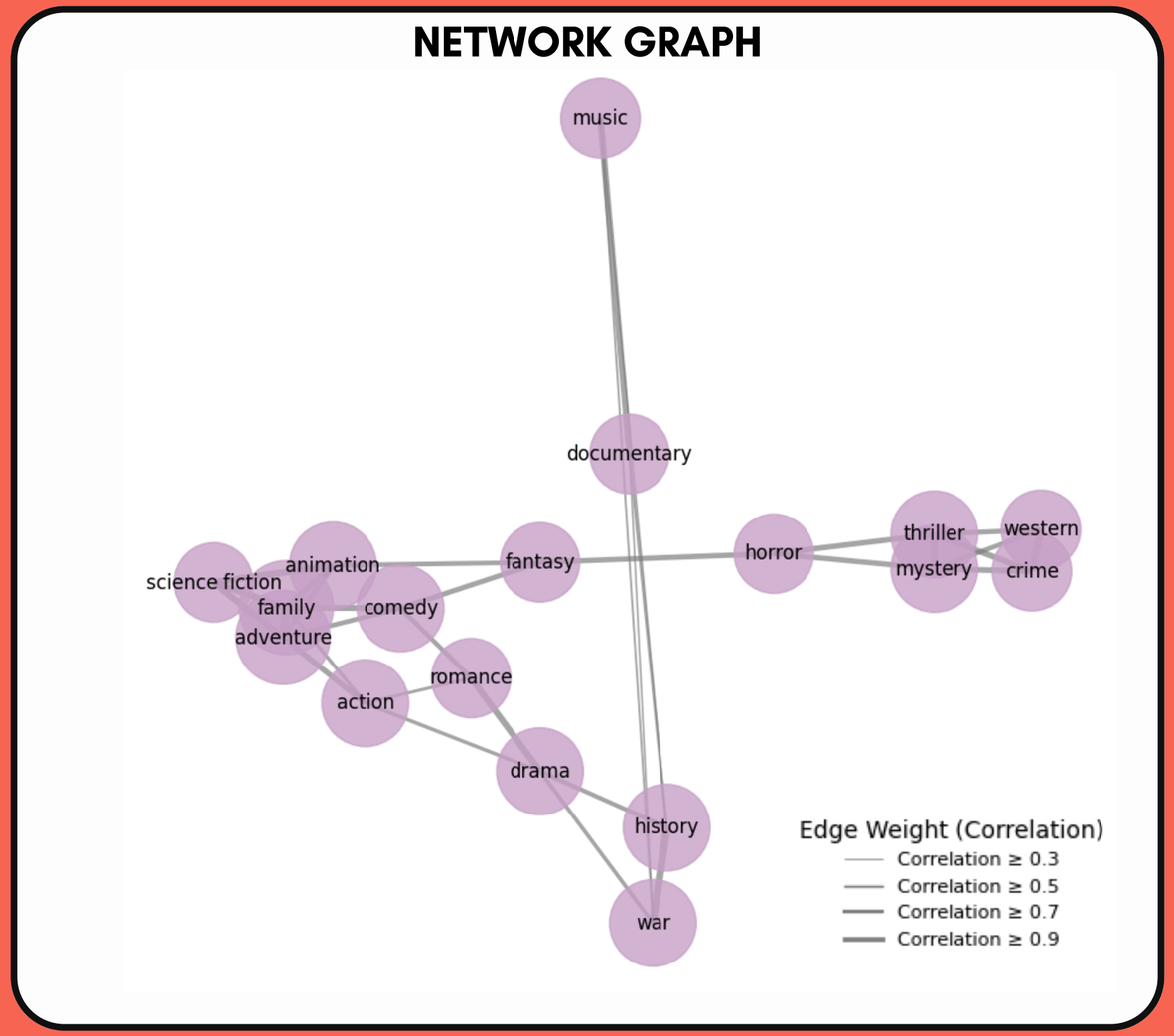
MODULARITY	0.5717
CLUSTERING COEFFICIENT	0.6648
GRAPH DENSITY	0.2157

HIERARCHICAL CLUSTERING (4)



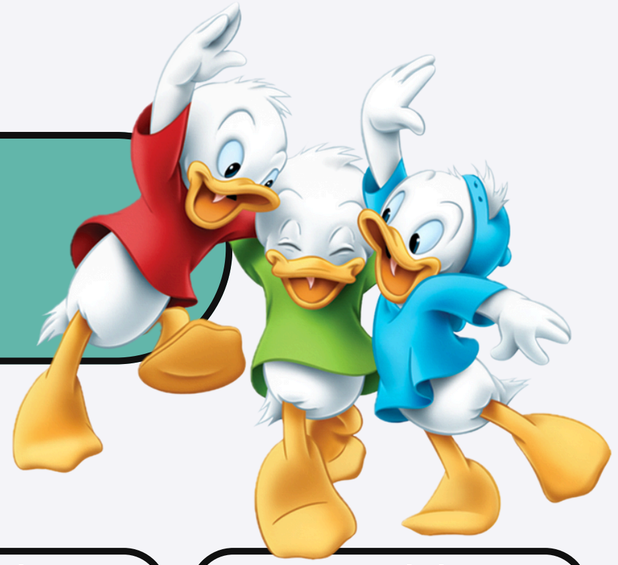
MODULARITY	0.3392
CLUSTERING COEFFICIENT	0.4722
GRAPH DENSITY	0.2418

HIERARCHICAL CLUSTERING (8)



MODULARITY	0.5123
CLUSTERING COEFFICIENT	0.5926
GRAPH DENSITY	0.2157

COMPARING RESULTS



	K-MEANS	FUZZY C-MEANS	HC (4)	HC (8)
MODULARITY	0.5792	0.5717	0.3392	0.5123
CLUSTERING COEFFICIENT	0.5796	0.6648	0.4722	0.5926
GRAPH DENSITY	0.2222	0.2157	0.2418	0.2157

Traffic light interpretation:

GOOD | MODERATE | AT RISK | BAD

COMPARING RESULTS

	K-MEANS	FUZZY C-MEANS	HC (4)	HC (8)
MODULARITY	0.5792	0.5717	0.3392	0.5123
CLUSTERING COEFFICIENT	0.5796	0.6648	0.4722	0.5926
GRAPH DENSITY	0.2222	0.2157	0.2418	0.2157

TOPIC MODELING

Data Cleaning

1

Remove Missing Values

Main columns, *overview* and *keywords*, must be complete.

2

Remove Adults Movies

The column *adult* is boolean and indicates if the movie is suitable only for adult audiences.

3

Split Composed Keywords

If a keyword consists of two different words, they should be treated separately.

Data Cleaning

Why remove **adults movies**?

SIMILAR KEYWORDS

They consist of more than 40.000 observations sharing similar keywords.

LDA POOR PERFORMANCE

The application of LDA (Latent Dirichlet Allocation) to keywords resulted in the identification of only **two topics**.

A SPECIFIC TOPIC FOR ADULTS MOVIES COMPOSED BY
[film, short, gay, sex, pornography, comedy, based, music, lgbt, love]

A GENERIC TOPIC THAT INCLUDES ALL THE OTHER MOVIES COMPOSED BY
[woman, director, relationship, war, based, sports, new, musical, world, documentary]

Data Cleaning

Why separate **composed keywords**?

LACK OF GENERALIZATION

Keywords such as *narcotics detective*, *buddy detective duo*, and *private detective* limit the system's ability to capture broader thematic connections.

LOST CONNECTIONS

Films with similar themes, such as investigations or collaborations, are not correctly grouped due to the emphasis on overly specific keywords.

Topic Modeling Results

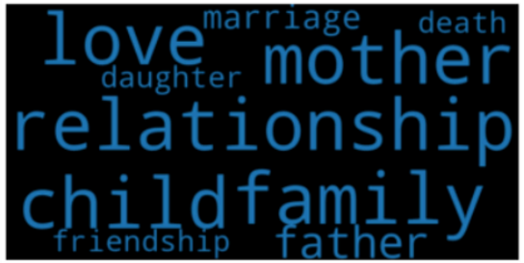
The application of **LDA** on cleaned keywords produces the following identified topics:

TOPIC 1

TOPIC 2

TOPIC 3

TOPIC 4



TOPIC 5

TOPIC 6

TOPIC 7

Movie Recommendation System

GOAL: To recommend movies based on **topics** and **cosine similarity** of **overviews**

HOW TO RECOMMEND MOVIES?

FEATURE EXTRACTION
Identify primary and secondary topics for each movie from the 7 identified with LDA, comparing certainty scores.

COSINE SIMILARITY
Calculate textual similarity between the input movie's overview and others.

TOPIC WEIGHTING
Boost similarity scores for movies with relevant topics.

Movie Recommendation System

HOW WERE TOPICS CONSIDERED?

CERTAINTY QUARTILES

- **QF**: Based on **certainty** values (**25**, **50**, and **75** percentiles)
- **QS**: Based on **second certainty** values (**25**, **50**, and **75** percentiles)

DECISION LOGIC

HIGH CERTAINTY (> QF[0.75])

- The dominant topic is selected (*first_only*)

MODERATE CERTAINTY (IN (QF[0.25] , QF[0.75]))

- **High second_certainty** (**QS[0.75]**): Both topics are considered (*both*)
- **Low second_certainty**: Only the primary topic is selected as moderate (*first_only_moderate*)

LOW CERTAINTY (QF[0.25])

- No topic is assigned (none)

Effectiveness Verification: Function Without considering Topics

Make a comparison of movie recommendation system performance for *Insurgent*, with topics consideration and without it

Without Considering Topics

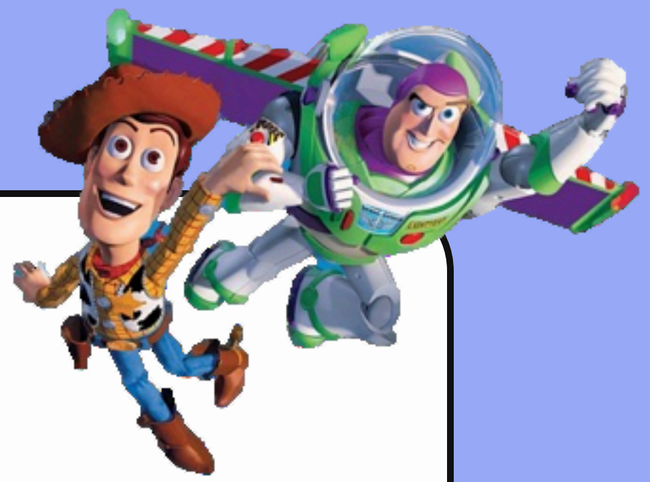
Recommendations based solely on overview similarity are thematically inconsistent, often including films from unrelated topics. These lack meaningful connections to the themes of “Insurgent.”

	title	overview	topic
223697	Antagonist	A story about a girl and her inner demons.	3
176210	Geißel der Menschheit	Venereal disease threatens to tear a young cou...	3
59112	Laura's Wedding	When love blossoms between a young Italian wom...	3

With Topic Consideration

Incorporating topic alignment results in more relevant recommendations, focusing on films within the same topic as “Insurgent.” Notably, Allegiant is recommended, being part of the same franchise. Other suggested films explore similar themes of conflict and resolution, significantly improving the relevance of the recommendations.

	title	overview	topic
601	Allegiant	Beatrice Prior and Tobias Eaton venture into t...	6
199115	Babangluksa	One year after the violent death of Carlos'; (...	6
16202	The Seven Deadly Sins: Grudge of Edinburgh Part 2	Reunited with Lancelot for the first time sinc...	6



Thank you for the attention!