



# Looking for film to watch?



By:  
Angélica Blanco & Juan Giussani



# Not knowing what to watch?

---



# Movie Recommender

## Google Colab Sheet

```
[ ] user_input = input("Enter your input: ")
recommendation = movie_recommendation2(user_input, loaded_model, movies_df)
#print("Suggested Movie:")
#print(recommendation)
```

Enter your input: mythical creature battle war

title	year	runtime	genre	rating	
Big Trouble in Little China	1986	99 min	Action, Adventure, Comedy	7.2	The true story of a heroic man, Hunter "Patch" Adams, determined to become an Egyptian Prince Moses learns of
Patch Adams	1998	115 min	Biography, Comedy, Drama	6.8	
The Prince of Egypt	1998	99 min	Animation, Adventure, Drama	7.2	

Movie Title: Big Trouble in Little China

YouTube Query Link: [https://www.youtube.com/results?search\\_query=Big%20Trouble%20in%20Little%20China](https://www.youtube.com/results?search_query=Big%20Trouble%20in%20Little%20China)

Movie Title: Patch Adams

YouTube Query Link: [https://www.youtube.com/results?search\\_query=Patch%20Adams](https://www.youtube.com/results?search_query=Patch%20Adams)

Movie Title: The Prince of Egypt

YouTube Query Link: [https://www.youtube.com/results?search\\_query=The%20Prince%20of%20Egypt](https://www.youtube.com/results?search_query=The%20Prince%20of%20Egypt)

# Dataset

## IMDB Movies - All Categories

- Dimension: 15 columns & 10.064 rows
- Source:
  - Author: [Dr. Shashikanth Vydyula](#)
  - Methodolgy: Webscraping with R from IMDB Website
  - Link to Dataset: <https://www.kaggle.com/datasets/drshashikanthvydyula/imdb-movies>

movies.csv (3.05 MB)



Detail Compact Column

15 of 15 columns

### About this file

IMDB Top Movies in All Categories with at least 25,000 votes

title	year	certificate	runtime	genre	# rating	# metacore	synopsis	director	# votes	gross	cast1	
movie name	movie released year	sensor certificate for the movie	movie run time	movie genre	IMDB rating	critics rating	movie overview	director of the movie	number of votes (minimum 25,000)			



# How does it work?

---



- Unsupervised ML model - NLP model based on the movie synopsis.
- BERTopic: Topic Modelling Technique (a state-of-the-art model language representation).

The BERTopic algorithm follows these main steps:

1. **Embedding:** Each document in the collection is transformed into a dense vector representation using BERT. This step captures the semantic meaning of the text.
2. **Dimensionality reduction:** The high-dimensional document embeddings are reduced to a lower-dimensional space using the UMAP algorithm. UMAP preserves the local structure and relationships between documents.
3. **Clustering:** The reduced embeddings are clustered using HDBSCAN, which is a density-based clustering algorithm.
  - HDBSCAN identifies clusters of documents based on their density and connectivity.
4. **Topic extraction:** Representative documents are selected for each cluster to serve as topic labels. These representative documents are usually the most representative or frequent documents within each cluster.

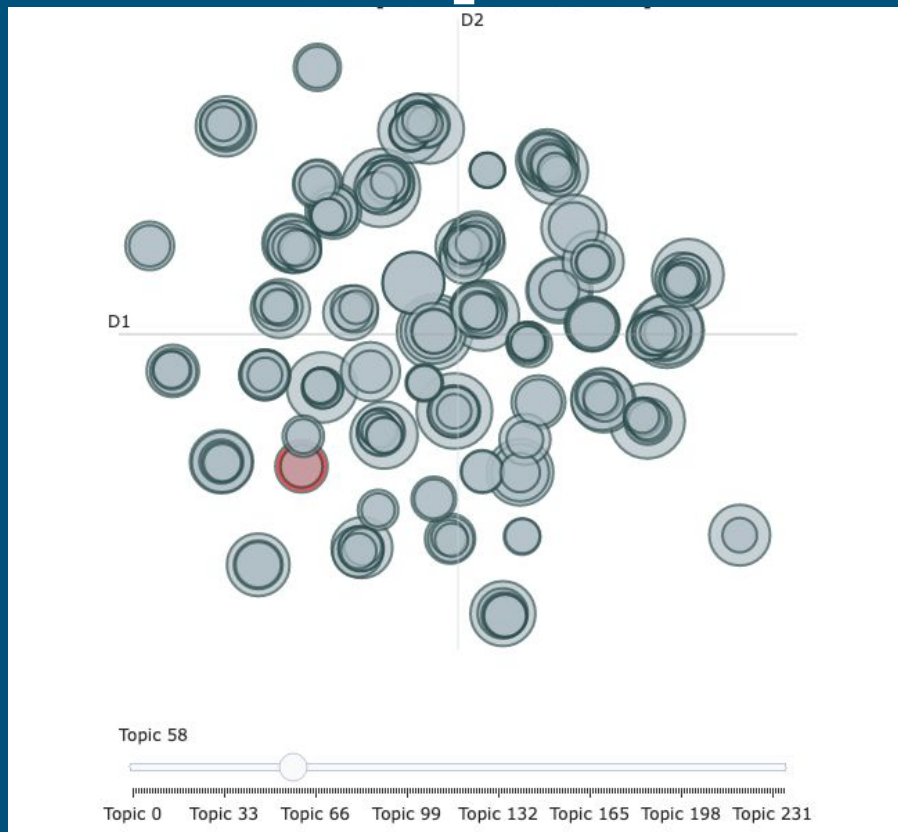
# Results

- 233 Topics
- Around 3.000 synopsis are classified as outliers
- Min cluster size: 15 synopsis

	Topic	Count	Name
0	-1	3058	-1_murder_crime_detective_encounter
1	0	97	0_magic lamp_lamp_youthful beauty_young artist
2	1	97	1_kidnapped daughter_kidnapped son_family kidn...
3	2	93	2_captured outlaw_outlaw_smalltime rancher_cat...
4	3	77	3_young couple_fall love_affair_relationship
...	...	...	...
228	227	15	227_time uncover_officer stuck_masked man_inve...
229	228	15	228_luck curtain_luck secretly_person luckier_...
230	229	15	229_survivor apocalypse_group survivor_survivo...
231	230	15	230_father disappearance_emily missing_discove...
232	231	15	231_holocaust jewishhungarian_auschwitz_fellow...

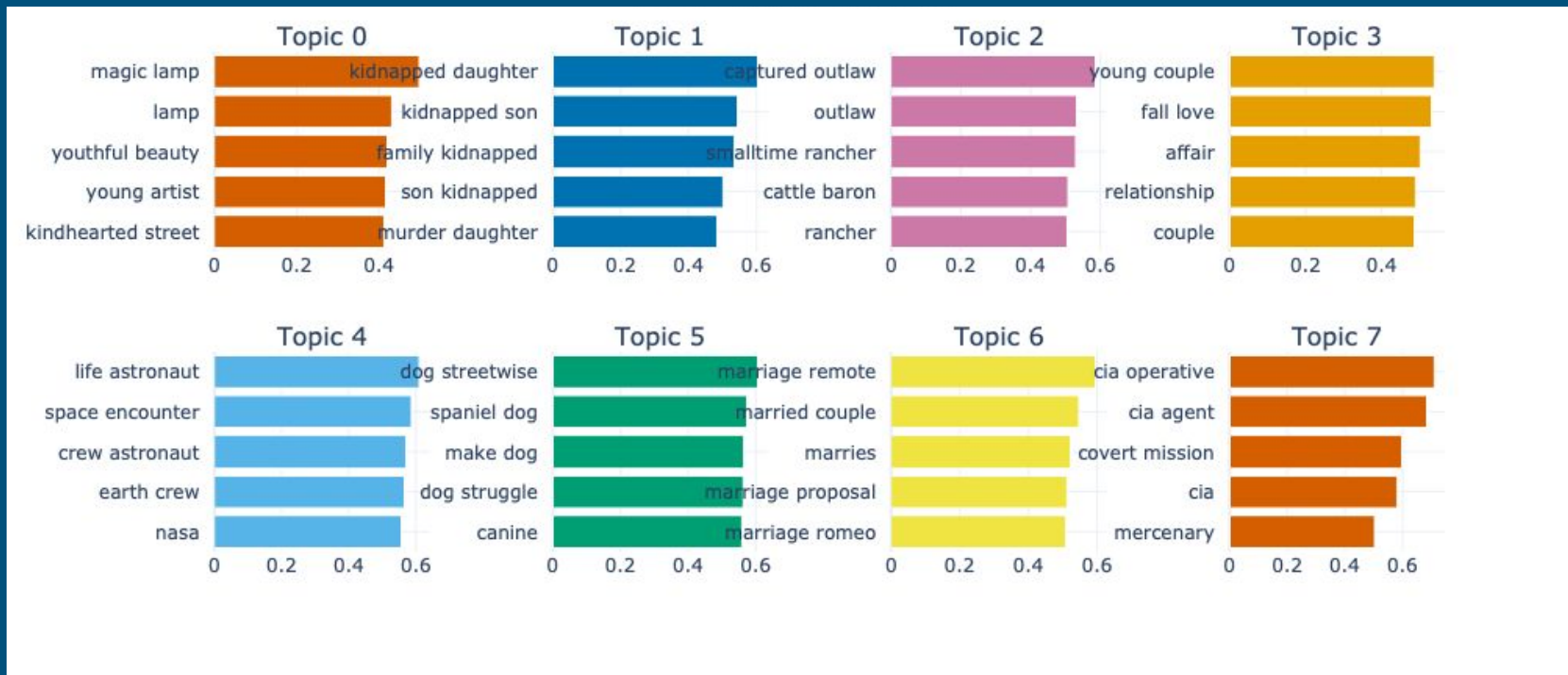
233 rows × 3 columns

# Intertopic Distance Map

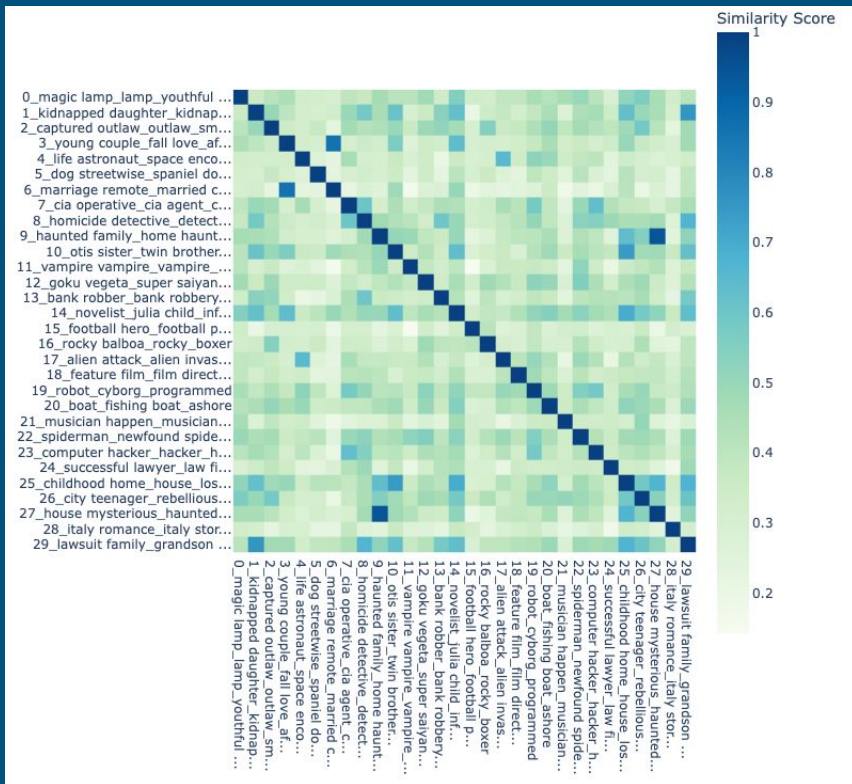




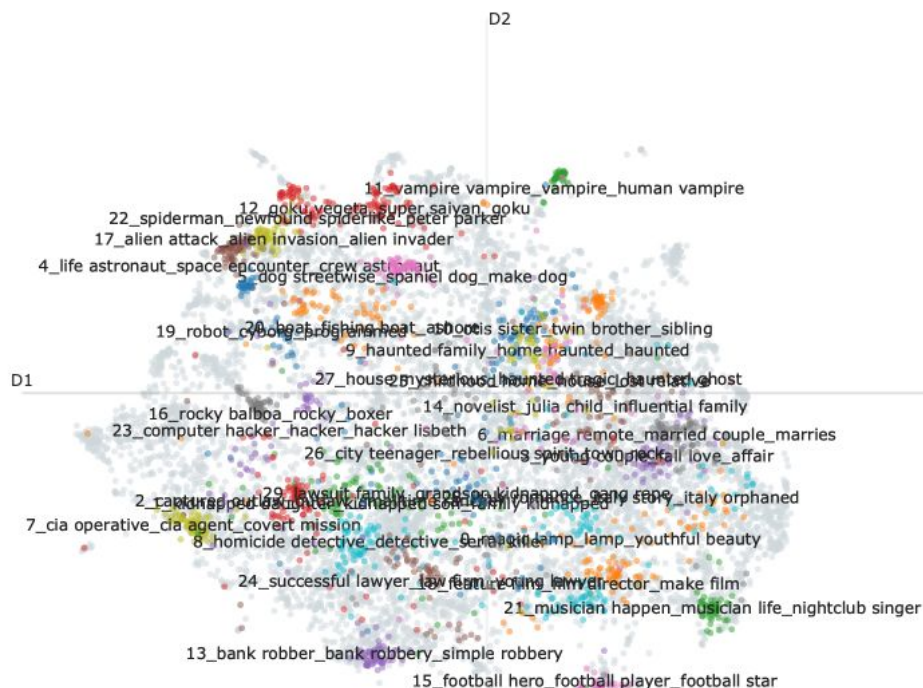
# Topic Word Scores



# Topic Similarity Heatmap



# Synopsis and clusters



- 0\_magic lamp\_lamp\_youthful beauty
- 1\_kidnapped daughter\_kidnapped son\_family kidnapped
- 2\_captured outlaw\_outlaw\_smalltime rancher
- 3\_young couple\_fall love\_affair
- 4\_life astronaut\_space encounter\_crew astronaut
- 5\_dog streetwise\_spaniel dog\_make dog
- 6\_marriage remote\_married couple\_marries
- 7\_cia operative\_cia agent\_covert mission
- 8\_homicide detective\_detective\_serial killer
- 9\_haunted family\_home haunted\_haunted
- 10\_otis sister\_twin brother\_sibling
- 11\_vampire vampire\_vampire\_human vampire
- 12\_goku vegeta\_super saiyan\_goku
- 13\_bank robber\_bank robbery\_simple robbery
- 14\_novelist\_julia child\_influential family
- 15\_football hero\_football player\_football star
- 16\_rocky balboa\_rocky\_boxer
- 17\_alien attack\_alien invasion\_alien invader
- 18\_feature film\_film director\_make film
- 19\_robot\_cyborg\_programmed
- 20\_boat\_fishing boat\_ashore
- 21\_musician happen\_musician life\_nightclub singer
- 22\_spiderman\_newfound spiderlike\_peter parker
- 23\_computer hacker\_hacker\_hacker lisbeth
- 24\_successful lawyer\_law firm\_young lawyer
- 25\_childhood home\_house\_lost relative
- 26\_city teenager\_rebellious spirit\_town rock
- 27\_house mysterious\_haunted tragic\_haunted ghost
- 28\_italy romance\_italy story\_italy orphaned
- 29\_lawsuit familv grandson kidnapped gang rape

# Why BERTopic?

---



- Model performance is subjective. Every model has different classifications
- Strengths:
  - Scales better with larger datasets than traditional models
  - Dimensionality map reduction embedded
  - Outliers: HDBSCAN leads more coherent and consistent topics
  - Finding automatically the optimal number of topics and also creating them
  - Performance is better with shorter documents - synopsis
- Weakness:
  - Longer training time
  - Expensive computational resources