Looking for film to watch?

By:
Angélica Blanco & Juan Giussani

Not knowing what to watch?

Movie Recommender

Google Colab Sheet

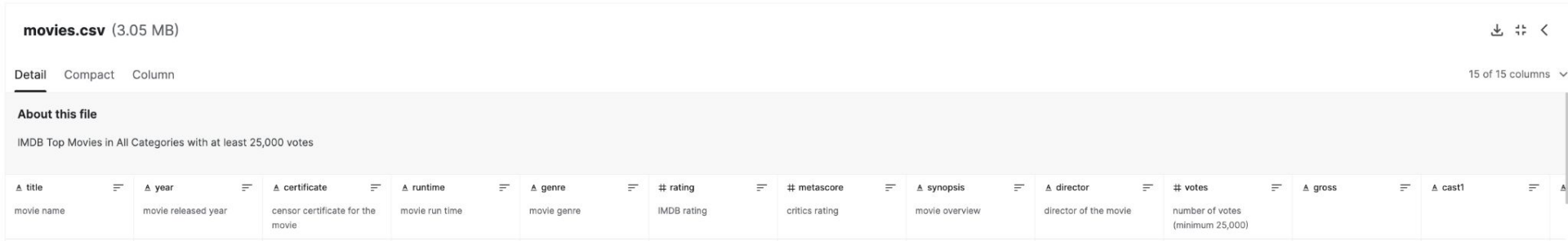
```
[ ] user_input = input("Enter your input: ")
    recommendation = movie_recommendation2(user_input, loaded_model, movies_df)
    #print("Suggested Movie:")
    #print(recommendation)
```

```
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   |                                       A rough-and-tumble trucker and his
|         Patch Adams         | 1998 | 115 min |  Biography, Comedy, Drama   |  6.8   | The true story of a heroic man, Hunter "Patch" Adams, determined to beco
|     The Prince of Egypt     | 1998 | 99 min  | Animation, Adventure, Drama |  7.2   |                                         Egyptian Prince Moses learns of
+-----------------------------+------+---------+-----------------------------+--------+--------------------------------------------------------------------------
Movie Title: Big Trouble in Little China
YouTube Query Link: 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

Movie Title: The Prince of Egypt
YouTube Query Link: 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




young
Two
man
life
woman
find
story
family
group
take
set
must
love
girl
Three
American
become
live
high school
world
look
come
sent
death
help
struggle
return
follow
small
murder
son
people
day
friend
young woman
young man
father
New York
former
teenager
discover
year old
team
work
begin
mysterious
couple
travel
fall
one
film
gets
meet
Four
mother
soldier
writer
Based
goes
scientist
wife
criminal
boy
cop
detective
daughter
living
attempt
forced
husband
tries
year
prison
small town
decide
brother
join
turn
college
agent
th
town
police
student
high

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.

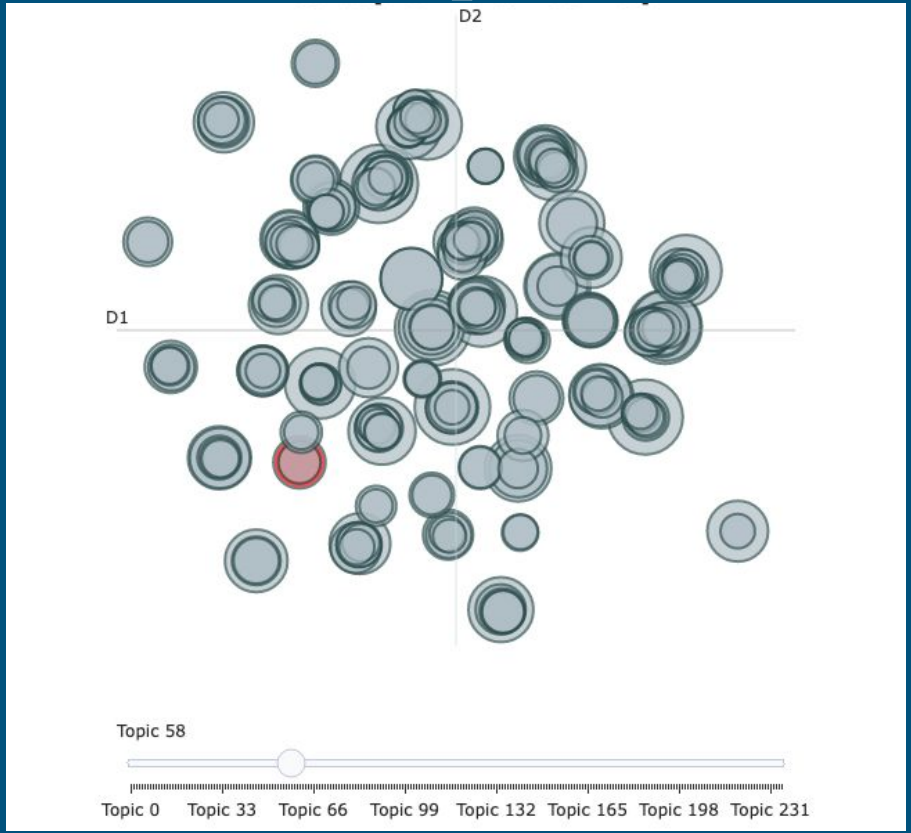
Results

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

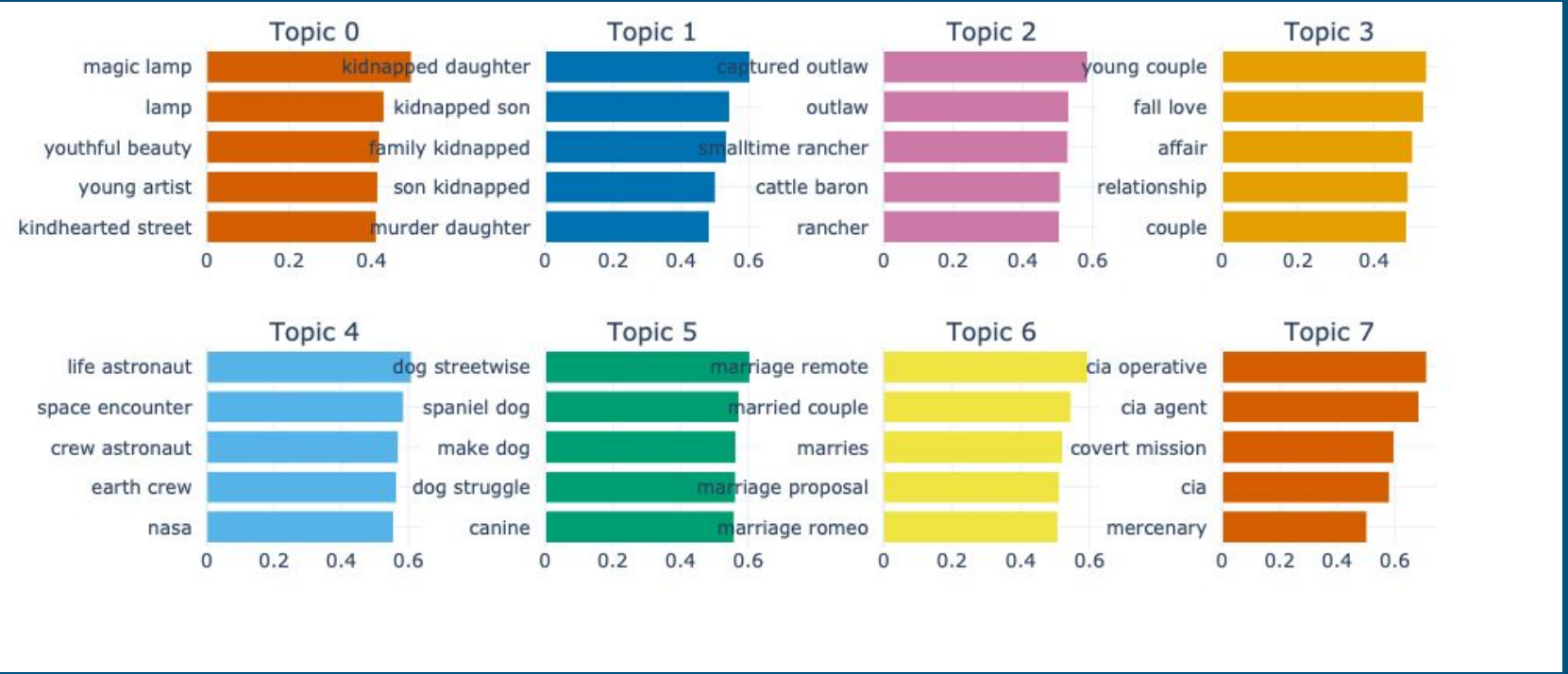
	Topic	Count	Name
0	-1	3058	-1_murder_crime_detective_encounter
1	0	97	0_magic lamp_lamp_youthful beauty_young artist
2	1	97	1_kidnapped daughter_kidnapped son_family kidn...
3	2	93	2_captured outlaw_outlaw_smalltime rancher_cat...
4	3	77	3_young couple_fall love_affair_relationship
...	...	...	...
228	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

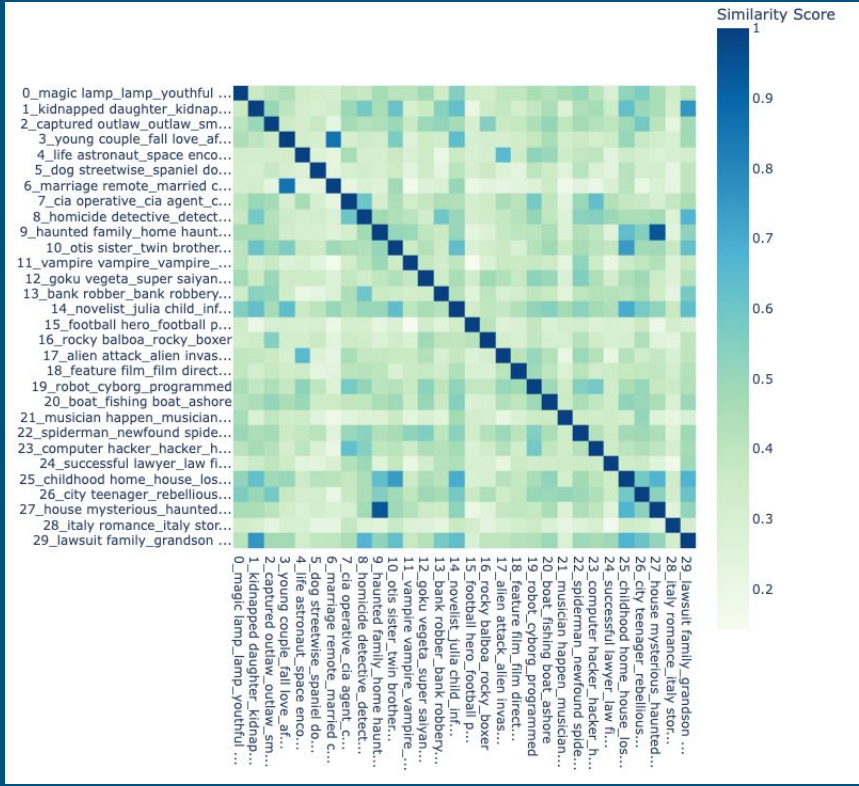
Intertopic Distance Map



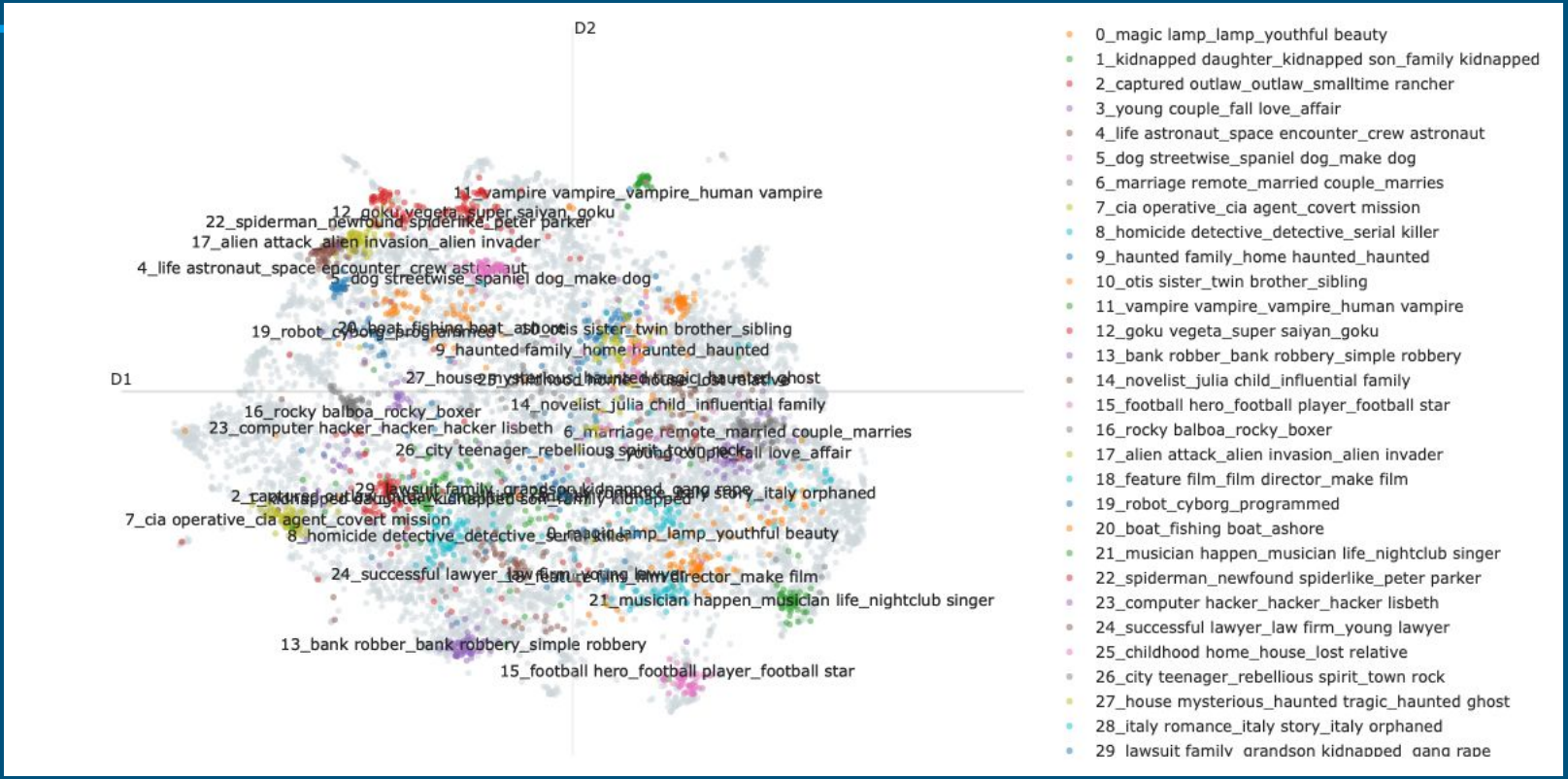
Topic Word Scores



Topic Similarity Heatmap



Synopsis and clusters



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 ressources