

# Exploring Netflix: Uncovering Patterns in Content Ratings, Genres, and Themes

Michael Wong

# Abstract

This project utilizes the Netflix Movies and TV Shows dataset from Kaggle, encompassing 8807 entries across 12 attributes. We aimed to investigate three key questions:

1. What are the most common genres for long-running TV shows (at least five seasons)?
2. Can we build a machine-learning model that predicts the content rating of a Netflix movie or TV show based on other attributes?
3. What are the most common themes or topics in Netflix TV shows and movie descriptions?

Descriptive statistics and machine learning techniques were employed to analyze the data.

Our findings reveal significant patterns in genres of long-running shows, demonstrate the feasibility of content rating prediction, and uncover recurring themes in descriptions.

# Motivation

In the era of streaming services, understanding content's nature and preferences is crucial for producers and consumers.

For producers and content creators, discerning the characteristics of long-running shows could guide creating new, successful content. Predicting content ratings could aid in targeted marketing and age-appropriate categorization.

For consumers, recognizing common themes can assist in choosing new shows and movies to watch. Additionally, this analysis can provide valuable insights for streaming platforms beyond Netflix, aiding in competitive analysis and strategic decision-making in the rapidly evolving entertainment industry.

# Dataset



The dataset I chose for this project is the Netflix Movies and TV Shows sourced from Kaggle (<https://www.kaggle.com/datasets/shivamb/netflix-shows>) with 8807 rows × 12 columns

This dataset includes detailed information about various movies and TV shows available on Netflix. It consists of key attributes such as type (movie or TV show), title, director, cast, country of production, date added to Netflix, release year, rating, duration, genre, and a concise show description.

	Variable	Type	Description
0	show_id	Nominal	Unique identifier
1	type	Nominal	'Movie' or 'TV Show'
2	title	Nominal	Title of show/movie
3	director	Nominal	Show/movie director
4	cast	Nominal	Cast members
5	country	Nominal	Production country
6	date_added	Interval	Date added to Netflix
7	release_year	Interval	Initial release year
8	rating	Ordinal	Content rating (e.g., PG-13)
9	duration	Mixed	Duration (minutes/seasons)
10	listed_in	Nominal	Show/movie genre(s)
11	description	Nominal	Show/movie description

# Data Preparation and Cleaning Methods I Chose

- Listwise Deletion: The simplest method involves removing all data for observation with one or more missing values.
- Mean/Median/Mode Imputation: A method to fill in the missing values with estimated ones. The objective is to employ known relationships that can be identified in the valid values of the data set to assist in estimating the missing values. It consists of replacing the missing data for a given attribute with the mean or median (quantitative attribute) or mode (qualitative attribute) of all known values of that variable.
- Prediction Models: Also a kind of imputation method, including regression models and machine learning models (like k-Nearest Neighbors and Random Forests). The missing values of an attribute are predicted with the help of other attributes.
- Last Observation Carried Forward & Next Observation Carried Backward: This technique is usually used in time-series data where missing values are filled with either the last or the next observed value.
- Interpolation and Extrapolation: Interpolation estimates a missing value between two known values. Extrapolation, on the other hand, estimates a value outside the range of known values.

# Data Preparation and Cleaning

- 'director': I chose to fill in missing values in the 'director' column based on the 'cast' because it's reasonable to assume a relationship between directors and the actors they frequently work with. Also, the 'director' column has a lower cardinality feature, making it manageable for this operation. If impossible, we'll fill the remaining nulls with 'Unknown.'
- 'country': I filled in the missing values in the 'country' column based on the 'director' because it's plausible that a director primarily works within a certain country. Also, the 'country' column also has a relatively lower cardinality, making it suitable for this operation. If impossible, we'll fill the remaining nulls with 'Unknown.'
- 'cast': This is more challenging due to its nature. It's a high cardinality feature (many unique actors), and many entries contain multiple actors. The relationship between the cast and other features isn't as direct or reliable, making assumptions for imputation less accurate. Therefore, I decided to fill the missing values with a placeholder 'Unknown.'

```
In [19]: # Check for missing values
missing_columns = df.columns[df.isnull().sum() > 0]
df[missing_columns].isnull().sum()
```

```
Out[19]: director      2634
cast                825
country             831
date_added          10
rating               4
duration            3
dtype: int64
```

# Data Preparation and Cleaning

- 'date\_added': This is a date column. A simple approach could be to fill in missing values with the most common date (mode); in this project, however, I chose a method called backfill or forward fill, where I fill missing values with the previous or next value in the column.
- 'duration': We can delete these rows since they are only a small portion of the dataset.
- 'rating': I will build a machine learning model to predict their values for the remaining rating columns. (Which is also one of my research questions :) )

```
In [19]: # Check for missing values
missing_columns = df.columns[df.isnull().sum() > 0]
df[missing_columns].isnull().sum()
```

```
Out[19]: director      2634
cast                825
country             831
date_added           10
rating                4
duration              3
dtype: int64
```

# Research Question 1

What are the most frequent genres for long-running TV shows (at least five seasons long)?

## Methods

Loaded and examined the Netflix titles dataset using pandas.

Filtered the data for TV shows with five seasons or more using pandas.

Extracted and counted these long-running TV shows' genres (listed\_in) using pandas.

Visualized the genre counts with a bar chart using matplotlib.

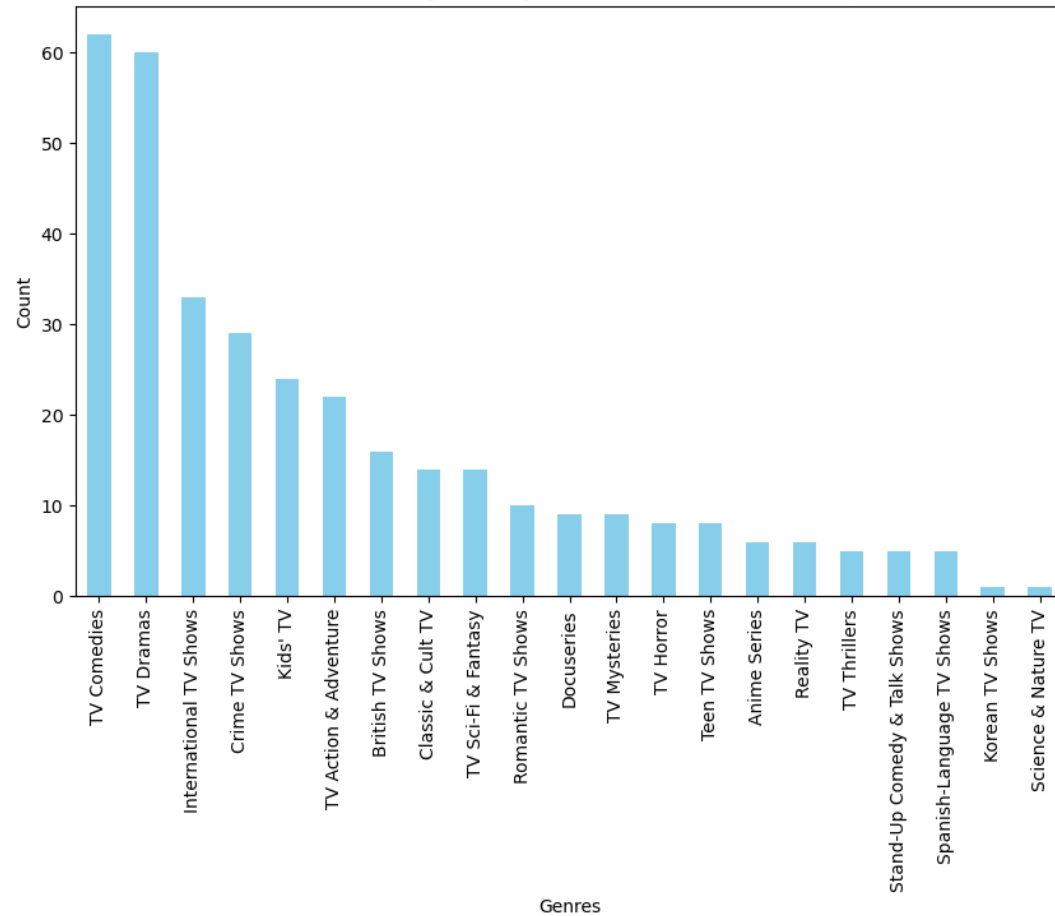


# Findings

Our findings suggest that "International TV Shows" is the most common genre among long-running TV shows, followed by "TV Dramas" and "TV Comedies. "

Lesser prevalent genres for long-running shows include "TV Action & Adventure," "Kids' TV," "Crime TV Shows," and "Reality TV." The other genres, like "Anime Series," "Docuseries," etc., have fewer long-running shows.

Genres of Long-Running TV Shows (5 Seasons or More)



# Limitations

The dataset only includes Netflix shows, so our conclusions may not apply to shows on other platforms.

The definition of 'long-running' as being five seasons or more is somewhat arbitrary. A different definition might yield different results.

Some shows may fall into multiple genres, which might affect the distribution of genres.

# Future Work to the Limitations

We can conduct a similar analysis with a different definition of 'long-running.'

We can compare the genre distribution of long-running shows on different streaming platforms.

We can explore the relationship between the number of seasons and other variables, such as viewer ratings or the country of origin.

# Conclusions

The results show that certain genres of TV shows will likely be long-runners; perhaps it can generate more topics and cause trends on social media. This can be a potential considering factor for TV show creators.

## Research Question 2

Can we use Scikit-Learn to create a predictive model that estimates the content rating of a Netflix movie or TV show based on its attributes such as 'type,' 'director,' 'cast,' 'country,' 'release\_year,' 'duration,' and 'listed\_in'?

# Methods

Approach: This can be treated as a multi-class classification problem to predict the exact rating or a binary classification problem to classify content as above or below a certain rating. If we convert the ratings into numerical scores, it could alternatively be approached as a regression problem.

The rows with missing 'rating' should neither be part of the training nor the test set during the model development and evaluation phase. Once the model is trained and evaluated, it can be used to predict the missing 'rating' values in the original dataset.

We trained a RandomForestClassifier to predict the 'rating' based on features like 'type,' 'director,' 'country,' 'date\_added,' 'release\_year,' 'duration,' and 'listed\_in.'

# Findings

The model achieved an accuracy of ~51% on the test set.

Classification Report:				
	precision	recall	f1-score	support
G	0.50	0.27	0.35	11
NR	0.00	0.00	0.00	11
PG	0.53	0.41	0.47	58
PG-13	0.46	0.35	0.39	84
R	0.58	0.46	0.51	154
TV-14	0.46	0.49	0.47	442
TV-G	0.33	0.07	0.11	45
TV-MA	0.58	0.72	0.64	662
TV-PG	0.30	0.15	0.20	168
TV-Y	0.67	0.62	0.64	65
TV-Y7	0.47	0.55	0.51	60
accuracy			0.52	1760
macro avg	0.44	0.37	0.39	1760
weighted avg	0.50	0.52	0.50	1760

Model Accuracy: 0.5244318181818182

Also, we used this model to fill in our original dataset's missing 'rating' values.

Rows with missing 'rating' before filling:

show_id	type	title \
5989	s5990 Movie	13TH: A Conversation with Oprah Winfrey & Ava ...
6827	s6828 TV Show	Gargantia on the Verdurous Planet
7312	s7313 TV Show	Little Lunch
7537	s7538 Movie	My Honor Was Loyalty

director	cast \
5989 Mark Ritchie	Oprah Winfrey, Ava DuVernay
6827 Unknown	Kaito Ishikawa, Hisako Kanemoto, Ai Kayano, Ka...
7312 Unknown	Flynn Curry, Olivia Deeble, Madison Lu, Oisín ...
7537 Alessandro Pepe	Leone Frisa, Paolo Vaccarino, Francesco Miglio...

country	date_added	release_year	rating	duration \
5989 United States	2017-01-26	2017	NaN	37
6827 Japan	2016-12-01	2013	NaN	1
7312 Australia	2018-02-01	2015	NaN	1
7537 Italy	2017-03-01	2015	NaN	115

listed_in \
5989 Movies
6827 Anime Series, International TV Shows
7312 Kids' TV, TV Comedies
7537 Dramas

description	year_added
5989 Oprah Winfrey sits down with director Ava DuVe...	2017
6827 After falling through a wormhole, a space-dwel...	2016
7312 Adopting a child's perspective, this show take...	2018
7537 Amid the chaos and horror of World War II, a c...	2017

Rows with missing 'rating' after filling:

show_id	type	title \
5989	s5990 Movie	13TH: A Conversation with Oprah Winfrey & Ava ...
6827	s6828 TV Show	Gargantia on the Verdurous Planet
7312	s7313 TV Show	Little Lunch
7537	s7538 Movie	My Honor Was Loyalty

director	country	date_added	release_year	rating \
5989 Mark Ritchie	United States	2017-01-26	2017	TV-MA
6827 Unknown	Japan	2016-12-01	2013	TV-14
7312 Unknown	Australia	2018-02-01	2015	TV-MA
7537 Alessandro Pepe	Italy	2017-03-01	2015	TV-MA

duration	listed_in \
5989 37	Movies
6827 1	Anime Series, International TV Shows
7312 1	Kids' TV, TV Comedies
7537 115	Dramas

description	year_added \
5989 Oprah Winfrey sits down with director Ava DuVe...	2017
6827 After falling through a wormhole, a space-dwel...	2016
7312 Adopting a child's perspective, this show take...	2018
7537 Amid the chaos and horror of World War II, a c...	2017

cast
5989 Oprah Winfrey, Ava DuVernay
6827 Kaito Ishikawa, Hisako Kanemoto, Ai Kayano, Ka...
7312 Flynn Curry, Olivia Deeble, Madison Lu, Oisín ...
7537 Leone Frisa, Paolo Vaccarino, Francesco Miglio...



MicroMasters®



UC San Diego

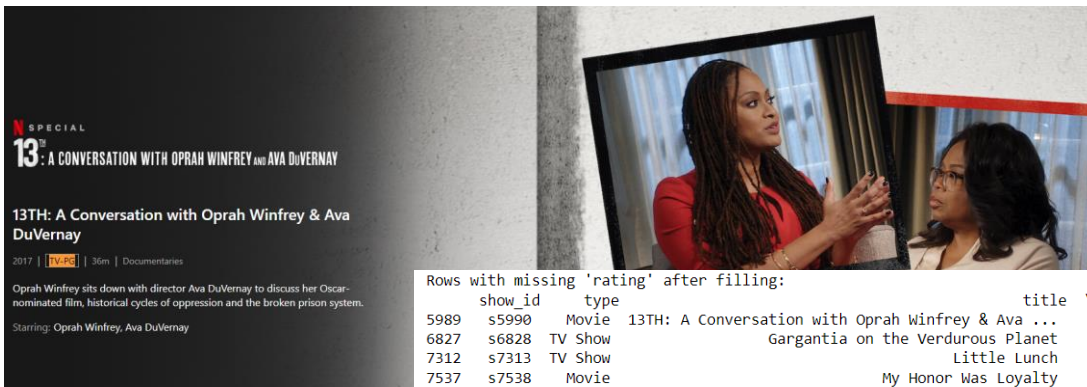
# Findings – How did the model perform?

13th: A Conversation with Oprah Winfrey & Ava DuVernay

The official rating: TV-PG

Our model rating: TV-MA

That's a miss.



Rows with missing 'rating' after filling:

	show_id	type	title \
5989	s5990	Movie	13TH: A Conversation with Oprah Winfrey & Ava ...
6827	s6828	TV Show	Gargantia on the Verdurous Planet
7312	s7313	TV Show	Little Lunch
7537	s7538	Movie	My Honor Was Loyalty

	director	country	date_added	release_year	rating \
5989	Mark Ritchie	United States	2017-01-26	2017	TV-MA
6827	Unknown	Japan	2016-12-01	2013	TV-14
7312	Unknown	Australia	2018-02-01	2015	TV-MA
7537	Alessandro Pepe	Italy	2017-03-01	2015	TV-MA

	duration	listed_in \
5989	37	Movies
6827	1	Anime Series, International TV Shows
7312	1	Kids' TV, TV Comedies
7537	115	Dramas

	description	year_added \
5989	Oprah Winfrey sits down with director Ava DuVe...	2017
6827	After falling through a wormhole, a space-dwel...	2016
7312	Adopting a child's perspective, this show take...	2018
7537	Amid the chaos and horror of World War II, a c...	2017

	cast
5989	Oprah Winfrey, Ava DuVernay
6827	Kaito Ishikawa, Hisako Kanemoto, Ai Kayano, Ka...
7312	Flynn Curry, Olivia Deeble, Madison Lu, Oisín ...
7537	Leone Frisa, Paolo Vaccarino, Francesco Miglio...



# Findings – How did the model perform?

13th: A Conversation with Oprah Winfrey & Ava DuVernay

The official rating: TV-14

Our model rating: TV-14

Great! That's a match!

The screenshot shows the IMDb page for the anime 'Gargantia on the Verdurous Planet'. The page includes the title, original title 'Suissei no Gargantia', and a TV-14 rating. It features a synopsis, a cast list with Michelle Ruff and Janice Kawaye, and a 'Watch on Hulu' button. The page also displays the IMDb rating of 7.4/10 and a 'Rate' button.

Rows with missing 'rating' after filling:					title \
	show_id	type			
5989	s5990	Movie	13TH: A Conversation with Oprah Winfrey & Ava ...		
6827	s6828	TV Show	Gargantia on the Verdurous Planet		
7312	s7313	TV Show	Little Lunch		
7537	s7538	Movie	My Honor Was Loyalty		
	director	country	date_added	release_year	rating \
5989	Mark Ritchie	United States	2017-01-26	2017	TV-MA
6827	Unknown	Japan	2016-12-01	2013	TV-14
7312	Unknown	Australia	2018-02-01	2015	TV-MA
7537	Alessandro Pepe	Italy	2017-03-01	2015	TV-MA
	duration		listed_in \		
5989	37		Movies		
6827	1	Anime Series, International TV Shows			
7312	1	Kids' TV, TV Comedies			
7537	115	Dramas			
		description	year_added		\
5989	Oprah Winfrey sits down with director Ava DuVe...		2017		
6827	After falling through a wormhole, a space-dwel...		2016		
7312	Adopting a child's perspective, this show take...		2018		
7537	Amid the chaos and horror of World War II, a c...		2017		
		cast			
5989	Oprah Winfrey, Ava DuVernay				
6827	Kaito Ishikawa, Hisako Kanemoto, Ai Kayano, Ka...				
7312	Flynn Curry, Olivia Deeble, Madison Lu, Oisín ...				
7537	Leone Frisa, Paolo Vaccarino, Francesco Miglio...				

# Findings – How did the model perform?

13th: A Conversation with Oprah Winfrey & Ava DuVernay

The official rating: TV-MA

Our model rating: TV-MA

Great! That's a match!

The screenshot shows the IMDb page for the TV series 'Little Lunch'. The page includes the title, genre (Comedy, Family), and a description: 'Short stories of what a few primary school friends get up to at little lunch, and in the classroom.' It also shows the stars Flynn Curry, Olivia Deeble, and Madison Lu. The IMDb rating is 8.4/10 with 566 votes. A yellow banner indicates it is available on Prime Video. The 'Episode guide' tab is selected, showing 28 episodes.

Awards 1 win & 3 nominations

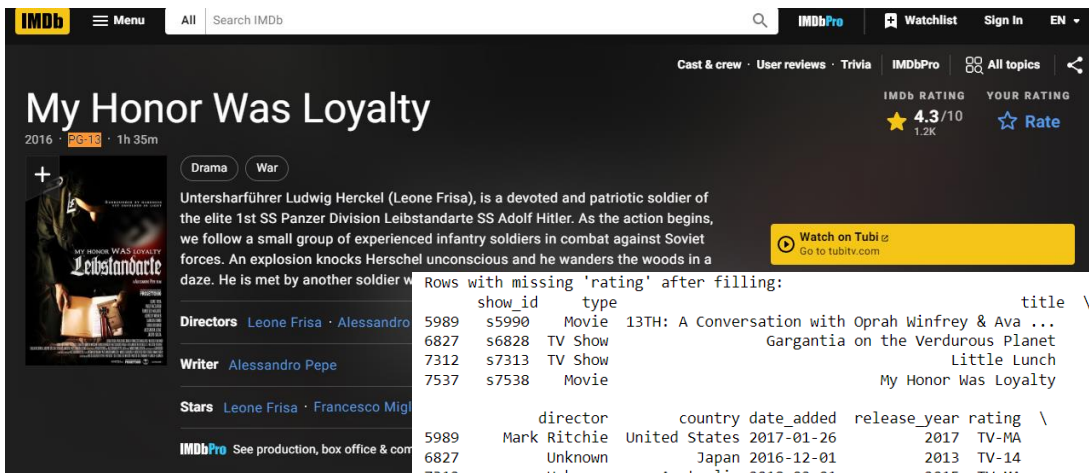
Rows with missing 'rating' after filling:							title \
	show_id	type					
5989	s5990	Movie	13TH: A Conversation with Oprah Winfrey & Ava ...				
6827	s6828	TV Show	Gargantia on the Verdurous Planet				
7312	s7313	TV Show	Little Lunch				
7537	s7538	Movie	My Honor Was Loyalty				
	director	country	date_added	release_year	rating \		
5989	Mark Ritchie	United States	2017-01-26	2017	TV-MA		
6827	Unknown	Japan	2016-12-01	2013	TV-14		
7312	Unknown	Australia	2018-02-01	2015	TV-MA		
7537	Alessandro Pepe	Italy	2017-03-01	2015	TV-MA		
	duration						listed_in \
5989	37						Movies
6827	1	Anime Series, International TV Shows					
7312	1	Kids' TV, TV Comedies					
7537	115						Dramas
						description	year_added \
5989	Oprah Winfrey sits down with director Ava DuVe...						2017
6827	After falling through a wormhole, a space-dwel...						2016
7312	Adopting a child's perspective, this show take...						2018
7537	Amid the chaos and horror of World War II, a c...						2017
	cast						
5989	Oprah Winfrey, Ava DuVernay						
6827	Kaito Ishikawa, Hisako Kanemoto, Ai Kayano, Ka...						
7312	Flynn Curry, Olivia Deeble, Madison Lu, Oisín ...						
7537	Leone Frisa, Paolo Vaccarino, Francesco Miglio...						

# Findings – How did the model perform?

13th: A Conversation with Oprah Winfrey & Ava DuVernay

The official rating: PG-13  
Our model rating: TV-MA

That's a miss.



Rows with missing 'rating' after filling:

show_id	type	title \
5989	s5990	Movie
6827	s6828	TV Show
7312	s7313	TV Show
7537	s7538	Movie

	director	country	date_added	release_year	rating \
5989	Mark Ritchie	United States	2017-01-26	2017	TV-MA
6827	Unknown	Japan	2016-12-01	2013	TV-14
7312	Unknown	Australia	2018-02-01	2015	TV-MA
7537	Alessandro Pepe	Italy	2017-03-01	2015	TV-MA

	duration	listed_in \
5989	37	Movies
6827	1	Anime Series, International TV Shows
7312	1	Kids' TV, TV Comedies
7537	115	Dramas

	description	year_added \
5989	Oprah Winfrey sits down with director Ava DuVe...	2017
6827	After falling through a wormhole, a space-dwel...	2016
7312	Adopting a child's perspective, this show take...	2018
7537	Amid the chaos and horror of World War II, a c...	2017

	cast
5989	Oprah Winfrey, Ava DuVernay
6827	Kaito Ishikawa, Hisako Kanemoto, Ai Kayano, Ka...
7312	Flynn Curry, Olivia Deeble, Madison Lu, Oisin ...
7537	Leone Frisa, Paolo Vaccarino, Francesco Miglio...

# Limitations

The mapping of 'cast' to 'director' and 'director' to 'country' might not be accurate for all entries, as one director or cast can work in different countries or with different casts.

The backfill method used for 'date\_added' assumes that the next value in the dataset is a good replacement for the missing value, which might not always be the case.

The model's performance is dependent on the chosen features and the RandomForestClassifier. Other models might yield different results.

Since we dropped the 'cast' column due to its complexity, some valuable information might have been lost.

# Future Work to the Limitations

The 'cast' column could be further analyzed and processed to include in the model.

Other models could be tried to see if they perform better.

Other feature engineering techniques could be applied to improve model performance.

More advanced techniques could be used for missing data imputation.

# Conclusions

The result shows that there can be a certain degree of correlation between content ratings and other attributes of a TV show or a movie, and we can infer this not just by using its description.

# Research Question 3

Can we identify the most frequent themes or topics in Netflix shows and movie descriptions using NLTK?

# Methods

Tokenization: Break down the description text into individual words or tokens.

Removing Stop Words: Remove common words that do not carry much meaning (known as "stop words").

Stemming and Lemmatization: Reduce words to their root form through stemming and lemmatization. These processes will convert words to their base form, considering the context for lemmatization.

Identifying themes or topics from text data is typically an unsupervised learning task because we don't have predefined labels or categories for the themes or topics. Instead, we want to discover these themes or topics from the data.

One common approach for this task is topic modeling, a statistical model for discovering the abstract "topics" that occur in a collection of documents. Latent Dirichlet Allocation (LDA) is an example of a topic model used to classify text in a document to a particular topic. Also, it is the model we will use in this project.



# Findings

Topic 0: This topic seems to be about high school life and teenage relationships, with words like 'school,' 'teen,' 'student,' 'friend,' 'high,' and 'girl.'

Topic 1: This topic could be about moving to a new place or returning home, with words like 'new,' 'friend,' 'home,' 'city,' and 'return.'

Topic 2: This topic seems to involve romantic and family relationships, with words like 'woman,' 'man,' 'love,' 'family,' 'fall,' and 'father.'

Topic 3: This topic appears to revolve around crime, law enforcement, and investigation, with words like 'murder,' 'crime,' 'police,' 'criminal,' 'cop,' and 'detective.'

Top 0 words for topic #0:  
["'s", 'young', 'family', 'woman', 'man', 'father', 'home', 'son', 'mother', 'daughter']

Top 1 words for topic #1:  
['-', 'friend', 'new', 'find', 'two', 'life', 'girl', 'boy', 'one', 'three']

Top 2 words for topic #2:  
['documentary', '...', '...', 'series', "'s", 'life', 'follows', 'film', 'history', 'explores']

Top 3 words for topic #3:  
["'s", 'crime', 'cop', 'help', 'criminal', 'detective', 'new', 'killer', 'police', 'case']

Top 4 words for topic #4:  
['story', 'true', 'life', 'love', 'special', 'stand-up', 'drama', 'comedy', 'tale', 'career']

Top 5 words for topic #5:  
['team', 'world', "'s", 'kid', 'adventure', 'game', 'save', 'rescue', 'friend', 'travel']

Top 6 words for topic #6:  
['school', 'student', 'high', 'college', "'s", 'teen', 'girlfriend', 'run', 'rise', 'band']

Top 7 words for topic #7:  
['war', 'world', 'battle', 'fight', 'force', 'evil', 'earth', 'take', 'power', 'art']



MicroMasters®



UC San Diego

# Findings

Topic 4: This topic seems to concern film and television production, with words like 'documentary,' 'series,' 'film,' 'star,' 'story,' and 'comedy.'

Topic 5: This topic could be about war, conflict, and power struggles, with words like 'war,' 'world,' 'fight,' 'power,' 'battle,' and 'earth.'

Topic 6: This topic might involve terrorism and marital relationships, with words like 'group,' 'couple,' 'attempt,' 'terrorist,' or 'married.'

Topic 7: This topic seems to be about gangs, drugs, and artists, with words like 'drug,' 'world,' 'gang,' 'artist,' and 'brother.'

Top 0 words for topic #0:  
["'s", 'young', 'family', 'woman', 'man', 'father', 'home', 'son', 'mother', 'daughter']

Top 1 words for topic #1:  
['-', 'friend', 'new', 'find', 'two', 'life', 'girl', 'boy', 'one', 'three']

Top 2 words for topic #2:  
['documentary', '``', '""', 'series', "'s", 'life', 'follows', 'film', 'history', 'explores']

Top 3 words for topic #3:  
["'s", 'crime', 'cop', 'help', 'criminal', 'detective', 'new', 'killer', 'police', 'case']

Top 4 words for topic #4:  
['story', 'true', 'life', 'love', 'special', 'stand-up', 'drama', 'comedy', 'tale', 'career']

Top 5 words for topic #5:  
['team', 'world', "'s", 'kid', 'adventure', 'game', 'save', 'rescue', 'friend', 'travel']

Top 6 words for topic #6:  
['school', 'student', 'high', 'college', "'s", 'teen', 'girlfriend', 'run', 'rise', 'band']

Top 7 words for topic #7:  
['war', 'world', 'battle', 'fight', 'force', 'evil', 'earth', 'take', 'power', 'art']

# Limitations

Our analysis is based on the words used in the descriptions, which might not fully capture the themes of the shows and movies.

Additionally, the LDA model is an unsupervised method, so the topics it identifies are based on patterns in word usage and not on the actual content or context of the shows and movies.

# Future Work to the Limitations

In the future, we could improve our topic modeling using more sophisticated methods, such as neural topic models.

We could also incorporate other information from the dataset, such as the genre or the director, to provide more context for the topics.

Additionally, we could try to validate the topics by comparing them with human-annotated topics or with external data sources.

# Conclusions

We used topic modeling to identify the most frequent themes in Netflix shows and movie descriptions. Our findings provide insights into the types of content available on Netflix and could be useful for content recommendation or understanding trends in the streaming industry.

# Acknowledgements

Data Source: the Netflix Movies and TV Shows from Kaggle  
(<https://www.kaggle.com/datasets/shivamb/netflix-shows>)

I would like to thank the Kaggle Community Notebook “Netflix Data: Cleaning, Analysis and Visualization” (<https://www.kaggle.com/datasets/ariyoomotade/netflix-data-cleaning-analysis-and-visualization>) for inspiring me on how to clean the data. (The original cleaning was carried out using PostgreSQL)

I would like to thank Dr. Leo Porter, Dr. Ilkay Altintas, and other staff working behind the scenes at UC San Diego for delivering this fantastic content for DSE200x.

Also, I would like to thank my peers from DSE200x at UCSanDiegoX for providing valuable feedback when evaluating my project.

# References

All work was performed independently for this analysis, using standard Python libraries for data analysis, namely pandas, sklearn, nltk, string, and gensim.

The dataset was sourced from Kaggle, a popular platform for data science competitions and datasets. No other external references or research papers were used.

The analysis was guided by my training and the information present in the dataset itself.