أكاديمية سدايا
SDAIA Academy

Movie Plots

Movie Recommender System

Genre Classification

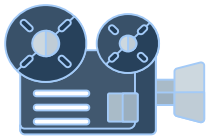
Plot Topic Modeling

Prepared by:
Sarah Alabdulwahab & Asma Althakafi

OBJECTIVE

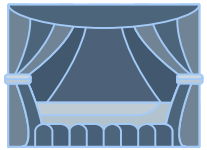
- **Find the degree of similarity between movie plots**
- **Classify the movie genres based on their plots**
- **Apply topic modeling on the plot of each movie to differentiate between them based on their genre**

DATA COLLECTION



The Movie Dataset

The Movie Database API



IMDB Web Scraping

Initially, the dataset contained **45,466 movies** and **24 features**

PRE-PROCESSING & CLEANING

1
English Movies

2
100+ Words per Plot

3
Fill Null Values

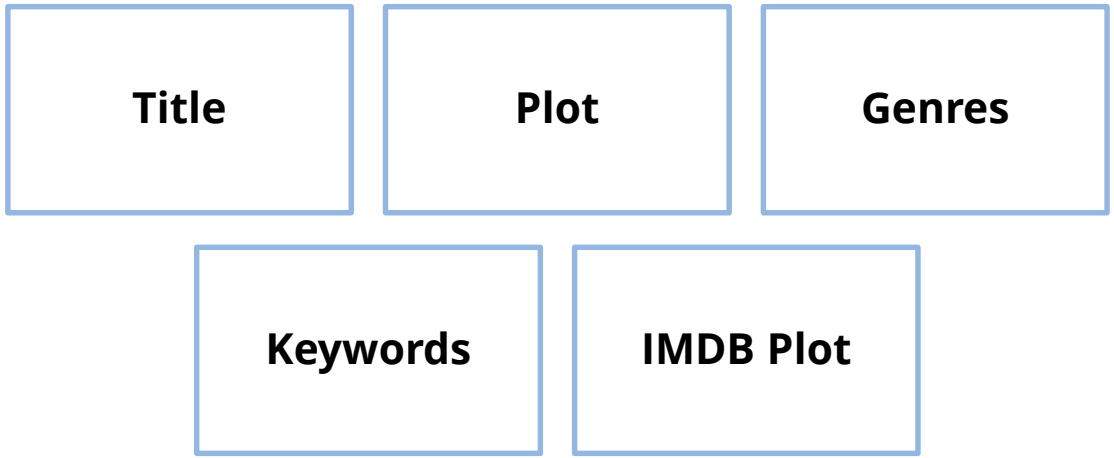
4
Merge Keywords with Plots

5
Named Entity Recognition

6
Stemming & Lemmatization

DATASET

The dataset contains **3,133 movies** and **5 features**



DATASET

The dataset contains **3,133 movies** and **4 features**

Title	**Plot**
Clean Plot	**Genres**

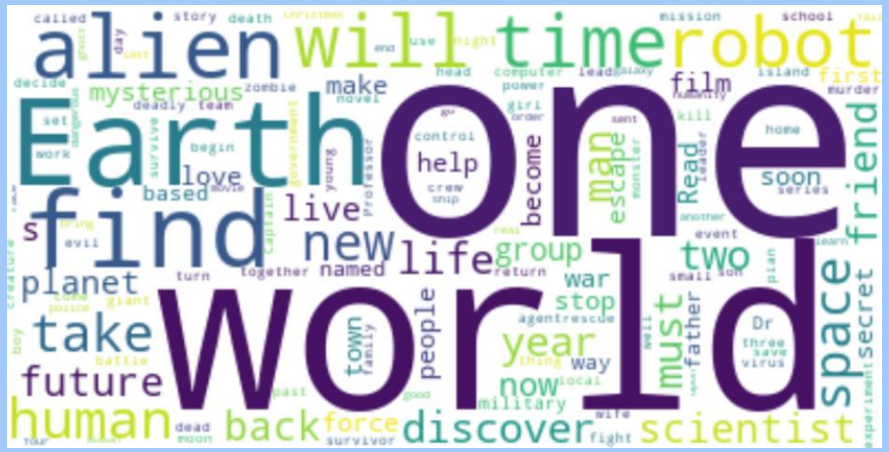
WORD CLOUD

ACTION

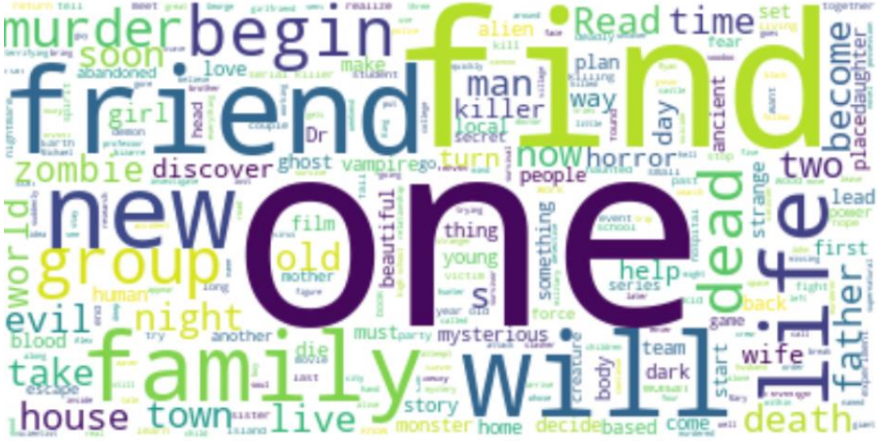
DRAMA

WORD CLOUD

SCI-FI



HORROR



METHODS

Recommender System

Classification

Topic Modeling

EXPERIMENTS

Count Vectorizer

- Original Plots
- Clean Plots

TF-IDF

- Original Plots
- Clean Plots

CONTENT-BASED RECOMMENDER SYSTEM

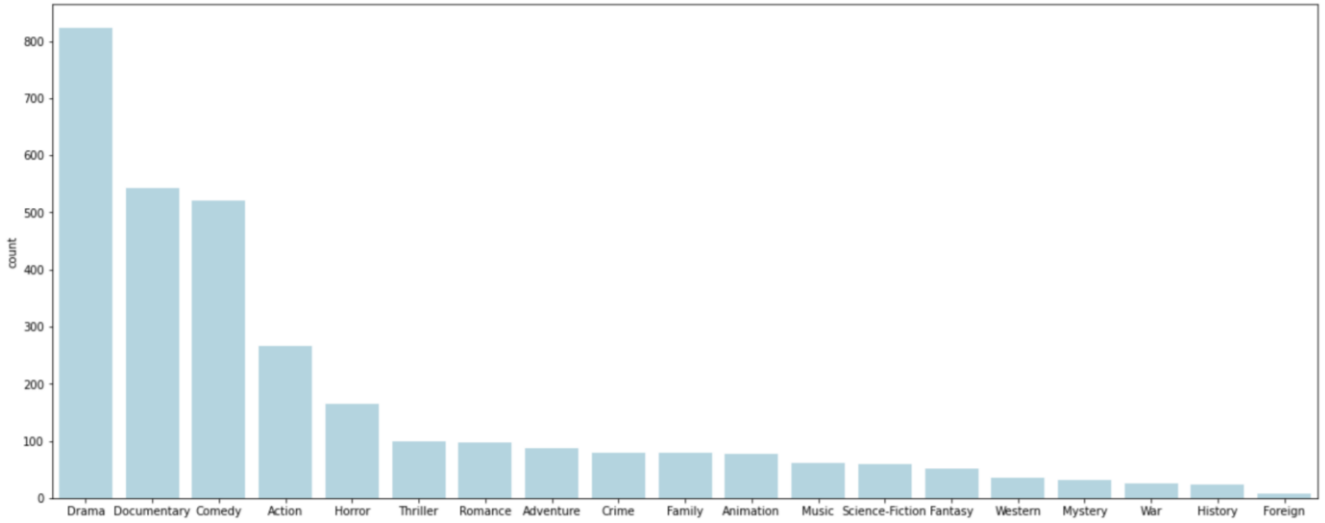
The best result: **clean plot with CountVectorizer**

The movie you chose is Twelve Monkeys

- Genres: **Science-Fiction, Thriller, Mystery**

Title	Similarity Score	Genres
Carriers	26%	Action, Drama, Horror, **Science-Fiction, Thriller**
Solos	23%	Horror, **Thriller, Science-Fiction**, Foreign
Day of the Dead 2: Contagium	22%	Horror, **Science-Fiction**

Classification using Multinomial Naive Bayes



Multilabel Classification

- With All Genres
- With Most Common Genres

Single Label Classification

- With All Genres
- With Most Common Genres

MULTILABEL CLASSIFICATION

The best result: **clean plot with CountVectorizer**

	Accuracy	**F1 Score**
All Genres	22.6%	46.2%
Most Common Genres	30.9%	55.4%

SINGLE LABEL CLASSIFICATION

The best result: **Original plot with CountVectorizer**

	Accuracy	**F1 Score**
All Genres	48.1%	40.5%
Most Common Genres	48.8%	43.3%

TOPIC MODELING

LATENT SEMANTIC ANALYSIS (LSA)

LATENT DIRICHLET ALLOCATION (LDA)

19 Genres -> 19 Topics

LATENT SEMANTIC ANALYSIS (LSA)

The best result: **Original plot with CountVectorizer**

Examples of Result:

Topic 3
war, world, ii, army, men, british, group, german, based, story

Topic 4
life, new, world, york, city, lives, journey, love, work, husband

LATENT DIRICHLET ALLOCATION (LDA)

The best result: **Original plot with CountVectorizer**

Now we will see the visualization

CONCLUSION

- **Unfortunately, results were not satisfying**
- **Need more data to add to the plot to improve the results**
- **Future work: Try BERT**

THANK YOU