Natural Language Processing

Abstract

In this NLP project, we explore and analyze a dataset obtained from Kaggle, specifically the "Netflix Titles" dataset. Our analysis encompasses three main modes: Exploratory Data Analysis (EDA), Sentiment Analysis, and Content Similarity.

Introduction

Netflix, being a global streaming platform, offers a diverse range of movies and TV shows. Understanding the distribution of content types, genres, release trends, and user sentiments can provide valuable insights for both the audience and content creators. By leveraging NLP techniques, we aim to uncover patterns and sentiments within the dataset.

Data

The dataset comprises information about Netflix titles, including details such as type (Movie or TV Show), title, director, cast, country, date added, release year, rating, duration, listed genres, and description. The dataset contains 8,807 entries.

Methods

We employ various methods to analyze the dataset:

Exploratory Data Analysis (EDA):

We start by exploring the basic characteristics of the dataset, including data types, missing values, and summary statistics. EDA involves visualizations such as type distribution, country distribution, release year trends, rating distribution, genre distribution, and popular genres.

Sentiment Analysis:

Utilizing Natural Language Processing (NLP) techniques, we perform sentiment analysis on the descriptions of the titles. We use the TextBlob library to calculate sentiment polarity scores, providing an understanding of the emotional tone conveyed in the descriptions.

Content Similarity:

To uncover content similarity, we use TF-IDF vectorization and cosine similarity. This approach allows us to identify how similar descriptions are to each other, aiding in content recommendation systems.

Results

Here are key findings from each mode:

Exploratory Data Analysis:

The dataset contains 6,131 movies and 2,676 TV shows.

United States, India, and the United Kingdom have the highest number of titles.

Release trends show an overall increase, with 2018 having the most titles.

TV-MA is the most common rating.

Sentiment Analysis:

Sentiment analysis reveals the emotional tone of descriptions. Descriptions tend to have a neutral to slightly positive sentiment.

Content Similarity:

Content similarity using cosine similarity provides a measure of similarity between descriptions.

Discussion

Interpreting these results involves considering the context of the streaming industry. The distribution of types, popular genres, and sentiment analysis can guide content creators in tailoring their productions. Content similarity can enhance recommendation algorithms, improving user experience.

Conclusion

This comprehensive analysis of the Netflix Titles dataset provides valuable insights into the streaming platform's content landscape. The findings contribute to understanding user preferences, content trends, and potential areas for future exploration. Further research and analysis can enhance the understanding of dynamic user preferences and content consumption patterns.