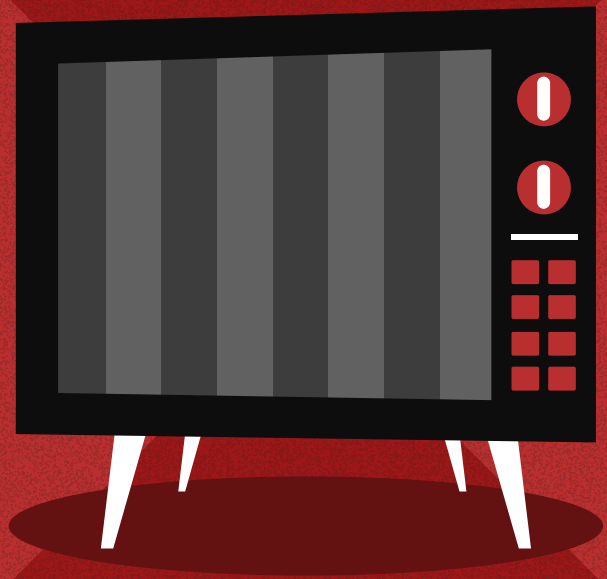
NETFLIX

RECOMMENDATION SYSTEM

Hijir Della Wirasti

Machine Learning
Unsupervised Learning: K-Means & Naive Bayes
Supervised Learning: KNN, Decision Tree, Logistic Regression, Random Forest

https://www.linkedin.com/in/hijirdella/ | https://github.com/hijirdella/Netflix-Recommendation-System



NETFLIX RECOMMENDATION SYSTEM

Dirancang untuk memberikan rekomendasi film yang dipersonalisasi berdasarkan preferensi pengguna dan atribut film. Dengan memanfaatkan berbagai model dan algoritma machine learning, sistem ini mengidentifikasi pola dalam data dan merekomendasikan film yang paling sesuai dengan preferensi pengguna.

Google Collab (Code):
https://colab.research.google.com/drive/1gzyUoEH4ii9060guBLO_2iVrjH9DCVoq?usp=sharing

TABLE OF CONTENTS

01

INTRODUCTION

About the Project, Bagaimana Sistem Bekerja, Penggunaan

ABOUT THE PROJECT

Fitur Utama

1. **Algoritma Rekomendasi**:
   - Mengintegrasikan model seperti **K-Nearest Neighbors (KNN)**, **Decision Trees**, **Logistic Regression**, **Random Forest**, dan **Naive Bayes** untuk rekomendasi film.
   - Menggunakan **K-Means Clustering** untuk mengelompokkan film ke dalam klaster yang serupa guna meningkatkan kualitas rekomendasi.
2. **Content-Based Filtering**:
   - Memanfaatkan atribut film seperti **director**, **cast**, **genre**, dan **rating** untuk menghitung kesamaan antar film menggunakan pendekatan **cosine similarity** dan **bag-of-words**.
3. **Pembelajaran Terawasi dan Tak Terawasi**:
   - Menggabungkan **Unsupervised Learning** (K-Means Clustering) dan **Supervised Learning**(model klasifikasi) untuk menyediakan strategi rekomendasi yang komprehensif.
4. **Evaluasi Model**:
   - Mengukur performa model menggunakan **Accuracy** untuk model klasifikasi dan **Davies-Bouldin Index** untuk kualitas klaster.

BAGAIMANA SISTEM BEKERJA

1. Persiapan Data:

Menangani nilai yang hilang dan mengubah data kategorikal menggunakan teknik seperti Label Encoding dan CountVectorizer.

2. Perhitungan Kesamaan:

Menghasilkan matriks kesamaan menggunakan cosine similarity untuk menemukan film dengan fitur yang serupa.

3. Proses Rekomendasi:

Memberikan rekomendasi untuk judul film tertentu dengan menemukan film dengan skor kesamaan tertinggi atau dari klaster yang sama.

4. Interaksi Pengguna:

Pengguna memasukkan judul film, dan sistem akan menyarankan film-film serupa berdasarkan model yang dipilih.

PENGGUNAAN

A

PLATFORM HIBURAN

Untuk merekomendasikan film atau acara TV kepada pengguna berdasarkan riwayat tontonan mereka

B

LAYANAN STREAMING

Memberikan saran yang dipersonalisasi untuk meningkatkan keterlibatan pengguna

C

SEGMENTASI PASAR

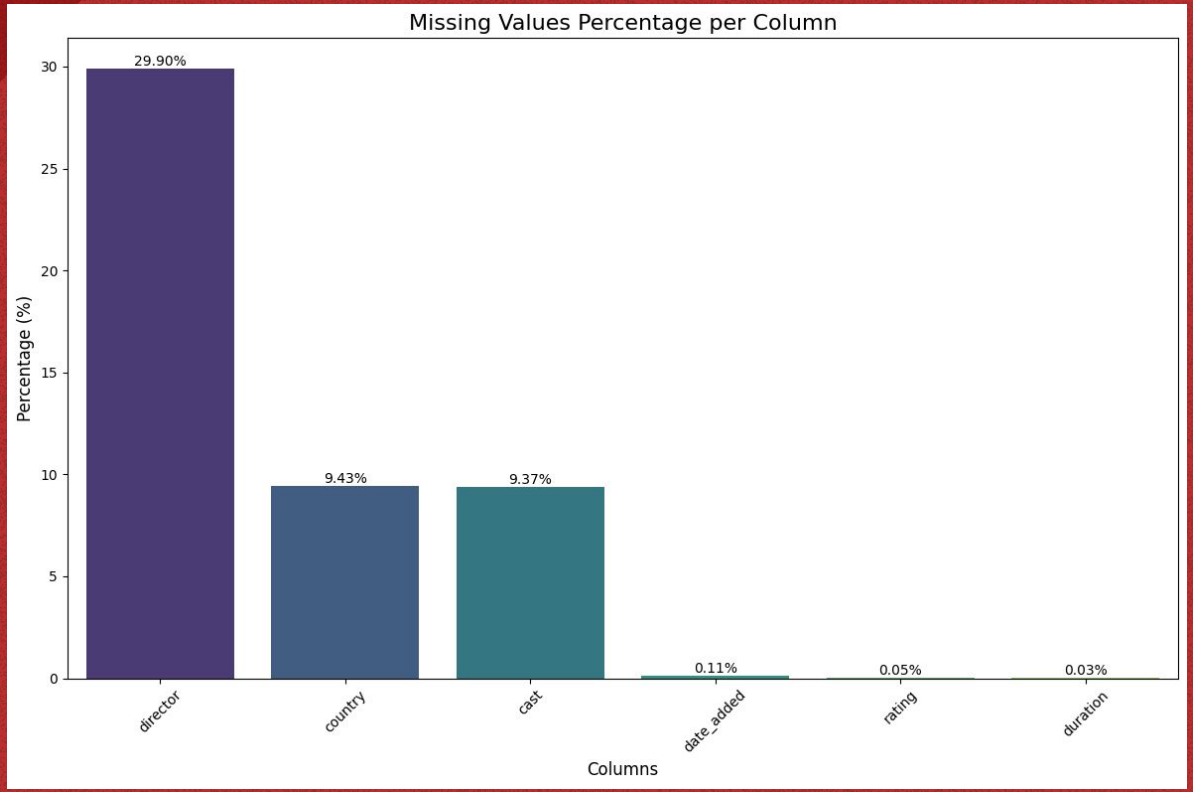
Mengidentifikasi pola preferensi pengguna dan menargetkan konten tertentu kepada segmen pengguna yang berbeda

02

DATA PREPROCESSING

Handling missing values, Renaming Rating Column, Bag Of Words

DATA PREPROCESSING



Handling Missing Values

Kolom **director**, **cast**, **country**, **date_added**, **rating**, dan **duration** memiliki nilai kosong. Kita dapat mengisi nilai kosong tersebut dengan **UnKnown**. Untuk kolom **rating**, kita bisa melakukan pencarian lebih lanjut untuk menemukan nilai yang benar. Selain itu, kolom **duration** dan **date_added** dianggap tidak diperlukan, sehingga dapat dihapus dari dataset.

Mengganti nama pada kolom rating agar mudah dipahami. Conoh: `TV-MA': 'Adults'.`

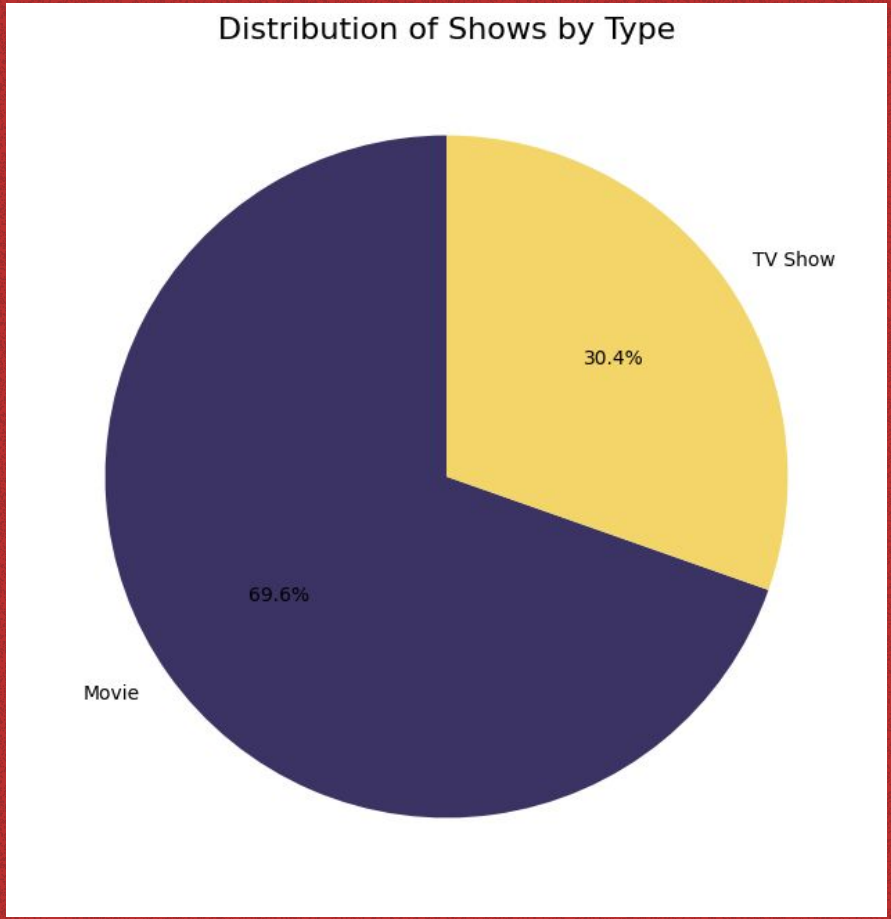
`Pendekatan Bag of Words digunakan untuk merepresentasikan data teks agar dapat dihitung, memungkinkan perbandingan kesamaan antar film menggunakan cosine similarity dan meningkatkan akurasi rekomendasi.`

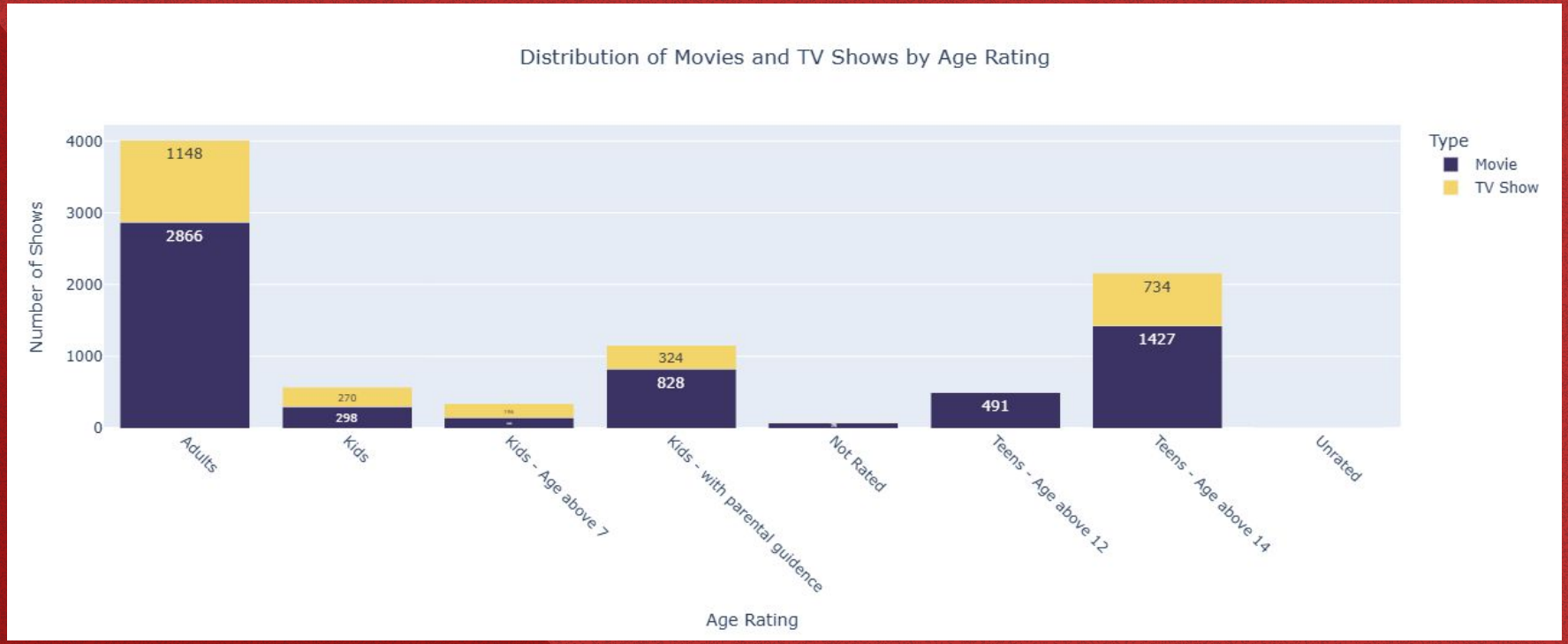
03

(EDA)
EXPLORATORY DATA ANALYSIS

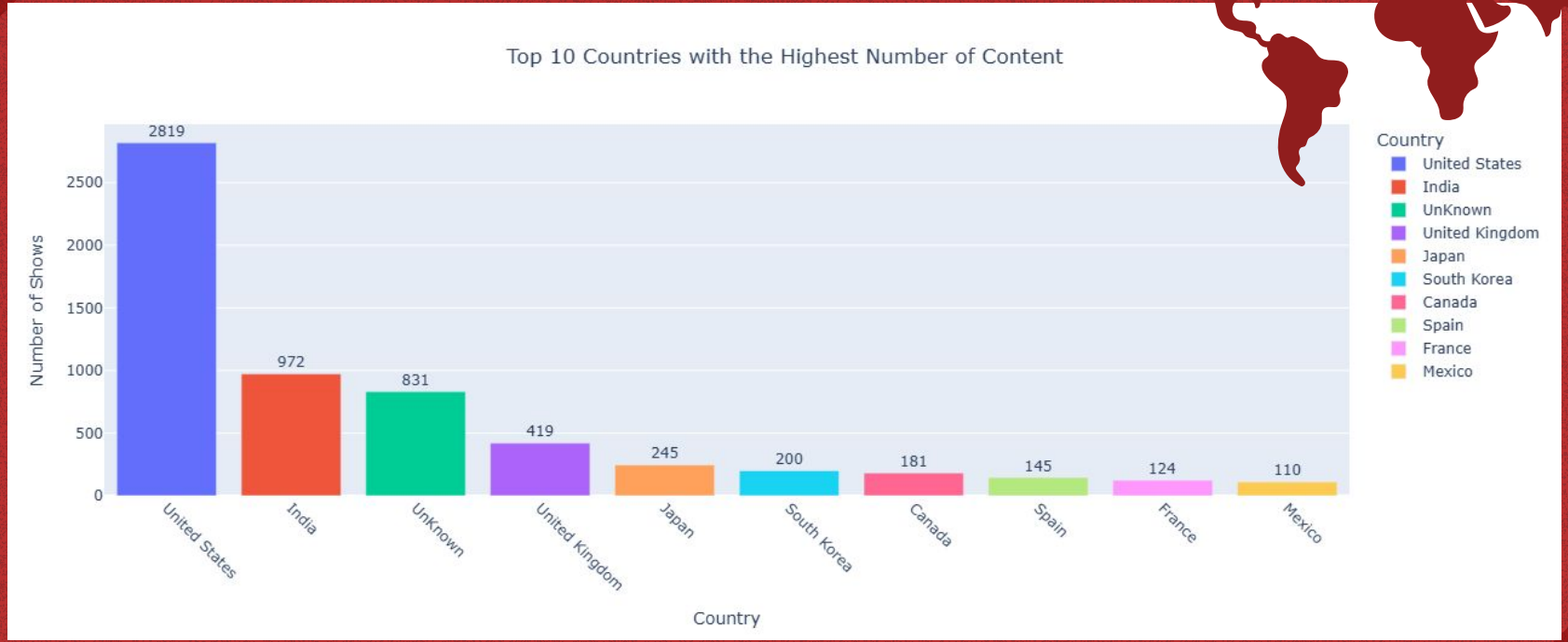
Dapat dilihat bahwa kategori **"Movie"** mencakup **69.6%** dari total konten, sementara kategori **"TV Show"** mencakup **30.4%.**

Hal ini menunjukkan bahwa Netflix memiliki lebih banyak film dibandingkan acara TV. Sebagian besar konten di Netflix terdiri dari film, sedangkan acara TV hanya merupakan bagian kecil dari keseluruhan konten, seperti yang ditunjukkan pada diagram lingkaran.

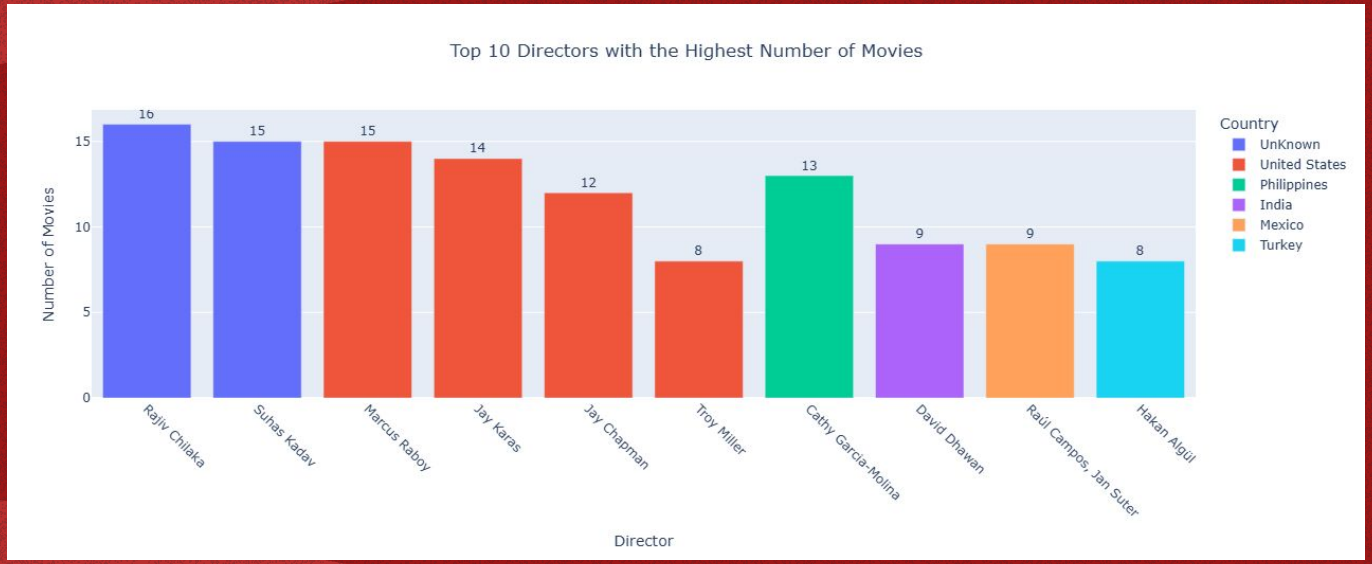




Sebagian besar orang dari semua rentang usia lebih memilih untuk menonton Movie dibandingkan dengan TV Show.

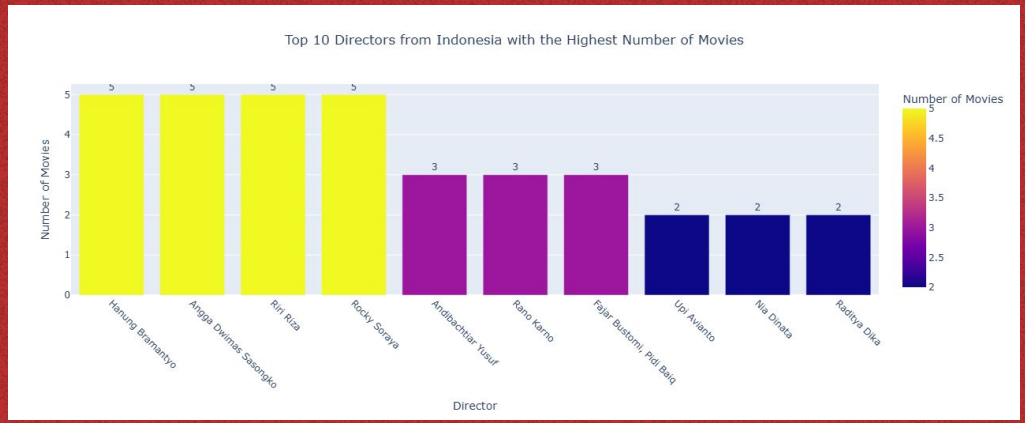


Amerika Serikat memproduksi sekitar **2819 konten**, yang merupakan jumlah yang jauh lebih besar dibandingkan dengan konten film yang diproduksi oleh negara lain.

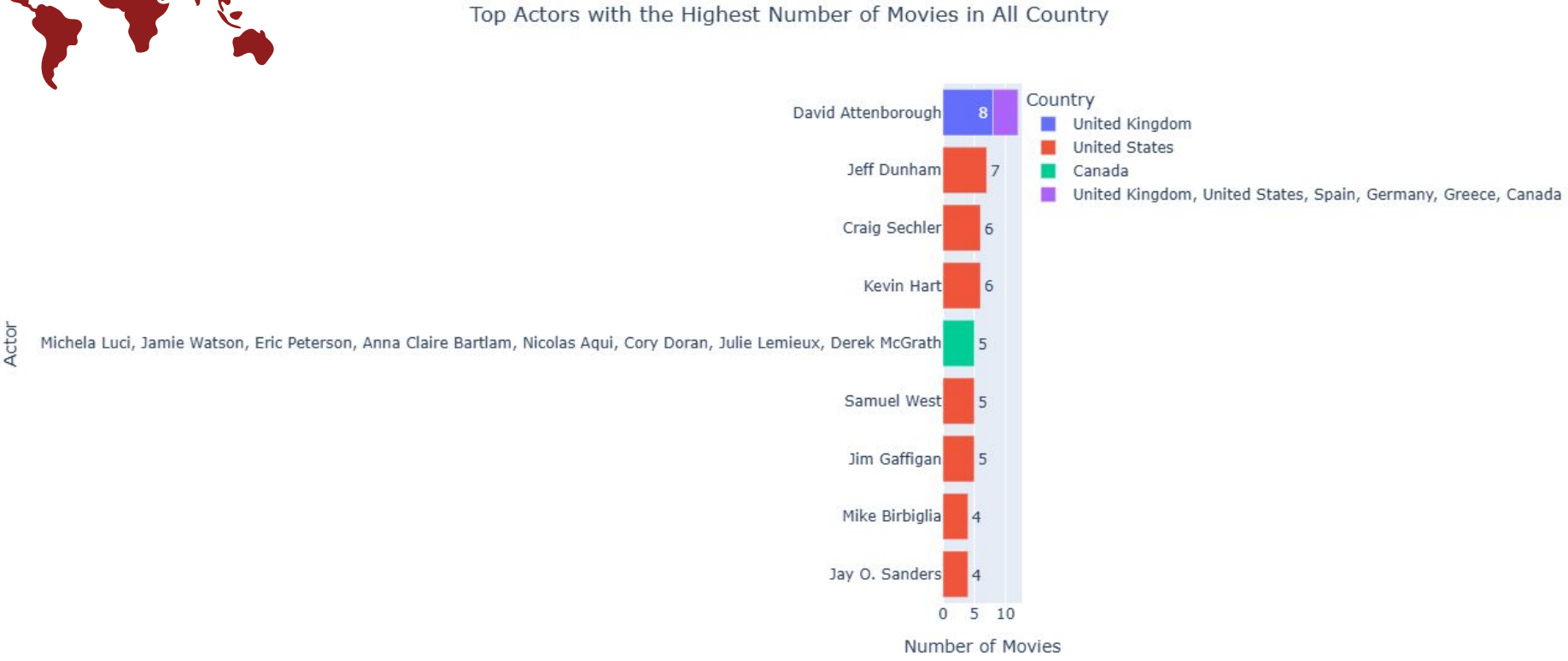


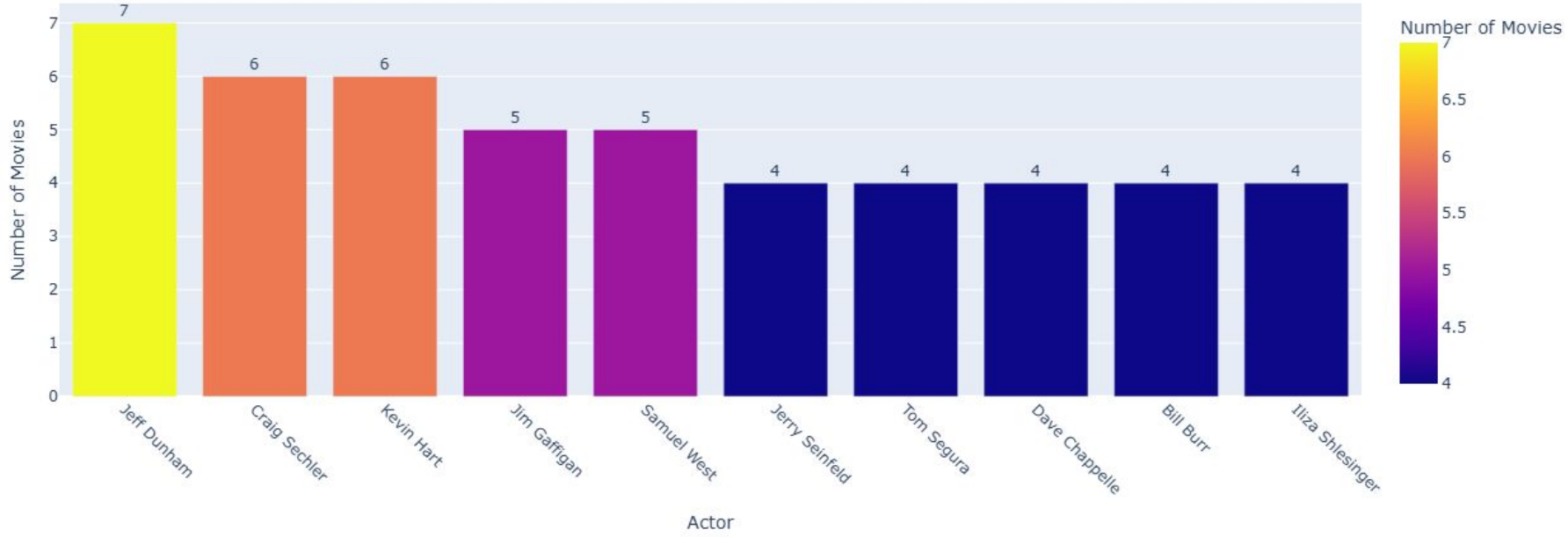
Rajiv Chilaka telah memproduksi konten film terbanyak di seluruh dunia.

Marcus Raboy telah memproduksi konten film terbanyak dari US

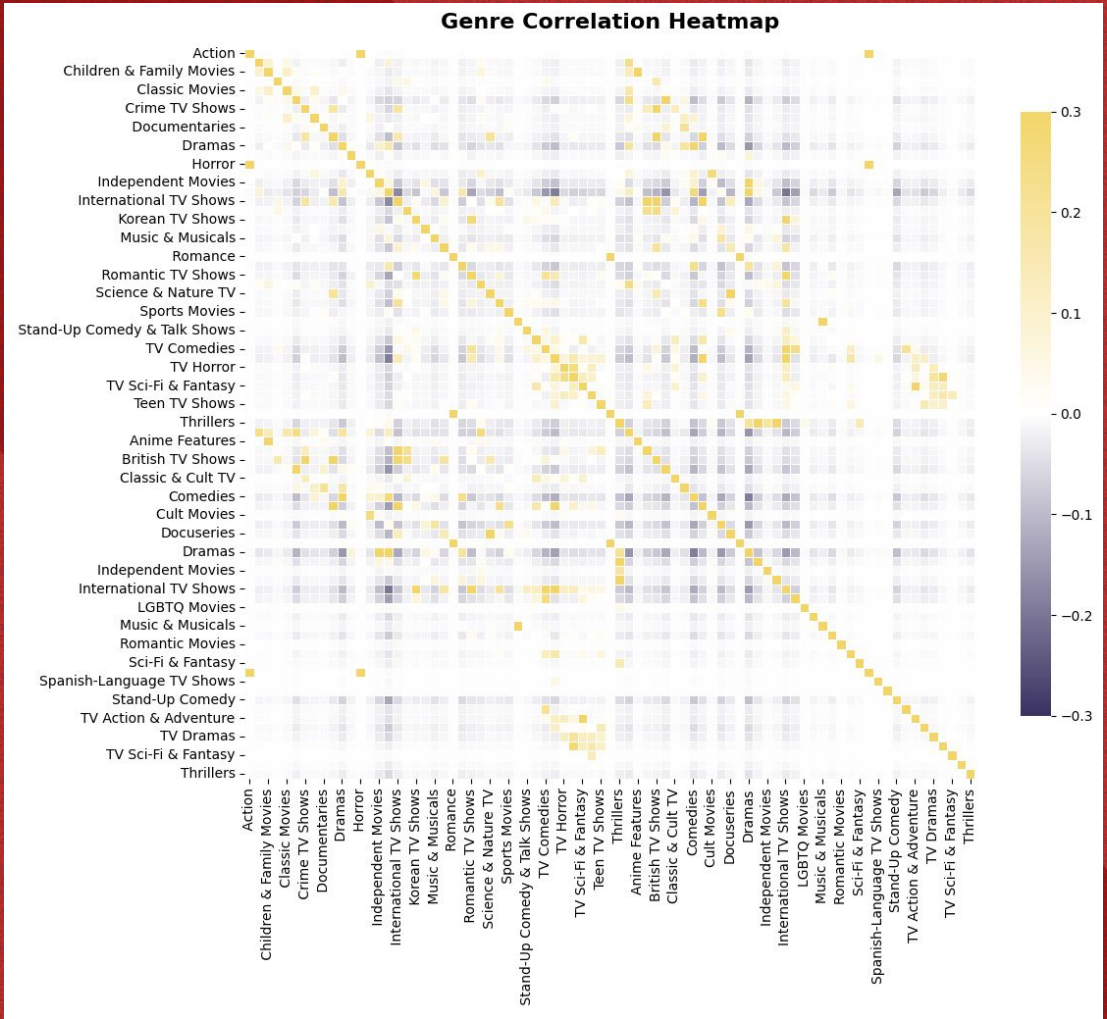


Hanung Bramantyo, Angga Dwimas Sasongko, Riri Riza, Rocky Soraya telah memproduksi konten film terbanyak dari Indonesia


Top Actors with the Highest Number of Movies in All Country
David Attenborough
Jeff Dunham
Craig Sechler
Kevin Hart
Michela Luci, Jamie Watson, Eric Peterson, Anna Claire Bartlam, Nicolas Aqui, Cory Doran, Julie Lemieux, Derek McGrath
Samuel West
Jim Gaffigan
Mike Birbiglia
Jay O. Sanders
8
7
6
6
5
5
5
4
4
Actor
0
5
10
Number of Movies
Country
United Kingdom
United States
Canada
United Kingdom, United States, Spain, Germany, Greece, Canada

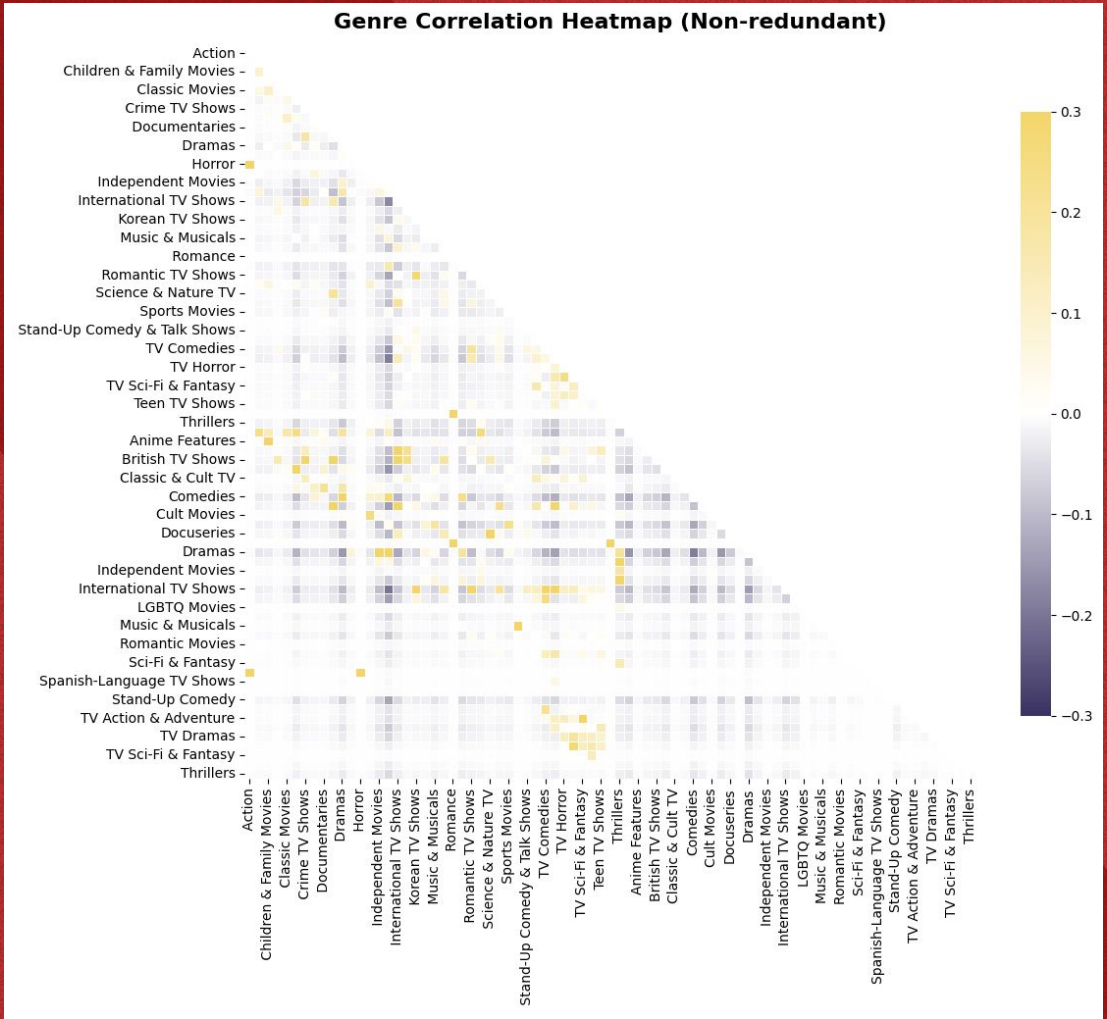

Top 10 Actors from the U.S. with the Highest Number of Movies
Number of Movies
Actor
Jeff Dunham
Craig Sechler
Kevin Hart
Jim Gaffigan
Samuel West
Jerry Seinfeld
Tom Segura
Dave Chappelle
Bill Burr
Iliza Shlesinger
7
6
6
5
5
4
4
4
4
4
0
1
2
3
4
5
6
7
Number of Movies
7
6.5
6
5.5
5
4.5
4


Top Actors from Indonesia with the Highest Number of Movies
Luna Maya, Herjunot Ali, Teuku Rifnu Wikana, Verdi Solaiman, Alex Abbad
Maudy Koesnaedi, Rano Karno, Cornelia Agatha, Mandra Naih, Aminah Tjendrakasih, Suty Karno, Adam Jagwani
Morgan Oey, Sheila Dara, Luna Maya, Kevin Julio, Anggika Bolsterli, Reza Nangin, Martin Anugrah, Fatih Unru
no Abraham, Anggika Bolsterli, Laura Theux, Christine Hakim, Unique Priscilla, Karina Suwandi, Mathias Muchus, Ayu Laksmi, I Made Sidia, Dominique Sanda, Bryan Domani, Wafda Saifan, Yayu Unru, Dave Hendrik
Nicholas Saputra, Sita Nursanti, Lukman Sardi, Indra Birowo, Wulan Guritno, Thomas Nawilis, Tutie Kirana, Robby Tumewu, Jonathan Mulia
Panji Zoni, Yayan Ruhian, Maizura, Cemal Farukh, Imelda Therinne, Surya Saputra, Jarot Superdj, Doyok Superdj, Hajra Romessa, Annette Edoarda
Prilly Latuconsina, Sandrinna Michelle, Shareefa Daanish, Indra Brotolaras, Kinaryosih, Inggrid Widjanarko, Asha Kenyeri Bermudez, Wesley Andrew, Kevin Bzezovski Taroreh, Gamaharitz
Prisia Nasution, Nyungsang Bungo, Rukman Rosadi, Nadhira Suryadi, Ines Somellera
Maudy Koesnaedi, Chicco Jerikho, Tutie Kirana, Olga Lydia, Joko Anwar, Nathania Angela Widjaja, Sendy Febrina, Suhartini, Atie Krisnawati, Ratnawati Sutandar, Ninik Winarni Utomo
Acha Septriasa, Ryan Delon, Makayla Rose Hilli, TJ
Actor
0
0.5
1
Number of Movies
Number of Movies
1.4
1.2
1
0.8
0.6



Dari heatmap genre correlation di atas, beberapa insight yang dapat diambil adalah:

1. **Hubungan Genre Positif**:
   - Genre seperti **Action** dan **TV Action & Adventure** memiliki korelasi positif yang cukup tinggi, menunjukkan bahwa konten dalam genre ini sering kali memiliki kesamaan elemen atau target audiens yang serupa.
   - Genre **Romantic Movies** cenderung berkorelasi dengan **Romantic TV Shows**, menandakan adanya preferensi lintas platform (film dan acara TV) untuk tema romantis.
2. **Genre dengan Korelasi Rendah**:
   - Genre seperti **Crime TV Shows** dan **Music & Musicals** menunjukkan korelasi yang rendah atau negatif, yang berarti konten dalam kategori ini cenderung berbeda secara signifikan dalam hal tema atau elemen.
   - Genre seperti **Horror** dan **Documentaries** juga tampak memiliki sedikit hubungan, menunjukkan bahwa mereka melayani audiens dengan minat yang sangat berbeda.



3. Pola Genre Khusus:

- Genre khusus seperti **Children & Family Movies** menunjukkan hubungan erat dengan genre **TV Shows for Kids**, mencerminkan konten yang sering kali ditujukan untuk anak-anak dan keluarga dalam berbagai format.

4. Cluster Genre:

- Kita bisa melihat ada beberapa cluster genre yang secara konsisten memiliki korelasi tinggi di antaranya, misalnya **Sci-Fi & Fantasy**, **TV Sci-Fi & Fantasy**, dan **Teen TV Shows**. Hal ini mengindikasikan bahwa genre ini memiliki elemen cerita yang sering tumpang tindih.

5. Strategi Peningkatan Konten:

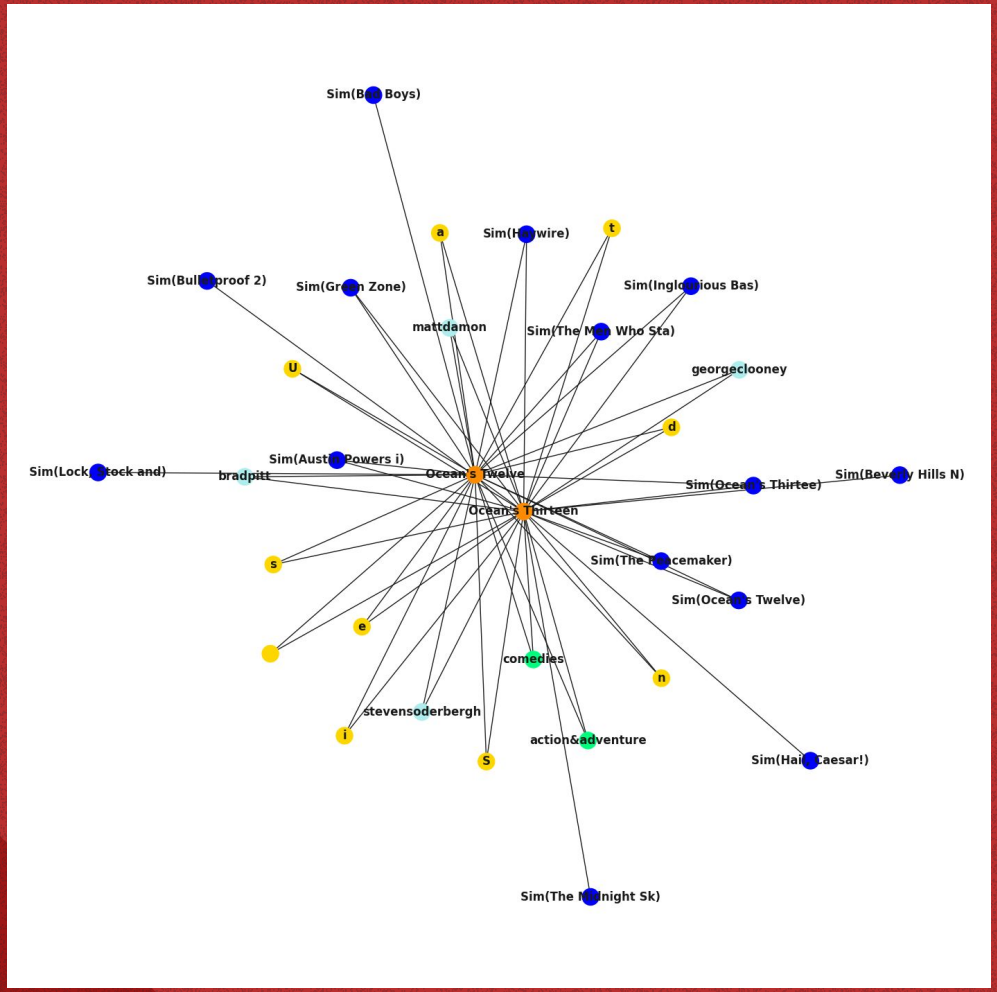
- Netflix dapat memanfaatkan pola ini untuk merekomendasikan konten lintas genre yang berkorelasi tinggi, seperti merekomendasikan **TV Action & Adventure** kepada pengguna yang menikmati **Action Movies**.

04

RECOMMENDATION SYSTEM

RECOMMENDATION CLUSTERING

Proyek ini, saya menggunakan **sistem rekomendasi berbasis konten** dengan **metode clustering**. Konten dari film (seperti pemeran, deskripsi, sutradara, genre, dll.) dianalisis untuk menentukan kesamaannya dengan film lainnya. Berdasarkan kesamaan tersebut, film yang paling mirip direkomendasikan.



K-MEANS CLUSTERING

K-Means Clustering:

- Algoritma ini mengelompokkan data tekstual (misalnya, deskripsi film) menggunakan K-Means dengan 200 cluster, dioptimalkan menggunakan MiniBatchKMeans untuk menangani dataset besar.
- Kolom `description` diubah menggunakan CountVectorizer, dan hasil clustering disimpan dalam kolom baru bernama `cluster`.
- Proses ini mengidentifikasi kelompok film berdasarkan kesamaan teks.

Sistem Rekomendasi Menggunakan Graf:

- Struktur graf dibuat dengan node yang merepresentasikan film, aktor, sutradara, kategori, dan negara, serta edge yang menggambarkan hubungan antar node.
- Pustaka NetworkX digunakan untuk menginisialisasi graf, menambahkan node, dan menghubungkannya berdasarkan hubungan yang ada.
- Fungsi `find_similar()` menghitung kesamaan film berdasarkan cosine similarity dan memperbarui graf dengan menambahkan edge kesamaan.

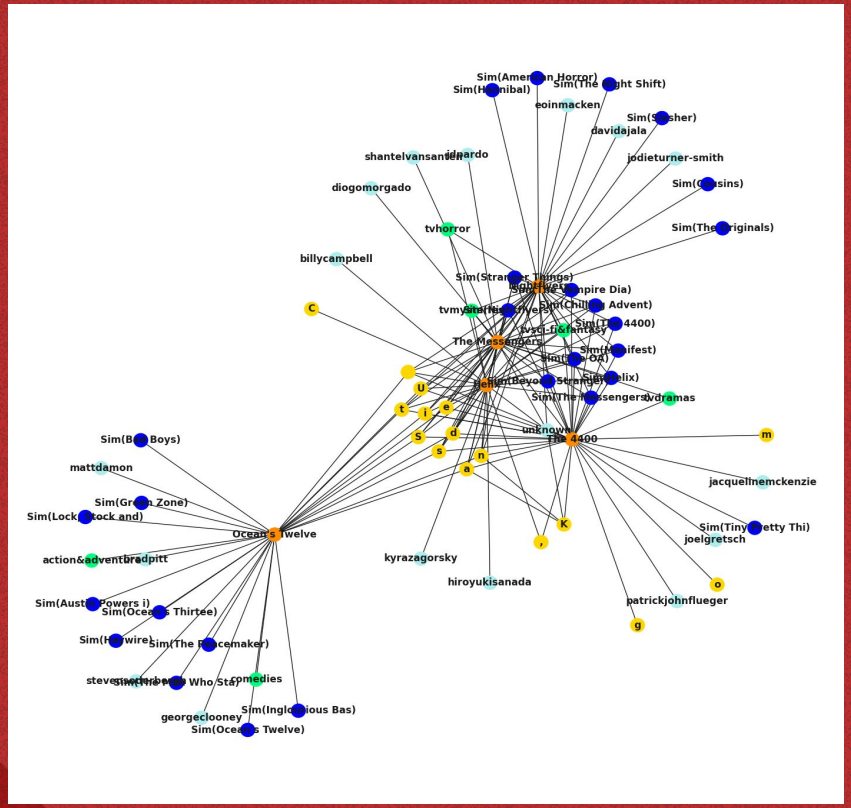
Visualisasi Subgraf:

- Fungsi `get_all_adj_nodes()` mengekstrak subgraf yang mencakup node tertentu beserta tetangganya.
- Fungsi `draw_sub_graph()` memvisualisasikan subgraf dengan node yang diberi warna sesuai jenisnya (misalnya, film, aktor).

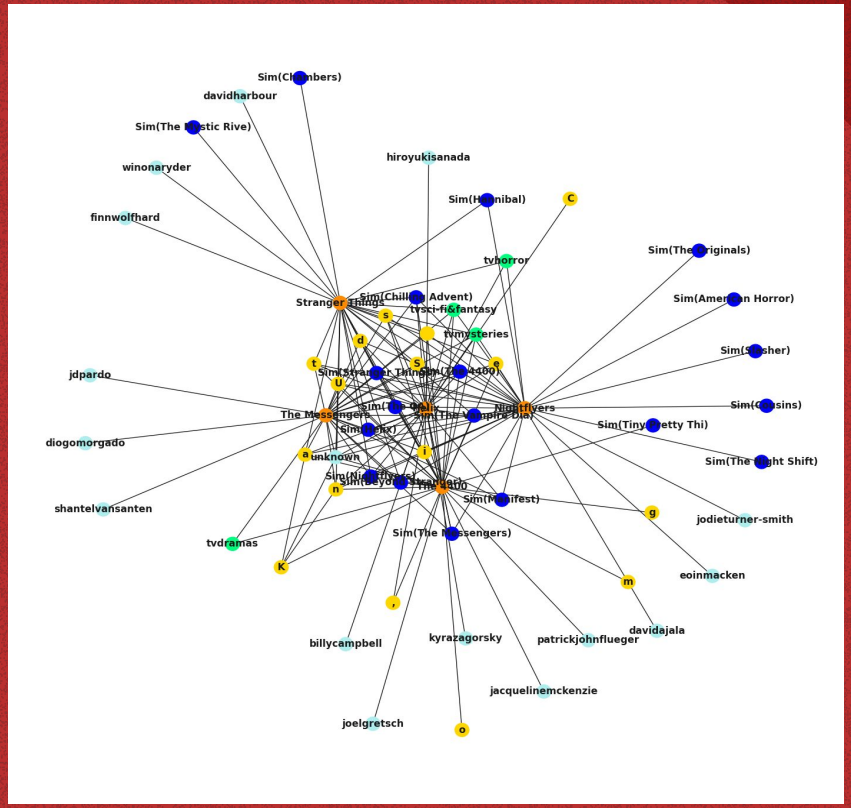
Tujuan dan Insight:

- Pendekatan graf dan clustering ini digunakan untuk merekomendasikan film dengan mengidentifikasi cluster yang mirip dan mengeksplorasi koneksi dalam graf.
- Metodologi ini memanfaatkan konten tekstual dan hubungan antar entitas (seperti aktor, genre) untuk menghasilkan rekomendasi film yang lebih relevan.

TOP RECOMMENDATION EXAMPLE



Ocean's Twelve



Stranger Things

05

MODEL EVALUATION

MODEL EVALUATION

Model	Metric	Value
KNN	Accuracy	1.00
Decision Tree	Accuracy	0.8638
Logistic Regression	Accuracy	0.3990
Random Forest	Accuracy	0.8383
Naive Bayes	Davies-Bouldin Index	1.310
K-Means Clustering	Davies-Bouldin Index	1.451

Insights

1. **KNN** menunjukkan akurasi tertinggi (100%), namun perlu diperiksa lebih lanjut untuk memastikan tidak terjadi overfitting.
2. **Decision Tree** dan **Random Forest** memberikan akurasi yang baik, masing-masing 86.38% dan 83.83%, menunjukkan kemampuan yang kuat dalam klasifikasi berbasis atribut.
3. **Logistic Regression** memiliki akurasi yang lebih rendah (39.90%), mungkin karena hubungan antar fitur tidak sepenuhnya linier.
4. Nilai **Davies-Bouldin Index** untuk **Naive Bayes** (1.310) dan **K-Means Clustering** (1.451) menunjukkan kualitas klaster yang cukup baik, dengan nilai lebih rendah pada **Naive Bayes** yang mengindikasikan klaster yang lebih terpisah.

06

REKOMENDASI BISNIS

REKOMENDASI BISNIS

PERSONALISASI KONTEN

Menggunakan KNN sebagai model utama dapat meningkatkan pengalaman pengguna dengan memberikan rekomendasi film atau acara TV yang sesuai dengan riwayat tontonan mereka.
Terapkan rekomendasi berbasis kesamaan aktor, sutradara, genre, atau negara asal untuk menarik minat pengguna.

SEGMENTASI PENGGUNA

Gunakan K-Means Clustering untuk mengelompokkan pengguna berdasarkan preferensi tontonan mereka. Ini akan membantu perusahaan untuk menargetkan penawaran promosi atau rekomendasi konten yang spesifik.

OPTIMASI PEMASARAN

Berdasarkan wawasan clustering, buat kampanye pemasaran yang terarah untuk kelompok pengguna tertentu, seperti penggemar film genre "Action & Adventure" atau penggemar aktor tertentu.

TERUS PANTAU PERFORMA MODEL DENGAN MENGGUNAKAN METRIK EVALUASI SEPERTI AKURASI, INDEKS DAVIES-BOULDIN, DAN FEEDBACK PENGGUNA UNTUK MEMASTIKAN SISTEM REKOMENDASI TETAP RELEVAN DAN BERKUALITAS TINGGI.

THANKS!

Do you have any questions?

hijirdw@gmail.com

https://www.linkedin.com/in/hijirdella/

https://github.com/hijirdella/Netflix-Recommendation-System