Rayah Mian   
July 14, 2023

Movie Recommendation System Proposal

**1. Introduction**

*1.1 Define the Scope*

I will leverage data from four prominent streaming platforms, namely Netflix, Amazon Prime, Disney Plus, and Hulu, to develop a movie recommendation system. This system will utilize content-based filtering techniques to generate personalized recommendations based on movie attributes. Users will have the ability to specify their preferences, whether it be a specific movie genre or their desired actor, enabling them to discover relevant titles within the system.

**2. Data Review**

*2.1 Gathering Data*

I will aggregate data from four distinct streaming platforms, each containing a collection of movies and TV shows with their respective attributes. By merging these datasets, I aim to create a comprehensive and consolidated dataset. The purpose of this consolidation is to establish a rich foundation from which I will develop and implement the recommendation system.

*2.2 Data Exploration*

The 4 datasets have the following number of entries:

* Amazon Prime: 9668
* Disney Plus: 1450
* Hulu: 3073
* Netflix: 8807

All 4 datasets have the following columns:

* show\_id
* type
* title
* director
* cast
* country
* date\_added
* release\_year
* rating
* duration
* listed\_in
* description

*2.4 Data Analysis*

I will aggregate the data from the 4 datasets to create a consolidated dataset. Then, I will proceed to clean the dataset by removing any duplicates, handling missing values, and ensuring the data is in a format suitable for analysis, extrapolation and coding.

By consolidating the dataset, the resulting columns are:

* show\_id
* platform
* type
* title
* director
* cast
* country
* date\_added
* release\_year
* rating
* duration
* listed\_in
* description

“Type” refers to whether the entry is a movie or a TV show.

“Country” refers to which country the movie or TV show was produced in. There are some entries that do not include the country of origin. The datasets include movies from the following countries: United States, India, United Kingdom, Canada, Japan, France, Germany, South Korea, Spain, Australia, Mexico, China, Italy, Egypt, Hong Kong, Turkey, Belgium, Brazil, Nigeria, Argentina, Taiwan, Indonesia, Philippines, Denmark, Thailand, Ireland, South Africa, Sweden, Colombia, Netherlands, Poland, Singapore, United Arab Emirates, New Zealand, Norway, Israel, Russia, Chile, Lebanon, Malaysia, Czech Republic, Switzerland, Pakistan, Austria, Luxembourg, Romania, Hungary, Uruguay, Bulgaria, Finland, Saudi Arabia, Iceland, Greece, Qatar, Peru, Jordan, Serbia, Kuwait, Vietnam, Portugal, Iran, Morocco, Kenya, Cambodia, West Germany, Venezuela, Syria, Ghana, Croatia, Malta, Bangladesh, Slovenia, Ukraine, Zimbabwe, Senegal, Namibia, Guatemala, Kazakhstan, Soviet Union, Afghanistan, Algeria, Georgia, Iraq, Mauritius, Cayman Islands, Montenegro, Nepal, Botswana, Kosovo, Albania, Panama, Angola, Jamaica, Cuba, Bahamas, Sri Lanka, Uganda, Sudan, Latvia, Liechtenstein, Somalia, Vatican City, Nicaragua, Dominican Republic, Armenia, Samoa, Azerbaijan, Lithuania, Mongolia, Slovakia, Ecuador, Bermuda, Palestine, Ethiopia, Burkina Faso, French Polynesia, Costa Rica, Tunisia, Mozambique, Belarus, Puerto Rico, Cyprus, Malawi, Paraguay, Tanzania, Monaco, Cameroon, East Germany.

Another important feature of these entries is “release\_year”. All the movies and TV shows were released between 1920 and 2021.

The last important feature is “listed\_in” which refers to the genre of the title. The following is a list of all the genres: Comedy, Drama, International, Action, Suspense, Documentary, Fantasy, Kids, Special Interest, Science Fiction, Adventure, Horror, Sports, Talk Show and Variety, Anime, Arts, Entertainment, and Culture, TV Shows, Animation, Music Videos and Concerts, Fitness, Faith and Spirituality, Military and War, Western, LGBTQ, Romance, Unscripted, Young Adult Audience, Arthouse, Historical, Family, Musical, Docuseries, Music, Biographical, Action-Adventure, Superhero, Reality, Survival, Animals & Nature, Coming of Age, Lifestyle, Movies, Concert Film, Crime, Anthology, Medical, Variety, Spy/Espionage, Buddy, Parody, Game Show / Competition, Romantic Comedy, Thriller, Police/Cop, Talk Show, Dance, Series, Mystery, Soap Opera / Melodrama, Disaster, Travel, Stand Up, Cooking & Food, Documentaries, Lifestyle & Culture, News, History, Teen, Health & Wellness, Black Stories, Latino, Late Night, Sketch Comedy, Classics, LGBTQ+, Adult Animation, Sitcom, Game Shows, Cartoons, Science & Technology, International TV Shows, TV Dramas, TV Mysteries, Crime TV Shows, TV Action & Adventure, Reality TV, Romantic TV Shows, TV Comedies, TV Horror, Children & Family Movies, Dramas, Independent Movies, International Movies, British TV Shows, Comedies, Spanish-Language TV Shows, Thrillers, Romantic Movies, Music & Musicals, Horror Movies, Sci-Fi & Fantasy, TV Thrillers, Kids' TV, Action & Adventure, TV Sci-Fi & Fantasy, Classic Movies, Anime Features, Sports Movies, Anime Series, Korean TV Shows, Science & Nature TV, Teen TV Shows, Cult Movies, Faith & Spirituality, LGBTQ Movies, Stand-Up Comedy, Stand-Up Comedy & Talk Shows, Classic & Cult TV.

* Implement the Algorithm: Use a programming language and machine learning libraries like Python and scikit-learn or TensorFlow to implement the chosen algorithm. Transform the data into a suitable format and train the model using your dataset.
* Evaluate the Model: Split your dataset into training and testing sets. Evaluate the performance of your recommendation system using appropriate evaluation metrics such as precision, recall, or mean average precision. This will help you assess how well your system is performing.
* Incorporate User Feedback: Collect user feedback to improve the recommendation system over time. Allow users to rate movies, provide reviews, or incorporate implicit feedback like watch history to enhance the recommendations.

**3. Software Design**

TF-IDF (Term Frequency-Inverse Document Frequency) is a numerical statistic used to evaluate the importance of a term within a document or a collection of documents. It is commonly used in information retrieval and text mining.

Cosine similarity is a metric used to determine the similarity between two vectors or documents. It measures the cosine of the angle between two vectors, where the vectors represent the term frequency (or TF-IDF) values of documents. Cosine similarity ranges from -1 to 1, with 1 indicating perfect similarity and -1 indicating perfect dissimilarity.

Both TF-IDF and cosine similarity are commonly used techniques in natural language processing (NLP) and text analysis to understand and compare textual data. Both techniques would be beneficial for the movie recommendation system because they will help identify similar titles according to their features and their scores

* Choose a Recommendation Algorithm: For content-based filtering, you can use techniques like TF-IDF (Term Frequency-Inverse Document Frequency) or cosine similarity.
* Implement the Algorithm: Use a programming language and machine learning libraries like Python and scikit-learn or TensorFlow to implement the chosen algorithm. Transform the data into a suitable format and train the model using your dataset.
* Evaluate the Model: Split your dataset into training and testing sets. Evaluate the performance of your recommendation system using appropriate evaluation metrics such as precision, recall, or mean average precision. This will help you assess how well your system is performing.
* Deploy the Recommendation System: Integrate your recommendation system into a user-friendly interface, such as a web or mobile application. Provide users with personalized movie suggestions based on their preferences and feedback.
* Incorporate User Feedback: Collect user feedback to improve the recommendation system over time. Allow users to rate movies, provide reviews, or incorporate implicit feedback like watch history to enhance the recommendations.

Since there are many countries that produce movies, we will have to display the most popular countries. Also, since there are many genres, we will have to merge some genres for conciseness.

**References**

<https://www.kaggle.com/datasets/shivamb/amazon-prime-movies-and-tv-shows>

<https://www.kaggle.com/datasets/shivamb/disney-movies-and-tv-shows>

<https://www.kaggle.com/datasets/shivamb/hulu-movies-and-tv-shows>

<https://www.kaggle.com/datasets/shivamb/netflix-shows>