|  |  |
| --- | --- |
|  | Content on Netflix  April Meyer  DSC680-T301 Applied Data Science (2215-1) |

# **Executive Summary**

There was a time when finding a TV show or movie to watch was not easily found. You would have to wait it to air and then record it. Nowadays there is a substantial amount of streaming content at our fingertips. One of the first and still main platforms to view content is on Netflix. Netflix, Inc. was founded in Scotts Valley, California by Reed Hastings and Marc Randolph in 1997 (3). In the beginning it provided DVD sales and rental through mail. By 2007 it had grown into one of the main online streaming subscription companies (3). It is now available worldwide except China (3). In 2013 Netflix has started to create its own production content including House of Cards (3).

A report was released in 2018 that showed the number of TV shows on Netflix has almost tripled since 2010, however, movies have decreased by more than 2,000. Movies and TV shows available on Netflix (1). The goal of this project is to suggest other streaming content to the viewer based on the current title. The data I will be exploring was collected from Flixable they are a third-party Netflix search engine (2). It contains movies and TV shows that are available on Netflix as of 2019.

The method I used to recommend new content was Cosine Similarity. It preformed well when tested using different titles. For example, when using ‘Breaking Bad’ it returned ‘Better Call Saul’ which is a spin-off of ‘Breaking Bad’. Also, when using ‘Transformers Prime’ it did return the other available Transformer Movies.

# **Preliminary Analysis**

## **Background of the problem**

A report was released in 2018 that showed the number of TV shows on Netflix has almost tripled since 2010, however, movies have decreased by more than 2,000. Movies and TV shows available on Netflix will be my domain for this first project (1). I wanted to investigate what other findings can be acquired from their streaming content.

## **Problem Statement**

The purpose of this project is to develop a recommendation system, so that viewers may identify what to watch next based on their current viewing.

## **Methods**

The method used for this project is a cosine similarity to create a recommender system. A recommender system uses current activity to recommend the suggest a similar item (4). Amazon, Facebook, and Netflix all use these machine learning algorithms to help the user (4). Cosine Similarity is a measurement used to establish the similarity between two items no matter the size (5). It calculates the cosine of the angle between two vectors in a multi-dimensional space (5). I imported cosine\_similarity from sklearn for the metric of similarity; from sklearn.metrics.pairwise import cosine\_similarity.

## **Data Sources**

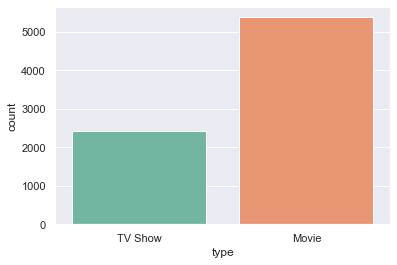
The data is from Kaggle. Link to data, https://www.kaggle.com/shivamb/netflix-shows.

## **Data Exploration**

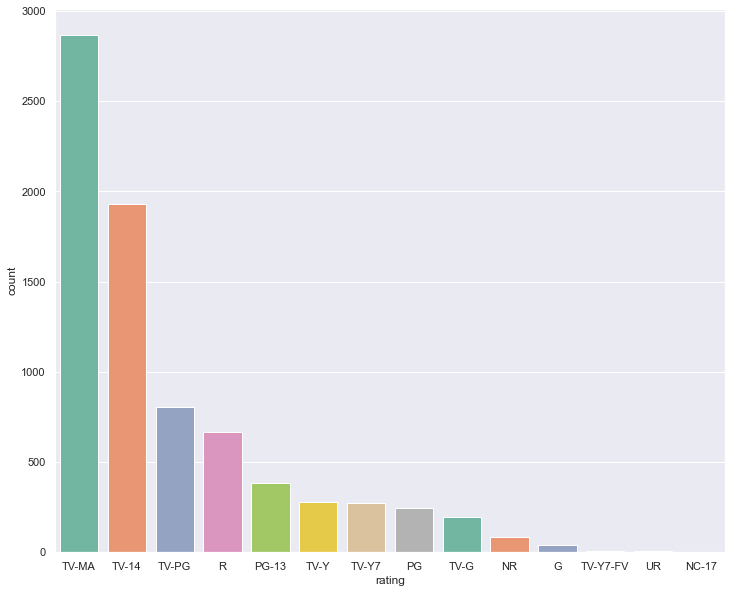
The initial analysis of the attributes for the dataset are below:

* show\_id: Unique ID for every Movie / Tv Show
* type: Identifier - A Movie or TV Show
* title: Title of the Movie / Tv Show
* director: Director of the Movie
* cast: Actors involved in the movie / show
* country: Country where the movie / show was produced
* date\_added: Date it was added on Netflix
* release\_year: Actual Release year of the move / show
* rating: TV Rating of the movie / show
* duration: Total Duration - in minutes or number of seasons
* listed\_in: Genere
* description: The summary description

**Comparison of Movies vs TV Shows:**

****

**Total by Ratings:**



## **Data Preparation**

For data preparation a new feature was created called combined\_fields. It contained cast, director, listed\_in, and description. Before combining cast and director all NAs were filled in with ‘Unknown’. Next the combined field required all words to be tokenized and preprocessed by changing to lower case, removing symbols, and stop words. I used English for the stop words. English stop words that were used most often were removed.

## **Modeling**

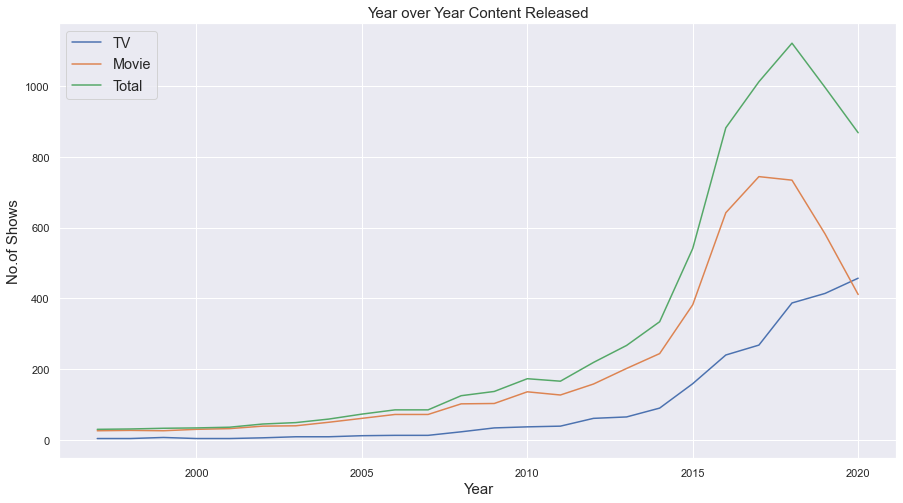
After the data preparation a function was built called get\_reommendations(title). It gets the pairwise similarity scores of all content with that content and returns the ten most similar.

## **Results**

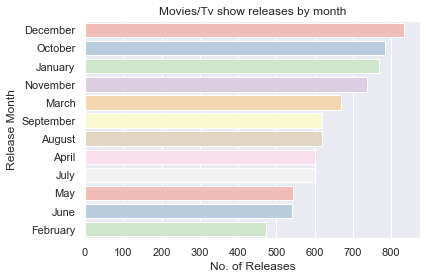
The get\_recommendations() performed well when tested with a couple titles. For example, ‘The Umbrella Academy’ returned ‘The Future of Water’, ‘The Staircase’, ‘Maharakshak Devi’, ‘The 4400’, ‘Mighty Little Bheem’, ‘Locke & Key’, ‘Nightflyers’, ‘Charmed’, ‘The Dragon Prince’, and ‘Biohackers’. Another example was with ‘Breaking Bad’ and it returned suggestions: ‘Better Call Saul’, ‘The Show’, ‘Have You Ever Fallen in Love, Miss Jiang?’, ‘Girlfriend's Day’, ‘W/ Bob & David’, ‘Get Shorty’, ‘Refresh Man’, ‘Extracurricular’, ‘Pyaar Tune Kya Kiya’, and ‘The Writer’.

## **Discussion/Conclusion**

Overall, I was happy with the recommendation system using cosine similarity. It suggested similar content based on the title entered. Some other insights gained from this data was how the content changed year over year. Recently movies have started to decrease while TV content has steadily increased.



Also, after looking at the content total release by month I looks like February would be the optimal time to release a new movie or TV show on Netflix.



## **Acknowledgements**

I would like to acknowledge the Kaggle Repository for providing this dataset I used for the project. I would also like to acknowledge Mahnoor Javed. Her code gave me a reference point and guided on using a cosine similarity for a recommendation system.

# **References:**

1. Bansal, S. (2021, January 18). Netflix movies and tv shows. Retrieved March 21, 2021, from <https://www.kaggle.com/shivamb/netflix-shows>

Kaggle, a subsidiary of Google LLC, is an online community of data scientists and machine learning practitioners. This reference is for the data set used.

1. Full list of movies and tv shows on Netflix. (n.d.). Retrieved March 21, 2021, from <https://flixable.com/>

Flixable is a search engine for video streaming services that offers a complete list of streaming content. Streaming services include Disney+, Hulu, HBO Max, and Netflix.

1. Netflix. (2021, March 21). Retrieved March 21, 2021, from <https://en.wikipedia.org/wiki/Netflix>
   1. From Wikipedia, the free encyclopedia. It provides an overview of Netflix and it’s history.
2. Javed, M. (2020, November 04). Using cosine similarity to build a movie recommendation system. Retrieved April 06, 2021, from <https://towardsdatascience.com/using-cosine-similarity-to-build-a-movie-recommendation-system-ae7f20842599>
3. Prabhakaran, S. (2020, October 11). Cosine similarity - understanding the math and how it works? (with python). Retrieved April 06, 2021, from <https://www.machinelearningplus.com/nlp/cosine-similarity/>
4. Yclaudel. (2020, February 10). Find similar articles with tf-idf. Retrieved March 21, 2021, from <https://www.kaggle.com/yclaudel/find-similar-articles-with-tf-idf>
   1. Walks through how to find what are the texts like a given text in a corpus of texts.