RCOMMENDATION SYSTEM BASED ON ARABIC SENTIMENT ANALYSIS

20/11/2022

Prepared By :

UData Team

TABLE OF CONTENT

Introduction

As we may all agree, innumerable of our lives decisions are effected by our moods and what we feel. This includes our biggest life decisions to the simple daily decisions such as what we watch or what we eat. Moreover, the variety of choices we all have nowadays made it challenging sometimes to pick a choice.

With all these difficulties, business tried to make it smoother for the customers to choose by building recommendation systems striving to meet customers' needs and satisfaction.

In the light of quality of life program, that was launched in 2018 under 2030 Saudi Arabia's vision that aims to expand the horizons of many sectors that directly impact citizens. We have built (shows recommendation system) that suggests film/series to watch based on what you feel in . This model have many appliances to use and it will surely contribute to elevate the level of entertaining in Saudi Arabia.

Data review

STC IP-TV is a Dataset collected in the time between 2017 to 2018. IP TV allows users to watch internet contents on the TV, This service was released in the middle of 2016 by STC one of the leading telecom companies in the gulf. This dataset have 10 variables and around 4 million records and it shows information about the movies and series the users have watched.

Data pre-processing

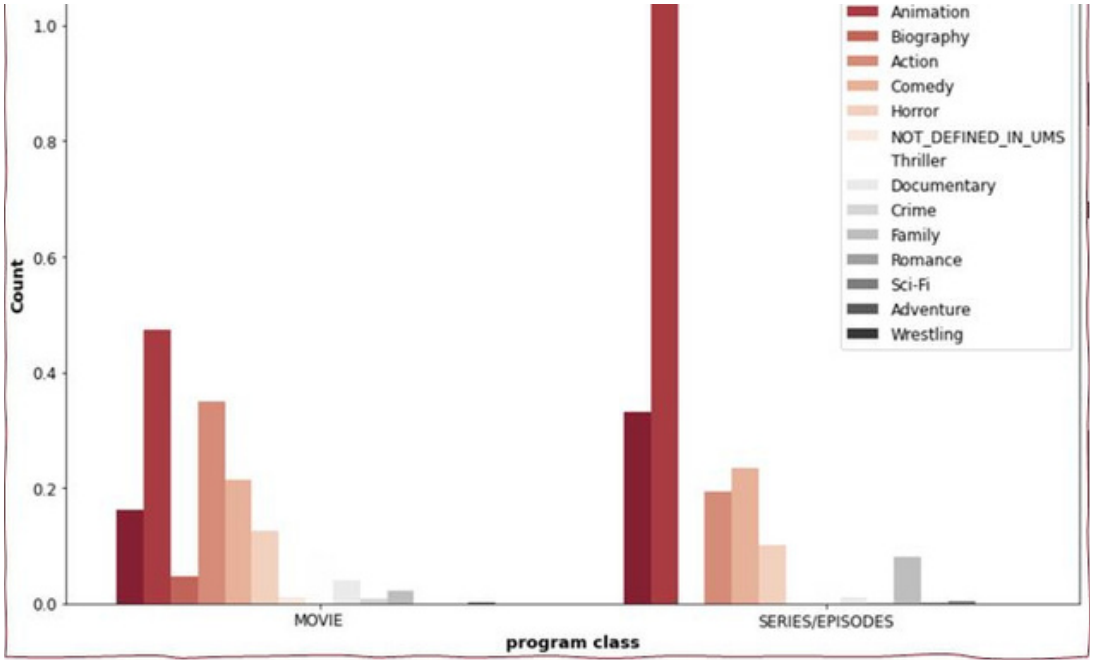
First of all, we started by striping rows and removing digits from (**program name**, **original name**). Afterward, we created 4 columns for **year**, **month**, **hour** and **minute** for a better exactness. Then, we used group by to find the count of users and the duration each user has spent in one day.

In addition, We initiated 3 new columns first columns is **commercial account:** based on the duration the user spent on the day, that if it was more than 8 hours and less than 16 hours a day it will be considered as a commercial account.

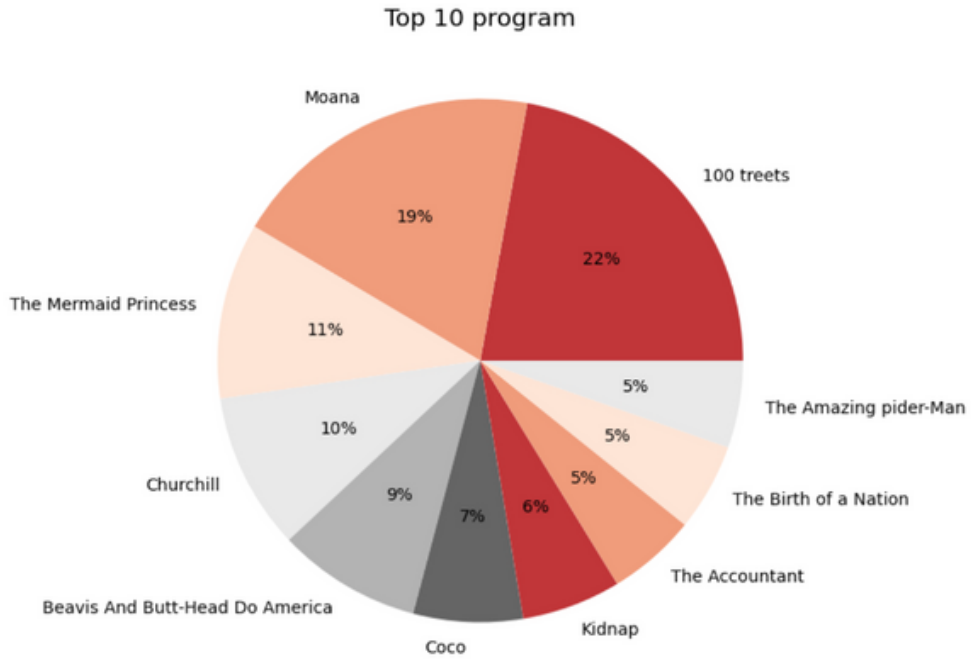
Family account: we have set a condition that if the using period hasn't exceeded 15 hours and more than 8 hours the account will be considered as family account.

Score: weighted score, it's calculated by dividing the mean duration in minutes of a movie by the total duration in minutes based on the genre, after that multiplying the result by the weight. following that, the sum the previous result with the result of dividing the count for each movie by the count of all movies in the same genre. Lastly, multiply the total result by 100. Worth to mention that duration form 20% of the weight.

Data exploration:

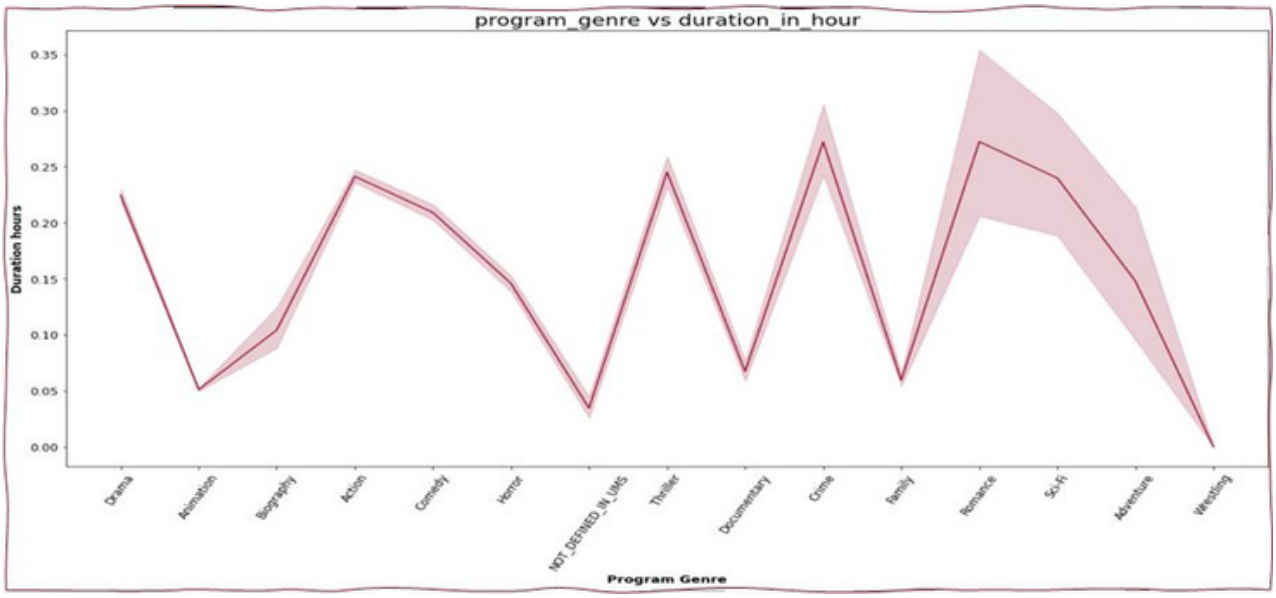


As we can see, Drama and animation genre upstaged all other genre in both movie and series.

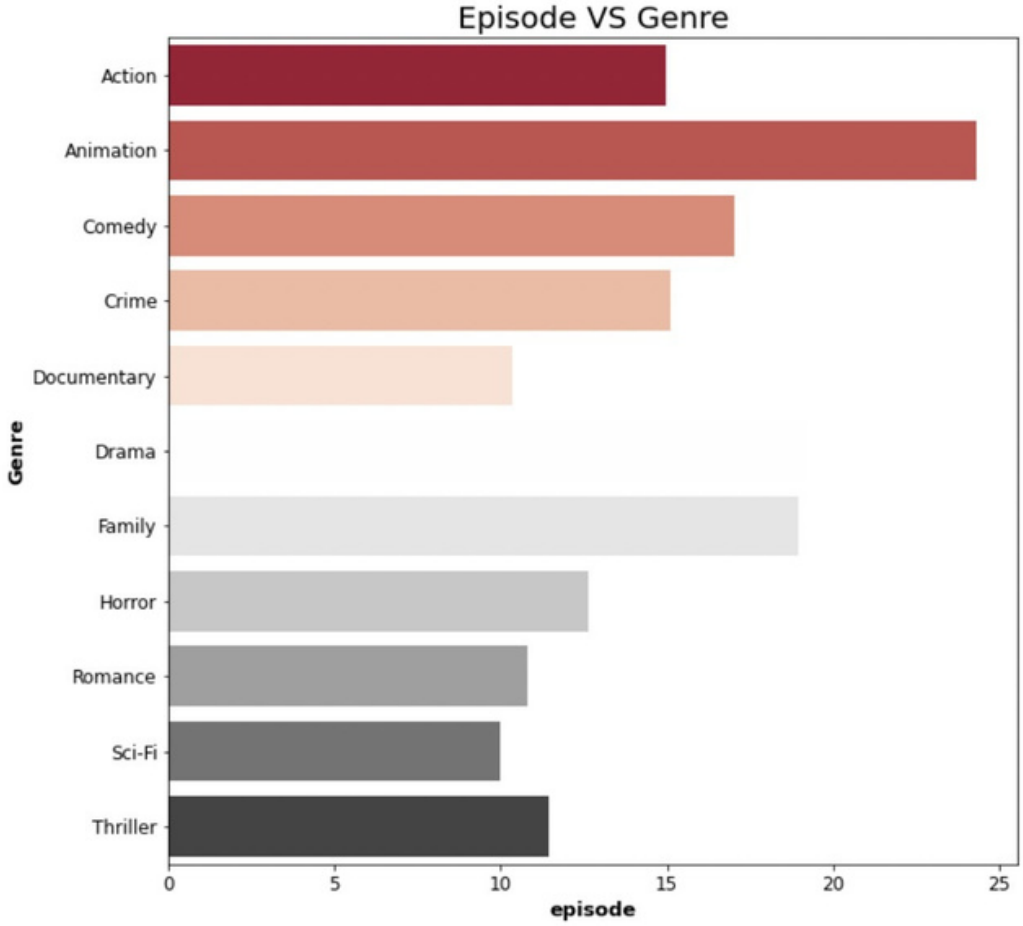


This pie chart illustrate Top 10 programs watched during 2017–2018. 100 streets, Moana and The mermaid princess were the most popular programs at that time.

Data exploration:

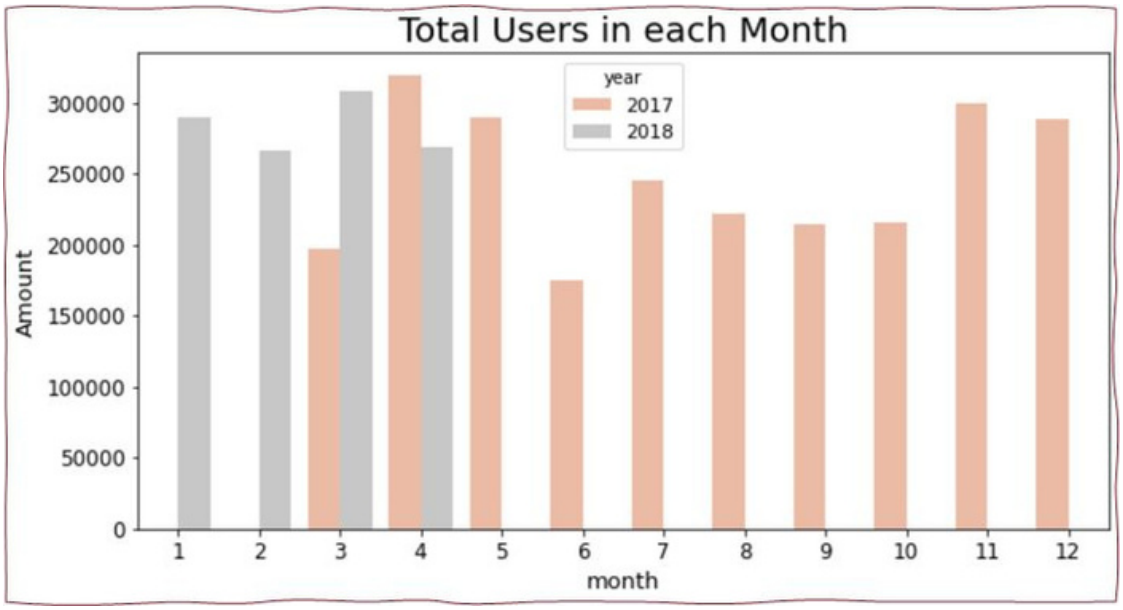


This line chart shows the trend in watching genre habits between 2017–2018. As this chart illustrate, unspecified genre series, romance and actions were the most shows watched at that time, Biography, documentary and wrestling were the least favorited to watch.

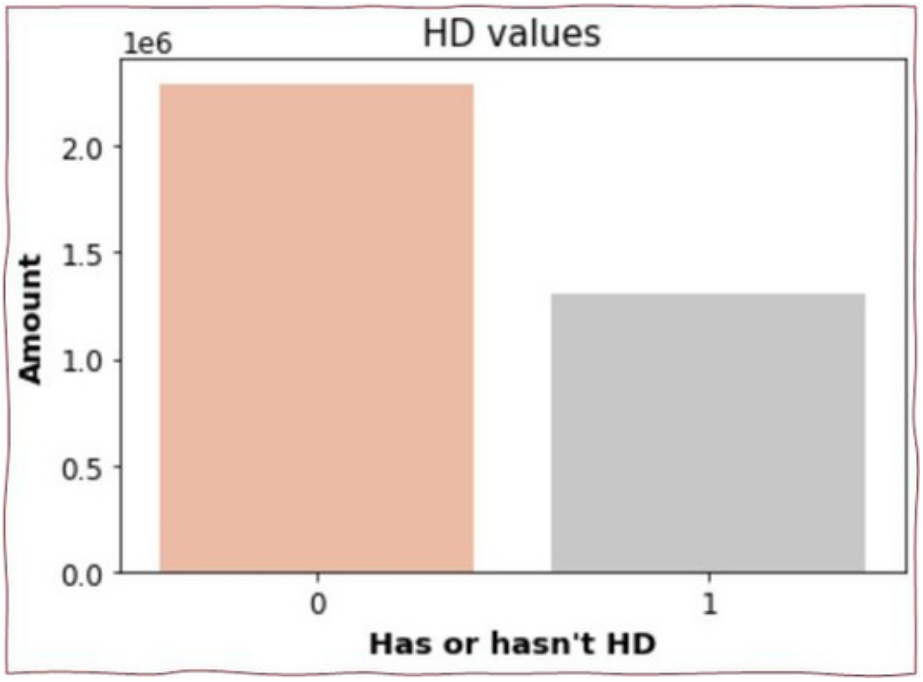


This chart shows that animation, Family and drama genre has the most episodes count respectively.

Data exploration:



As we can see, there's a significant increase in the number of users in March 2018 than it was in march 2017. However, it has slightly decreased in April 2018 than it was in April 2017.



This bar chart shows that the majority of contents haveHD resolutions as expected.

Building the model

we started by setting our targeted emotions, There were 9 emotions we wanted our model to recognize. we believe that those are the main emotions that can drive someone to watch a movie, Those emotions are: anger, boredom, surprise, sadness, fear, sympathy, joy, love and not feeling something is a considered feeling too.

we labeled Arabic words with the emotions that indicate to it by using a dataset that contains around 3000 row of different emotions . Afterwards, we used MARBERT's pre-trained model which uses BERT. BERT is a applying the bidirectional training of Transformer, a popular attention model, to language modeling.

Afterward, we started to build and train our model and test the dataset, define training arguments and build a trainer.

Lastly, we connected the model with our recommendation system and it succeeded to show the top 3 movies from each genre as expected.

Result and evaluation

Based on the training model we have used, The results were: 77% macro average,76% weighted average in precision and 76% in both in recall.

Epoch	Training Loss	Validation Loss	Macro F1	Macro F1 Pos Neg	Macro Precision	Macro Recall	Accuracy	Runtime	Samples Per Second
1	0.978300	0.713265	0.761335	0.671266	0.774841	0.762248	0.755888	444.279900	6.021000
2	0.519100	0.703555	0.779534	0.694961	0.782036	0.778945	0.769346	443.866700	6.027000

Conclusion

In this project, We explored users behaviors and watching habits and uses during 2017-2018 . In fact, we observed the way they used the newly service -at that time-.

This model was developed striving to contribute enriching Arabic contents and support organizations and business who are looking for improvement in their performance or starting new business in Saudi Arabia .

Future work

This project has been mainly focused on the use of sentiment analysis to identify emotions, and there are many other emotions Arabic is prosperous with. The following ideas could be tested:

1- Add new emotions, examine their effect on an individual.

2- It could be interesting to compare results with other streaming services

3- Apply the recommendation system to different uses.

4- Instead of using popularity recommendation, we can make a model that learns the users behavior.