2021

Content-based Recommendation system

BY:

- REYAD MELIES
- SALMA MOUSA
- SHAHENDA YOUSSEF
- TOKKA HASSAN

CONTENT-BASED RECOMMENDATION SYTEM

OVERVIEW

> ℹ *Nowadays, businesses tend to make their products the closest to user preferences in order for the user to feel as a customized service is offered to him. One of the ways to increase user satisfaction is recommendation systems where the user is offered several suggestions related to his chosen preference. This led us to think of recommendation system as a final project.*

OUR PROPOSAL

According to this course's content, we believe we have taken the essentials that provide us the ability to engage in a real use case that is widely used today and has high impact on the service quality.

We propose a content-based recommendation system for movies and series titles. The system recommends different movie titles similar to the entered one, not only semantically similar, but also, related to the movie type and content. The recommender uses different features to determine the suggestions in order to produce the best results.

Technical/Project Approach

Techniques:

- similarity measures that are used to compute the similarity*:* ***cosine similarity***, ***Euclidean Distance***.
- text preprocessing: ***remove stop words***, ***stemming*** and ***lemmatization.***
- feature engineering: ***Bag of words (BOW), word embedding, or TF-IDF***.

Python libraries:

- Pandas ➔ for data manipulation and analysis
- Sklearn ➔ metrics, feature_extraction
- Numpy
- Surprise ➔ for building and analyzing recommender systems that deal with explicit rating data.
- Nltk

Data Sources and Description

Our dataset should have movie titles, actors, genre and many more attributes that facilitates the recommendation of movies and series.

We chose different open data sources for different reasonings that could assist us to achieve our goal. Here are the datasets with their reasoning to be picked:

- **MovieLens 25M Dataset:**
  - ✓ table benchmark dataset. 20 million ratings and 465,000 tag applications applied to 27,000 movies by 138,000 users.
  - ✓ Includes tag genome data with 12 million relevance scores across 1,100 tags.
- **Netflix Prize Dataset:**
  - ✓ consists of about 100,000,000 ratings for 17,770 movies given by 480,189 users.
  - ✓ Each rating in the training dataset consists of four entries: user, movie, date of grade, grade.

EVALUATION

Different Evaluation metrics will be used to give us different insights to help us increase our system's efficiency and maintain the expected results. Here are some metrics that will be used:

- Mean Absolute Error
- Root Mean square Error
- Precision, Recall and F-score
- Mean Average Precision
- ROC curve
- Mean Reciprocal Rank
- Spearman Rank Correlation Coefficient
- Normalized Discounted Cumulative Gain
- Coverage

EXPECTED RESULTS

Enhanced user experience through personalized recommendations based on the user's input data.

We expect our proposed system to work in this manner:

- ✓ When the user search for a movie, the top similar movies are recommended.
- ✓ The system will recommend movies in the same genre, starring the same actor, the highest rate movies.