Institute of Technology of Cambodia

Programming For Data Science
2020-2021

3rd year Engineer's Degree in Data Science

Department of Applied Mathematics and Statistics

Project Guideline: Movie Recommendation System

Group members:

Name	ID
Tang Piseth	e20201634
Set Mongkol	e20201255
Thornthea Gechhai	e20201321
You Phakkorn	e20200727
Sreng Seangleng	e20200840
Thong Chhunher	e20200711

Lecturers:

Prof.Chan Sophal

Introduction

This project will develop a movie recommendation system that uses a content-based filtering approach. The system will recommend movies to users based on their past Movie name, ratings, overview, cast/star, gross collection and vote and who have similar tastes.

The project will be implemented using the following steps:

1. Scraping data from a movie website.
2. Cleaning the data.
3. Exploratory data analysis (EDA).
4. Feature engineering.
5. Choosing a model (KNN).
6. Running a system using Streamlit.

1. Scraping Data from a Movie Website

The first step is to scrape data from a movie website. This can be done using a variety of tools, such as BeautifulSoup. The data that is scraped should include the following information:

- ID
- MovieName
- Overview
- Genre
- Vote
- Gross collection
- Rating
- Cast
- Director

2. Cleaning the Data

Once the data has been scraped, it needs to be cleaned. This includes removing any duplicate rows, fixing any errors in the data, and converting the data to the correct format.

3. Exploratory Data Analysis (EDA)

The next step is to perform exploratory data analysis (EDA) on the data. This involves visualizing the data to understand the distribution of the data and to identify any trends and finding multicollinearity.

4. Feature Engineering

Feature engineering is an essential step in preparing data for machine learning models. In our movie analysis project, we carefully selected relevant features such as Movie Title, Overview, Genre, Vote, Gross Collection, Rating, Cast, and Director. To enhance the data representation, we merged certain features together. This approach allowed us to capture diverse aspects of movies and create a more comprehensive dataset for our model. By merging features like Cast and Director, and Genre and Overview, we aimed to uncover relationships and patterns that could contribute to accurate predictions and valuable insights. Additionally, merging features like Vote, Gross Collection, and Rating provided a consolidated measure of popularity, financial success, and critical acclaim. Overall, our feature engineering process aimed to optimize the data for effective machine learning analysis.

5. **Text Preprocessing for Model**

   To prepare the text data for analysis, we apply the following preprocessing steps:
   - Text Cleaning

     -Convert the text to lowercase to ensure consistency.
   - Stemming

     -By using the PorterStemmer library from python we can apply stemming techniques to reduce words to their base form. This helps in standardizing the text data and capturing the essence of words.
6. **Bag of Words Representation**
   - Document-Term Matrix

     -Utilize the CountVectorizer function to convert the preprocessed text data into a document-term matrix representation.

     -Each row in the matrix represents a document, and each column represents a unique word from the tag column.

     -The values in the matrix indicate the frequency of each word in each document.
7. **Choosing a Model**

There are a variety of machine learning models that can be used for movie recommendation systems. In this project, we will build a model which is KNN with using metric.

8. **Running the Recommendation System by deploying it on Streamlit**

The final step is to run a recommendation system via Streamlit. Streamlit is a Python library that allows you to create interactive web applications. The Movie recommendation system will allow users to choose their favorite movie and will then recommend movies to them based on the movie similarity of the chosen movie.

Conclusion

This project developed a movie recommendation system using content-based filtering. The process involved scraping movie data, cleaning and exploring the data, performing feature engineering, selecting models (KNN), and implementing a user interface using Streamlit. The system allows users to input their favorite movie and receives personalized recommendations based on their preferences and similarities of the movie. Overall, the project successfully created an interactive movie recommendation system using content-based filtering techniques.