charlesot / phase-4

Code Issues Pull requests Actions Projects Wiki Security Insights Settings

phase-4 Public

1 Branch 0 Tags Go to file t Go to file Add file Code

charlesot updating index notebook f76bb73 · 4 minutes ago

Data	add data files	3 days ago
.gitignore	Initial commit	3 days ago
README.md	Update README.md	5 hours ago
index.ipynb	updating index notebook	4 minutes ago

README

DSF-PT08P3- Phase 4 project

Novelle Movies Recommendation system Project Summary

About

No description, website, or topics provided.

Readme
Activity
0 stars
1 watching
0 forks

Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

Languages

1. project overview

Unlike established streaming services, small businesses in Kenya that sell movies have high customer churn and dissatisfaction rates because they cannot recommend movies to their customers. The business owners have to watch all the movies to give recommendations, this not only wastes time but also increases missed opportunities to sell movies that the vendor has never watched. The Novelle Movies recommendation system project will develop a personalized movie recommendation system that will enable small business owners to improve sales and retention of their customers through improved customer experience

2. Business understanding

- Small businesses in Kenya that sell movies have high customer churn and dissatisfaction rates.
- They spend a lot of time watching all the movies to give recommendations.
- They miss opportunities to sell movies that vendors have never watched.
- They have limited resources to develop a personalized movie recommendation system.

Project objectives

- Increase customer engagement by recommending movies based on user preference
- Increase sales by supporting customers to find movies of their taste
- Be able to make recommendations to new customers

Key features of Novelle Movie recommendation system

- Collaborative filtering (SVD) for personalized recommendations for existing active users
- Content-based filtering using movie genre and tags to handle new users
- Hybrid system that combines both methods for the best recommendation

3. Data understanding

The project will use the MovieLens small dataset from the GroupLens research lab at the University of Minnesota. It contains 100,863 ratings and 3683 tag applications on 9742 movies. It was created by 610 users between March 29, 1996, and september 24, 2018. The dataset was generated on September 26, 2018, and is available for download at http://grouplens.org/datasets/.

Dataset citation

F. Maxwell Harper and Joseph A. Konstan. 2015. The MovieLens Datasets: History and Context. ACM Transactions on Interactive Intelligent Systems (TiiS) 5, 4: 19:1–19:19. https://doi.org/10.1145/2827872

Variables

The MovieLens small dataset has the following

- Movies.csv which contains the movie details (movieId, title, genres)
- ratings.CSV which contains user ratings (userId, movieId, rating, timestamp)
- tags.CSV contains user-generated movie tags (userId, movieId, tag, timestamp)

4. Data preparation

Import the Python libraries to use

import pandas as pd import seaborn as sns import numpy as np import matplotlib.pyplot as plt from surprise import SVD, Dataset, Reader, SVDpp from surprise.model_selection import train_test_split, cross_validate from surprise import accuracy from collections import defaultdict from sklearn.feature_extraction.text import TfidfVectorizer from sklearn.metrics.pairwise import linear_kernel from sklearn.pipeline import Pipeline

Load the dataset

Dataset if downloaded from http://grouplens.org/datasets/ and loaded into pandas data frames movies_df = pd.read_csv(r'./Data\movies.csv') tags_df = pd.read_csv(r'./Data\tags.csv') ratings_df = pd.read_csv(r'./Data\ratings.csv')

Exploration

- The ratings_df, tags_df, and movies_df data frames were explored to see the first five rows, the information, shape, missing entries, and the number of unique users and movies.
- The column timestamp dropped because it was not being used
- All tags converted to lowercase
- All the tags per movie aggregate
- The movies_df and the tag_df merged on movieID column
- All Nan features filled with an empty string
- "genre" and "tags" combined to form one feature a "combined_feature"

- The modified movies_df is merged with the ratings_df to form merged_df dataframe which will be used in development of the model
- merged_df data frame is further explored to identify the top ten rated movies, the distribution of the movie rating, the 10 most rated movies, and the distribution number of ratings per user

5. Modelling

Collabotarive filtering (SVD) modeling

- Used Surprise SVD to predict user preferences: data split into training and testing with test_size of 0.2 and random _state of 42 . training of the model used n_factor of 50
- Model evaluates using root mean square error (RMSE) and mean absolute error (MAE)
- function developed to get movie recommendations using SVD MAE: 0.5781 RMSE: 0.7500

Content-based filtering

- The TF_IDF and cosine similarity to recommend movies based on combined features

Hybrid recommendation

- Hybrid recommendation system combines collaborative filtering (SVD) and content-based filtering. If the user has rated at least 5 movies, use collaborative filtering (SVD). If the user has rated less than 5 movies, use content-based filtering and recommend the top-rated movie. If the user has

not rated any movies, recommend the top-rated movies to address the cold start problem

6. Model improvement: Hyperparameter Tuning