Project B1: Anime recommendation database

Building anime rating predictor algorithm and analysing anime reviews

Task 1

Github repository: https://github.com/Tyrkson/AnimeDSProject

Task 2

Identifying your business goals

We don't have a business goal nor do we work for a business. So our goals are individual but also as the database name already says, anime watchers may benefit from this project. Our personal goal is to improve our understanding of data mining and write our first data-mining program based on real data.

Right now our experiences and knowledge are limited to this course's material. It might seem a lot but we have to take into account that we probably have forgotten most of the topics by now and have to re-understand them.

We'd say our success criteria is to pass this project successfully and accomplish our main goals related to this project (algorithms, graphs, implementations).

Assessing your situation

We are doing this project as a 2 member team Ahto Türkson and Bogdan Mihhailjuk. Also we can ask for advice from our instructors, may it be Anna Aljanaki or Victor Pinheiro. We also have the data from Kaggle (https://www.kaggle.com/CooperUnion/anime-recommendations-database?select=anime.csv) an are planning to use our every day computers to write the code. Also we both are members of the student organization which has an access to a premium account on canva.com. So we can use it to create a poster for our project.

Our requirements are pretty simple. We meet every Monday and work on this project together. We do our research if needed on other days, so we wouldn't waste too much time on it on Mondays. In order to say that our project is finished we need to create a poster which should showcase our doings visually. Also we have to implement all our goals and finally make sure that we fulfill the requirements made by our instructors.

Our biggest risk is our lack of needed knowledge to implement our goals. We plan to create 2 Machine learning (ML) algorithms which later on would predict outputs based on user input. For that we need to understand how ML algorithms work and how to clean the data for the specific algorithm. Another risk is time consumption. If one of our planned Mondays shouldn't be very productive then it may happen that we have to hurry in order to accomplish the criterias.

Terminology:

Machine learing (ML) algorithm - An algorithm where computer learns to predict the output itself by doing calculations on a training data inputs and outputs.

Anime - An animation movie which is usually made in Japan.

We don’t have any financial costs planned. Our only cost is time.

Our plan to present our data-mining goals is to create visual graphs and tables, and for ML algorithms our initial plan is to show numbers of how accurate our ML algorithm’s predictions are and some examples of the predicted outputs. If we have time then we may create a simple website where everyone could give their input to our ML algorithm and therefore test and see our ML algorithm’s predictions themselves.

Our data-mining success criteria is very wide. If we accomplish our main project goals then automatically we accomplish data-mining criteria too. We haven’t set any expectations to our ML algorithm’s accuracy besides that we hope it works and actually predicts something.

Task 3

Gathering data

We don’t have to worry about finding the correct data because our project goal was set on the data we already found on Kaggle.

The data really exists. If to open the Kaggle link mentioned above then we can see that there are 2 csv files and there's plentiful data in there for our project. We also have downloaded the data which means that we don’t have to worry about the data removal from the Kaggle page.

Both csv files (anime.csv and rating.csv) are needed for our project and also every field is needed too because our ML algorithms need as many fields as possible to predict accurately. We only don’t need the rows where fields aren’t fully completed or are missing some data, and due to that we are also doing data cleaning in our project.

Describing data

anime.csv file has 12292 entires and 7 fields

- anime_id - myanimelist.net's unique id identifying an anime.
- name - full name of anime.
- genre - comma separated list of genres for this anime.
- type - movie, TV, OVA, etc.
- episodes - how many episodes in this show. (1 if movie).
- rating - average rating out of 10 for this anime.
- members - number of community members that are in this anime's "group".

rating.csv has 7.81 million entries and 3 fields

- user_id - non identifiable randomly generated user id.
- anime_id - the anime that this user has rated.
- rating - rating out of 10 this user has assigned (-1 if the user watched it but didn't assign a rating).

Exploring data

We already are quite familiar with the data because we have implemented 1 ML algorithm, and what we noticed is that in anime.csv file on fields genre, type, episodes are cases where one of these field values is nan or unknown. It's a little annoying that these fields aren't defined the same way like with 'NaN' but instead they have done this in two different ways which therefore means we have to do more data validation checks.

But in file rating.csv everything seems to be clean, no NaNs or unknowns so every field has some value and the value is in the expected type.

Our data is already verified, and we can be 100% sure that it exists and it works. Plus we have downloaded it into our computers which means that even if Kaggle removes the Anime dataset for some reason then we still can continue with our project and accomplish our goals.

Task 4

Tasks

1. Create an algorithm, that takes looks at user's watch history and tries to recommend an anime to this user, based on other similar user's watch history. [6 hours per face]
2. Create an algorithm, that takes an anime, looks at it's genres, type, number of episodes and size of fanbase and tries to predict, what it's rating could be. [6 hours per face]
3. Make a list of animes, that got good rating from the users, who usually rate very critically. [3 hours per face]

   (All of the above tasks are done together, for more efficient cooperation and quicker results)

4. Make a graph with the correlation between genres and rating. [1 hour Ahto]
5. Make a pie chart of anime genres. [2 hours Bogdan]
6. Find out how many % of users don't rate any animes. [1 hour Ahto]
7. Find out how many times are animes watched without being rated. [1 hour Bogdan]
8. Find out what % of animes have rating above a certain number. [1 hour Ahto]
9. Finish the poster with all the facts/graphs from above [6 hours per face]

Methods

Our plan is to use python3 and we currently aren't using Jupyter notebook

Our main Python libraries which we are using or are planning to use are Pandas, Numpy, Matplotlib, Seaborn, scikit-learn