business_understanding

August 11, 2025

1 Anime Recommendation System

Author: Delano Francis
Student Id: 620109194
Course Code: COMP6830
Course Title: Capstone Project II

1.1 Domain - Anime

Anime is a style of animation which originated in Japan. It has become a popular form of entertainment among children and young adults. There is a variety of genre and because of this it provides a source of entertainment for all interests.

It has been around for a number of years with the earliest documented creation being in 1917 (Yamaguchi, 2013). Initially, anime or Japanimation as it was referred to was not popular outside of Japan, but over time this changed. The production has risen to such levels that every season of the year sees new adaptations from manga, the source material for many anime, or outright anime originals – those that have no manga from which they are adapted. With the sheer volume of content out there, it can be difficult to know which anime would suit each individual.

So, even though there are many streaming platforms or many avenues by which one can view anime, it can be very time-consuming to try to filter through all the content to find the best anime for the individual. When he does find one, it is only a temporary reprieve until he must begin the search again. Thus, we see a cycle of searching, trying, liking and searching once that series has been completed. Note that this only includes the happy path. Undoubtedly, there are many times when an iteration leads to frustration instead of satisfaction.

Therefore, the need arises for a solution that would bring the individual closer to content more suitable for his particular taste and limit the amount of time spent searching for new content. It would also provide the additional benefit of having the individual spend more time engaged in viewing the right content which leads to a better experience overall. This recommendation system would also benefit the providers of the anime as their users would be more satisfied whenever they use the service and more likely to recommend the service to others.

1.2 Business Understanding

1.2.1 Business Objectives

1. To provide a more streamlined experience for anime consumers looking to choose the next series to pick up based on their own preferences and preferences of users who liked similar anime.

2. To enable recommendations to be made to potential anime converts who may be new to the medium and wish to watch something with a particular storyline as their gateway anime.

1.2.2 Assessment of the situation

Data availability

1. There are 3 data sources taken from Kaggle containing anime reviews, anime data and ratings of the anime by users. These all cite myanimelist.net as the source of the data.
2. There are a number of attributes describing each anime including title, genre, episode count among others.

3. The anime available only goes up to the year 2020.

4. The unoffical API for the website (myanimelist.net) will be used to fill in the gaps for any missing data points that cannot be reconciled across the 3 datasets.

5. The reviews data was corrupted. It had to be collected directly from the site using the unofficial API.

Users of the output The output will be 3 models used to make recommendations for the user. These will be consumed by the application developers for the recommendation website.

Technical Constraints

1. Deployment will be done using AWS and Vercel.
   - This includes the deployment of the microservice responsible for using the models to make the recommendations
   - the models themselves which will be stored using AWS S3
   - the application server which will manage the business logic of the website and make calls to the microservice in order to make the recommendations
   - the user interface which will be accessed using a browser
2. Training and modelling will be constrained by the physical computer resources
   - 1 TB SSD
   - 32 GB RAM
   - 12th Gen Intel(R) Core(TM) i5-12400
   - AMD Radeon RX 6600
3. Flat files (csv), will be used in storing the datasets

Risks and Contingencies

1. The physical resources outlined above may not be enough to process the data. University resources or cloud based solutions may need to be utilized in order to supplement the physical resources.

1.2.3 Data Mining Goals

1. Use user-based collaborative filtering to recommend the 5 top rated anime in the user's cluster that he has not yet seen.

2. Use association rule mining to recommend anime based on the ones the user already enjoyed
3. Use Latent Semantic Analysis to recommend the 5 most relevant anime whose plots match a prompt provided by a user

1.2.4 Project Plan

Milestones

Data Exploration Explore the various datasets to see the useful features for the recommendation system.
Determine the size of the datasets
Determine the amount of missing data

Data Cleaning and Preparation Remove unnecessary features from the dataset.
For the collaborative filtering, the important features will be the ratings supplied by users for the anime they have seen.
For the topic modelling the useful features will be the synopsis and the reviews data. The reviews data will be used to supplement the synopsis data. This is possible as reviews by necessity need to mention the plot of the anime that is being reviewed.

Database creation The database will be created mainly to store the data retrieved from the datasets in a more permanent form. They will also be beneficial for the creation of the web application as the anime data is what the user will interact with.
A relational model will be used to store the anime data and the rating data with the exception of the reviews provided.
The scripts will be created and then used to populate the tables.

Data modelling

1. Association Rule Mining - this will be done with frequent itemsets at different support values. The rules will then be sorted by confidence values.
2. Clustering (user-based) - this will be used to group users according to anime they rated. In particular, k-means clusters will be used.
3. Latent Semantic Analysis will be used for topic modelling

Models will be evaluated based on precision and recall. Precision will be the more important metric as the focus is on making recommendations the user will be more likely to enjoy.

Microservice for recommendations Once the models are created, a microservice will be developed to make the recommendations via its APIs.
The microservice will be built using flask. Flask gives easy access to the many data science utilities in the python ecosystem and because it is minimalistic, it can be kept as light-weight as possible.

Application server The business logic of the website will be achieved through a spring boot application server. It will be responsible for relaying all requests to the microservice as well as retrieving data from the database for the application. It will facilitate the creation of users as well as the rating of an anime.

User Portal It will be built using nextjs. It will allow the user to see the top picks based on anime the user already rated. The user will be able to see the anime catalogue. He will see the anime he has already rated. He will also be able to see recommendations based on a prompt he provided.

Tools and Platforms

1. VSCode, Spyder and IntelliJ will be used for all the tasks.
   - VSCode and Spyder will be used for preliminary exploration as well as the data preparation and cleaning
   - VSCode will be used to prepare the jupyter notebook, code the microservice as well as the user portal
2. MySQL and mongodb are the databases used in storing the structured and unstructured data.
3. MySQL is also the database used for the website that is to be developed.
4. Python, Java and JavaScript are the programming languages to be used throughout the project.

Evaluation Metrics

Association Rule Mining Support - the fraction of the total transactions that include the itemset. This will be used to determine the frequent itemsets.
Confidence - the fraction of times the items in the consequents appear in transactions that contain the antecedents. These will be used to rank the relevance of the recommendation.
Lift - used to determine the association among itemsets. A value greater than 1 implies a positive association while a value less than one implies a negative association and exactly 1 means no association. We will start with a threshold of 1 for the lift.

Clustering K-means Inertia measures how internally coherent a cluster is - how far away from the centroid the points are. As the number of clusters increases, the inertia will fall so this will need to be balanced with other metrics.

The silhouette score will also be used as it balances out inertia well. Silhouette score also assesses how far the points are from other clusters in addition to how close the points in a cluster are to its centroid.
A score of 1 is the ideal as it indicates that the points are well clustered.
0 means there are overlapping clusters.
A negative score means a lot of points are not in the correct cluster.

The Davies-Bouldin Index will also be used to assess the quality of the clusters. This measure is used to determine how well spaced a cluster is in comparison to its size. A lower value is better as it means clusters are more compact and the distance from the centroid is appropriate. Higher values may imply that the clusters are too spread out or not well separated.

Topic Modelling Domain knowledge will be used when manually checking the recommendations made by the model based on the input provided.

The topic space will also be looked at to determine the quality of the model. Better topics will lead to better recommendations.

Similarity distribution plots will also be used to assess the quality of the recommendations. If the plots are too flat this indicates that the model is not able to distinguish between similar content and disparate content. The plot should show bunching around the lower positive values and tapering towards the higher positive values. This shows that only a few data points match the prompt provided.

Quality of Recommendation Precision, recall and coverage will be used where possible to evaluate the recommendations made by the models.
Precision gives the percentage of relevant recommendations out of all recommendations made.
Recall gives the percentage of relevant recommendations out of all relevant anime.
Coverage gives the percentage of the entire catalogue that is recommended. This can be used to see if the recommendations are just focussed on the more popular anime.

Precision and coverage will be the more important measures as we want to maximize the relevant recommendations the user sees and not just focus on the most popular anime in our recommendations.

Reporting Plan Results of the process will be documented and shared in pdf form. Check-ins will be done based on the accomplishment of a milestone.