Anime Recommendation Engine

Yilu Chen, Andrew Gatchalian, Sara Hart, Hsuan-Yi Lin, Rakesh Venkata Subramaniyan

BANA 212 – Group 12A

University of California, Irvine

Introduction

Over the past few decades, Japanese animation has successfully emerged as one of the most influential and rapidly growing entertainment industries in the world. This global phenomenon has captivated audiences of all ages and backgrounds with staples like *One Piece* and *Naruto*, to hit newcomers like *Attack on Titan* and *Jujutsu Kaisen*. From streaming platforms, gaming, merchandising, live-action films, and conventions, the anime industry today is worth over $20 billion dollars.[1] This remarkable growth has naturally led to a surge of online communities dedicated to the artform.

This project focuses on analyzing data on anime and viewer preferences. One of the largest online anime forums is MyAnimeList, a community-driven platform which allows users to track, rate, and review anime and manga. The platform tracks data on thousands of anime titles and nearly 14 million user profiles, their ratings, reviews, and more.[2] Despite anime's mainstream success in recent years, many outsiders still consider the artform a niche form of entertainment. Western audiences, in particular, may not be accustomed to the foreign media due to its unfamiliarity, language, and intimidating barrier of entry. This project aims to explore the data within the world of anime to reveal general trends amongst shows and users on the MyAnimeList platform. Additionally, we will attempt to give accurate recommendations based on historical viewing data. By conducting this project, we seek to gain insights into the preferences of anime enthusiasts, promote accessibility for newer viewers, and offer existing fans a more personalized anime viewing experience.

Related Works

We found the dataset "Anime Recommendations Database" on Kaggle that includes detailed information for over 12,000 anime titles.[3] This dataset could be a blueprint to how we conduct our own data mining for the project. It includes essential attributes like genres, viewer ratings, and popularity, which are crucial for building a robust recommender system. By analyzing this dataset, we can identify key factors that influence anime preferences, aiding in the refinement of our recommendation algorithms.

Another related project found on Kaggle is "Content Based Anime Recommender!". This project used KNN to find similar anime titles based on an input anime title.[4] The recommender returns a list of animes that are modeled to be similar to the given input. This project demonstrates how specific attributes of an anime can be used to find titles with similar characteristics. In our project, we will explore a similar machine learning approach using KNN to arrive at a similarity-based learning model.

Data Retrieval

To commence the project, we retrieved data by utilizing the MyAnimeList Application Programming Interface (API). The API enables developers to access and interact programmatically with the MAL database and its functionalities, allowing external applications or services to retrieve information from MAL, such as anime details, manga details, user lists, reviews, and more. Although, multiple datasets from MyAnimeList can be found on the Kaggle platform encompassing similar data, it is important to acknowledge that a significant portion of these datasets have not been updated for a duration spanning 5 to 7 years. Understanding this discrepancy prompts us to conduct a more contemporary analysis, inclusive of recently released shows and users who have exhibited recent activity. This approach ensures a more current and comprehensive perspective in our analysis, accounting for the latest developments within the anime community.

Our strategy is to pull the data of users and anime from the website and store the collected data in a data frame which is the anime list and user ratings. We later stored them into our local file to process the data and perform machine learning algorithms. The data acquisition process involved three distinct procedures. First, we accessed a comprehensive list of available anime titles from the MyAnimeList website. Subsequently, we gathered usernames by scraping data from the website to ensure a sufficiently large sample size. Finally, we use the scraped usernames in conjunction with MyAnimeList API to pull the user ratings.

To begin, each of our team members created an account on MAL and obtained a Client ID to use the OAuth 2.0 protocol to authenticate tokens and access the data. Due to our team's composition of five members, we opted to allocate 10,000 segments of data to each member. This division was chosen to facilitate data processing within our designated timeframe, considering the substantial number of anime IDs, which approximate around 50,000 according to the website's records. Throughout the process, the API encounters 403 errors, commonly known as "forbidden" errors if calls are made too frequently. To mitigate this issue, we implemented a 3-second delay within our code, which enabled us to circumvent encountering this problem. Also, if there were titles with missing information, we also designated “-1” as well as changing the dataframe “studio” to “studios” for continuity. After each team member ran their code, saving the resulting file into our shared folder and subsequently appending it to consolidate the data, we then exported the CSV file. From this, we were able to obtain a comprehensive list of all the available titles on the MyAnimeList site.

Secondly, we conducted web scraping to gather usernames from the website, aiming to generate a substantial sample size. Our project's objective involves predicting anime recommendations based on other users' ratings, thus, we also needed the user list. However, utilizing the entire user base of 14 million on MyAnimeList would result in excessive time consumption and storage requirements for this project. Consequently, we made the strategic choice to employ a smaller sample size consisting of the most recently active users, ensuring more efficient and accurate decision-making processes. We utilized the “BeautifulSoup” library to navigate and extract information from the website. By utilizing the URL provided within our code, which generates a random list of recent users upon refreshing the site, we gathered approximately 20 users with each page refresh. By using the BeautifulSoup library, we parsed through the HTML content, specifically targeting the sections where usernames were listed within the HTML structure. This process allowed us to collect as many usernames as required. After the process and removing the duplicates to make sure each user was unique, we stored each member's lists in a CSV file with approximately 70,000 usernames.

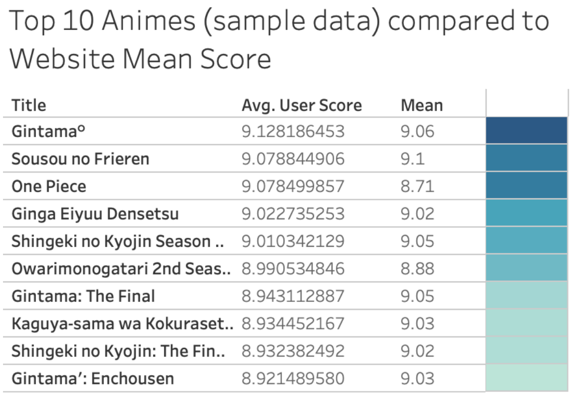
Now that we have obtained a list of usernames and all available anime titles through web scraping, we proceed to leverage this data in conjunction with the MyAnimeList API to extract user ratings. We imported the CSV file of usernames that we created, then ran through a loop to iterate through each “username”. Then, used the API to call on the rating lists for each user. We repeated this process until all the usernames were processed. Considering the limitations in time and storage capacity, we initially incorporated 20,000 usernames into our model and exported the data as a file. However, due to privacy settings and restricted access to certain user accounts that were named as “private accounts”, our final dataset comprised only 18,145 unique usernames.

Data Cleaning

Extensive cleaning was undertaken before implementing our model to enhance the quality of our datasets. Given the size of some of our data, we were required to chunk them into smaller files initially, before later combining them into one comprehensive data frame (and CSV). We formatted each of the columns by removing brackets and quotations (that were included in the raw data) and replaced spaces with ‘_’ to ensure readability. We dropped duplicate rows and any NA values from our data and created dummy variables for several features. For consistency, we made specific assumptions during the cleaning process. Since MyAnimeList tracks the progress of a user on a given show, we opted to only keep rows that were marked as “Completed” and “Dropped”. Additionally, we removed any instances of “0” ratings as we assumed this could mean either the show was not watched, or the user simply forgot to give the show a rating. As we continued with the cleaning process, we noticed another discrepancy. Multiple ratings of the same show would appear from a single user. The site happened to also track every instance of ratings that users would place on shows, thus, if someone decided to rate a show multiple times, each instance would appear in the dataset. To address this, we only kept the most updated rating for each user-anime pair using the “updated_at” column that provided a timestamp of when each rating was made.

Exploratory Analysis

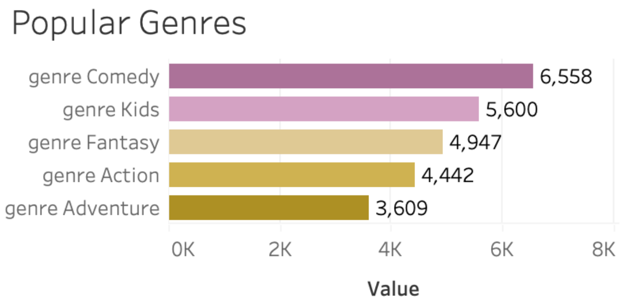
Following the completion of data pre-processing, our final data set comprises of over 44 million rows of anime-related data, representing interactions from 18,145 unique users and 16,135 unique shows. Before implementing our model, we undergo an exploratory analysis to gain initial insights on the data’s characteristics and ensure its robustness, considering the sample size limitation.



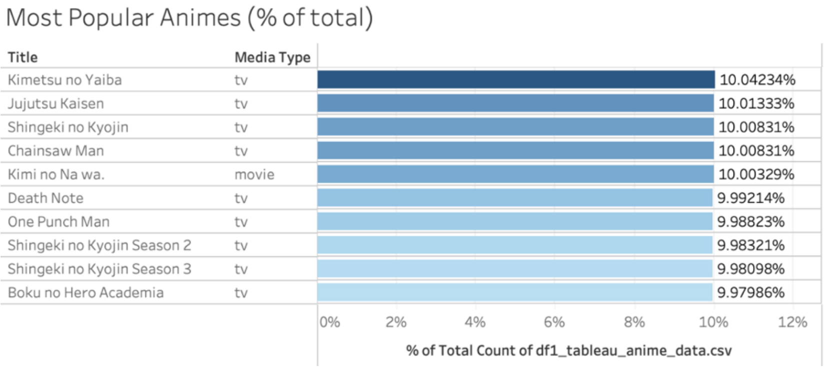
Top 10 Animes (sample data) compared to Website Mean Score

Title	Avg. User Score	Mean	
Gintama°	9.128186453	9.06	■
Sousou no Frieren	9.078844906	9.1	■
One Piece	9.078499857	8.71	■
Ginga Eiyuu Densetsu	9.022735253	9.02	■
Shingeki no Kyojin Season ..	9.010342129	9.05	■
Owarimonogatari 2nd Seas..	8.990534846	8.88	■
Gintama: The Final	8.943112887	9.05	■
Kaguya-sama wa Kokuraset..	8.934452167	9.03	■
Shingeki no Kyojin: The Fin..	8.932382492	9.02	■
Gintama’: Enchousen	8.921489580	9.03	■

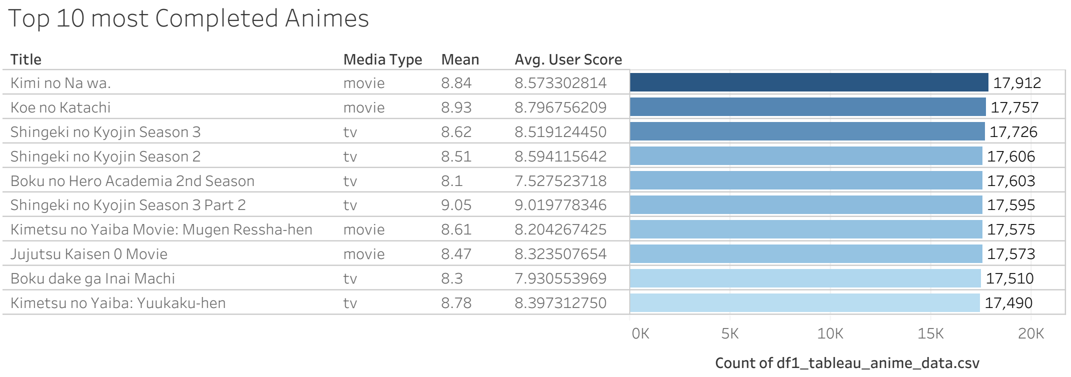
By extracting the top 10 animes based on average user ratings (minimum of 1,000 ratings), we find several noteworthy titles from classic shows like *Gintama* (2015) and *One Piece* (1999), to highly acclaimed newcomers such as *Shingeki no Kyojin* (2019) and *Sousou no Frieren* (2023). A common characteristic between these shows being that all happen to rank amongst the top 50 shows on the MyAnimeList site. In fact, by comparing the top 10 shows in our sample data to their mean aggregate scores on the website, we find that each score is actually relatively similar. The close proximity of our sample scores to the site's scores suggests that despite our sample size being limited, the dataset still manages to capture the essence of user preferences with some level of consistency.



As we continue to explore the data, we find that the most popular genres fell under Comedy, Kids, Fantasy, Action, and Adventure; which make sense since a large demographic of anime viewers are teens and young boys. The most popular titles in our data seem to have around a 10% share between all user ratings. These titles include many of anime's biggest shows such as *Kimetsu no Yaiba (*2019), *Jujutsu Kaisen* (2020), and *Death Note* (2006). The most recently released show on this list *Chainsaw Man* was completed in late 2022. Although the show is popular, it can also be noted that the method in which we have collected this data (based on recent users) may influence how recently released shows are perceived.

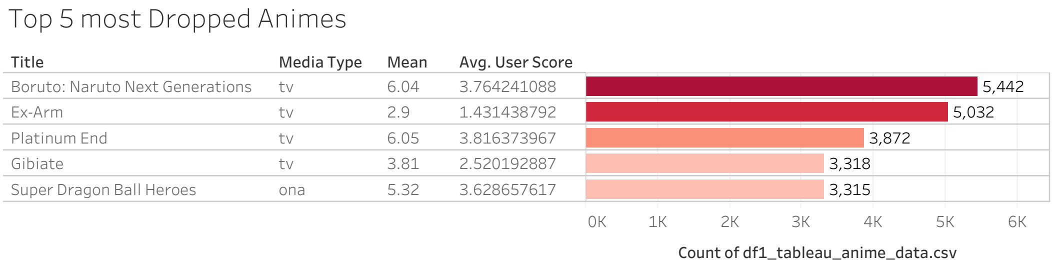


As we discussed previously in our data cleaning section, we opted to only keep shows that were marked as "Completed" and "Dropped", as an indication that user had some conscious familiarity with the show on their lists. The most completed titles naturally were some of the most popular animes in our data. We see more instances of movies in the figure below, likely because movies are easier to "complete" as opposed to television series. By comparing the average user score to the mean, we once again find some consistency in the user ratings. However, when examining some of the most "dropped" animes in our data, we find a large discrepancy between ratings. A dropped show indicates that a user had attempted to complete the show and for whatever reason decided to stop watching it. The average user score for some of the most dropped shows is consistently below their mean score on the website, indicating that users in our data have much stronger opinions about shows that they dislike.



Top 10 most Completed Animes

Title	Media Type	Mean	Avg. User Score	Count of df1_tableau_anime_data.csv
Kimi no Na wa.	movie	8.84	8.573302814	17,912
Koe no Katachi	movie	8.93	8.796756209	17,757
Shingeki no Kyojin Season 3	tv	8.62	8.519124450	17,726
Shingeki no Kyojin Season 2	tv	8.51	8.594115642	17,606
Boku no Hero Academia 2nd Season	tv	8.1	7.527523718	17,603
Shingeki no Kyojin Season 3 Part 2	tv	9.05	9.019778346	17,595
Kimetsu no Yaiba Movie: Mugen Ressha-hen	movie	8.61	8.204267425	17,575
Jujutsu Kaisen 0 Movie	movie	8.47	8.323507654	17,573
Boku dake ga Inai Machi	tv	8.3	7.930553969	17,510
Kimetsu no Yaiba: Yuukaku-hen	tv	8.78	8.397312750	17,490



Top 5 most Dropped Animes

Title	Media Type	Mean	Avg. User Score	Count of df1_tableau_anime_data.csv
Boruto: Naruto Next Generations	tv	6.04	3.764241088	5,442
Ex-Arm	tv	2.9	1.431438792	5,032
Platinum End	tv	6.05	3.816373967	3,872
Gibiate	tv	3.81	2.520192887	3,318
Super Dragon Ball Heroes	ona	5.32	3.628657617	3,315

KNN Analysis

We implemented K-Nearest Neighbors (KNN), a machine learning algorithm designed to predict instances based on the majority class or average of its k-nearest neighbors. For our project, KNN becomes a tool for personalized recommendations by leveraging the historical preferences (ratings) of users on the MyAnimeList site. By targeting a specific anime, KNN can identify other titles with similar user ratings and recommend animes that are frequently rated highly by the nearest neighbors of the target anime. Using our cleaned data, we create a user-item matrix with anime IDs as rows, user IDs as columns, and user scores as values. We then utilize the 'scikit-learn' library and 'NearestNeighbors' with cosine similarity as the metric and the brute-force algorithm to train our model. Additionally, we experiment with Z-score normalization to improve our model's performance and incorporate genres as dummy variables to give us more nuanced results.

Our analysis revealed intriguing insights into the functionality of our model. Z-score normalization seemed to improve accuracy, however, incorporating genres as dummy variables did not have a significant effect due to the nature of our model. Popular titles with a substantial number of ratings seemed to exhibit more stable and consistent neighbors, suggesting that the recommendations for extremely popular shows were less sensitive to changes. On the contrary, titles with mid-low popularity demonstrated greater sensitivity due to the impact of normalization. After multiple trial runs (as documented in our code), our model was able to achieve some level of accuracy from the recommendations, indicating a consistency with the way that users rated animes. For example, for the classic anime film *Akira* (1988), recommended titles include: *Cowboy Bebop* (1998), *Koukaku Kidoutai* (1955), and *Perfect Blue* (1997); all critically acclaimed animes that fall under genres related to *Akira* such as cyber-punk, sci-fi, and psychological. Though the accuracy of these recommendations is subjective, it should be noted that the titles being recommended are not necessarily similar by their content, but similar in how our model has perceived how users enjoy them. In fact, any resemblance in the content being recommended simply implies that users who enjoy the targeted anime tend to also enjoy other titles that are similar in content.

Conclusion

By leveraging data from MAL and implementing a KNN model, we are able to derive valuable insights into viewer preferences on the MyAnimeList platform, providing personalized anime recommendations based on historical user data. However, despite this success, it is still crucial to acknowledge our model’s limitations. The dataset, constrained by time and memory, only captures a sample of around 18 thousand users. Sensitivity in our model due to the sheer popularity of certain titles, alongside recency bias, can potentially lead to inaccurate results. Future improvements can involve increasing our sample size, exploring more sophisticated algorithms, and expanding the training data to include additional features such as episode count and animation studio. The model, which currently recommends titles based on a target anime, can inversely be applied to users to recommend shows based on other users who have similar taste. In summary, this project allowed us to explore one of the methods of how recommendation engines are built using a fun and interesting topic. Our model’s effectiveness in generating recommendations exhibits the potential for leveraging machine learning to further enhance the viewing experience within the world of anime.

Bibliography

1. Kim, S. (2023, July 13). Crunchyroll Eyes India as Japanese anime becomes $20 billion industry. Bloomberg.com. https://www.bloomberg.com/news/articles/2023-07-13/crunchyroll-eyes-india-for-growth-as-japanese-anime-becomes-20-billion-industry
2. Happy New Year &amp; 2022 wrap-up!. MyAnimeList.net. (2023). https://myanimelist.net/forum/?topicid=2068385
3. CooperUnion. (2016). Anime recommendations database. Retrieved from https://www.kaggle.com/datasets/CooperUnion/anime-recommendations-database/data
4. Bertmanila. (2017). Retrieved from https://www.kaggle.com/code/bertmanila/content-based-anime-recommender