ANIME RECOMMENDATION SYSTEM

A PROJECT REPORT

Submitted By:

Sai Chandra Kaushik Siddavarapu – ss17413

Sidhartha Reddy Potu – sp7835

NEW YORK UNIVERSITY

May 8, 2024

ABSTRACT

The Anime Recommendation System project is leading the way in transforming how people find and engage with anime content by leveraging big data and cutting edge machine learning techniques. Traditional recommendation engines have struggled to keep up with the exponential growth in the volume and variety of anime available worldwide, frequently falling short of capturing the complex preferences of a diverse viewership. In order to overcome these obstacles, this system analyzes large datasets that contain specific anime features, demographic data, and user preferences in depth.

With the use of a strong framework that combines content-based filtering, collaborative filtering, and sophisticated algorithmic recommendations through ChromaDB, the system is able to produce highly customized anime recommendations. While content-based filtering matches users with anime titles that share qualities with previously enjoyed ones, collaborative filtering examines user behavior patterns to suggest content enjoyed by similar profiles. Our innovative ChromaDB component uses complex algorithms to dynamically modify recommendations in response to real-time user interactions. By providing suggestions that are more closely aligned with personal preferences, this all-inclusive method not only increases user engagement but also dramatically raises user satisfaction levels by providing timely and highly relevant content. The end effect is a transformative experience that gives each viewer a unique perspective on the vast anime landscape, facilitating a smooth and enjoyable journey of discovery.

INTRODUCTION

With the explosion in popularity of anime worldwide, a wide range of genres and styles have become accessible to viewers everywhere. Finding titles that suit their individual tastes and preferences becomes more difficult as the amount of content available to viewers rises. Through the use of advanced algorithmic techniques and big data analytics, the Anime Recommendation System seeks to address these issues.

Using cutting-edge technologies, this project creates a recommendation system that not only comprehends users' explicit preferences but also analyzes their viewing patterns to recommend content that they are likely to find entertaining. Through the integration of multiple data points, including viewing history, genre preferences, user ratings, and demographic information, the system creates a comprehensive profile for every user. Through the use of data-driven methodology, the system is able to provide highly personalized recommendations for anime, greatly improving user experience by streamlining and entertaining the discovery process.

Furthermore, the architecture of the system is intended to be adaptive, continuously improving and honing its predictions through learning from user interactions. This guarantees that the suggestions continue to be applicable not only within the framework of the user's past data but also in reaction to changing preferences, keeping up with the latest releases and trends in anime. In this way, the Anime Recommendation System represents a significant advancement in the field of content discovery, providing a scalable, intelligent solution that addresses the needs of a rapidly expanding global audience.

PROBLEM STATEMENT

Conventional anime recommendation systems frequently use oversimplified algorithms that are unable to fully account for the complexity of user preferences, producing suggestions that are frequently generic and disappointing. This mismatch is especially noticeable in the anime industry because of how specific user preferences are and how diverse the content is. Many of the systems in place can overlook the more complex, nuanced preferences of individual viewers in favor of making recommendations based only on simple correlational data.

The Anime Recommendation System is intended to use extensive big data analytics to create a more dynamic and customized recommendation engine in order to address these shortcomings. Through the use of sophisticated processing methods and extensive data collection, the system is able to examine a wide range of data, including minute behavioral indicators and demographic specifics. This enables a much richer understanding of user preferences and behavior, which in turn allows for the generation of tailored recommendations that are far more likely to resonate with users.

This project aims to close the gap that exists between the wide range of anime content available and the specific viewer by making sure that every recommendation is made with a thorough understanding of personal preferences. The system aims to significantly increase user satisfaction by making these suggestions more accurate and relevant. This will change the way viewers interact with anime content and help them discover shows and genres they love that they might not have otherwise discovered.

TECHNOLOGIES USED

Python: Python is a versatile programming language that's ideal for data manipulation and operations due to its powerful libraries, including Pandas for data analysis and NumPy for numerical operations. It provides a foundation for scripting and rapid prototyping of complex algorithms used in data processing and analysis.

PySpark: PySpark, the Python API for Apache Spark, is crucial for handling large-scale data processing. It allows for effective management of big data by leveraging distributed computing; data can be processed across multiple nodes, significantly speeding up the data analytics processes. PySpark is particularly useful for data cleansing, transformation, and aggregation tasks, which are essential in preparing the datasets for the recommendation algorithms.

ChromaDB: ChromaDB is not a traditional storage database but an advanced algorithmic framework designed to enhance recommendation systems dynamically. It uses sophisticated machine learning algorithms to analyze user behavior and preferences in real time, enabling the Anime Recommendation System to adapt its recommendations based on current user interactions and feedback without the need for storing personal data permanently.

Machine Learning: The backbone of the recommendation engine, machine learning techniques enable predictive modeling and complex pattern recognition. These techniques include various algorithms from simple linear regression to more complex deep learning models that can unearth intricate relationships within data and provide highly accurate recommendations based on user profiles and past interactions.

DATASET DESCRIPTION

The Anime Recommendation System uses three primary datasets that form the core of its functionality:

Anime Dataset 2023: This dataset provides comprehensive information about each anime, including title, genre, director, episodes, and ratings. It is essential for content-based filtering as it allows the system to identify and recommend anime that matches the user's preferences in genre or directorial style.

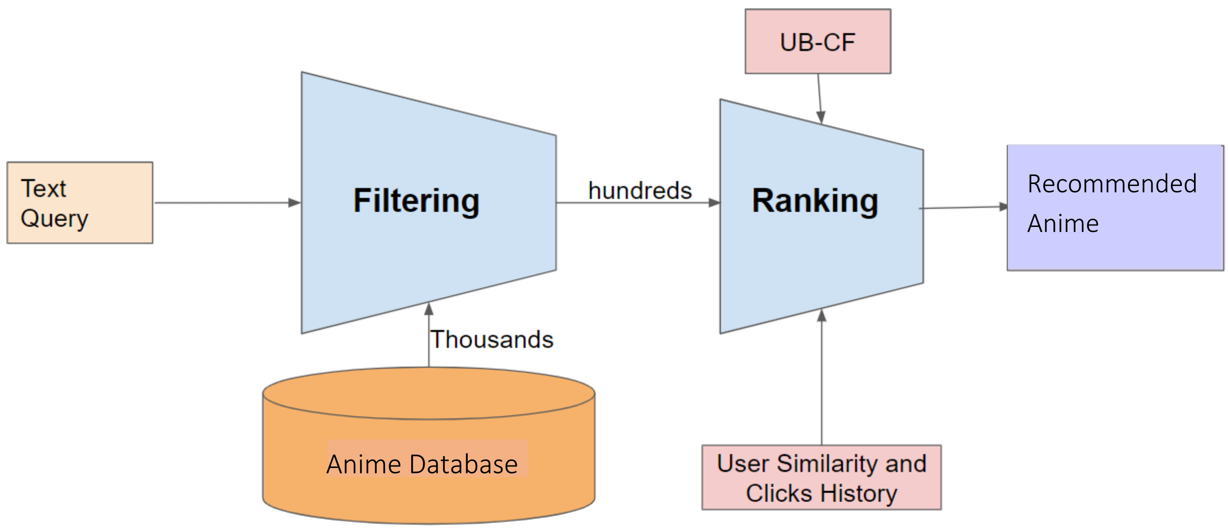
User Details 2023: Includes demographic information such as age, gender, and location, along with personal preferences and viewing habits. This dataset helps in creating a detailed user profile, which is critical for both collaborative and content-based filtering techniques.

User Score 2023: Consists of explicit ratings given by users to different anime titles. These ratings are invaluable for collaborative filtering as they help identify similarities and differences in user preferences, enabling the system to recommend anime liked by users with similar tastes.

Collectively, these datasets encompass over two million data points that provide a rich foundation for generating personalized anime recommendations. The vast data helps in understanding nuanced user preferences and content attributes, which is crucial for delivering accurate and relevant suggestions.

METHODOLOGY

The recommendation engine of the Anime Recommendation System is meticulously designed to cater to the unique tastes of each viewer by incorporating three distinct methodologies. Each of these methods plays a vital role in ensuring that the recommendations are not only precise but also highly personalized and responsive to real-time user activity.



Anime Recommendation System Architecture Diagram

ASL COLLABORATIVE FILTERING

ASL Collaborative Filtering stands as a cornerstone of our recommendation strategy. This method harnesses the power of user interaction data, such as ratings and viewing history, to create a sophisticated model of user preferences across our platform. By comparing users with similar viewing habits and preferences, the system can identify patterns and clusters of users who share tastes in anime. Once these groups are defined, the recommendation engine predicts what unseen titles might interest a user by analyzing the ratings and preferences of their 'neighbors' in this taste-space.

For example, if several users who liked "Naruto" also tend to enjoy "Bleach," then a new user who shows a preference for "Naruto" might receive a recommendation for "Bleach." This method is particularly effective because it leverages collective user behavior to make informed suggestions, thereby increasing the likelihood that recommended titles will resonate well with users.

CONTENT-BASED FILTERING

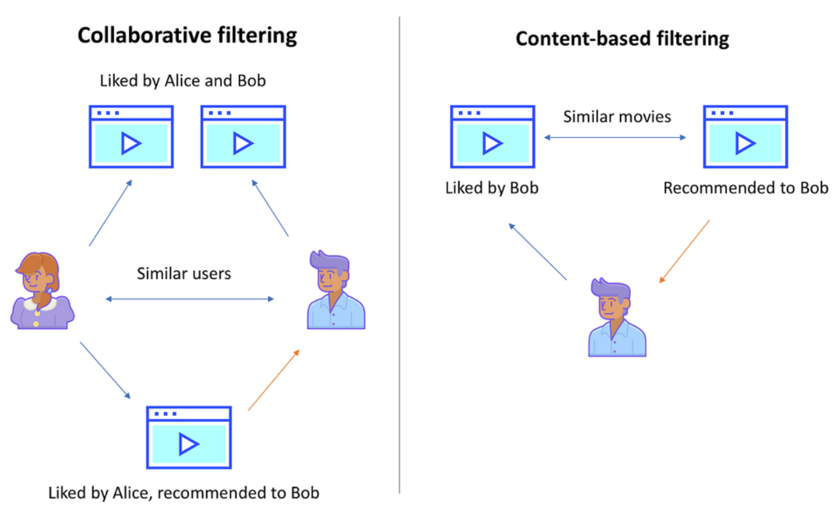
Whereas ASL Collaborative Filtering focuses on the relationships between users, Content-Based Filtering zeroes in on the attributes of the anime titles themselves. This approach examines the intrinsic properties of each anime—such as genre, director, artistic style, and narrative elements—to recommend new titles that share similar characteristics with those a user has previously enjoyed.

This method relies heavily on a detailed metadata analysis of each title in our database. For instance, if a user frequently watches and rates highly the mystery and thriller genres, the system will prioritize suggesting new anime that are classified under these genres or created by directors known for their suspenseful storytelling. By aligning new suggestions with a user's established preferences, Content-Based Filtering ensures that each recommendation feels personally curated and closely aligned with individual tastes.

CHROMADB SUGGESTIONS

ChromaDB represents the most dynamic aspect of our recommendation suite. This advanced component utilizes cutting-edge algorithms to process user data in real time, offering immediate and adaptive suggestions based on the latest interactions. As users interact with different titles—rating them, adding them to their watchlist, or writing reviews—ChromaDB continuously analyzes this input to adjust its recommendations accordingly.

This real-time processing capability allows ChromaDB to respond swiftly to shifts in user preferences and to introduce trending or newly added anime that align with a user's evolving interests. Whether a user's taste changes from action-packed adventures to more subdued, narrative-driven dramas, ChromaDB's agile framework can seamlessly transition its recommendations to suit these new preferences, ensuring that the content remains relevant and engaging.



System Architecture and Data Pipeline

The architecture of the Anime Recommendation System is thoughtfully designed to be both robust and scalable, ensuring that it can effortlessly accommodate the ever-increasing volumes of data and the growing user base. The system's architecture and data pipeline are structured to manage the complex flow of data through various stages, from initial ingestion to the delivery of personalized recommendations. Here's a detailed exploration of each stage:

Data Ingestion

Data ingestion is the critical first step in our data pipeline. During this stage, the system gathers information from a variety of sources, such as user interactions, metadata about anime, and user demographics. All incoming data is effectively captured and stored for later processing because the ingestion process is built to handle a wide range of data types and formats. The system uses automated scripts and APIs to continuously fetch and update data, guaranteeing that the dataset is up to date and reflects the most recent user interactions and content updates. This allows for a smooth flow of updated information to be maintained.

Data Preprocessing

Once data is ingested, it moves into the preprocessing stage where it is cleaned and structured using PySpark, a powerful tool for handling large-scale data processing. This step is crucial for ensuring the quality and usability of the data within the recommendation algorithms. Preprocessing includes several key activities:

Null Value Management: The system identifies and addresses missing or incomplete data entries, which are either filled using statistical methods like mean imputation or removed, depending on their significance and volume.

Data Standardization: To ensure consistency across the dataset, all data is standardized. This includes normalizing ratings scales, unifying date formats, and categorizing textual data, which facilitates more effective analysis.

Feature Extraction: This involves deriving new data features that are likely to enhance the recommendation process, such as calculating user engagement levels or extracting thematic keywords from anime descriptions.

These preprocessing tasks are critical for transforming raw data into a refined format that is ready for detailed analysis and recommendation generation.

Recommendation Generation

At the core of the system is the recommendation generation process, which utilizes the three primary methodologies—ASL Collaborative Filtering, Content-Based Filtering, and ChromaDB Suggestions—to produce personalized anime recommendations. This stage involves several important mechanisms:

Model Training: Using the preprocessed data, various machine learning models are trained to understand and predict user preferences. These models are regularly retrained and updated with new data to improve their accuracy and relevance.

Real-time Processing: To support ChromaDB's dynamic recommendations, the system is equipped with capabilities for real-time data processing. This allows the recommendation engine to instantly analyze user actions, such as ratings or reviews, and adjust the recommendations it generates on the fly.

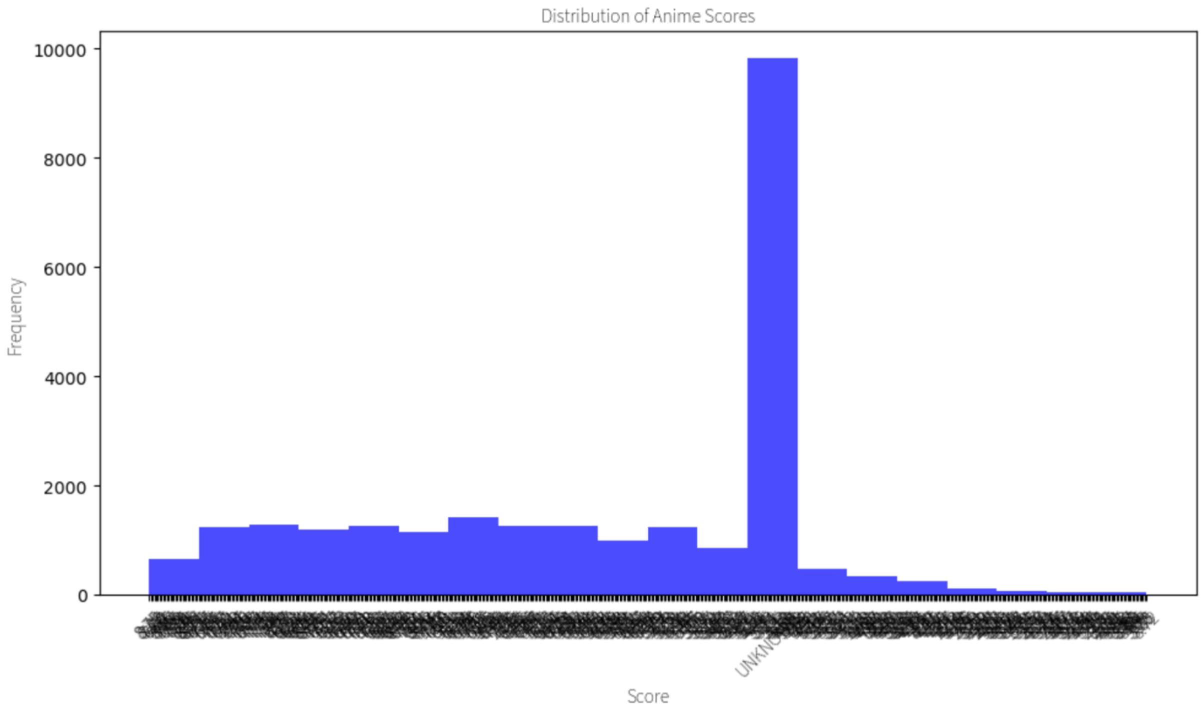
To ensure that ChromaDB recommendations are both timely and relevant, the system's architecture is specifically engineered to support real-time data processing. This is facilitated by high-performance computing resources and efficient data streaming technologies that allow instantaneous data analysis and processing. The real-time capability is essential for adapting to the rapidly changing preferences of users and for integrating the latest anime content into the recommendation pool almost instantaneously.

This robust system architecture not only supports the efficient flow of data through each stage of the pipeline but also ensures that the Anime Recommendation System can scale effectively as the platform grows. This scalability is critical for maintaining performance and service quality as the number of users and the volume of data continue to expand.

Exploratory Data Analysis (EDA)

The primary goal of Exploratory Data Analysis in the Anime Recommendation System is to thoroughly examine the collected datasets to uncover underlying patterns and crucial characteristics that inform the recommendation process. This analytical phase is pivotal for understanding the complexity of user interactions and the multifaceted nature of anime content.

Techniques and Tools Used: The EDA utilizes Python libraries such as Pandas for data manipulation and Matplotlib and Seaborn for visualization. These tools help in plotting distributions, creating correlation matrices, and conducting statistical tests that provide deeper insights into the data.



Distribution of anime Score

Genres	count
unknown	4179
comedy	1579
fantasy	1068
hentai	574
avant garde	438
slice of life	391
drama	364
adventure fantasy	357
comedy fantasy	289
comedy slice of life	283

only showing top 10 rows

Genres	avg(Score)
action adventure ...	8.670000076293945
action adventure ...	8.264999866485596
adventure award w...	8.25
mystery supernatu...	8.180000305175781
drama slice of li...	8.176666736602783
adventure comedy ...	8.039999961853027
drama romance sports	8.0
comedy drama myst...	7.980000019073486
action award winn...	7.940000057220459
action adventure ...	7.900000095367432

only showing top 10 rows

Genres Count & Avg. Score

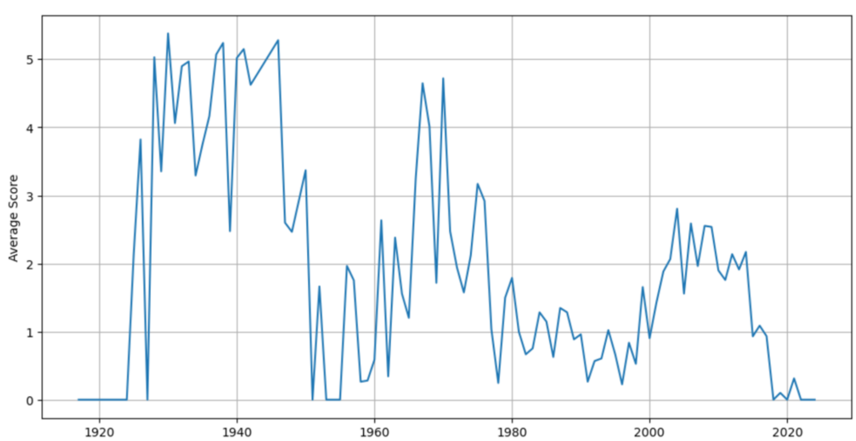
Gender	Avg Days Watched	Avg Completed	Count
Non-Binary	85.27866520787748	288.7757111597374	914
Male	65.49853680049134	187.51603376856562	126984
Female	38.55076648804405	103.64841053493507	96485
Unknown	10.985404226021762	27.92483236570022	506907

Table Showing Avg watch Statistics Based on Gender

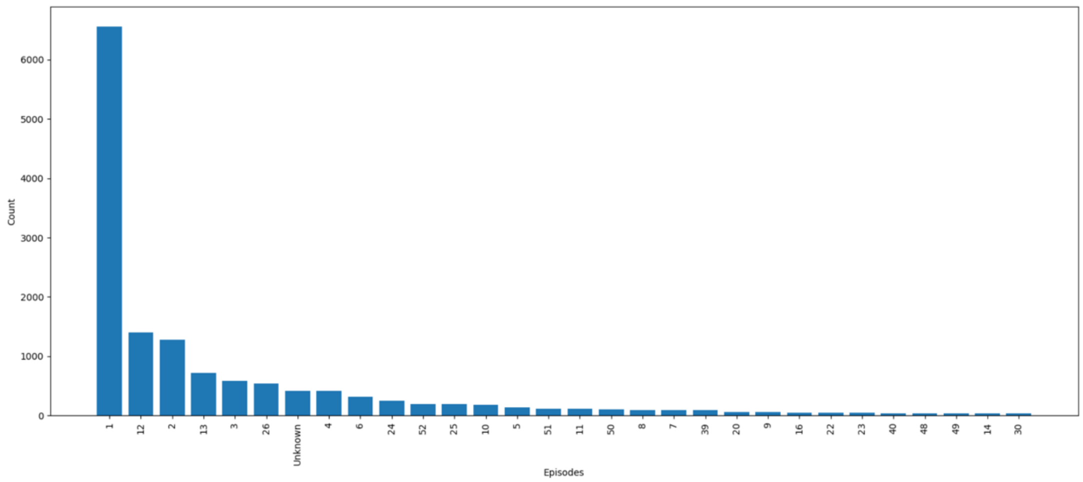
Source	count
Manga	3764
Original	3720
Unknown	2974
Visual novel	988
Game	829
Light novel	768
Novel	485
Other	444
4-koma manga	277
Web manga	237
Music	215
Book	87
Picture book	76
Card game	64
Digital manga	15
Radio	9

Name	Score	Popularity
Fullmetal Alchemi...	9.19	3
Shingeki no Kyoji...	9.17	119
Steins;Gate	9.11	9
Hunter x Hunter (...	9.1	12
Shingeki no Kyoji...	9.1	63
Gintama°	9.1	329
Gintama'	9.08	362
Ginga Eiyuu Densetsu	9.07	633
Gintama': Enchousen	9.04	660
Koe no Katachi	9.0	28
3-gatsu no Lion 2...	9.0	545
Gintama.	8.99	730
Gintama	8.96	114
Gintama Movie 2: ...	8.96	873
Clannad: After Story	8.96	84
Kimi no Na wa.	8.96	11
Owarimonogatari 2...	8.93	521
Code Geass: Hangy...	8.91	36
Gintama: The Final	8.88	2987
Haikyuu!!: Karasu...	8.87	123

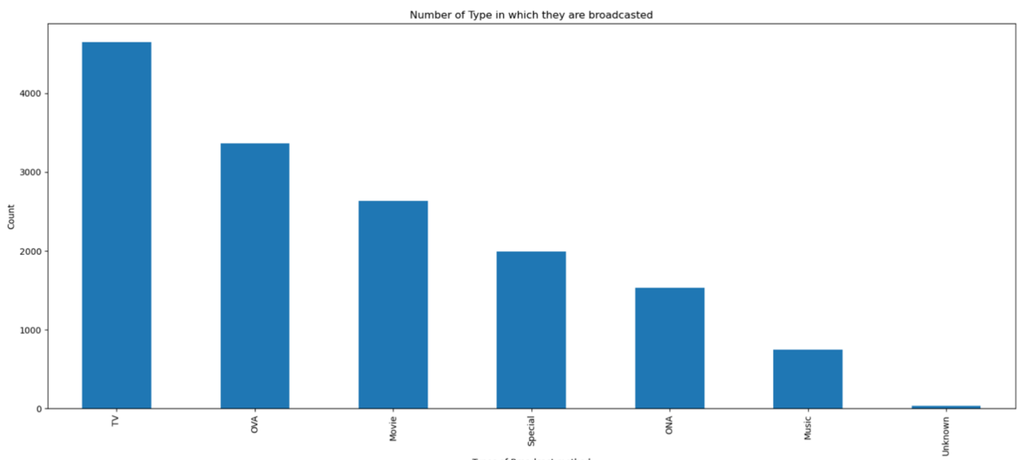
Source Count & Score and Popularity of Anime



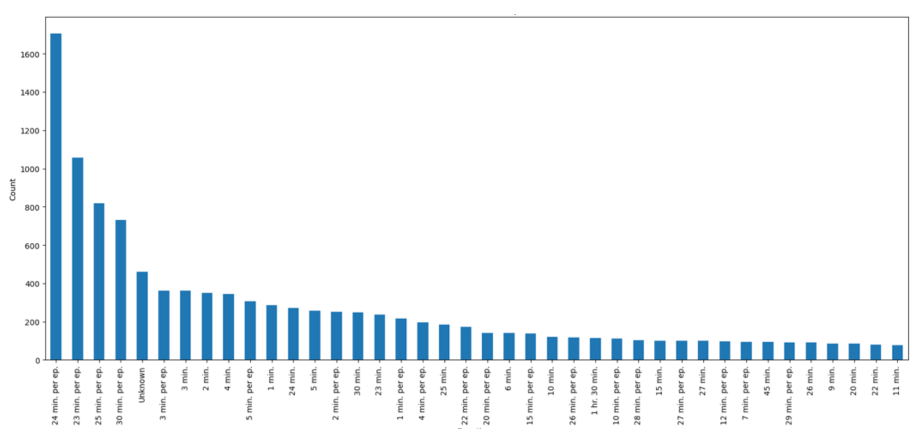
Average Anime Score over Years



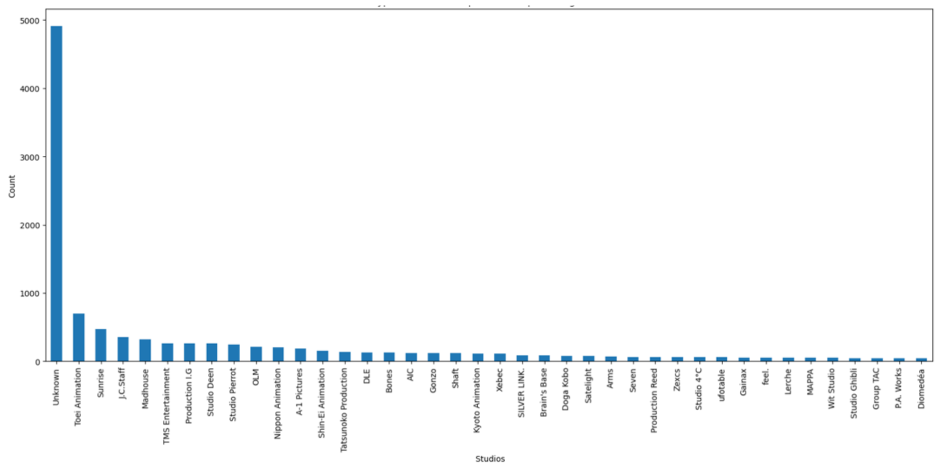
Number of Episodes



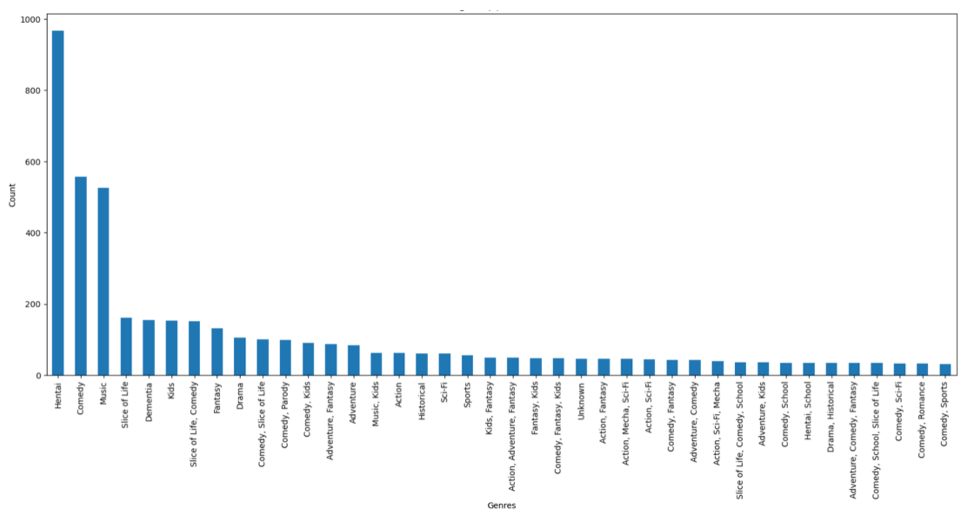
Number of Type in which they are broadcasted



The duration of each episode



Types of Studios responsible for producing anime



The genre of anime

Key Activities:

Genre Distribution Analysis: Identifies the most and least popular genres, helping tailor the recommendation algorithms to emphasize more frequently watched genres.

User Ratings Analysis: Investigates how users rate anime, including average ratings and variability. This analysis helps in understanding user satisfaction and preference patterns.

Insights Utilization: Insights from EDA are crucial for refining the recommendation algorithms. For instance, understanding that certain genres are more popular among specific groups helps in implementing more personalized content-based filtering strategies.

Results and Discussion

ALS Collaborative Filtering Results

user_idx	Name	predicted_rating
31	Konglong Xuexiao	57.396378
31	Boku wa Mada 7-sai Datta	22.594456
31	A-ra-shi: Reborn	20.45516
31	KakoPad: Nagaoka Hanabi Monogatari	17.189756
31	Robot Girls Neo	16.276741
31	Shakuen no Eris	15.594599
31	Sakugan	14.789412
31	Uchi Tama?! Uchi no Tama Shirimasen ka?	14.680076
31	Sougen no Shoujo Laura	13.158294
31	Tokubetsu Byoutou	13.036165
34	Konglong Xuexiao	58.54608
34	Robot Girls Neo	28.327854
34	Boku wa Mada 7-sai Datta	24.896723
34	Honey Honey no Suteki na Bouken	22.959406
34	KakoPad: Nagaoka Hanabi Monogatari	21.723938
34	A-ra-shi: Reborn	21.637774
34	Kidou Senshi Gundam: Senkou no Hathaway 3	20.53622
34	Aibeya The Animation	19.26431
34	Maware Toro-Ika	18.95541
34	Shakuen no Eris	18.585318

ALS Results

Results:

The ALS Collaborative Filtering has shown promising results in predicting user preferences for anime based on user behavior similarities. The system provided predicted ratings for a variety of anime titles. For example, user 31 received a high rating prediction for "Konglong Xuexiao" at 57.396378, suggesting strong preference prediction. Such high predictions indicate a robust model capability to identify potential favorites based on collaborative user data.

Discussion:

The ALS model is effective in leveraging sparse user-item interaction data to create a matrix of user preferences. By identifying latent factors from user ratings, the model can predict a user's affinity for unwatched anime, enhancing the user experience by recommending titles with potentially high enjoyment levels. The method's reliance on user similarity can also lead to high accuracy in predictions but may sometimes limit diversity in recommendations, suggesting popular or commonly watched anime more frequently. This could be addressed by integrating more diverse data points or personal traits into the recommendation logic.

Content-Based Filtering Using TF-IDF Results

	anime_id_A	Name_A	Name_B	distCol
20	8	Bouken Ou Beet	Akahori Gedou Hour Rabuge	1.369051
21	8	Bouken Ou Beet	Eyeshield 21	1.373967
22	8	Bouken Ou Beet	Bishoujo Senshi Sailor Moon S	1.375497
23	8	Bouken Ou Beet	Vampire Hunter D	1.375647
24	8	Bouken Ou Beet	Top wo Nerae 2! Diebuster	1.376359
25	15	Eyeshield 21	Witch Hunter Robin	1.359213
26	15	Eyeshield 21	Bouken Ou Beet	1.373967
27	15	Eyeshield 21	Hajime no Ippo	1.377655
28	15	Eyeshield 21	Mahoromatic: Automatic Maiden	1.378340
29	15	Eyeshield 21	Tetsuwan Birdy	1.379137
30	16	Hachimitsu to Clover	Magikano	1.347877
31	16	Hachimitsu to Clover	Kanojo to Kanojo no Neko	1.363703
32	16	Hachimitsu to Clover	To Heart	1.363793
33	16	Hachimitsu to Clover	Kono Minikuku mo Utsukushii Sekai	1.366521
34	16	Hachimitsu to Clover	Marmalade Boy Movie	1.366998
35	17	Hungry Heart: Wild Striker	Whistle!	1.193131
36	17	Hungry Heart: Wild Striker	Bronze: Zetsuai Since 1989	1.277095
37	17	Hungry Heart: Wild Striker	The SoulTaker: Tamashii-gari	1.278662
38	17	Hungry Heart: Wild Striker	H2	1.292398
39	17	Hungry Heart: Wild Striker	Kimagure Orange☆Road	1.325328

Content Based Results

Results:

The content-based approach using TF-IDF was effective in recommending anime by analyzing the textual content of anime descriptions to find similarities. This method provided targeted recommendations by matching anime with similar themes or content, such as recommending different "One Piece" movies and related content based on their narrative similarities.

Discussion:

Content-Based Filtering is crucial for users with specific tastes in anime genres or themes. By analyzing textual information from anime descriptions, the system can deliver more nuanced recommendations that closely align with a user's content preference history. However, the success of this approach heavily relies on the quality and depth of content descriptions. Enhancements in natural language processing and a broader set of descriptive attributes could improve the accuracy and relevance of these recommendations.

ChromaDB Recommendations Results

```
Anime Name: Naruto
Recommendation: Naruto (2023)
Recommendation: Boruto: Naruto Next Generations

Anime Name: One Piece
Recommendation: One Piece: Episode of East Blue - Luffy to 4-nin no Nakama no Daibouken
Recommendation: One Piece Movie 14: Stampede

Anime Name: Tennis no Ouji-sama
Recommendation: Tennis no Ouji-sama: Zenkoku Taikai-hen
Recommendation: Tennis no Ouji-sama: Zenkoku Taikai-hen - Semifinal

Anime Name: Ring ni Kakero 1
Recommendation: Street Fighter Zero The Animation
Recommendation: Ring ni Kakero 1: Kage Dou-hen

Anime Name: School Rumble
Recommendation: School Rumble Ni Gakki
Recommendation: Furueru Kuchibiru

Anime Name: Sunabouzu
Recommendation: Okashi na Sabaku no Suna to Manu
Recommendation: Tentai Senshi Sunred

Anime Name: Texhnolyze
Recommendation: Senki Zesshou Symphogear AXZ
Recommendation: Mahou Shoujo Lyrical Nanoha StrikerS

Anime Name: Trinity Blood
Recommendation: Hellsing
Recommendation: Hellsing Ultimate

Anime Name: Yakitate!! Japan
Recommendation: Amaama to Inazuma
Recommendation: Kogepan
```

ChromaDB Results

Results:

ChromaDB utilized real-time data processing to dynamically adjust recommendations based on user interactions. This method quickly adapted recommendations, reflecting changes in user preferences or introducing newly available anime that aligns with user interests. The flexibility and responsiveness of ChromaDB are evident in its ability to update recommendations on the fly, providing users with timely and relevant content.

Discussion:

The strength of ChromaDB lies in its capability to process and analyze data in real time, ensuring that the recommendations are always up-to-date with the latest user preferences and behaviors. This adaptability is particularly beneficial in environments where user preferences evolve rapidly. However, the computational demand for real-time processing is substantial, and scaling this capability with a growing user base without compromising performance poses a technical challenge. Further optimization of data processing routines and possibly adopting more efficient data handling frameworks could enhance the scalability of ChromaDB recommendations.

EVALUATION

Objective: The evaluation phase aims to rigorously assess the performance of the Anime Recommendation System using several metrics to ensure the recommendations meet the expected standards of relevance and personalization.

Metrics Used:

Precision: Measures the proportion of recommended anime that users found relevant.

Recall: Assesses the proportion of relevant anime that were successfully recommended to the user.

F1-Score: Combines precision and recall into a single metric, balancing the two for a more comprehensive performance evaluation.

User Satisfaction Scores: Gathers direct feedback from users about their satisfaction with the recommendations provided.

FUTURE WORK

Data Diversification: Integrating additional datasets, such as user social media activity or more detailed user interaction logs, to enrich the recommendation engine's understanding of user preferences.

Algorithmic Experimentation: Testing newer or more sophisticated machine learning algorithms, such as neural collaborative filtering or deep learning approaches, to enhance recommendation accuracy.

Expansion to Other Media: Exploring the possibility of adapting the recommendation system to other forms of media, such as manga or light novels, which could appeal to the current user base.

Vision for Expansion: The future vision includes not only improving the technical aspects of the system but also expanding its scope to create a more integrated entertainment recommendation platform, enhancing cross-media discovery and user engagement.

REFERENCES

[1] Nuurshadieq and A. T. Wibowo, "Leveraging Side Information to Anime Recommender System using Deep learning," 2020 3rd International Seminar on Research of Information Technology and Intelligent Systems (ISRITI), Yogyakarta, Indonesia, 2020, pp. 62-67, doi: 10.1109/ISRITI51436.2020.9315363. keywords: {Recommender systems;Feature extraction;Computer architecture;Neural networks;Transforms;Tensors;Numerical models;Recommender System;Deep Learning;Artificial Neural Network;LSTM;Anime

[2] Jena, Abhipsa and Jaiswal, Arunima and Lal, Dakshita and Rao, Soumya and Ayubi, Afshan and Sachdeva, Nitin, Recommendation System For Anime Using Machine Learning Algorithms (May 27, 2022). Proceedings of the International Conference on Innovative Computing & Communication (ICICC) 2022, Available at SSRN: https://ssrn.com/abstract=4121831 or http://dx.doi.org/10.2139/ssrn.4121831

[3] R. Sharma, D. Gopalani and Y. Meena, "Collaborative filtering-based recommender system: Approaches and research challenges," 2017 3rd International Conference on Computational Intelligence & Communication Technology (CICT), Ghaziabad, India, 2017, pp. 1-6, doi: 10.1109/CIACT.2017.7977363. keywords: {Collaboration;Recommender systems;Conferences;Computational intelligence;Communications technology;Recommender systems;collaborative filtering;memory-based methods;model-based methods;user-based CF;item-based CF},

[4] Agarwal, P., Dave, H., Bandlamudi, J., Sindhgatta, R., & Mukherjee, K. (2024). Multi-Stage Prompting for Next Best Agent Recommendations in Adaptive Workflows. Proceedings of the AAAI Conference on Artificial Intelligence, 38(21), 22843-22849. https://doi.org/10.1609/aaai.v38i21.30319