Kenan Kayembe & Furkan Salman

ANIME RECOMMENDER FINAL REPORT

INTRODUCTION

Definition of Task:

The primary objective of this project is to develop an advanced anime recommendation model. This model is designed to provide anime recommendations, by inputting the name of an anime, the model will generate a curated list of highly relevant and appealing anime suggestions.

About Dataset:

Context

This data set contains information on user preference data from 73,516 users on 12,294 anime. Each user can add anime to their completed list and give it a rating; this data set is a compilation of those ratings.

Content

Anime.csv

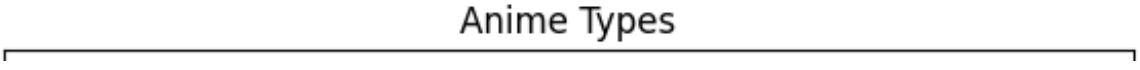
- anime_id - myanimelist.net's unique id identifying an anime.
- name - full name of anime.
- genre - comma-separated list of genres for this anime.
- type - movie, TV, OVA, etc.
- episodes - how many episodes in this show? (1 if movie).
- rating - an average rating out of 10 for this anime.
- members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

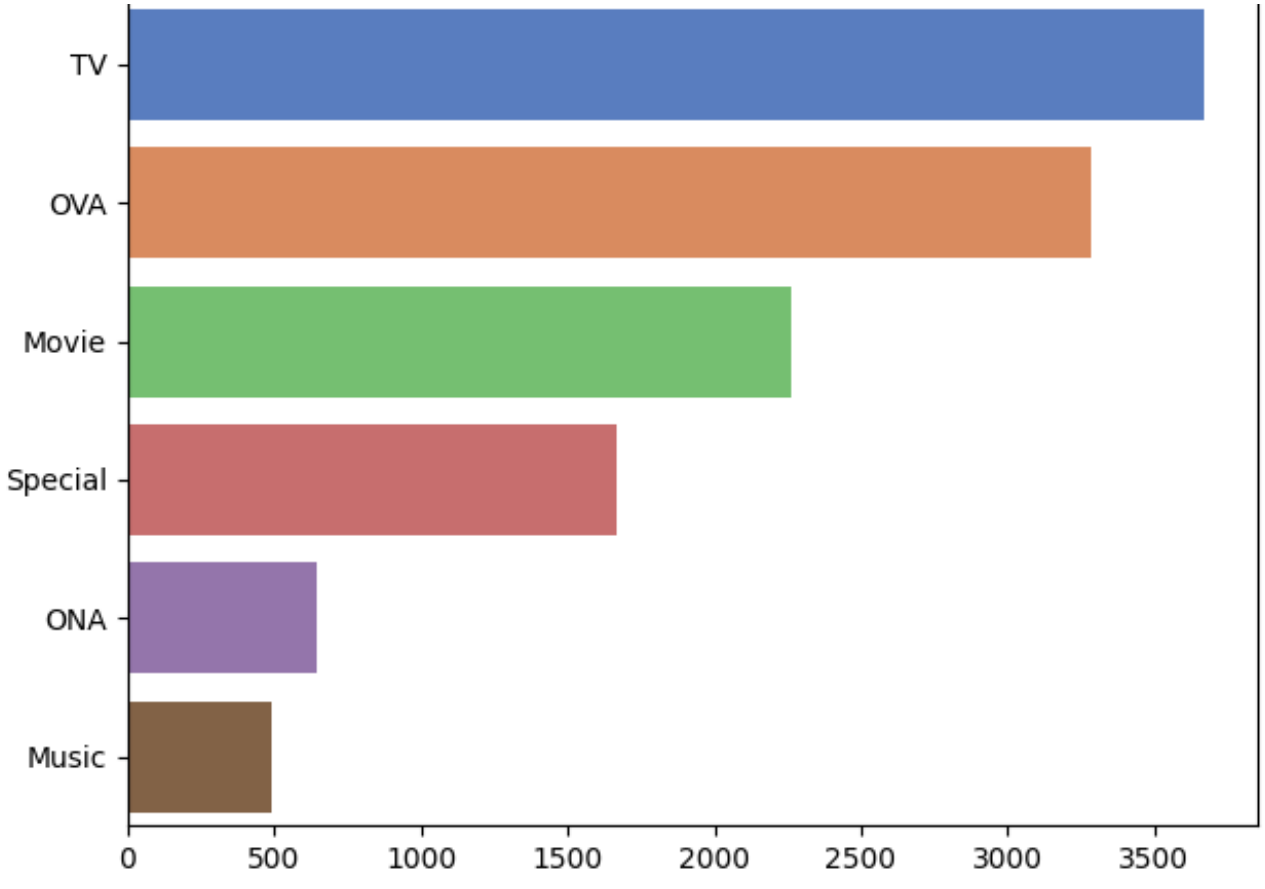
Rating.csv

- 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).

Based on the given task, our model should generate a list of the most relevant animes for a given anime name. It can be inferred that the demographic information of the user will not be taken into consideration for this prediction.

Anime Types



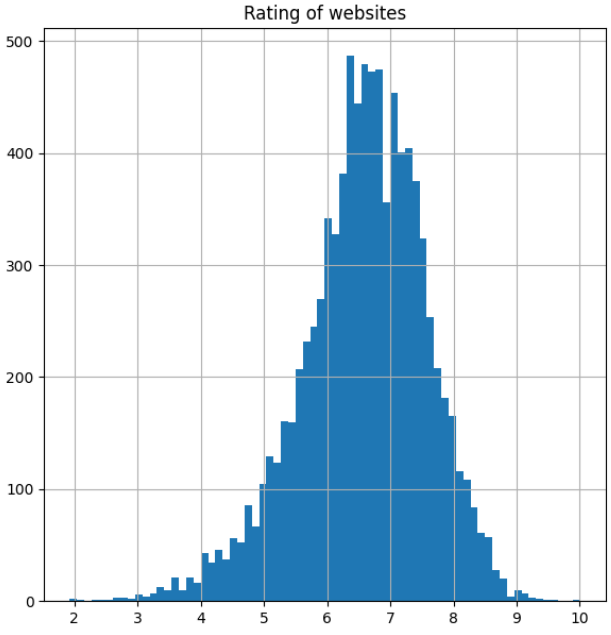


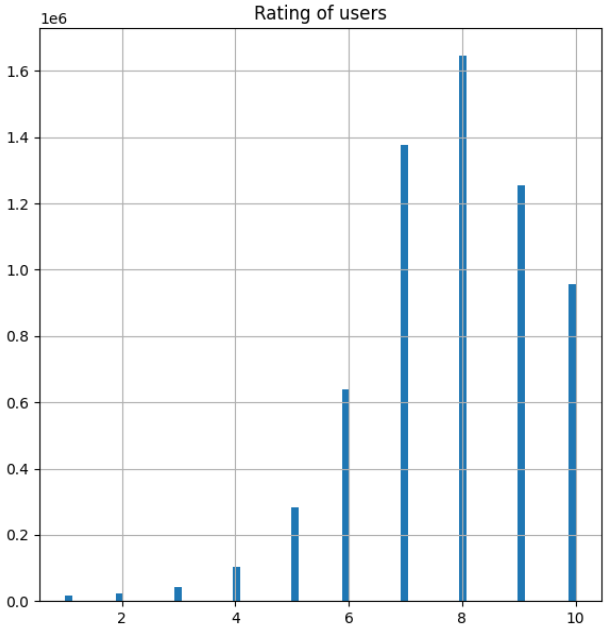
Here, we have the anime types, we only kept TV series, Movies, Ova and dropped the rest.

Word-cloud of Genres

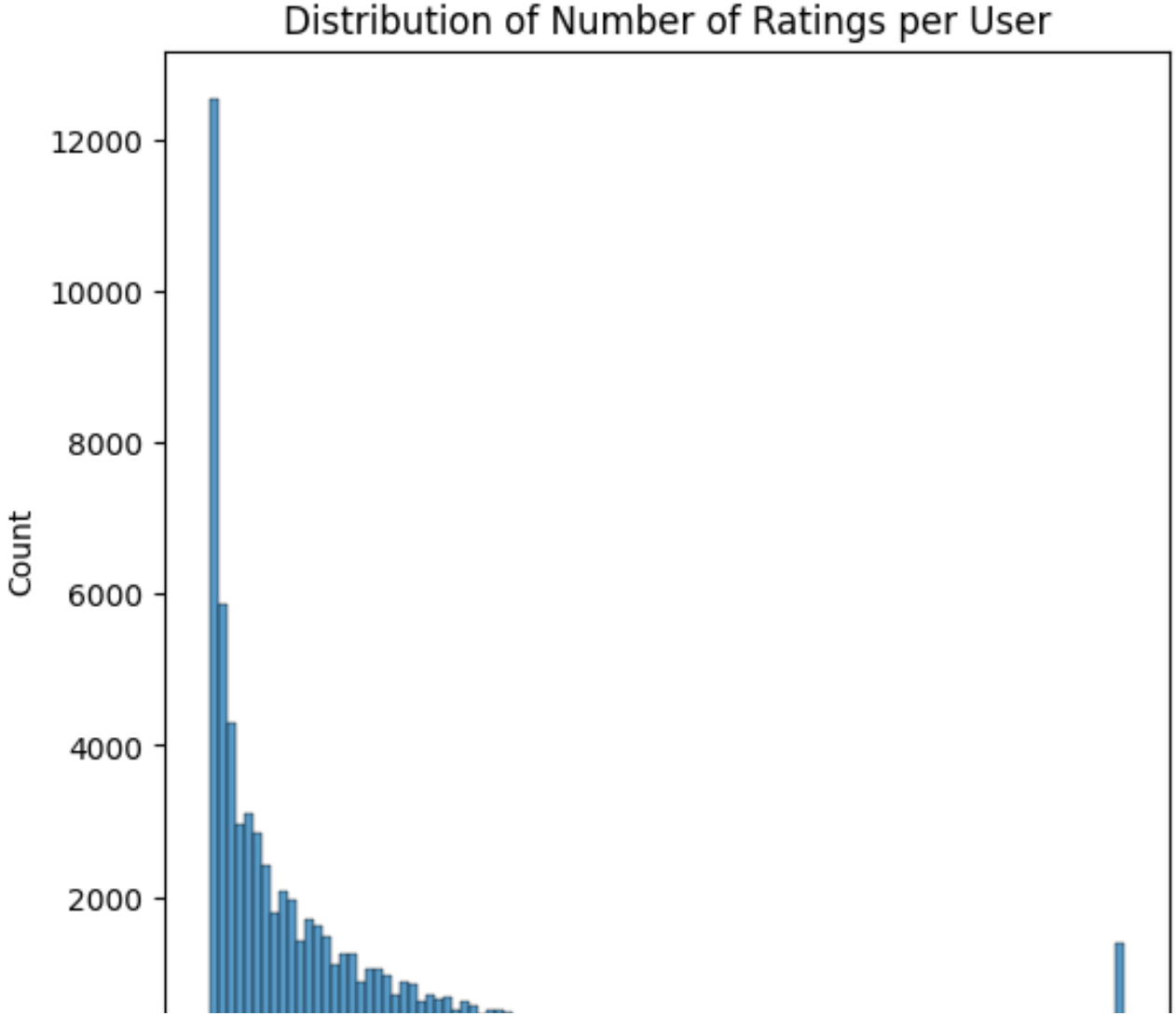
Here, we can visualize the genres in our dataset, we can see that the bigger the name of the genre is shown, the more it is present in our dataset. For example, animes with Adventure, Comedy and Action are the most present in the dataset.

Distribution of Ratings





On these charts, we can see the anime's ratings on websites and the ratings of users.



On this chart, we can see the distribution of the Number of Ratings per user. We can see that the vast majority of users have less than 100 ratings. This is a piece of crucial information as it helped us make a decision in our model construction.

Algorithms Used

For this project, we employed two key algorithms: Collaborative Filtering, Content-Based Filtering, and Hybrid Recommendation system.

Collaborative Filtering

Collaborative Filtering is an ideal choice due to its user-item-based approach, which suits our project requirements. Given a large number of users and anime titles, each user has rated only a small subset of the available anime, resulting in a sparse user-item matrix. Collaborative Filtering effectively manages this sparsity and generates accurate recommendations based on user behavior and preferences.

We decided to remove users' ratings with less than 100 ratings in order to make our models more accurate because users with a few ratings will be biased. A pivot table, consisting of animes as rows and user ratings as columns, is used to calculate the cosine similarity matrix that consists of similarity scores for each anime pair. Then Nearest Neighbors algorithm is used on the cosine similarity matrix to retrieve and output the most similar animes for a given anime.

Content-Based Filtering

The Content-Based Filtering algorithm is also well-suited for our project objectives. It leverages the features or attributes of items to recommend similar items to users. In our dataset, we have detailed descriptions of anime genres, themes, and tags, making Content-Based Filtering an appropriate choice. This approach is particularly valuable in scenarios where user interaction data is limited, such as when dealing with new users or newly released anime that have not yet been extensively rated. In such cases, Content-Based Filtering relies solely on the characteristics of the anime itself to provide recommendations.

TF-IDF Vectorizing is used to get numerical values for different combinations of genres each anime has. Using this TF-IDF matrix we got a cosine similarity matrix which includes the similarity between each anime pair. By sorting these similarities for given anime we've been able to retrieve most similar animes and recommend them to the user.

Hybrid Model - Content-Based + Collaborative

The Hybrid recommendation systems combine multiple recommendation techniques to provide more accurate and diverse recommendations to users. The hybrid recommendation can be useful because it can leverage the strengths of multiple recommendation techniques and overcome their individual limitations. By combining collaborative filtering and content-based filtering techniques in a hybrid

recommendation system, the system can provide more diverse and unexpected recommendations while still providing personalized recommendations based on user preferences.

The Hybrid Model's implementation consists of getting recommendations from both Content-Based and Collaborative Models and sorting the recommendations we got by similarities each scaled according to its group's minimum and maximum values.

INTERMEDİATE RESULTS

Content-Based Filtering

For content-based filtering, the only modification we could do was change the number of genre combinations. We have experimented with up to 3, 4, and 5 combinations in TF-IDF vectorization.

We have used custom-defined measures such as similarity and diversity measures to assess the model's performance.

Experiment-1: Four-Genre Combinations

For 'Death Note' example, we the have following recommendations:

	Anime	Similarity	Type	Rating
0	Mousou Dairinin	0.951487	TV	7.74
1	Higurashi no Naku Koro ni Kai	0.798029	TV	8.41
2	Higurashi no Naku Koro ni	0.759862	TV	8.17
3	Higurashi no Naku Koro ni Rei	0.754682	OVA	7.56
4	Shigofumi	0.734848	TV	7.62
5	Himitsu: The Revelation	0.725537	TV	7.42
6	Hikari to Mizu no Daphne	0.706674	TV	6.87
7	Monster	0.702523	TV	8.72
8	AD Police	0.675006	OVA	6.47
9	Jigoku Shoujo Mitsuganae	0.659810	TV	7.81

We got 0.26 from our diversity measure and 0.91 from our similarity measure for this variant of model.

Experiment-2: Five-Genre Combinations

For 'Death Note' example, we the have following recommendations:

	Anime	Similarity	Type	Rating
0	Mousou Dairinin	0.951487	TV	7.74
1	Higurashi no Naku Koro ni Kai	0.798029	TV	8.41
2	Higurashi no Naku Koro ni	0.759862	TV	8.17

2	Higurashi no Naku Koro ni	0.759862	TV	8.17
3	Higurashi no Naku Koro ni Rei	0.754682	OVA	7.56
4	Shigofumi	0.734848	TV	7.62
5	Himitsu: The Revelation	0.725537	TV	7.42
6	Hikari to Mizu no Daphne	0.706674	TV	6.87
7	Monster	0.702523	TV	8.72
8	AD Police	0.675006	OVA	6.47
9	Jigoku Shoujo Mitsuganae	0.659810	TV	7.81

We got 0.21 from our diversity measure and 0.90 from our similarity measure for this variant of the model.

Diversity has dropped because this variant tends to recommend even more similar genres to the input anime's genres.

Experiment-3: Three-Genre Combinations

For 'Death Note' example, we the have following recommendations:

	Anime	Similarity	Type	Rating
0	Mousou Dairinin	0.951487	TV	7.74
1	Higurashi no Naku Koro ni Kai	0.798029	TV	8.41
2	Higurashi no Naku Koro ni	0.759862	TV	8.17
3	Higurashi no Naku Koro ni Rei	0.754682	OVA	7.56
4	Shigofumi	0.734848	TV	7.62
5	Himitsu: The Revelation	0.725537	TV	7.42
6	Hikari to Mizu no Daphne	0.706674	TV	6.87
7	Monster	0.702523	TV	8.72
8	AD Police	0.675006	OVA	6.47
9	Jigoku Shoujo Mitsuganae	0.659810	TV	7.81

We got 0.23 from our diversity measure and 0.92 from our similarity measure for this variant of the model.

Since there are no significant changes between experimented different number of genres, we have used 4 as our value for ngrams in TF-IDF vectorizer.

Collaborative Filtering

For 'Death Note' example, we have the following recommendations:

	Anime	Similarity	Rating	Type

0	Code Geass: Hangyaku no Lelouch	0.752515	8.83	TV
1	Code Geass: Hangyaku no Lelouch R2	0.730489	8.98	TV
2	Elfen Lied	0.704465	7.85	TV
3	Fullmetal Alchemist: Brotherhood	0.698455	9.26	TV
4	Shingeki no Kyojin	0.697792	8.54	TV
5	Angel Beats!	0.677316	8.39	TV
6	Sword Art Online	0.672457	7.83	TV
7	Fullmetal Alchemist	0.670533	8.33	TV
8	Naruto	0.665845	7.81	TV
9	Toradora!	0.655238	8.45	TV

We got 0.73 from our diversity measure and 0.4 from our similarity measure for this variant of the model.

From the recommendations above we can conclude that the Collaborative Filtering model tends to recommend popular animes. Regarding the measures we can say the Collaborative Filtering model has really diverged recommendations compared to the Content-Based but the similarity is lower in general. This makes us think that we should benefit from the advantages of both Content-Based and Collaborative Filtering Models.

As an extra example, recommendations for 'Ao Haru Ride:

	Anime	Similarity	Rating	Type
0	Ookami Shoujo to Kuro Ouji	0.584047	7.47	TV
1	Tonari no Kaibutsu-kun	0.533387	7.77	TV
2	Gekkan Shoujo Nozaki-kun	0.527068	8.24	TV
3	Sukitte Ii na yo.	0.524460	7.71	TV
4	Shigatsu wa Kimi no Uso	0.508573	8.92	TV
5	Noragami	0.503133	8.17	TV
6	Ao Haru Ride OVA	0.501029	7.76	OVA
7	Tokyo Ghoul	0.488985	8.07	TV
8	Golden Time	0.488717	7.92	TV
9	Nisekoi	0.477469	7.91	TV

When we review the recommendations from an anime fan perspective without regarding any measures it still seems to recommend similar animes.

FINAL RESULTS

From what we've seen on intermediate results we have decided to use a hybrid model approach to leverage the advantages of both Content-Based and Collaborative Filtering models in order to get more diverse and accurate recommendations.

For 'Death Note' example, we have the following recommendations:

	Anime	Similarity	Type	Rating	Recommendation Type
0	Mousou Dairinin	0.939736	TV	7.74	Content-based
10	Code Geass: Hangyaku no Lelouch	0.752515	TV	8.83	Collaborative
11	Code Geass: Hangyaku no Lelouch R2	0.730489	TV	8.98	Collaborative
12	Elfen Lied	0.704465	TV	7.85	Collaborative
13	Fullmetal Alchemist: Brotherhood	0.698455	TV	9.26	Collaborative
14	Shingeki no Kyojin	0.697792	TV	8.54	Collaborative
1	Higurashi no Naku Koro ni Kai	0.734143	TV	8.41	Content-based
15	Angel Beats!	0.677316	TV	8.39	Collaborative
2	Higurashi no Naku Koro ni	0.697524	TV	8.17	Content-based
16	Sword Art Online	0.672457	TV	7.83	Collaborative

We got 0.62 from our diversity measure and 0.61 from our similarity measure for this variant of the model.

We also included the resource of recommendation as a column which is helpful to see what two different approaches prioritize in these recommendations. From the recommendations and measures we can clearly see that we have a more diverse and accurate recommender.

As an extra example, recommendations for 'Ao Haru Ride:

	Anime	Similarity	Type	Rating	Recommendation Type
0	Kareshi Kanojo no Jijou	1.000000	TV	7.66	Content-based
10	Ookami Shoujo to Kuro Ouji	0.584047	TV	7.47	Collaborative
11	Tonari no Kaibutsu-kun	0.533387	TV	7.77	Collaborative
1	Kimi ni Todoke	0.906350	TV	8.19	Content-based
3	Ao Haru Ride OVA	0.906350	OVA	7.76	Content-based
2	Kimi ni Todoke 2nd Season	0.906350	TV	8.17	Content-based
12	Gekkan Shoujo Nozaki-kun	0.527068	TV	8.24	Collaborative
13	Sukitte Ii na yo.	0.524460	TV	7.71	Collaborative
6	Nijiiro Days OVA	0.863007	OVA	6.73	Content-based
7	Chou Kuse ni Narisou	0.863007	TV	6.59	Content-based

CONCLUSİON

To conclude, both Content-Based and Collaborative have their own advantages. The Content-Based Filtering model is able to recommend similar animes regarding the content, and themes of animes whereas Collaborative Filtering is able to recommend animes that are popular among users with similar tastes.

Although these algorithms present great advantages, they also have their disadvantages.

Content-Based Filtering relies heavily on content features, if the data does not have comprehensive and accurate information about animes, its recommendation may suffer. In our dataset this seems to be the case, relying solely on genres is not a good way to associate animes but our dataset makes it hard for us to find another way. Therefore the things we could improve related to this problem could be adding more up-to-date and detailed data.

Collaborative Filtering, on the other hand, relies heavily on the assumption that every user belongs to a group that has similar tastes. For some users, this might not be the case, and since popular animes are rated high by most of the users this model tends to recommend popular animes. It is also computationally more costly compared to the Content-Based Filtering model and that's because there are lots of user ratings to process in order to get the similarity scores for animes. Even though this is a problem, having more user ratings could be helpful to build a more accurate model and this is also something we could change about our solution to improve it.

Regarding the advantages and disadvantages of both solutions, we have come up with a Hybrid Model utilizing the advantages of both models. According to our similarity and diversity measures Hybrid Model presents the best balance. That helps Hybrid Model to avoid recommending only popular animes or animes with the same genres.

To go a little bit further, we could use deep-learning models that are going to be better choices in case we have a vast amount of user and anime data.

If it's possible we could try to ask the user to select his interested genres in order to assign weight to these genres or filter out the uninterested genres for more accurate recommendations. If it's possible we could also fetch the user's data from a website like myanimelist which includes the animes user has watched and the user's ratings for them and use this additional data in our Collaborative Filtering Model.