

Anime Recommendation System

Erina K., Fay F., Kristine U., Soumeng C.

Agenda

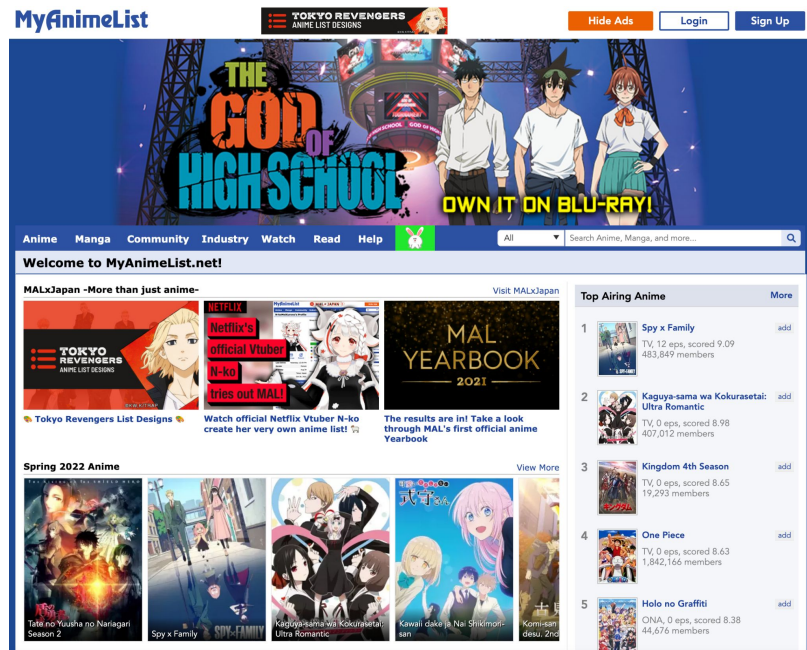
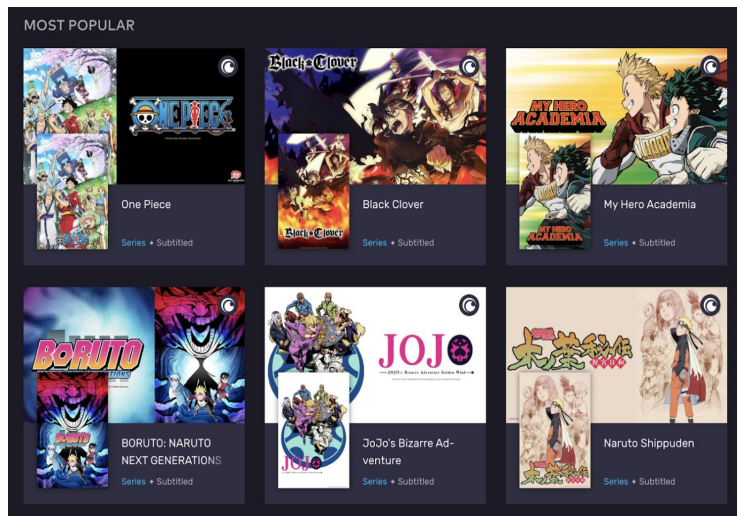
- Introduction
- Recommendation Filtering Concepts

- Data Preprocessing
- Methodologies and Results
- Conclusion

Introduction

The image features a dark blue background with the word "Introduction" in a large, white, sans-serif font. The text is centered horizontally and vertically. Surrounding the text are various decorative elements: small white stars, pink and teal circles of different sizes, and small white dots. These elements are scattered across the background, creating a cosmic or abstract feel. The overall composition is clean and modern.

The Problem with Anime Recommendations

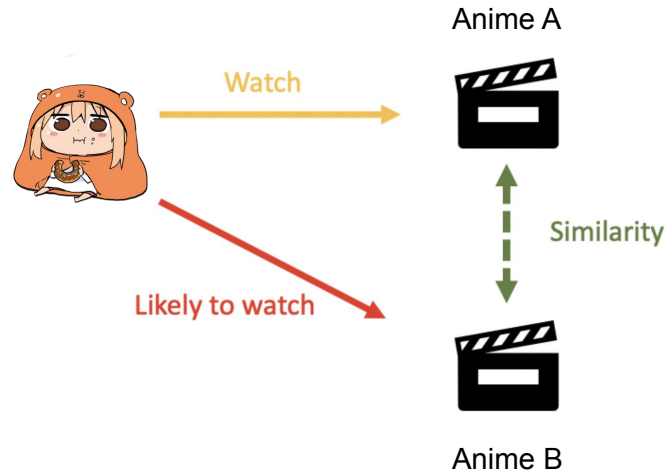


The background is a dark navy blue rectangle. It is decorated with various small, stylized elements: white four-pointed stars, small white dots, and larger dots in shades of pink, teal, and light blue. These elements are scattered around the perimeter and in the corners of the frame.

Recommendation Filtering Concepts

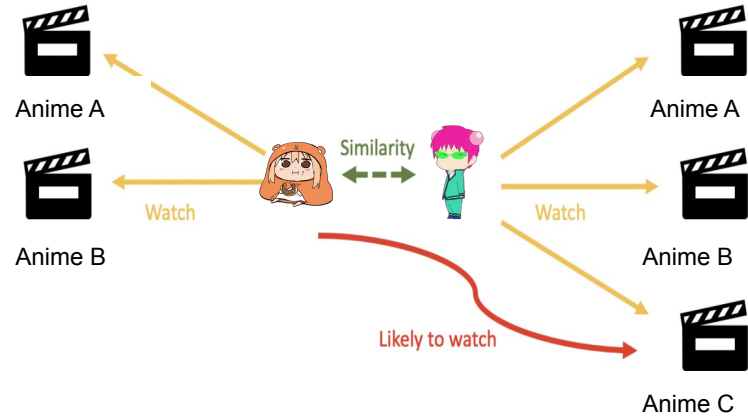
Content-Based Filtering

- Based on contents of items
- Item features used to recommend other similar items
- Useful when smaller user interactions present or new service deployed
- Could lack novelty as outliers unlikely to be included



Collaborative Filtering

- Based on similarities between users and items
- Make recommendation for user based on another user with similar interests
- Can diversify recommendation process
- Large user preference dataset needed



Data Preprocessing

The background is a dark navy blue rectangle. It is decorated with various celestial-themed elements: small white dots, four-pointed white stars, and colored circles in pink, teal, and light grey. These elements are scattered around the perimeter and in the background, creating a cosmic or space-like atmosphere.

Anime User Dataset


- Preprocessed from user_data.csv
- Extracted essential information such as user_id, mal_id, rating, and indication for favorited
- Favorites indicated in binary (1: favorited, 0: not favorited)
- Combined to provide more data points together

user_id	mal_id	rating
1	39764	6
1	628	9
2	9682	3

User Score Dataset Example

user_id	mal_id	favorited
2	237	1
2	82525	1
2	407	1

User Favorited Dataset Example

The background is a dark navy blue rectangle. It is decorated with various celestial motifs: small white dots, four-pointed white stars, and solid-colored circles in pink, teal, and light grey. These elements are scattered around the perimeter of the slide, creating a cosmic or space-themed aesthetic.

Methodologies And Results

Linear Regression

- Linear regression can be used to predict missing anime rating data
- Train and fit linear regression model using training sets to fit missing anime ratings on correlated variables to make predictions
- Predicted values filled in place of missing values

mal_id	rating	favorited
1	9.0	0.0
3	NaN	1.0
5	6.0	0.0
6	6.0	0.0
11	NaN	1.0



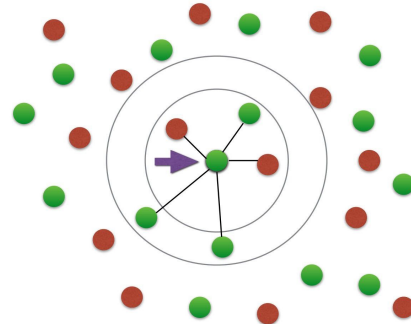
mal_id	rating	favorited
1	9.0	0.0
3	5.99	1.0
5	6.0	0.0
6	6.0	0.0
11	5.99	1.0



Nearest Neighbors

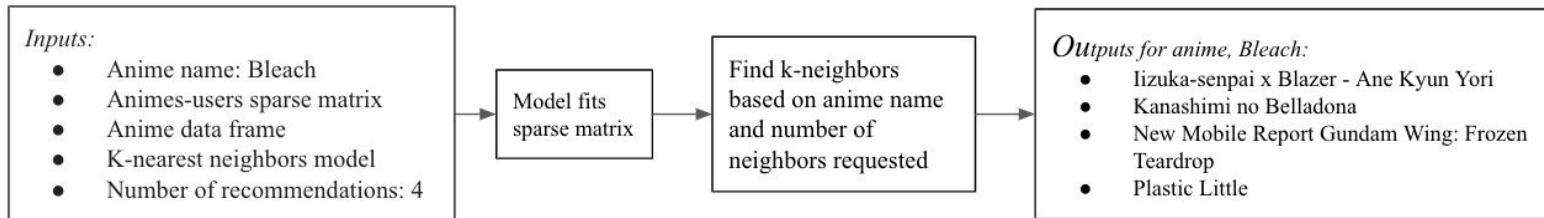
K-Nearest Neighbors

- Collaborative filtering
- Generate recommendations based on specified anime
- 2 Classes in Python
 - KNN Classifier class
 - Supervised
 - NearestNeighbors class
 - Unsupervised



K-Nearest Neighbors

- KNN classifier had accuracy, recall, F1 scores in 20 - 30 percentile for train and test data
- Used NearestNeighbors instead
 - returns the distances and indices of the most similar animes to the given anime



Clustering

The image features a dark blue background with the word "Clustering" in a large, white, sans-serif font. The text is centered horizontally and vertically. Surrounding the text are several decorative elements: small white stars, pink dots of varying sizes, and teal dots. These elements are scattered across the background, with a higher concentration of dots and stars near the corners and edges, creating a cosmic or abstract feel. The overall composition is clean and modern.

K-Mean Clustering

- K-Means is one of the most well-known unsupervised clustering algorithms, in which the algorithm can categorize the input dataset into different category groups.
- The clustering based on algorithm to categorize users interests into groups of the same interest through K-Means.
- For this specific algorithm, the recommendation system will presents user most watch anime genres.

K-Mean Clustering

- The recommendation system also works with comparing two different user genres

Input:

Assign a variable that access for specific user ID in the fav_movie.csv, then the recommendation function



Iteration:

The assign variable will run through a comparison loop to match the specific user ID with fav_movie.csv and a merge of three dataframe on 'mal_id'. The loop will continue until it finds the most watch genres for that specific users.



Output:

- Action
- Fantasy
- Game



Categorical

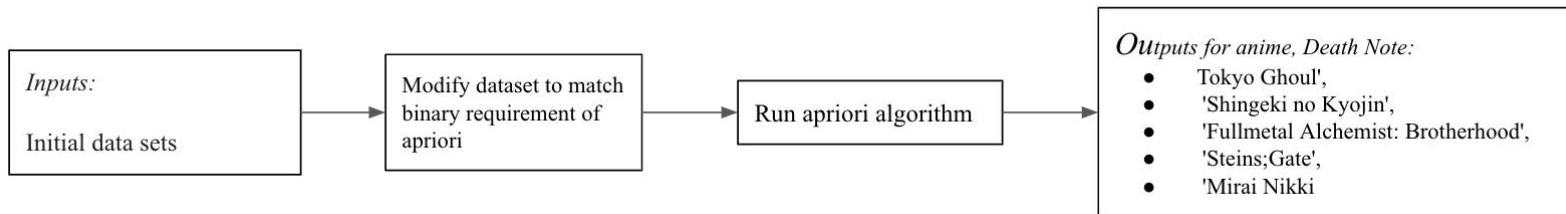
Apriori Algorithm

- Apriori is a technique in Market Basket Analysis used to discover items that are frequently sold together.
- In our case, movies can be viewed as a set of items
- Using the apriori technique enable us to investigate an underlying relationship within the movies by calculating support, confidence and lift

Apriori Algorithm

The algorithm consists of the following steps

- Find each movie watched by user.
- Calculate the support of the movie.
- If support is less than minimum support, discard the movie. Else, insert it into frequent itemset.
- Calculate confidence for each non- empty subset.
- If confidence is less than minimum confidence, discard the subset. Else, it into strong rules.



Apriori Algorithm

Support val > 0.4

Lift > 1

support	itemsets
0.51	(Akame ga Kill!)
0.61	(Angel Beats!)
0.52	(Ano Hi Mita Hana no Namae wo Bokutachi wa Mad...

antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
(Akame ga Kill!)	(Boku no Hero Academia)	0.51	0.74	0.44	0.87	1.17	0.07	2.01
(Boku no Hero Academia)	(Akame ga Kill!)	0.74	0.51	0.44	0.59	1.17	0.07	1.22

Matrix Factorization

The background is a dark navy blue rectangle. It is decorated with various celestial motifs: small white four-pointed stars, larger pink and teal circles representing planets, and a cluster of three small white dots in the top right corner. The text 'Matrix Factorization' is centered in a large, white, bold, sans-serif font.

Singular Value Decomposition

What is SVD?

1. Decomposition into 3 distinct matrices represented by:

$$Data_{m \times n} = U_{m \times m} \Sigma_{m \times n} V^T_{n \times n}$$

2. Dimensionality Reduction

For data that is sparse, SVD with LSA is used directly on the dataset.

Finding the best model

Number of Latent Variables	RSME
1	8.560858
2	6.684418
3	6.020406
4	18.636003
5	16.708803
6	15.280429
7	14.172834
8	13.297916
9	12.638448
10	13.807466
11	14.447258
12	14.894947
13	14.350341
14	12.944814
15	12.529722
16	13.083031
17	11.889878
18	11.641128
19	12.216349
20	11.919922

Singular Value Decomposition

Inputs:

- Anime name: Bleach
- Animes-users sparse matrix
- Anime dataframe
- TruncatedSVD
- Number of recommendations: 10

SVD Model fit
on transposed
matrix

Find highest Pearson
R Correlation values
between animes

Outputs for anime, Bleach:

Rank	Anime Title	R Correlation
1	Gintama Season 5	1.000000
2	Welcome to the N.H.K.	1.000000
3	Psycho-Pass 2	1.000000
4	The Future Diary	0.999999
5	The Rising of the Shield Hero	0.999998
6	Prison School	0.999998
7	Toradora!	0.999998
8	Gintama	0.999995
9	The Rose of Versailles	0.999995
10	Hitorijime My Hero	0.999994

Alternating Least Squares

What is ALS?

1. Latent Factor Model

$$R \approx U^T V$$

2. Iterative Approach

$$\arg \min_{U, V} \sum_{\{i, j | r_{ij} \neq 0\}} (r_{ij} - u_i^T v_j)^2 + \lambda \left(\sum_i n_{u_i} \|u_i\|^2 + \sum_j n_{v_j} \|v_j\|^2 \right)$$

Finding the best model



```
# View the predictions
test_predictions = best_model.transform(test)
RMSE = evaluator.evaluate(test_predictions)
print(RMSE)
```

1.7448107340371852

Alternating Least Squares

Inputs:

- Anime name: Bleach
- Animes-users sparse matrix
- Anime dataframe
- PySpark ALS
- Number of recommendations: 5

RMSE used to
evaluate 16
different ALS
models

Cross validated and fit
to best model

Outputs for anime, Bleach:

Anime	Rating
High School DxD	10.006992
Shattered Angels	10.000742
Rozen Maiden	9.954891
A-Real	9.88063
Gitama	9.830604

Conclusion



Conclusion

- Recommendation help users find new animes to watch that they will likely enjoy
- Collaborative and Content-based filters were used
- The relationships discovered could further be used not only to make anime recommendations but also to create new marketing campaigns and research customer's behavior
- Future improvements:
 - Collect more data from varied sources like including implicit data like user watch times from streaming services
 - Combine top performing algorithms to improve recommendations



Thank you!