Northeastern University

Anime Recommendation System

Kachun Lee (Tim)

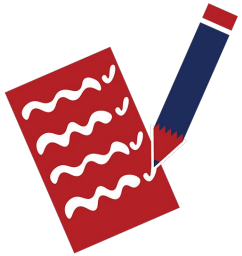
Introduction

Objectives

- Provide insights for the anime community
- Recommender System
  - Help creating a better streaming platform
  - Help Anime lovers find animes that interest them
  - Boost playing time and revenue for anime streaming platforms

Northeastern University

Agenda



- Data Description
- Data Cleaning
- Exploratory data analysis
- Data Preprocessing
- Clustering
- Recommender System

Northeastern University

Data Description

	anime_id	name	genre	type	episodes	rating	members
0	32281	Kimi no Na wa	Drama, Romance, School, Supernatural	Movie	1	9.37	200630
1	5114	Fullmetal Alchemist Brotherhood	Action, Adventure, Drama, Fantasy, Magic, Mili...	TV	64	9.26	793665
2	28977	Gintama	Action, Comedy, Historical, Parody, Samurai, S...	TV	51	9.25	114262
3	9253	SteinsGate	Sci-Fi, Thriller	TV	24	9.17	673572
4	9969	Gintama039	Action, Comedy, Historical, Parody, Samurai, S...	TV	51	9.16	151266

Dimensionality

(12294, 7)

	user_id	anime_id	rating
0	1	20	-1
1	1	24	-1
2	1	79	-1
3	1	226	-1
4	1	241	-1

Number of unique user: 73515

Dimensionality

(7813737, 3)

- The raw anime dataset contains 12,294 animes with 7 features
- The raw rating dataset contains 7,813,737 ratings from 73,515 unique users

Source: https://www.kaggle.com/lavanyaanandm/recommending-anime-s-using-all-recommendation-sys#4.-Exploratory-Data-Analysis-

Northeastern University

Data Cleaning

Two tables were cleaned separately and later merged for future applications.

Anime Table:

- Text cleaning
- Missing value treatment
- Check and remove any duplicate rows
- Index reset
- Determining data type

Dimensionality
(12294, 7) ⟶ (11830, 7)

Rating Table:

- Replacing -1 ratings
- Missing value treatment
- Check and remove duplicate rating
- Index reset
- Determining data type

Dimensionality
(7813737, 3) ⟶ (6337241, 3)

Unique users
73515 ⟶ 69600

Northeastern University

Data Cleaning

Two tables were first merged and duplicated {user_id, anime_id} pairs were then removed

	anime_id	name	genre	type	episodes	average_rating	members	user_id	user_rating
0	32281	Kimi no Na wa	Drama, Romance, School, Supernatural	Movie	1	9.37	200630	99	5
1	32281	Kimi no Na wa	Drama, Romance, School, Supernatural	Movie	1	9.37	200630	152	10
2	32281	Kimi no Na wa	Drama, Romance, School, Supernatural	Movie	1	9.37	200630	244	10
3	32281	Kimi no Na wa	Drama, Romance, School, Supernatural	Movie	1	9.37	200630	271	10
4	32281	Kimi no Na wa	Drama, Romance, School, Supernatural	Movie	1	9.37	200630	278	-1
...	...	...	...	...	...	...	...	...	...
7813602	6133	Violence Gekiga Shin David no Hoshi Inma Densetsu	Hentai	OVA	1	4.98	175	39532	-1
7813603	6133	Violence Gekiga Shin David no Hoshi Inma Densetsu	Hentai	OVA	1	4.98	175	48766	-1
7813604	6133	Violence Gekiga Shin David no Hoshi Inma Densetsu	Hentai	OVA	1	4.98	175	60365	4
7813605	26081	Yasuji no Pornorama Yacchimae	Hentai	Movie	1	5.46	142	27364	-1
7813606	26081	Yasuji no Pornorama Yacchimae	Hentai	Movie	1	5.46	142	48766	-1

7813607 rows × 9 columns

```
There are 7 duplicates in this dataset
```

After removing the 7 duplicates the final dataset has 7,813,600 records and 9 features with no missing values and no duplicate data

Northeastern University

Exploratory Data Analysis

	episodes	average_rating	members	user_rating
count	6337137.00000	6337137.00000	6337137.00000	6337137.00000
mean	18.75274	7.67501	184576.39191	7.80854
std	35.20937	0.66990	190952.79433	1.57244
min	1.00000	2.00000	33.00000	1.00000
25%	3.00000	7.29000	46803.00000	7.00000
50%	12.00000	7.70000	117091.00000	8.00000
75%	24.00000	8.15000	256325.00000	9.00000
max	1818.00000	9.37000	1013917.00000	10.00000

Numerical Data:

- The minimum average rating is 2 while the maximum average rating is 9.37
- The community size of each anime ranges from 33 to 1,013,917
- The number of episodes ranges from 1 to 1,818
- User rating ranges from 1 to 10 with an average of 7.81

Exploratory Data Analysis

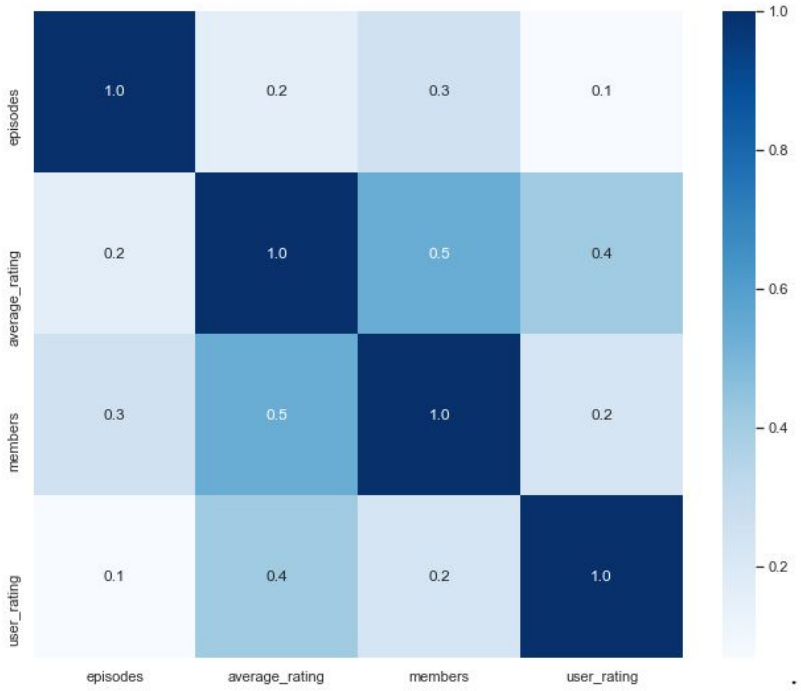
	anime_id	name	genre	type	user_id
count	7813600	7813600	7813600	7813600	7813600
unique	11158	11135	3154	6	73515
top	1535	Death Note	Hentai	TV	48766
freq	39340	39340	62435	5283586	10223

Categorical Data:

- There are 11,158 unique anime id
- Death Note is the most popular anime. It's watched and rated by 39,340 unique users
- There are 3,154 unique genre and Hentai is the most repeated genre with a frequency of 62,435
- There are 6 unique anime type and TV being the most repeated type with a frequency of 5,283,586
- There are 73,515 unique users and the user id 48,766 have watched 10,223 animes

Exploratory Data Analysis

Correlation heatmap



We find a positive correlation with community size and average rating. It is expected because as the community size increase i.e anime becomes more popular, the average rating of the anime will also like to increase.

There is no strong correlation observed between other attributes.

Northeastern University

Exploratory Data Analysis

Hottest anime

By # of users

	name	users
1	Death Note	39340
2	Sword Art Online	30583
3	Shingeki no Kyojin	29584
4	Code Geass Hangyaku no Lelouch	27718
5	Elfen Lied	27506
6	Angel Beats	27183
7	Naruto	25925
8	KOn	25597
9	Fullmetal Alchemist	25032
10	Fullmetal Alchemist Brotherhood	24574

By # of members

	name	members
1	Death Note	1013917
2	Shingeki no Kyojin	896229
3	Sword Art Online	893100
4	Fullmetal Alchemist Brotherhood	793665
5	Angel Beats	717796
6	Code Geass Hangyaku no Lelouch	715151
7	Naruto	683297
8	SteinsGate	673572
9	Mirai Nikki TV	657190
10	Toradora	633817

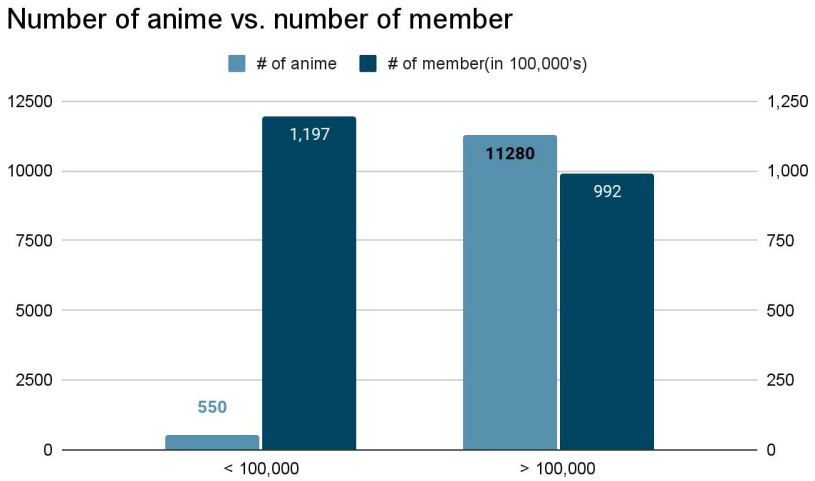
By average rating (>500)

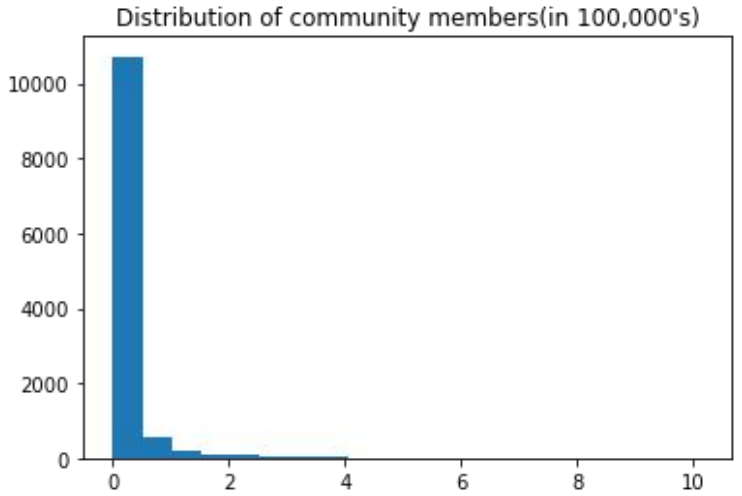
	name	average_rating
1	Kimi no Na wa	9.37
2	Fullmetal Alchemist Brotherhood	9.26
3	Gintama	9.25
4	SteinsGate	9.17
5	Gintama039	9.16
6	Haikyuu Karasuno Koukou VS Shiratorizawa Gakue...	9.15
7	Hunter x Hunter 2011	9.13
8	Gintama039 Enchousen	9.11
9	Gintama Movie Kanketsuhen Yorozuya yo Eien Nare	9.10
10	Clannad After Story	9.06

Northeastern University

Exploratory Data Analysis

Anime communities





Most of the members come from only a few anime communities

Northeastern University

Exploratory Data Analysis

Word Cloud

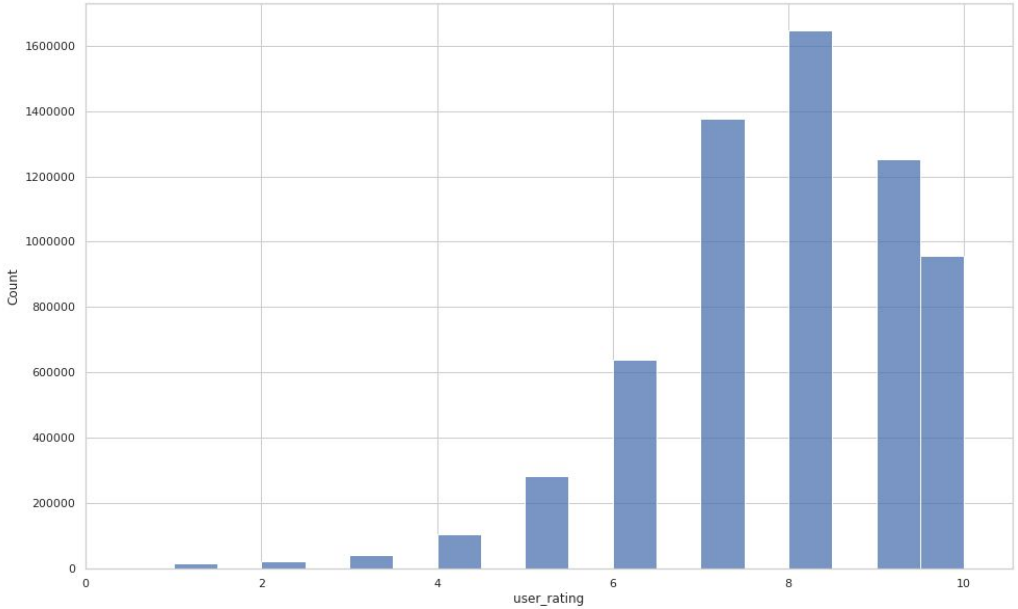
Most watched genres
- Comedy
- Action
- Fantasy
- Romance
- Supernatural

Least watched genres
- Parody
- Horror
- Demons
- Martial Arts
- Historical

Northeastern University

Exploratory Data Analysis

Distribution plot of user- rating

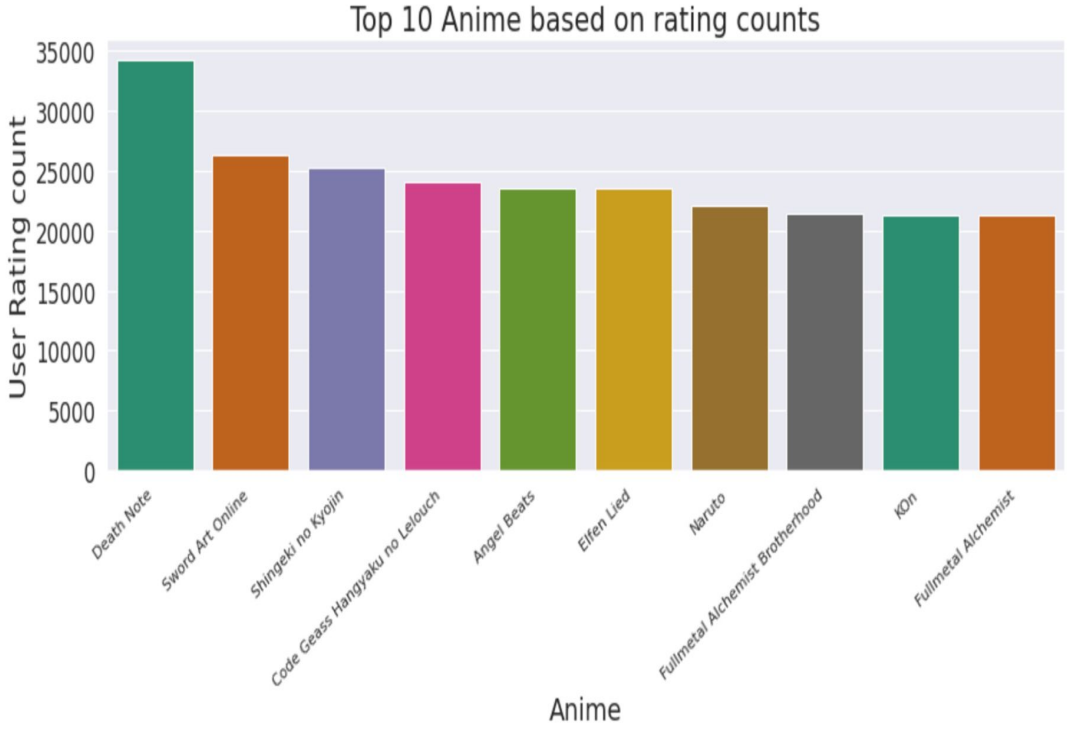


- Heavily left-skewed distribution routhly 8/10 rating is the most frequent (1.6 million)
- The second frequent is roughly 7/10 rating
- The least frequent is 1/10 rating
- User normally satisfied anime since people inclined to rate between 7-9 rating.

Northeastern University

Exploratory Data Analysis

Distribution plot of user- rating count and Anime



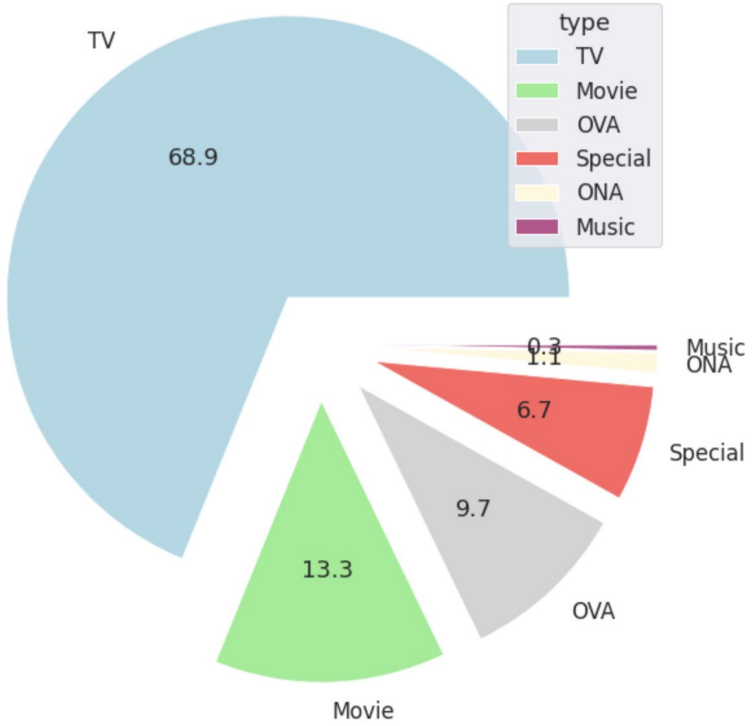
-Death note is the most popular Anime since it is the highest Anime based on rating count

-The second highest is the Sword art online

Northeastern University

Exploratory Data Analysis

Type pie chart



-TV is the type that user to watch Anime the most (68.9%)

-Movie is the second highest one (13.3%)

Northeastern University

Clustering

Data Preprocessing

- Expand genre column
- One Hot Encoding
- Standard Scale Numerical Variables

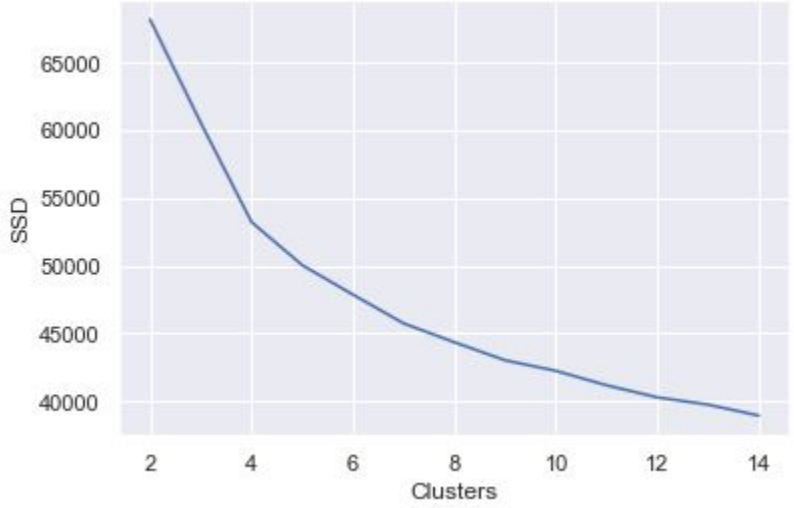
	episodes	rating	members	type_Movie	type_Music	type_ONA	type_OVA	type_Special	type_TV	0_Action	...	9_ School	9_ Sci-Fi	9_ Shounen	9_ Space
0	-0.243905	2.831301	3.289181	1	0	0	0	0	0	0	...	0	0	0	0
1	1.093813	2.723363	13.999758	0	0	0	0	0	1	1	...	0	0	0	0
2	0.817776	2.713551	1.729322	0	0	0	0	0	1	1	...	0	0	0	0
3	0.244468	2.635050	11.830805	0	0	0	0	0	1	0	...	0	0	0	0
4	0.817776	2.625238	2.397637	0	0	0	0	0	1	1	...	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
11825	-0.243905	-2.290843	-0.330509	0	0	0	1	0	0	0	...	0	0	0	0
11826	-0.243905	-2.163280	-0.331015	0	0	0	1	0	0	0	...	0	0	0	0
11827	-0.180204	-1.574528	-0.330365	0	0	0	1	0	0	0	...	0	0	0	0
11828	-0.243905	-1.476403	-0.331159	0	0	0	1	0	0	0	...	0	0	0	0
11829	-0.243905	-1.005401	-0.331755	1	0	0	0	0	0	0	...	0	0	0	0

11830 rows × 308 columns

Northeastern University

Clustering

Choosing number of cluster using silhouette score

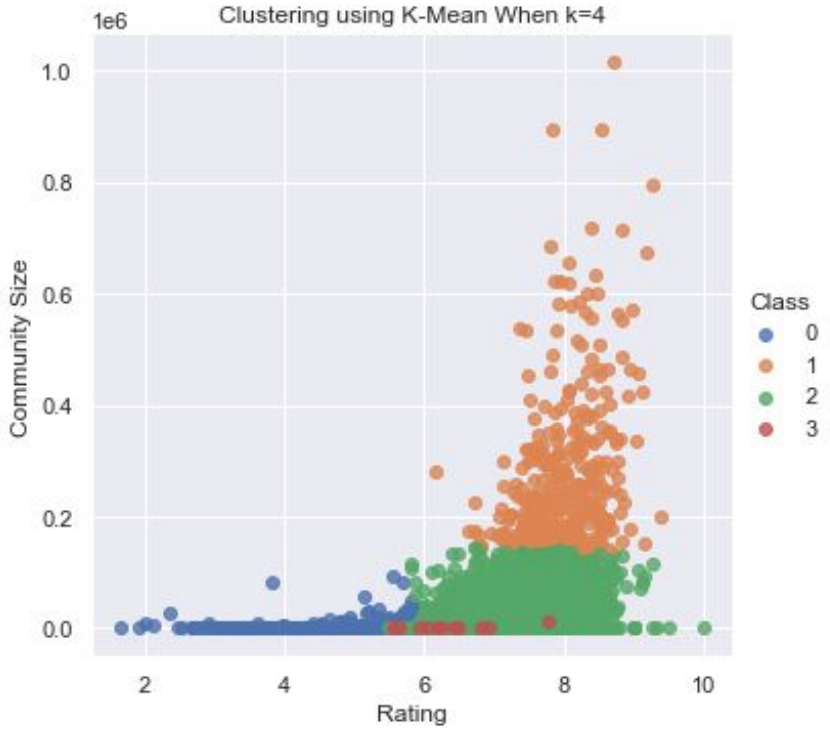


There is a clear elbow when k=4, therefore our data will have 4 clusters

Northeastern University

Clustering

Visualizes the clusters on a 2-D grid



Northeastern University

Clustering

Class 0

	name	rating	Class
10984	Gakuen 3	6.63	0
10987	Ryoujoku Hitozuma Onsen	6.63	0
10983	Bikyaku Seido Kaichou Ai	6.63	0
10992	Rennyuu Tales The Animation	6.62	0
10993	Rin x Sen Hakudaku Onna Kyoushi to Yaroudomo	6.62	0

Class 1

	name	rating	Class
0	Kimi no Na wa	9.37	1
1	Fullmetal Alchemist Brotherhood	9.26	1
3	SteinsGate	9.17	1
4	Gintama039	9.16	1
6	Hunter x Hunter 2011	9.13	1

Class 2

	name	rating	Class
10277	Taka no Tsume 8 Yoshidakun no XFiles	10.00	2
9446	Mogura no Motoro	9.50	2
8958	Kahei no Umi	9.33	2
2	Gintama	9.25	2
10589	Yakusoku Africa Mizu to Midori	9.25	2

Class 3

	name	rating	Class
926	Doraemon 1979	7.76	3
3637	Ninja Hattorikun	6.92	3
9464	Monoshiri Daigaku Ashita no Calendar	6.80	3
5332	Manga Nippon Mukashibanashi 1976	6.48	3
9107	Kirin Ashita no Calendar	6.43	3

Northeastern University

Recommender

Popularity-based filtering

$$Weighted Rating(\mathbf{WR}) = \left(\frac{\mathbf{v}}{\mathbf{v}+\mathbf{m}} \cdot \mathbf{R}\right) + \left(\frac{\mathbf{m}}{\mathbf{v}+\mathbf{m}} \cdot \mathbf{C}\right)$$

- v is the number of votes for the anime (‘total rating count’ in our dataset)
- R is the average rating of the anime;(‘average rating’ in our dataset)
- m is the minimum votes required to be listed in the chart (animes having total rate count greater than 85%(quantile function))

  m= 987.6499999999996
- C is the mean vote across the whole report. (mean of ‘average rating’)

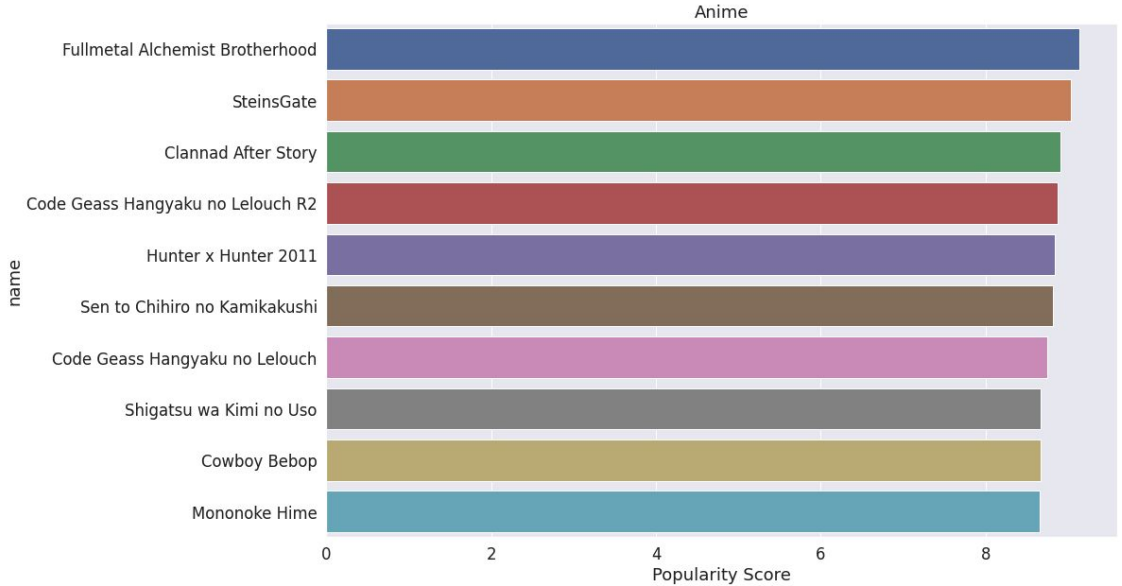
  C= 6.592642062689574
- TotalRatingcount>=m(987.65)->total 1484 anime are qualified to have score(8406 anime are not qualified to have score)
- Getting score by applying the WR function
  example:((21494/(21494+987.65)*9.26)+(987.65/(987.65+21494)*6.59)=9.14

	name	totalRatingCount	avg_rating	score
1	Fullmetal Alchemist Brotherhood	21494	9.260000	9.142819
3	SteinsGate	17151	9.170000	9.029663
10	Clannad After Story	15518	9.060000	8.912360
13	Code Geass Hangyaku no Lelouch R2	21124	8.980000	8.873365
6	Hunter x Hunter 2011	7477	9.130000	8.833943
15	Sen to Chihiro no Kamikakushi	19481	8.930000	8.817218
19	Code Geass Hangyaku no Lelouch	24125	8.830000	8.742007
16	Shigatsu wa Kimi no Uso	8271	8.920000	8.671733
22	Cowboy Bebop	13449	8.820000	8.667620
24	Mononoke Hime	13679	8.810000	8.660683

Northeastern University

Recommender

Popularity-based filtering (result)



Northeastern University

Recommender

Content-based filtering

Preprocessing
- Title text cleaning
- Tf-idf vectorization of genres

Genre

Anime

Tf-idf
12,294 X 1,554

Using cosine similarity to construct item similarity matrix

	0	1	2	3	4	5	6	7	8	9	...	12284	12285	12286	12287	12288	1
0	1.000000	0.023127	0.000000	0.000000	0.000000	0.072155	0.000000	0.033067	0.000000	0.000000	...	0.0	0.054530	0.0	0.0	0.0	
1	0.023127	1.000000	0.025494	0.000000	0.025494	0.040651	0.086116	0.072591	0.025494	0.025494	...	0.0	0.000000	0.0	0.0	0.0	
2	0.000000	0.025494	1.000000	0.069244	1.000000	0.028357	0.033364	0.057629	1.000000	1.000000	...	0.0	0.000000	0.0	0.0	0.0	
3	0.000000	0.000000	0.069244	1.000000	0.069244	0.000000	0.000000	0.107605	0.069244	0.069244	...	0.0	0.000000	0.0	0.0	0.0	
4	0.000000	0.025494	1.000000	0.069244	1.000000	0.028357	0.033364	0.057629	1.000000	1.000000	...	0.0	0.000000	0.0	0.0	0.0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
12289	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	1.0	0.226196	1.0	1.0	1.0	
12290	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	1.0	0.226196	1.0	1.0	1.0	
12291	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	1.0	0.226196	1.0	1.0	1.0	
12292	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	1.0	0.226196	1.0	1.0	1.0	
12293	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	1.0	0.226196	1.0	1.0	1.0	

12294 rows × 12294 columns

Northeastern University

Recommender

Content-based filtering

Top 10 recommendations examples

Case 1
‘Pokemon’

	Anime name	Rating	Similarity	Genre
0	Pokemon XYandZ	7.91	1.0	Action, Adventure, Comedy, Fantasy, Kids
1	Digimon Adventure	7.89	1.0	Action, Adventure, Comedy, Fantasy, Kids
2	Pokemon XY: Mega Evolution	7.70	1.0	Action, Adventure, Comedy, Fantasy, Kids
3	Pokemon XY	7.52	1.0	Action, Adventure, Comedy, Fantasy, Kids
4	Pokemon	7.43	1.0	Action, Adventure, Comedy, Fantasy, Kids
5	Pokemon Diamond and Pearl	7.23	1.0	Action, Adventure, Comedy, Fantasy, Kids
6	Pokemon Advanced Generation	7.20	1.0	Action, Adventure, Comedy, Fantasy, Kids
7	Pokemon XYandZ Specials	7.05	1.0	Action, Adventure, Comedy, Fantasy, Kids
8	Pokemon Best Wishes! Season 2: Episode N	7.04	1.0	Action, Adventure, Comedy, Fantasy, Kids
9	Pokemon Crystal: Raikou Ikazuchi no Densetsu	7.04	1.0	Action, Adventure, Comedy, Fantasy, Kids

Case 2
‘Kimi no Nawa’ (Your name)

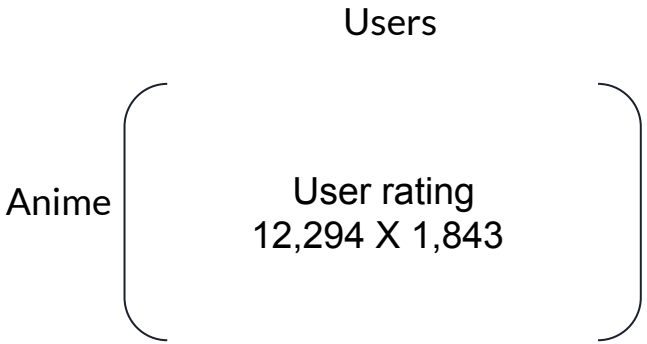
	Anime name	Rating	Similarity	Genre
0	Wind: A Breath of Heart OVA	6.35	1.000000	Drama, Romance, School, Supernatural
1	Wind: A Breath of Heart (TV)	6.14	1.000000	Drama, Romance, School, Supernatural
2	Aura: Maryuuin Kouga Saigo no Tatakai	7.67	0.893057	Comedy, Drama, Romance, School, Supernatural
3	Kokoro ga Sakebitagatterunda.	8.32	0.709423	Drama, Romance, School
4	Clannad: After Story - Mou Hitotsu no Sekai, K...	8.02	0.709423	Drama, Romance, School
5	True Tears	7.55	0.709423	Drama, Romance, School
6	Bungaku Shoujo Memoire	7.54	0.709423	Drama, Romance, School
7	Kimikiss Pure Rouge	7.48	0.709423	Drama, Romance, School
8	Myself; Yourself	7.41	0.709423	Drama, Romance, School
9	Koi to Senkyo to Chocolate	7.30	0.709423	Drama, Romance, School

Recommender

Collaborative filtering

Preprocessing
- Filter out animes with less than 500 users

Matrix factorization
- With *surprise* library *dataset*



Model selection for prediction
- Cross validated

	Test RMSE	Train time	Test time
SVG	1.930	46.173	2.061
KNNWithZScore	2.005	21.771	98.629
SlopeOne	2.022	23.476	103.182
NMF	2.165	54.137	2.255

Northeastern University

Recommender

Collaborative filtering

KNN recommender ‘Pokemon’

	Anime name
1	Pokemon Advanced Generation
2	Pokemon Kesshoutou no Teiou Entei
3	Pokemon Mewtwo no Gyakushuu
4	Pokemon Maboroshi no Pokemon Lugia Bakutan
5	Pokemon Celebi Toki wo Koeta Deai
6	Digimon Adventure
7	Pokemon Advanced Generation Mew to Hadou no Yu...
8	Pokemon Mizu no Miyako no Mamorigami Latias to...
9	Pokemon Advanced Generation Rekkuu no Houmonsh...
10	Pokemon Advanced Generation Nanayo no Negaibos...

SVD recommender ‘Pokemon’

	Anime name
1	Digimon Adventure
2	Pokemon Advanced Generation
3	Pokemon Advanced Generation Mew to Hadou no Yu...
4	Pokemon Advanced Generation Rekkuu no Houmonsh...
5	Pokemon Celebi Toki wo Koeta Deai
6	Pokemon Diamond amp Pearl
7	Pokemon Kesshoutou no Teiou Entei
8	Pokemon Maboroshi no Pokemon Lugia Bakutan
9	Pokemon Mewtwo no Gyakushuu
10	Pokemon Mizu no Miyako no Mamorigami Latias to...

Recommender

Collaborative filtering

Baseline RMSE = 3.951
- Error = |average_rating - user_rating|

SVD test RMSE = 1.095

Prediction results

	user_id	anime_id	rating
0	917	18	8.0
1	2243	63	7.0
2	6164	184	7.0
3	6525	189	8.0
4	11594	295	8.0
...	...	...	...
221898	56971	7812481	8.0
221899	57995	7812520	7.0
221900	58736	7812556	7.0
221901	66118	7812792	8.0
221902	68787	7812872	8.0

Challenges

- Topic change
  - No user data
  - Extremely messy data
  - Excessive NA
- Unused feature for EDA and recommender system: episodes
- No publication time for anime

Northeastern University

Thank you!

Questions?

Xianrui (Ray) She
Young Zhang
Ka Chun (Tim) Lee