MyAnimeList User Clustering Based on Anime Ratings

Francis dela Cruz, Luis Flores, and Alexandre Mariano
Department of Computer Science
College of Engineering
University of the Philippines, Diliman

***Abstract*—Given a dataset of around 12,000 anime and 72,000 users in MyAnimeList, the userbase was clustered and sorted into cluster groups, by first performing dimensionality reduction to 200 components using truncated SVD, then performing K-means on the transformed data, using 13 clusters as determined by the elbow method. The clusters revealed groups of users that seemed to predominantly prefer Action genres, with some showing preferences for Rom-Com anime. Half of the userbase was grouped as a cluster of those who didn't seem to rate their favorite shows a lot, which seems to imply that many people didn't rate the anime they watched. The clusters For future iterations of the project, other clustering methods like subspace clustering may be used, or other metrics like the Pearson's R coefficient or the cosine distance for clustering may be considered.**

I. Introduction

We live in the age of data. All kinds of information are collected from people by institutions both public and private, and while it can be used for malicious intent, it can also be used to optimize user experience. Given the advances in technology especially in the field of artificial intelligence, various techniques now exist to better interpret the massive amounts of data we are now capable of collecting. This is especially evident in the entertainment industry such as streaming sites recommending content to users based on what they previously viewed. The model created for this paper is our attempt to make sense of data collected from anime viewers, a large and continuously growing population.

Using the ratings given by users for anime titles, the model applies K-Means to cluster the viewers according to taste. This opens the door for user profiling and interaction. Users would be able to talk to fellow users who share in their taste. Companies can now determine which users prefer which anime and can look for patterns in age, nationality, gender, and so on. It is a simple means to interpret raw data in order to optimize the user experience of anime viewers.

II. Short Review of Related Studies

A. Clustering in High-Dimensional Spaces

Clustering high-dimensional data is the cluster analysis of data with dozens of dimensions or more. According to Wikipedia, clustering high-dimensional data is affected by what is known as the curse of dimensionality. "Multiple dimensions are hard to think in, impossible to visualize, and, due to the exponential growth of the number of possible values with each dimension, complete enumeration of all subspaces becomes intractable with increasing dimensionality". According to Quora, this means the distance between any two points in the dataset converges as the number of dimnesions tends toward infinity. The maximum and minimum distances between any two points will be the same, which is problematic for K-means since it uses euclidean distance. "One possible solution if you want to still use K-Means is to change the distance metric you use. I've used spherical K-Means, based on the cosine distance with thousands of features without any problems. It is a method that can be used to extract features from images and has results comparable to Deep Learning methods".

B. Dimensional Reduction

According to scikit-learn.org, TruncatedSVD implements a variant of singular value decomposition that only computes the k largest singular values where k is user-defined. "When truncated SVD is applied to term-document matrices (as returned by CountVectorizer or TfidfVectorizer), this transformation is known as latent semantic analysis (LSA), because it transforms such matrices to a semantic space of low dimensionality. In particular, LSA is known to combat the effects of synonymy and polysemy (both of which roughly mean there are multiple meanings per word), which cause term-document matrices to be overly sparse and exhibit poor similarity under measures such as cosine similarity".

C. K-Means algorithm

The K-means algorithm is a clustering algorithm defined by scikit-learn.org as clustering by attempting to separate samples into n groups of equal variance, minimizing the inertia or within-cluster sum-of-squares. "The k-means algorithm divides a set of N samples X into K disjoint clusters C, each described by the mean

$$\mu_j$$

of the samples in the cluster. The means are commonly called the cluster centroids; note that they are not, in general, points from X, although they live in the same space. The K-means algorithm aims to choose centroids that minimise the inertia, or within-cluster sum of squared criterion".

$$\sum_{i=0}^{n} \min_{\mu_j \in C}(||x_j - \mu_i||^2)$$

In basic terms, it is a three-step algorithm. First would be to choose initial centroids, with the most basic method being to choose k samples from the dataset X. After this, it simply loops between the next two steps. The first assigns the samples to the nearest centroid. The second creates new centroids by taking the mean value of all of the samples assigned to each previous centroid. The difference between the old and the new centroids are computed and the algorithm repeats these last two steps until this value is less than a threshold. In other words, it repeats until the centroids do not move significantly.

Although K-Means will always converge given enough time, it may be to a local minimum. ”This is highly dependent on the initialization of the centroids. As a result, the computation is often done several times, with different initializations of the centroids. One method to help address this issue is the k-means++ initialization scheme, which has been implemented in scikit-learn (use the init=’k-means++’ parameter). This initializes the centroids to be (generally) distant from each other, leading to provably better results than random initialization, as shown in the reference”. While K-Means does scale well to a large amount of data, it requires the number of clusters, or number of centroids, to be specified.

D. Determining the number of centroids to use

According to Wikipedia, the correct choice of k (number of centroids) is often ambiguous, with interpretations depending on the shape and scale of the distribution of points in a data set and the desired clustering resolution of the user. ”In addition, increasing k without penalty will always reduce the amount of error in the resulting clustering, to the extreme case of zero error if each data point is considered its own cluster (i.e., when k equals the number of data points, n). Intuitively then, the optimal choice of k will strike a balance between maximum compression of the data using a single cluster, and maximum accuracy by assigning each data point to its own cluster”. Two of the methods indicated were the elbow method and the silhouette method.

The elbow method shows that at a certain number of clusters, the elbow point, the increase in pecentage of variance explained will show diminishing returns. ”The elbow method looks at the percentage of variance explained as a function of the number of clusters: One should choose a number of clusters so that adding another cluster doesn’t give much better modeling of the data. More precisely, if one plots the percentage of variance explained by the clusters against the number of clusters, the first clusters will add much information (explain a lot of variance), but at some point the marginal gain will drop, giving an angle in the graph. The number of clusters is chosen at this point, hence the ”elbow criterion”...Percentage of variance explained is the ratio of the between-group variance to the total variance, also known as an F-test. A slight variation of this method plots the curvature of the within group variance”.

The silhouette method determines the number of clusters by assessing a criterion known as the silhouette of the data. ”The silhouette of a data instance is a measure of how closely it is matched to data within its cluster and how loosely it is matched to data of the neighbouring cluster, i.e. the cluster whose average distance from the datum is lowest.[7] A silhouette close to 1 implies the datum is in an appropriate cluster, while a silhouette close to 1 implies the datum is in the wrong cluster...It is also possible to re-scale the data in such a way that the silhouette is more likely to be maximised at the correct number of clusters”.

E. MyAnimeList

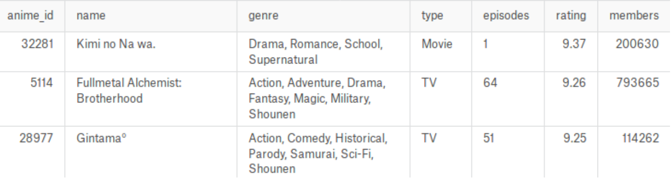
MyAnimeList is a database of anime TV shows, movies, and other forms of Japanese media. Users of the website can rate these anime, as well as give reviews to state their sentiments of the anime. The website doubles as a forum for discussion. The primary use, however, is for users to track down anime they completed, are watching, dropped, and are planning to watch.

III. Methodology

The methodology includes the preprocessing of the data into feature vectors, dimensionality reduction of the data space to allow for better clustering, finding the number of clusters to use, the actual clustering (care of the K-Means algorithm), and the interpretation of the clusters and centroids. Note that for randomized algorithms, a random seed of 180 was given, for reproducility’s sake.

A. Preprocessing

The dataset obtained from Kaggle is composed of 2 sets. The first is simply a list of animes, which would help map the anime IDs used to the actual animes they represent.



anime_id	name	genre	type	episodes	rating	members
32281	Kimi no Na wa.	Drama, Romance, School, Supernatural	Movie	1	9.37	200630
5114	Fullmetal Alchemist: Brotherhood	Action, Adventure, Drama, Fantasy, Magic, Military, Shounen	TV	64	9.26	793665
28977	Gintama°	Action, Comedy, Historical, Parody, Samurai, Sci-Fi, Shounen	TV	51	9.25	114262

Fig. 1. Data Header and three samples

The second is a list of records containing user IDs, the anime IDs of the anime they watched, and the scores that they gave that anime. Note that a score of -1 means it was unrated. For our purposes, we treated these ratings as if they didn’t exist, and that the user didn’t watch the anime, since it would be hard to infer their sentiment otherwise.

These datasets were preprocessed into feature vectors, where each vector / row represented a user, and each feature / column represented an anime. The value of the feature of that vector corresponds to the rating that user gave for that anime. This value is 0 if no rating is available. For ease of use, the anime ID, shifted to a 0-index, was used as the column or feature index of the vector. This made it easy to read the feature vector, but made it even sparser, since the distribution of anime IDs was not dense.

In the end, the sparse matrix contained 73516 users, and at most 34475 anime, although the dataset only contained 12295 anime in actuality.

B. Dimensional Reduction

Next, the sparse, high-dimensional matrix was moved to a lower dimensional space to allow clustering to proceed easier. Truncated SVD was used, as it worked well with sparse data. 200 was chosen for the number of resulting components (and thus dimensions), and 30 iterations were performed, as suggested by the documentation of the method. The resulting matrix had 73516 samples and 200 features

For validation, the total explained variance ratio of the 200 eigenvalues was 55%, which was fair enough given that 2 anime fans are unlikely to have the exact same tastes.

C. Determining a Good Number of Centroids to Use

K-means will be used as the clustering algorithm for the methodology. However, this raises the question of finding the number of clusters to use. One method is to look at the score of the K-Means model and how it varies with the number of clusters, and find the elbow point, or where it begins to experience diminishing returns. The silhouette scores will act as validation for this as well, as both their elbow points are in similar spots.

10% of the total dataset is used for repeated runs of K-means, which is done to graph the K-means score with respect to the increase of clusters.

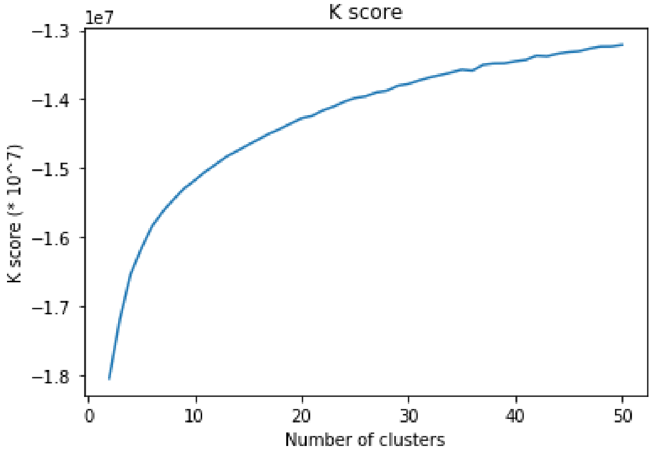


Fig. 2. Graph of K-means score vs number of clusters.

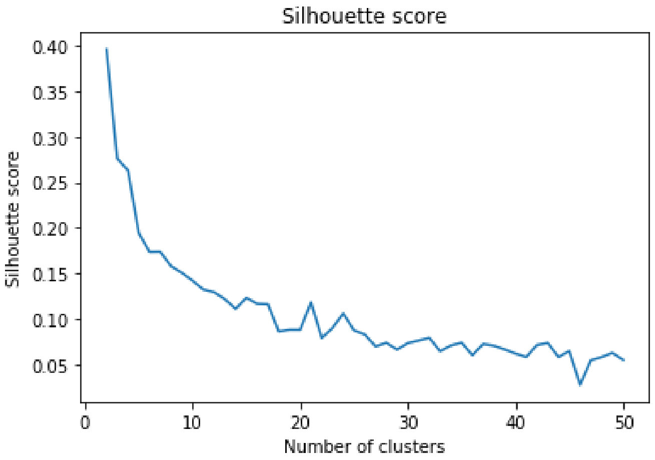


Fig. 3. Graph of Silhouette score vs number of clusters.

The elbow of the K-means score curve was then obtained. The elbow is defined as the further point of the curve from the line segment running from the first to the last point of the curve. In terms of vector algebra, it is defined as such:

$$v_{elbow} = \underset{1\leq i\leq n}{\arg\max}\, v_i \cdot v_{normal}$$

$$v_{normal} \perp (v_n - v_1)$$

Using these computations, a cluster count of 13 was obtained. The computed elbow points for both the K-means score and silhouette score curves can be found below.

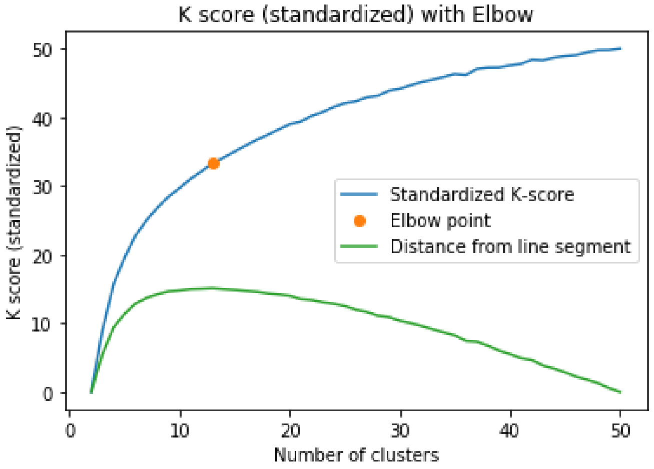


Fig. 4. Graph of K-means score vs number of clusters, with the elbow point.

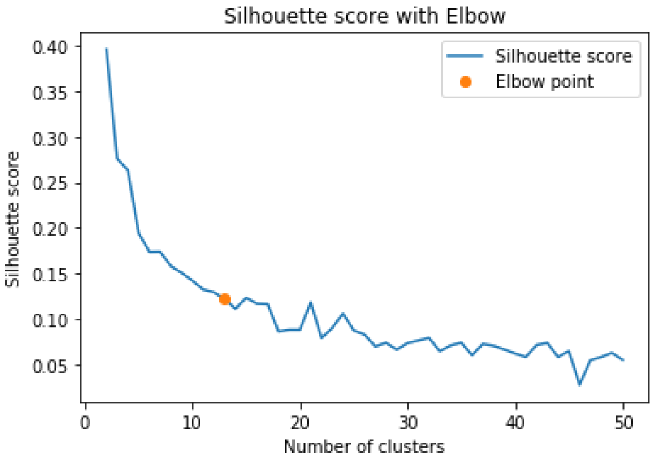


Fig. 5. Graph of Silhouette score vs number of clusters, with the elbow point.

D. Clustering

The actual clustering used a K-means model with 13 centroids, and which was trained on 60% of the dataset. After training, the entire dataset was labeled.

IV. Results

The 13 centroids were transformed back to the original space, representing anime ratings. Each of the 13 centroids were treated as Users whose tastes are supposedly representative of their clusters. Their top 15 anime, as well as their supposed ratings, are shown below.

Rank	Score	Anime
1	8.31	Code Geass: Hangyaku no Lelouch
2	8.16	Code Geass: Hangyaku no Lelouch R2
3	7.87	Fullmetal Alchemist: Brotherhood
4	7.77	Steins;Gate
5	7.61	Death Note
6	7.55	Tengen Toppa Gurren Lagann
7	7.49	Angel Beats!
8	7.42	Bakemonogatari
9	7.28	Durarara!!
10	7.22	Darker than Black: Kuro no Keiyakusha
11	7.21	Toradora!
12	7.10	Sen to Chihiro no Kamikakushi
13	7.09	Suzumiya Haruhi no Yuuutsu
14	7.07	Fate/Zero
15	7.01	Baccano!

Fig. 6. Top 15 Anime and ratings for User 1.

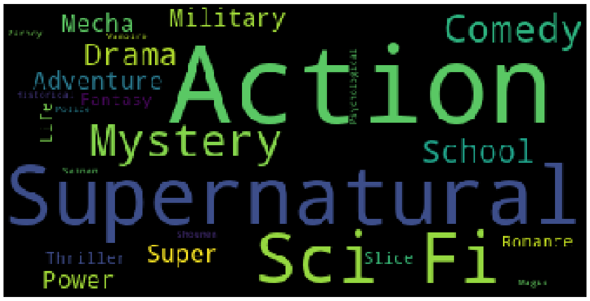


Fig. 7. Wordcloud of User 1's genres.

Rank	Score	Anime
1	1.48	Death Note
2	0.85	Naruto
3	0.82	Shingeki no Kyojin
4	0.79	Ouran Koukou Host Club
5	0.79	Sword Art Online
6	0.71	Angel Beats!
7	0.66	Elfen Lied
8	0.66	Sen to Chihiro no Kamikakushi
9	0.62	Fullmetal Alchemist
10	0.61	Fullmetal Alchemist: Brotherhood
11	0.52	Toradora!
12	0.51	Kuroshitsuji
13	0.50	Clannad
14	0.50	Code Geass: Hangyaku no Lelouch
15	0.46	Mahou Shoujo MadokaMagica

Fig. 8. Top 15 Anime and ratings for User 2.

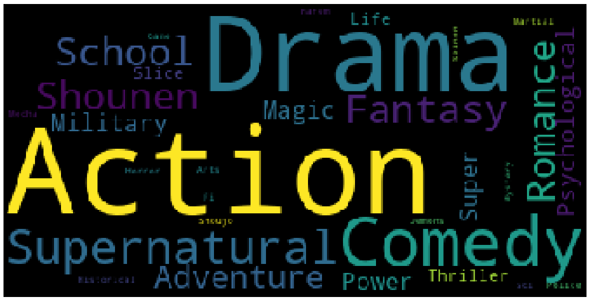


Fig. 9. Wordcloud of User 2's genres.

Rank	Score	Anime
1	7.77	Shingeki no Kyojin
2	7.09	Death Note
3	7.06	Sword Art Online
4	6.64	No Game No Life
5	6.63	Fullmetal Alchemist: Brotherhood
6	6.49	Tokyo Ghoul
7	6.11	One Punch Man
8	5.90	Steins;Gate
9	5.80	Code Geass: Hangyaku no Lelouch
10	5.53	Code Geass: Hangyaku no Lelouch R2
11	5.53	Kiseijuu: Sei no Kakuritsu
12	5.40	Akame ga Kill!
13	5.34	Mirai Nikki (TV)
14	5.05	Noragami
15	5.02	Angel Beats!

Fig. 10. Top 15 Anime and ratings for User 3.

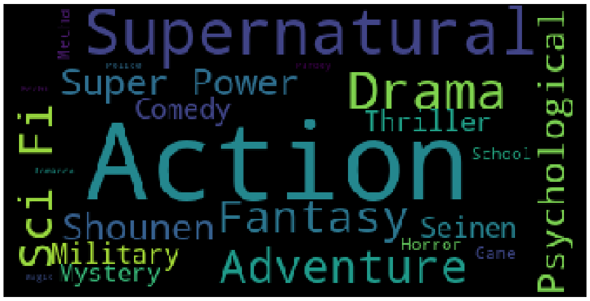


Fig. 11. Wordcloud of User 3's genres.

Rank	Score	Anime
1	6.79	Sen to Chihiro no Kamikakushi
2	5.86	Mononoke Hime
3	5.64	Howl no Ugoku Shiro
4	5.49	Death Note
5	5.31	Cowboy Bebop
6	4.98	Neon Genesis Evangelion
7	4.64	Fullmetal Alchemist
8	4.40	Tonari no Totoro
9	4.06	FLCL
10	3.78	Elfen Lied
11	3.77	Samurai Champloo
12	3.75	Tengen Toppa Gurren Lagann
13	3.69	Akira
14	3.67	Code Geass: Hangyaku no Lelouch
15	3.62	Fullmetal Alchemist: Brotherhood

Fig. 12. Top 15 Anime and ratings for User 4.

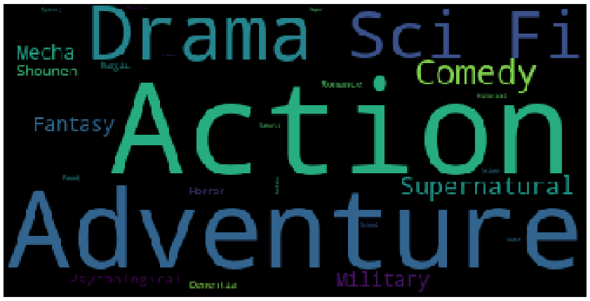


Fig. 13. Wordcloud of User 4's genres.

Rank	Score	Anime
1	8.67	No Game No Life
2	8.54	Angel Beats!
3	8.30	Sword Art Online
4	8.24	Toradora!
5	8.22	Yahari Ore no Seishun Love Comedy wa Machigatteiru.
6	8.16	Shingeki no Kyojin
7	8.10	Steins;Gate
8	7.97	Chuunibyou demo Koi ga Shitai!
9	7.90	Noragami
10	7.89	Log Horizon
11	7.80	Bakemonogatari
12	7.78	Hataraku Maou-sama!
13	7.55	Kami nomi zo Shiru Sekai
14	7.54	Sakurasou no Pet na Kanojo
15	7.54	Kyoukai no Kanata

Fig. 14. Top 15 Anime and ratings for User 5.

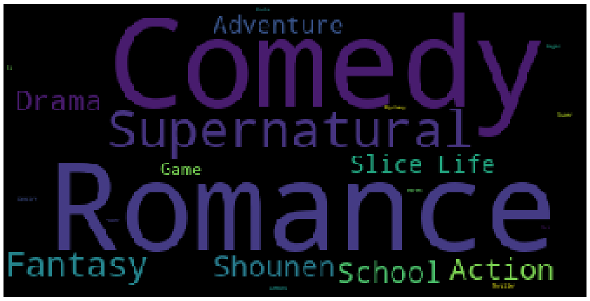


Fig. 15. Wordcloud of User 5's genres.

Rank	Score	Anime
1	7.40	Code Geass: Hangyaku no Lelouch
2	7.10	Death Note
3	6.71	Code Geass: Hangyaku no Lelouch R2
4	6.62	Elfen Lied
5	6.34	Fullmetal Alchemist
6	6.00	Suzumiya Haruhi no Yuuutsu
7	5.74	Toradora!
8	5.66	Full Metal Panic!
9	5.65	Clannad
10	5.37	Darker than Black: Kuro no Keiyakusha
11	5.33	Tengen Toppa Gurren Lagann
12	5.29	Fate/stay night
13	5.06	Fullmetal Alchemist: Brotherhood
14	5.05	Neon Genesis Evangelion
15	4.98	Full Metal Panic? Fumoffu

Fig. 16. Top 15 Anime and ratings for User 6.

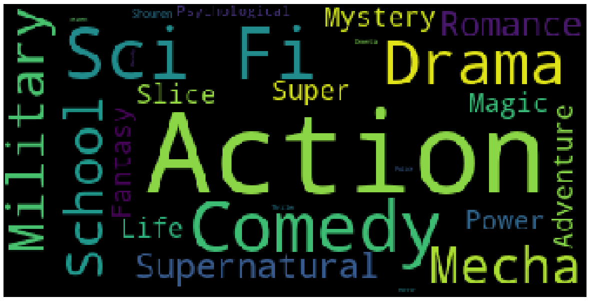


Fig. 17. Wordcloud of User 6's genres.

Rank	Score	Anime
1	8.41	No Game No Life
2	8.04	Sword Art Online
3	7.62	Shingeki no Kyojin
4	7.57	Angel Beats!
5	7.35	Dungeon ni Deai wo Motomeru no wa Machigatteiru Darou ka
6	7.25	Yahari Ore no Seishun Love Comedy wa Machigatteiru.
7	7.22	One Punch Man
8	7.19	Nisekoi
9	7.08	Akame ga Kill!
10	6.94	Sakurasou no Pet na Kanojo
11	6.92	Noragami
12	6.89	Toradora!
13	6.79	Hataraku Maou-sama!
14	6.75	Tokyo Ghoul
15	6.71	Steins;Gate

Fig. 18. Top 15 Anime and ratings for User 7.

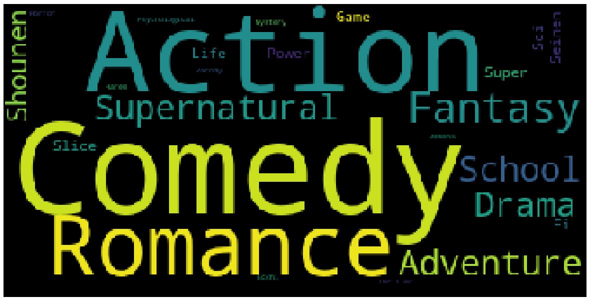


Fig. 19. Wordcloud of User 7's genres.

Rank	Score	Anime
1	6.27	Ouran Koukou Host Club
2	5.61	Death Note
3	5.44	Kuroshitsuji
4	5.02	Kaichou wa Maid-sama!
5	4.75	Vampire Knight
6	4.07	Sen to Chihiro no Kamikakushi
7	3.99	Shingeki no Kyojin
8	3.90	Vampire Knight Guilty
9	3.85	Ao no Exorcist
10	3.79	Toradora!
11	3.67	Kuroshitsuji II
12	3.62	Kimi ni Todoke
13	3.49	Howl no Ugoku Shiro
14	3.37	Sword Art Online
15	3.33	Durarara!!

Fig. 20. Top 15 Anime and ratings for User 8.

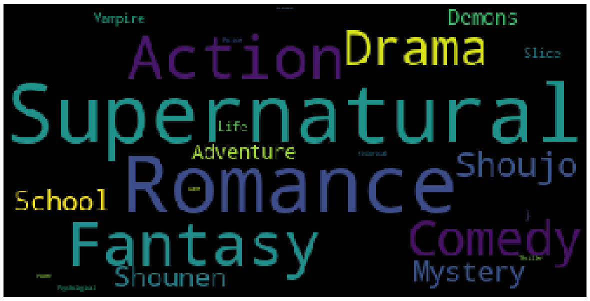


Fig. 21. Wordcloud of User 8's genres.

Rank	Score	Anime
1	6.81	Sword Art Online
2	6.73	Angel Beats!
3	5.80	Toradora!
4	5.12	High School DxD
5	5.09	Boku wa Tomodachi ga Sukunai
6	4.99	Mirai Nikki (TV)
7	4.97	Shingeki no Kyojin
8	4.86	Highschool of the Dead
9	4.83	Sakurasou no Pet na Kanojo
10	4.72	Clannad
11	4.61	Chuunibyou demo Koi ga Shitai!
12	4.44	Death Note
13	4.38	Clannad: After Story
14	4.36	Ore no Imouto ga Konnani Kawaii Wake ga Nai
15	4.33	Kore wa Zombie Desu ka?

Fig. 22. Top 15 Anime and ratings for User 9.

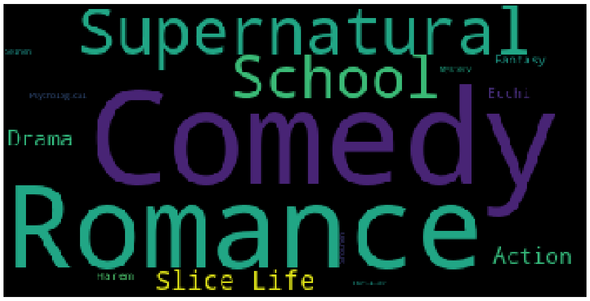


Fig. 23. Wordcloud of User 9's genres.

Rank	Score	Anime
1	8.02	Shingeki no Kyojin
2	7.36	Ouran Koukou Host Club
3	7.07	Ao no Exorcist
4	7.05	Free!
5	6.72	Death Note
6	6.72	Kuroshitsuji
7	6.53	Noragami
8	6.49	Durarara!!
9	6.46	Kuroko no Basket
10	6.39	Kaichou wa Maid-sama!
11	6.31	Tokyo Ghoul
12	6.29	Sen to Chihiro no Kamikakushi
13	6.20	Tonari no Kaibutsu-kun
14	6.10	Kamisama Hajimemashita
15	5.97	Sword Art Online

Fig. 24. Top 15 Anime and ratings for User 10.

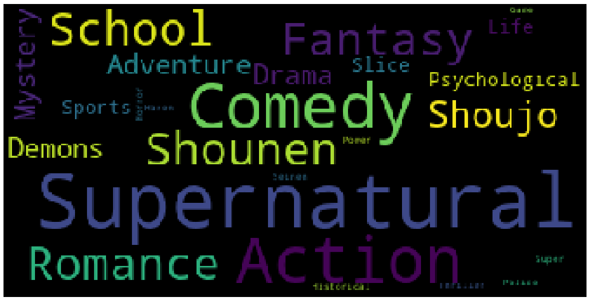


Fig. 25. Wordcloud of User 10's genres.

Rank	Score	Anime
1	5.99	Death Note
2	4.53	Shingeki no Kyojin
3	4.43	Sword Art Online
4	4.13	Fullmetal Alchemist: Brotherhood
5	3.89	Code Geass: Hangyaku no Lelouch
6	3.56	Code Geass: Hangyaku no Lelouch R2
7	3.54	Angel Beats!
8	3.28	Naruto
9	3.07	Elfen Lied
10	2.99	Mirai Nikki (TV)
11	2.90	Fullmetal Alchemist
12	2.71	Steins;Gate
13	2.69	Highschool of the Dead
14	2.53	Ao no Exorcist
15	2.42	Another

Fig. 26. Top 15 Anime and ratings for User 11.

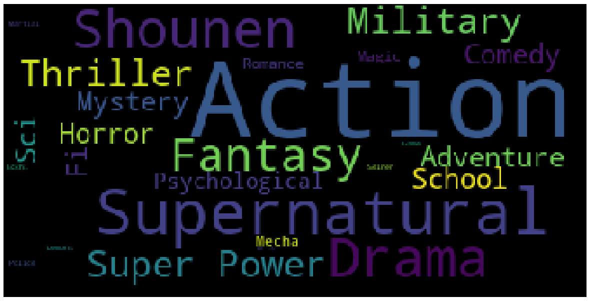


Fig. 27. Wordcloud of User 11's genres.

Rank	Score	Anime
1	8.34	Steins;Gate
2	7.07	Mahou Shoujo MadokaMagica
3	7.03	Bakemonogatari
4	6.75	Angel Beats!
5	6.51	Toradora!
6	6.46	Code Geass: Hangyaku no Lelouch
7	6.45	Death Note
8	6.43	Shingeki no Kyojin
9	6.39	Fullmetal Alchemist: Brotherhood
10	6.36	Tengen Toppa Gurren Lagann
11	6.12	Code Geass: Hangyaku no Lelouch R2
12	6.11	Ano Hi Mita Hana no Namae wo Bokutachi wa Mada Shiranai.
13	5.95	Suzumiya Haruhi no Yuuutsu
14	5.90	Clannad: After Story
15	5.87	Fate/Zero

Fig. 28. Top 15 Anime and ratings for User 12.

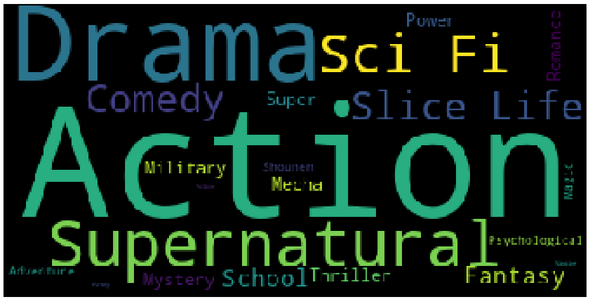


Fig. 29. Wordcloud of User 12's genres.

Rank	Score	Anime
1	8.02	Sword Art Online
2	7.73	Angel Beats!
3	7.50	Toradora!
4	7.46	Boku wa Tomodachi ga Sukunai
5	7.39	High School DxD
6	7.37	Kore wa Zombie Desu ka?
7	7.33	Zero no Tsukaima
8	7.28	Kami nomi zo Shiru Sekai
9	7.12	Highschool of the Dead
10	6.92	IS: Infinite Stratos
11	6.87	Hataraku Maou-sama!
12	6.81	Clannad
13	6.70	Code Geass: Hangyaku no Lelouch
14	6.67	Zero no Tsukaima: Futatsuki no Kishi
15	6.65	Kami nomi zo Shiru Sekai II

Fig. 30. Top 15 Anime and ratings for User 13.

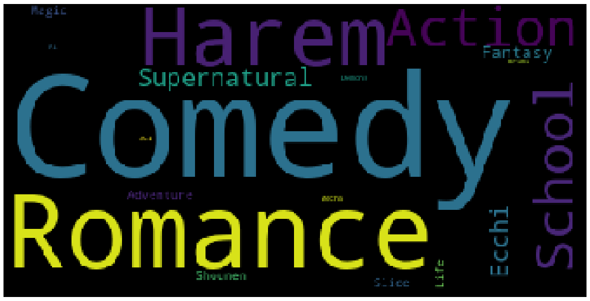


Fig. 31. Wordcloud of User 13's genres.

V. Discussion

A. Centroid Analysis

Each centroid can be considered a representative member of their group or label, with each having a set of genre preferences that define the entire group.

An important thing to mention is the uniqueness of centroid / User 2, where its ratings only range from 0-1, compared to the normal users who rates at least an average of 3 or above. This could be due to the fact that some MAL users only rate a few anime or none at all, so this may be the cluster that captures those users.

The list below shows each of the centroids as representative users, and their top three top-rated anime genres, arranged in decreasing order.

- **User 1:** Action, Supernatual, Sci-Fi
- **User 2:** Action, Drama, Comedy
- **User 3:** Action, Supernatural, Drama/Adventure
- **User 4:** Adventure, Action, Drama/Sci-Fi
- **User 5:** Romance, Comedy, Supernatural
- **User 6:** Action, Sci-Fi, Comedy
- **User 7:** Comedy, Romance, Action
- **User 8:** Supernatural, Romance, Action/Fantasy
- **User 9:** Comedy, Romance, School/Supernatural
- **User 10:** Supernatural, Comedy, Action
- **User 11:** Action, Supernatural, Shounen
- **User 12:** Action, Drama, Supernatural
- **User 13:** Comedy, Romance, Harem

It is interesting to note that a lot of these centroids have Action as the most frequent genre. These centroids also share a lot of common genres (albeit in a somewhat different order) due to the fact that there are a lot of common and shared anime between centroids, such as Death Note, Attack on Titan, and others. Aside from Action, there are also many clusters that mostly enjoy the Rom-Com genre.

While a lot of centroids have similar tastes, some are of particular interest. User 4, for one, seems to be a fan of Miyazaki films (Top 1: Spirited Away, Top 2: Princess Mononoke, Top 3: Howl's Moving Castle, Top 8: My Neighbor Totoro). User 8 seems to be a fan of Romance and Fantasy anime such as Ouran High School Host Club, Black Butler, Vampire Knight, and Kimi ni Todoke. User 13 seems to be a fan of Rom-Com and Harem animes like Sword Art Online, Toradora, Haganai, The World God Only Knows, High School DxD, and others.

B. Cluster Analysis

The cluster sizes have a fair amount of variety, with an overwhelming 42.85% of users in the dataset belonging to Cluster 2, which may imply that a lot of users don't rate the anime they watch. The proportions also suggest that a lot of users seem to fall under fans of the Action, Supernatural, and/or Shounen genres. Granted, there is a large overlap between cluster groups in terms of taste, so this may not be an accurate enough generalization.

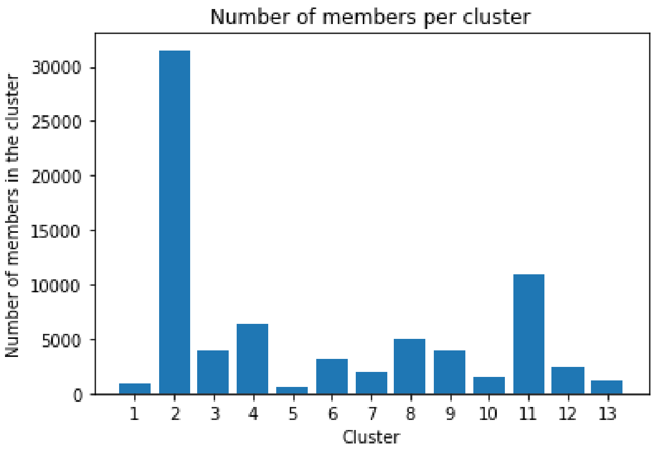


Fig. 32. A graph showing the cluster sizes.

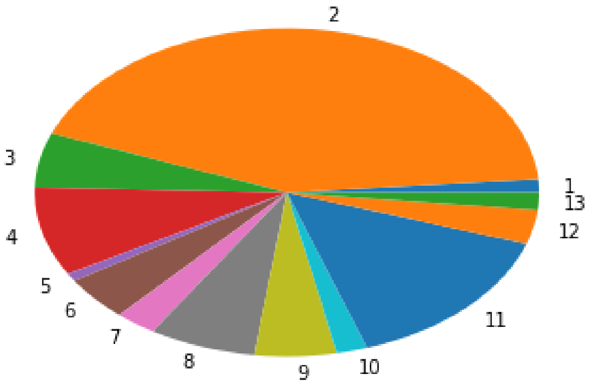
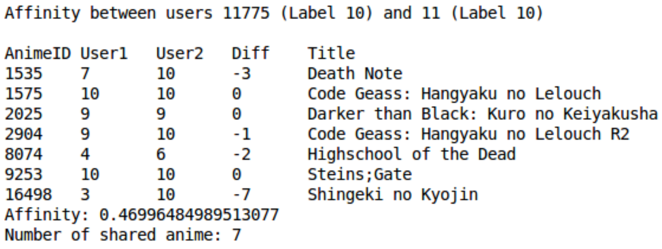


Fig. 33. A pie chart showing the proportions of the cluster sizes.

The silhouette score of the K-means model was computed to be 0.1059, which may imply a large amount of overlap. However, this value was calculated using only 30% of the dataset, so its reliability as a metric is up for debate.

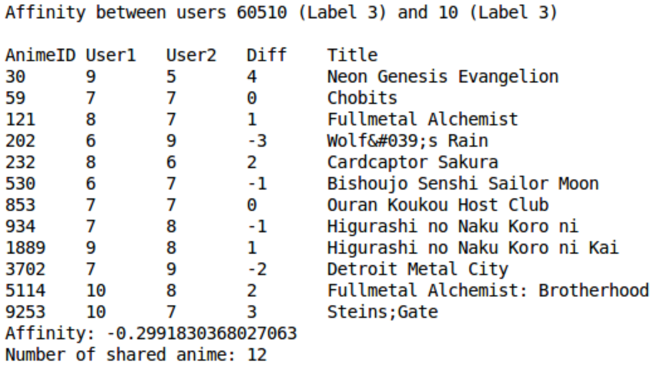
The affinity between two users can be measured by finding the Pearson's R coefficient of their rated anime. While some users of the same label share similar tastes, some pairs can have low affinity scores, as opposed to pairs from different clusters. This could be due to the fact that the methodology favors pairs with more shared anime, regardless of their scores and affinity.



```
Affinity between users 11775 (Label 10) and 11 (Label 10)

AnimeID User1   User2   Diff    Title
1535    7       10      -3      Death Note
1575    10      10      0       Code Geass: Hangyaku no Lelouch
2025    9       9       0       Darker than Black: Kuro no Keiyakusha
2904    9       10      -1      Code Geass: Hangyaku no Lelouch R2
8074    4       6       -2      Highschool of the Dead
9253    10      10      0       Steins;Gate
16498   3       10      -7      Shingeki no Kyojin
Affinity: 0.46996484989513077
Number of shared anime: 7
```

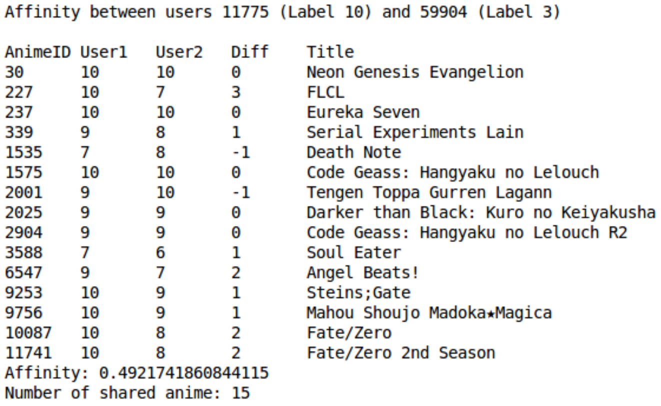
Fig. 34. An example of 2 users from the same cluster group, with high affinity.



```
Affinity between users 60510 (Label 3) and 10 (Label 3)

AnimeID User1   User2   Diff    Title
30      9       5       4       Neon Genesis Evangelion
59      7       7       0       Chobits
121     8       7       1       Fullmetal Alchemist
202     6       9       -3      Wolf&#039;s Rain
232     8       6       2       Cardcaptor Sakura
530     6       7       -1      Bishoujo Senshi Sailor Moon
853     7       7       0       Ouran Koukou Host Club
934     7       8       -1      Higurashi no Naku Koro ni
1889    9       8       1       Higurashi no Naku Koro ni Kai
3702    7       9       -2      Detroit Metal City
5114    10      8       2       Fullmetal Alchemist: Brotherhood
9253    10      7       3       Steins;Gate
Affinity: -0.2991830368027063
Number of shared anime: 12
```

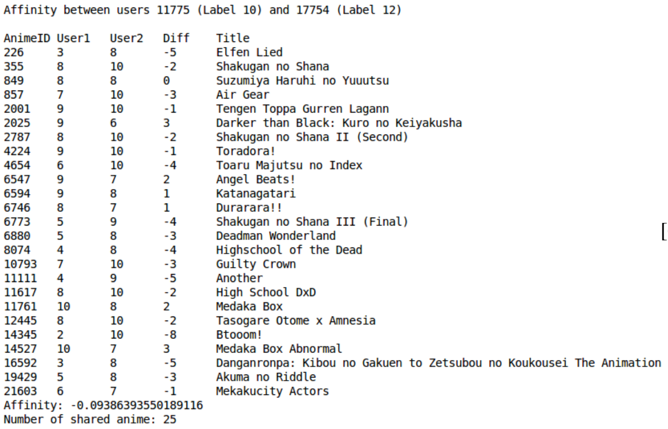
Fig. 35. An example of 2 users from the same cluster group, with low affinity.



```
Affinity between users 11775 (Label 10) and 59904 (Label 3)

AnimeID User1   User2   Diff    Title
30      10      10      0       Neon Genesis Evangelion
227     10      7       3       FLCL
237     10      10      0       Eureka Seven
339     9       8       1       Serial Experiments Lain
1535    7       8       -1      Death Note
1575    10      10      0       Code Geass: Hangyaku no Lelouch
2001    9       10      -1      Tengen Toppa Gurren Lagann
2025    9       9       0       Darker than Black: Kuro no Keiyakusha
2904    9       9       0       Code Geass: Hangyaku no Lelouch R2
3588    7       6       1       Soul Eater
6547    9       7       2       Angel Beats!
9253    10      9       1       Steins;Gate
9756    10      9       1       Mahou Shoujo Madoka★Magica
10087   10      8       2       Fate/Zero
11741   10      8       2       Fate/Zero 2nd Season
Affinity: 0.4921741860844115
Number of shared anime: 15
```

Fig. 36. An example of 2 users from different cluster groups, with high affinity.



```
Affinity between users 11775 (Label 10) and 17754 (Label 12)

AnimeID User1   User2   Diff    Title
226     3       8       -5      Elfen Lied
355     8       10      -2      Shakugan no Shana
849     8       8       0       Suzumiya Haruhi no Yuuutsu
857     7       10      -3      Air Gear
2001    9       10      -1      Tengen Toppa Gurren Lagann
2025    9       6       3       Darker than Black: Kuro no Keiyakusha
2787    8       10      -2      Shakugan no Shana II (Second)
4224    9       10      -1      Toradora!
4654    6       10      -4      Toaru Majutsu no Index
6547    9       7       2       Angel Beats!
6594    9       8       1       Katanagatari
6746    8       7       1       Durarara!!
6773    5       9       -4      Shakugan no Shana III (Final)
6880    5       8       -3      Deadman Wonderland
8074    4       8       -4      Highschool of the Dead
10793   7       10      -3      Guilty Crown
11111   4       9       -5      Another
11617   8       10      -2      High School DxD
11761   10      8       2       Medaka Box
12445   8       10      -2      Tasogare Otome x Amnesia
14345   2       10      -8      Btooom!
14527   10      7       3       Medaka Box Abnormal
16592   3       8       -5      Danganronpa: Kibou no Gakuen to Zetsubou no Koukousei The Animation
19429   5       8       -3      Akuma no Riddle
21603   6       7       -1      Mekakucity Actors
Affinity: -0.09386393550189116
Number of shared anime: 25
```

Fig. 37. An example of 2 users from different cluster groups, with low affinity.

VI. Conclusion

The K-means model was able to cluster the users in the dataset, with some clusters being very similar in taste and having similar top choices, and others being more distinct, like clusters 2, 4, 8, and 13. The model also revealed that a lot of users don't rate a lot of the anime they watch.

However, the Pearson's R coefficient was not optimized as a metric, as their are pairs in the same group with low affinity, and vice-versa. Overall, it can be said that clustering, especially with K-means, can be performed in similar datasets, with interpretable results. However, it can be said that there exist better, albeit more difficult, methods.

VII. Recommendations

For future iterations of this project, it is recommended to balance out the scores by applying an exponential scale to give more weight to higher scores. This is needed to be done because users have a varying rating scales, with some using 0-10, some using 6-10, and so on.

It is also recommended to use better methods to cluster a sparse and high-dimenstional space (like subspace clustering). If dimensional reduction is still to be used in conjunction with clustering, it would be better to use a more reasonable dimension or distance metric, as the metric for choosing the number of components used in this paper was not evaluated.

Finally, a clustering method that will use Pearson's R coefficient (or something similar) as a criterion may be used in the future, as this is the metric chosen for determining the affinity of 2 users.

References

[1] Argerich, L. (2015, May 12). What happens when you try clustering data with higher dimensions using k-means? For example, if the dimensionality of the data set is 1000, number of clusters is 10, and there are 50 000 samples set. How effective is K-means for this kind of data? Retrieved May 30, 2018, from https://www.quora.com/What-happens-when-you-try-clustering-data-with-higher-dimensions-using-k-means-For-example-if-the-dimensionality-of-the-data-set-is-1000-number-of-clusters-is-10-and-there-are-50-000-samples-set-How-effective-is-K-means-for-this-kind-of-data

[2] Clustering high-dimensional data. (2018, January 23). Retrieved from https://en.wikipedia.org/wiki/Clustering_high-dimensional_data

[3] Clustering. (n.d.). Retrieved from http://scikit-learn.org/stable/modules/clustering.html

[4] Determining the number of clusters in a data set. (2018, May 16). Retrieved from http://scikit-learn.org/stable/modules/clustering.html

[5] Singh, M. (2017, October 18). Finding the elbow or knee of a curve. Retrieved May 30, 2018, from https://dataplatform.ibm.com/analytics/notebooks/54d79c2a-f155-40ec-93ec-ed05b58afa39/view?access$_{token} = 6d8ec910cf2a1b3901c721fcb94638563cd646fe14400fecbb76cea6aaae2fb1$

[6] L. (n.d.). Anime recommendation based on user clustering. Retrieved May 30, 2018, from https://www.kaggle.com/tanetboss/user-clustering-for-anime-recommendation