



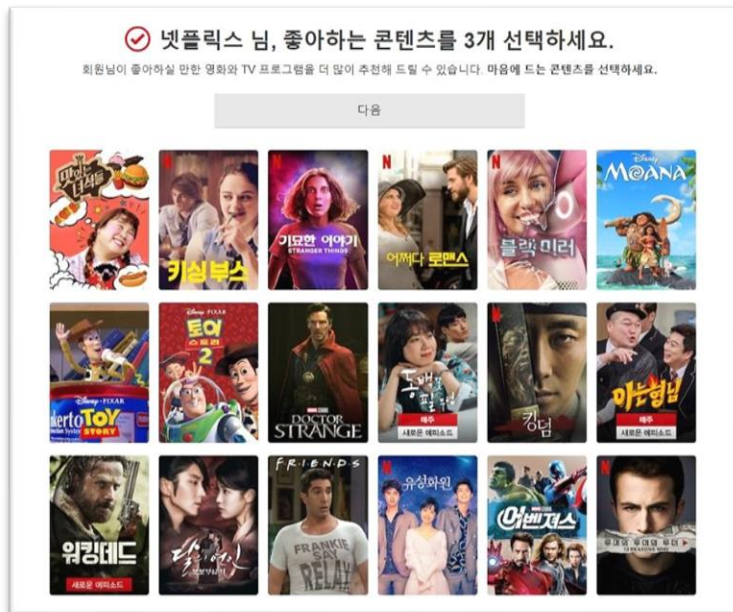
# Animation Recommendation System

---

- 20191314 Jihyuk Choi
- 20201181 Jihwan Oh

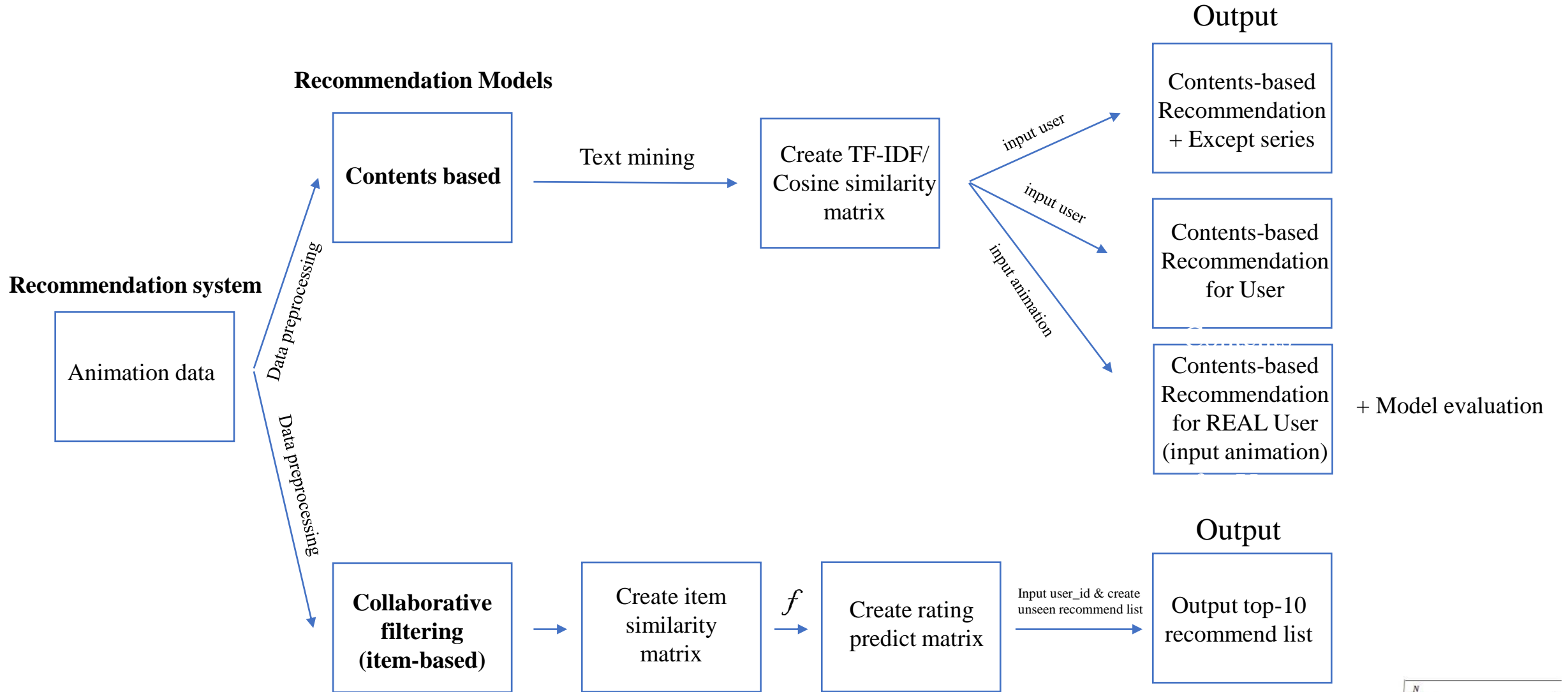
# Introduction

1. What is **Recommendation System**?  
: As a kind of information filtering technology, information that a specific user may be interested in is recommended. The recommendation system is divided into a contents-based model and a collaborative filtering model.
2. **Motivation**: We want to understand the principles of various recommendation systems, develop our own recommendation systems, and actually recommend them to people. Explore and find out recommended models other than the late factor model learned in class.
3. How do video **streaming services** such as Netflix recommend contents?





# Project process



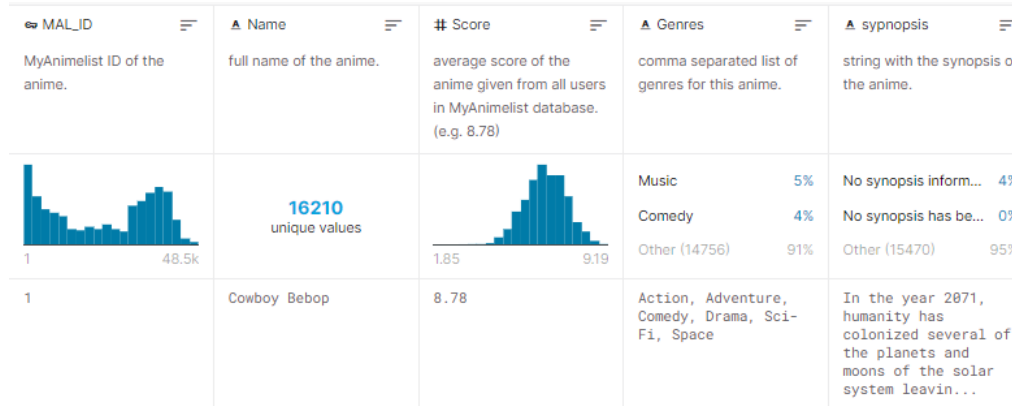
$$RMSE = \sqrt{\frac{\sum_{i=1}^N (Predicted_i - Actual_i)^2}{N}}$$

$f$ : Find the best forecasting model with the smallest RMSE(Root Mean Squared Error) value.

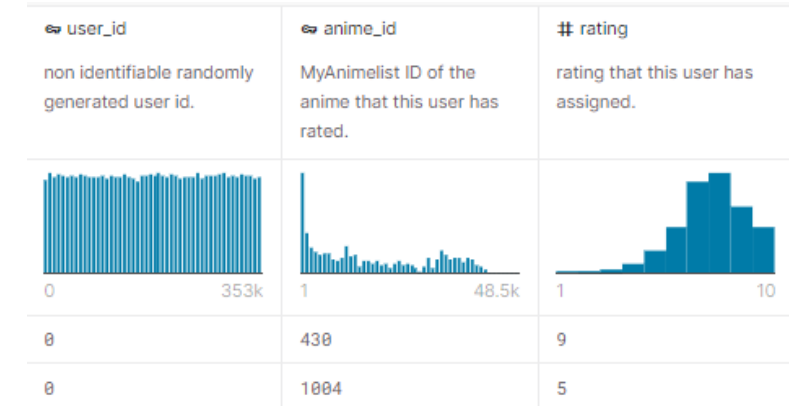
# Contents based – Problem & Used data

## Data:

<Animation information with genres and synopsis>



<User id with animation rating>



We got this data from <https://www.kaggle.com/hernan4444/anime-recommendation-database-2020>

## Data preprocessing:

```
df=df.rename(columns={'MAL_ID':'anime_id'})
```



```
[ ] 1 df.isnull().sum()
```

```
anime_id    0
Name         0
Score        0
Genres       0
synopsis     8
dtype: int64
```

```
[ ] 1 df.dropna(inplace = True)
    2 df.drop("Score", axis = 1, inplace = True)
```

```
[ ] 1 df.isnull().sum()
```

```
anime_id    0
Name         0
Genres       0
synopsis     0
dtype: int64
```



```
1 df.shape
(16206, 4)
```

```
1 df.head()
```

anime_id	Name	Genres	synopsis
0	1	Cowboy Bebop	Action, Adventure, Comedy, Drama, Sci-Fi, Space
1	5	Cowboy Bebop: Tengoku no Tobira	Action, Drama, Mystery, Sci-Fi, Space
2	6	Trigun	Action, Sci-Fi, Adventure, Comedy, Drama, Shounen
3	7	Witch Hunter Robin	Action, Mystery, Police, Supernatural, Drama, ...
4	8	Bouken Ou Beet	Adventure, Fantasy, Shounen, Supernatural

## Problem definition:

**Service 1:** First of all, a service that receives input as an animation using only content characteristics (genre, synopsis) and recommends related animations through output.

**Service 2:** Then, considering user-rating data, a service that provides customized recommendations to users.

# Contents based – Result analysis 1

## Text-mining with TF-IDF:

$$w_{x,y} = \text{tf}_{x,y} \times \log\left(\frac{N}{\text{df}_x}\right)$$

**TF-IDF**

Term  $x$  within document  $y$

$\text{tf}_{x,y}$  = frequency of  $x$  in  $y$   
 $\text{df}_x$  = number of documents containing  $x$   
 $N$  = total number of documents

## Cosine similarity

$$\text{Sim}(u, v)^{\text{COS}} = \frac{\vec{r}_u \cdot \vec{r}_v}{\|\vec{r}_u\| \cdot \|\vec{r}_v\|} = \frac{\sum_{i \in I_u \cap I_v} r_{ui} \cdot r_{vi}}{\sqrt{\sum_{i \in I_u \cap I_v} r_{ui}^2} \cdot \sqrt{\sum_{i \in I_u \cap I_v} r_{vi}^2}}$$

```
array([[1.          , 0.67192397, 0.59656493, ..., 0.          , 0.20889977,
        0.17826774],
       [0.67192397, 1.          , 0.42077551, ..., 0.14595344, 0.          ,
        0.16450843],
       [0.59656493, 0.42077551, 1.          , ..., 0.          , 0.22814966,
        0.1864261 ],
       ...,
       [0.          , 0.14595344, 0.          , ..., 1.          , 0.          ,
        0.          ],
       [0.20889977, 0.          , 0.22814966, ..., 0.          , 1.          ,
        0.00330979],
       [0.17826774, 0.16450843, 0.1864261 , ..., 0.          , 0.00330979,
        1.          ]])
```

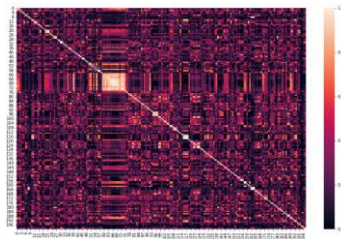
## Genre Text-mining

```
1 vectorizer = TfidfVectorizer(stop_words = "english", analyzer = "word",)
2 Gen_TF_IDF_matrix = vectorizer.fit_transform(df[["Genres"]])
3 Gen_TF_IDF_matrix.shape
```

(16206, 45)

```
1 |vectorizer.vocabulary_
```

```
{'action': 0,
 'adventure': 1,
 'ai': 2,
 'arts': 3,
 'cars': 4,
 'comedy': 5,
 'dementia': 6,
 'demons': 7,
 'drama': 8,
```



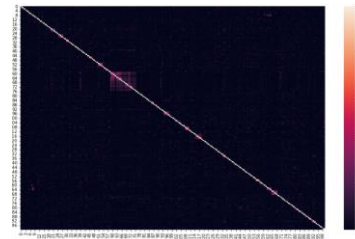
<Genre cosine similarity – heatmap>

## Synopsis Text-mining

```
1 vectorizer = TfidfVectorizer(stop_words = "english", analyzer = "word")
2 Syp_TF_IDF_matrix = vectorizer.fit_transform(df[["synopsis"]])
3 Syp_TF_IDF_matrix.shape
```

(16206, 45064)

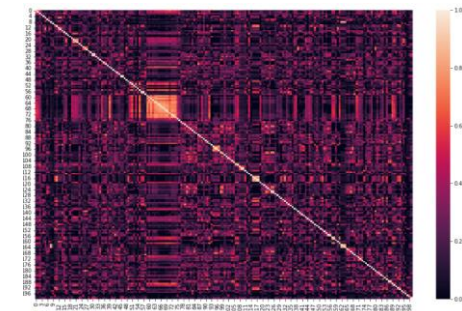
```
{'year': 44035,
 '2071': 384,
 'humanity': 17426,
 'colonized': 7306,
 'planets': 30048,
 'moons': 25742,
 'solar': 36746,
 'leaving': 22460,
 'uninhabitable': 41682,
 'surface': 38335,
 'planet': 30041,
 'earth': 10924,
 'inter': 18640,
 'police': 30268,
 'attempts': 3278,
 'peace': 29421,
 'galaxy': 14064,
 'aided': 1425,
 'outlaw': 28756,
 'bounty': 5037,
 'hunters': 17472,
 'referred': 32113,
 'combustion': 8198,
```



:Synopsis cosine similarity – heatmap>

## Combination

```
1 Fin_cos_sim = 0.8*Gen_cos_sim+0.2*Syp_cos_sim
2 Fin_cos_sim
```



<Final cosine similarity – heatmap>

# Contents based – Result analysis 2

## Contents-based recommendation

```
1 content_based_recommender(df, anime_name = "Naruto", rec_count = 10)
```

	anime_id	Name	Genres	synopsis	similarity_score
10	20	Naruto	Action, Adventure, Comedy, Super Power, Martial Arts, ...	Events prior to Naruto Uzumaki's birth, a huge...	1.000000
1508	1735	Naruto: Shippuuden	Action, Adventure, Comedy, Super Power, Martial Arts, ...	It has been two and a half years since Naruto ...	0.872604
11346	34566	Boruto: Naruto Next Generations	Action, Adventure, Super Power, Martial Arts, ...	Following the successful end of the Fourth Shi...	0.856797
6158	13667	Naruto: Shippuuden Movie 6 - Road to Ninja	Action, Adventure, Super Power, Martial Arts, ...	Returning home to Konohagakure, the young ninja c...	0.853405
8831	28755	Boruto: Naruto the Movie	Action, Comedy, Martial Arts, Shounen, Super P...	The spirited Boruto Uzumaki, son of Seventh Ho...	0.827924
4598	8246	Naruto: Shippuuden Movie 4 - The Lost Tower	Action, Comedy, Martial Arts, Shounen, Super P...	Assigned on a mission to capture Mukade, a miss...	0.819015
5518	10686	Naruto: Honoo no Chuunin Shiken! Naruto vs. Ko...	Action, Adventure, Martial Arts, Shounen, Supe...	Naruto faces off against his old pupil Konoham...	0.812497
10244	32365	Boruto: Naruto the Movie - Naruto ga Hokage ni...	Action, Comedy, Super Power, Martial Arts, Sho...	Bundled with the limited edition of Blu-ray/DV...	0.807435
3904	6325	Naruto: Shippuuden Movie 3 - Hi no Ishi wo Tsu...	Action, Comedy, Martial Arts, Shounen, Super P...	Ninjas with bloodline limits begin disappearin...	0.797498
11640	35072	Boruto: Jump Festa 2016 Special	Action, Adventure, Comedy, Super Power, Martia...	The special anime adaptation of Boruto will be...	0.200624

## Contents-based recommendation(Except series)

```
1 content_based_recommender_except_Series(df, anime_name = "Naruto", rec_count = 10)
```

	anime_id	Name	Genres	synopsis	similarity_score
11	21	One Piece	Action, Adventure, Comedy, Super Power, Drama, ...	Gol D. Roger was known as the "Pirate King." L...	0.554792
11666	35104	Christmas Carol	Kids, Supernatural	anime adaptation of Charles Dickens' A Christm...	0.479910
5537	10717	Towa no Quon 6: Towa no Quon	Action, Sci-Fi, Super Power, Supernatural	The story follows a boy named Quon and others ...	0.390219
1515	1742	Souryuuden	Mystery, Super Power, Supernatural, Drama	The four Ryudo brothers hold an ancient secret...	0.328704
11372	34629	MiniGARO	Comedy	short web-anime created for the "miniGARO" pro...	0.155219
8855	28853	Pokemon: Pikachu to Pokemon Ongakutai	Adventure, Kids, Fantasy	Pikachu and the rest of the group practice sin...	0.121063
3917	6372	Higashi no Eden Movie I: The King of Eden	Comedy, Drama, Mystery, Romance, Slice of Life...	After preventing Japan's destruction, Akira Tak...	0.037810
6179	13795	The Green Wind	Fantasy	Promotion video for 21st Century Museum of Con...	0.000000
4613	8312	Piece	Music	music video to the song Piece by singer Aragak...	0.000000
10269	32420	Ken-chan	Music, Kids	music video for the song "Ken-chan" by Yuuki K...	0.000000

## Contents-based recommendation for User

```
df2.drop(df2[df2['user_id']>10000].index, axis=0, inplace=True)
df2.drop(df2[df2['rating']<10].index, axis=0, inplace=True)
```

user_id	anime_id	rating
6	0	578
8	0	1571
20	0	415
21	0	2236
136	1	11577
...	...	...
1048531	6745	33255
1048529	6745	30727
1048526	6745	28621
1048550	6745	32951
1048573	6745	35466

125446 rows x 3 columns

```
1 # 10 recommendations each other(Some of them might be the same.)
2 content_based_recommender(df, anime_id = 578, rec_count = 10)
3 content_based_recommender(df, anime_id = 1571, rec_count = 10)
4 content_based_recommender(df, anime_id = 415, rec_count = 10)
5 content_based_recommender(df, anime_id = 2236, rec_count = 10)
```

Duplicate recommended items are deleted except one, and a recommended score is created by combining the simplicity score. If the recommendation score is more than 1, at least two recommendations are combined.

$$\text{Recommendation\_score} = \sum \text{similarity\_score}$$

```
1 def user_recommender_top_n(df2, user_id=0, top_n=10):
```

```
1 user_recommender_top_n(df2, user_id=73, top_n=10)
```

```
** Here is user73 information **
number of animations that user have seen: 498
number of animations that the user gave 10 points: 29
number of whole recommendation for user: 231
number of animations in a recommendation that user have seen: 62
number of animations in a recommendation that user have not seen: 169
```

	recommendation_score
Kagewani	2.093175
Mi Yu Xing Zhe	1.805161
Code Geass: Hangyaku no Lelouch I - Koudou	1.622466
Yamada-kun to 7-nin no Majo: Mou Hitotsu no Suzaku-sai	1.618745
The Samurai	1.618589
Denpa Kyoushi (TV)	1.609665
Code Geass: Hangyaku no Lelouch III - Oudou	1.598034
Code Geass: Hangyaku no Lelouch II - Handou	1.595694
Code Geass: Fukkatsu no Lelouch	1.592043
Code Geass: Boukoku no Akito 3 - Kagayaku Mono Ten yori Otsu	1.546247

# Contents based – Evaluate our system

## 1. Quantitative evaluation

### Personalization :

It is an indicator for evaluating whether the recommendation system recommends the same product to other users equally or differently. That is, the dissimilarity between the recommended product lists of users is obtained.

user – item matrix

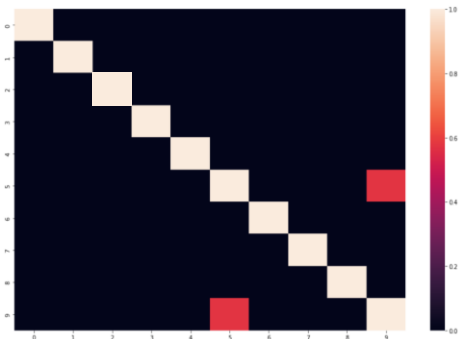
	0	1	2	3	4	5	6	7	8	9	10	...	190	191	192	193	194	195	196	197	198	199
0	0	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	1	...	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0

10 rows × 200 columns

Recommended item:1 , Not recommended item: 0

```
1 evaluation_cos_sim = cosine_similarity(evaluation_matrix_df)
2 evaluation_cos_sim.shape
```

(10, 10)



Cosine similarity =  $0.6/45 = 0.013$

Dissimilarity =  $1 - \text{Cosine similarity}$

Dissimilarity = 0.987



Our recommend system is recommended differently depending on the individual.

## 2. Qualitative evaluation

Contents-based recommendation for REAL User (Input animation)

```
def anime_input_recommendation_except_Series(df, top_n=10):
```

```
1 anime_input_recommendation_except_Series(df, top_n=10)
Enter number of your best animations: 3
```

```
1 anime_input_recommendation_except_Series(df, top_n=10)
Enter number of your best animations: 3
Enter your best animations name: Naruto
Enter your best animations name: One Piece
Enter your best animations name: Bleach
```

```
1 anime_input_recommendation_except_Series(df, top_n=10)
```

Enter number of your best animations: 3  
Enter your best animations name: Naruto  
Enter your best animations name: One Piece  
Enter your best animations name: Bleach  
Here are your best animations: ['Naruto', 'One Piece', 'Bleach']

number of whole recommendation for user: 30

\*\* Here is top 10 recommendation \*\*

recommendation_score	
Name	
Shaman King	0.804254
Erementar Gerad	0.634507
Christmas Carol	0.479910
Towa no Quon 6: Towa no Quon	0.390219
Souryuuden	0.328704
Tennis no Ouji-sama	0.257246
Yu☆Gi☆Oh! Duel Monsters	0.249106
Kaze no Stigma	0.213254
Ai no Senshi Rainbowman	0.191297
Goku Sayonara Zetsubou Sensei	0.189183

My best animations




Top-3 recommended animations



# Item-based Collaborative Filtering - Data Preprocessing

- Filtering recommendations based on behaviors such as ratings and purchase history left by users.
- The similarity between items is measured and the rating is predicted using the evaluation of similar items.

User ID	Item ID	Rating
User 1	Item 1	3
User 1	Item 3	4
User 2	Item 2	2
User 3	Item 3	5



	Item 1	Item 2	Item 3
User 1	3		4
User 2		2	
User 3			5

	User 1	User 2	User 3	User 4
다크 나이트	5	5	3	4
테넷	5	4	3	
인터스텔라	4	5	2	

We use 2 dataset, animelist.csv and anime.csv.

animelist.csv : rating file

anime.csv : animation information file

But there was a problem.

The item\_id of animelist.csv and anime.csv is different.  
So, we need to leave the item\_id that exists in anime\_csv through the preprocessing process.

However, animelist.csv has 109 million rows and anime.csv has 16000 animation data.

As a result of checking through the computer, it takes about 76 days for all processing to take place, so we decided to proceed by selecting three types of data by setting conditions.

$$109 \times 10^6 \div 1000 \div 60 \div 24 =$$

75.6944444444

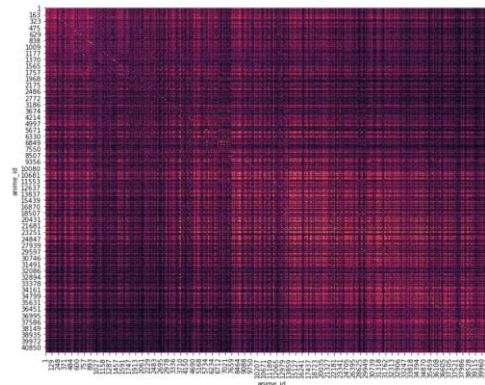


# Item-based Preprocessing and Similarity and Predict method

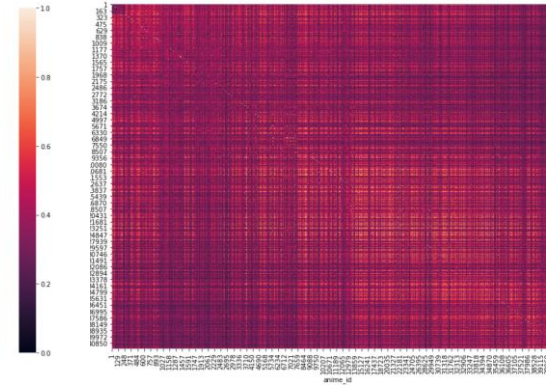
We created a total of 3 datasets based on how many animations the user evaluated.

1. A group that rated 100 animations
2. A group that rated between 2000-2300 animations
3. Group that rated over 5000 animations

Based on our experience, we divided the data into a group that watched animation for 3-4 years and a group that watched about 20% or 50% of the total animation.



Cosine similarity



Adjusted Cosine similarity



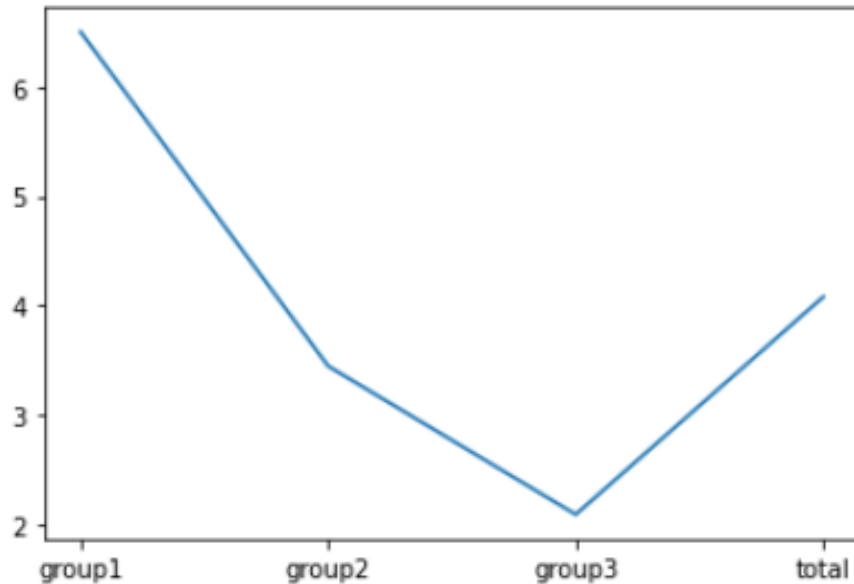
Deformed Cosine similarity

Reference : <https://data-science-hi.tistory.com/150>

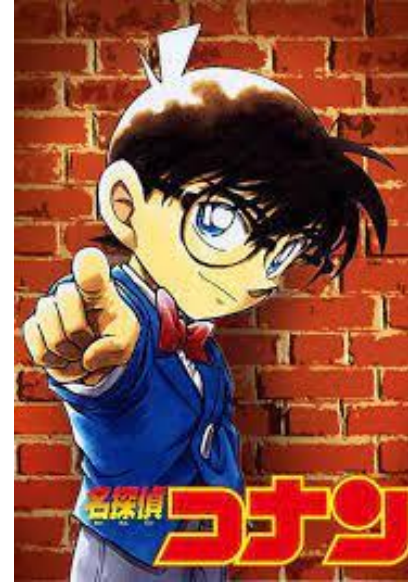
Predict method

$$\widehat{R}_{u,i} = \sum^N (S_{i,N} * R_{u,N}) / \sum^N (|S_{i,N}|)$$

# Item-based Result & analyze



It can be seen that the more the group watched the animation, the lower the RMSE level. And even if the RMSE is low, going beyond 2 gives bad results for this model with a range of 0-10. Cause: In some cases, the pred value is 0. If the user has never seen an animation that enters the top 20 similarity, it is calculated as 0 and the value is different.



	pred_score
Pokemon XY&Z: Subete no Nazo wo Tokiakase!	9.148946
Choujigen Kakumei Anime: Dimension High School	9.014642
Eievui to Colorful Friends	9.011052
Servamp Specials	9.002866
Popful Mail	8.984061
Hakyuu Houshin Engi: Kou-ke no Chi	8.947726
Pokemon: Pikachu no Summer Bridge Story	8.919728
Pokemon Fushigi no Dungeon: Magnagate to Mugendai Meikyu	8.898751
Pokemon: Utae Meloetta - Rinka no Mi wo Sagase!	8.857176
Megumi to Taiyou II: Kajuu Gummi Tweet Mystery - Kieta Sapphire Roman no Nazo	8.849227

# Additional goals and improvements

---

## Content-based Filtering

1. Application of Jaccard similarity, which is often used in a set of words.
2. Personality analysis for all data.
3. Measure accuracy based on data gathered by surveying real people.
4. Try using a text mining technique other than TF-IDF.

## Item-based Collaborative Filtering

1. Try Jaccard or Pearson or any other similarity in the reference.
2. Measure it using all rating data.
3. Comparison of user-based nearest collaborative filtering rather than item-based collaborative filtering.
4. Apply by changing the prediction method.

**Thank you**