
Project 4 - Group 03
Awesome Anime Answers
Our Anime Recommender

Our Goal:

To use Machine Learning and the K-nearest neighbor algorithm to build a model to read a broad dataset and recommend new animes to users based on their preferences.

Our Team

Daniel Purrier

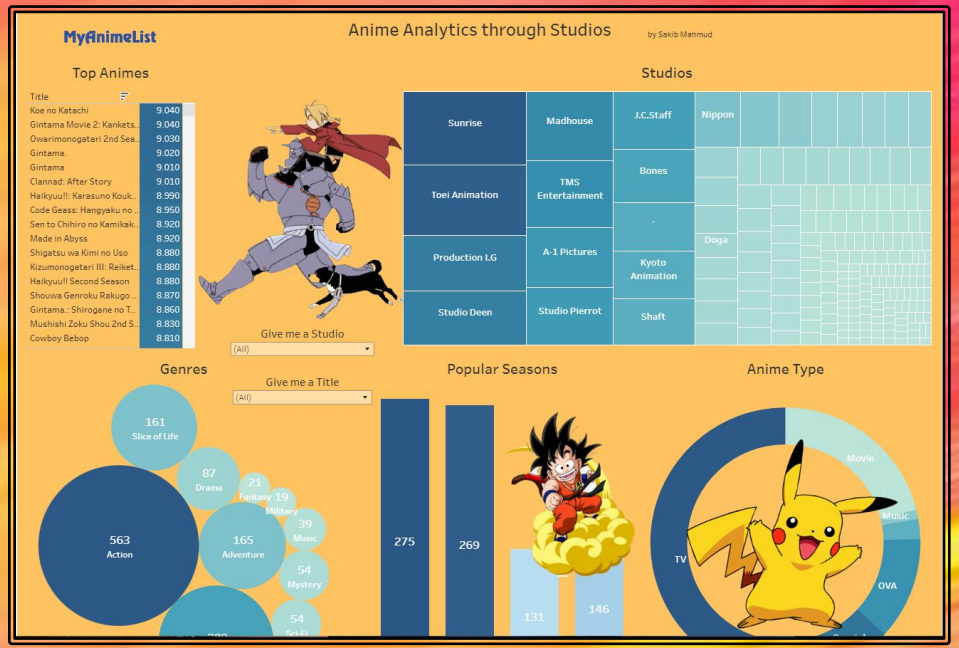
Carlos Ruiz

Steven Madden

Amar Patil

Our Inspiration

Anime is an incredibly popular form of entertainment. In the U.S. alone, approximately 72% of the population watches some form of anime.* However, with so many choices of titles and genres, it can be overwhelming for someone new to Anime or even an Anime veteran to find new shows and movies they might like. After finding other recommenders and our own team's enthusiasm for Anime, we thought it would be fun to build our own anime recommender.



*According to a 2024 servy by https://worldpopulationreview.com/country-rankings/anime-popularity-by-country

https://public.tableau.com/app/profile/sakib.mahmud1560/viz/AnimeAnalyticsthroughStudios/AnimeAnalyticsDashboard

Data and cleaning

We started by choosing our dataset,

https://www.kaggle.com/datasets/CooperUnion/anime-recommendations-database

based on file size, number of rows, and perceived amount of cleaning necessary. We estimated we might only lose around 3.5% of rows when cleaning was done.

We started by reading in our dataset looking at its shape and searching for nulls.

Since the impact of dropping null values is less than 1% in genre and type columns, we can drop all rows with null values. We lost another 2.5% when we found entries in the episode column contained non-numeric values.

We found some strange symbols in some of the names, so we removed them. Then it was a matter of taking all the unique entries in the genre and type columns and turning them into their own Boolean columns, and then dropping the genre and type columns.

Once done, we dropped just a few more columns, when realized several genre columns were of an 18+ plus variety, so we decided to drop them.

Machine Learning

Since we are building a recommender we knew we would use k-nearest neighbor. Let's explain why:

Primary Components of a recommender System

1. **Candidate generation**

In this first state, the system starts from huge dataset and generates a smaller subset of data. Example in Youtube reduces large amount of videos down to hundreds or thousands. This is first stage of Recommendation. There are two common candidate generation approaches:

Content-based Filtering

Uses similarity between items to recommend items similar to what user likes

Ex: if User watches dog videos, then the system can recommend cute animals videos to that user.

Collaborative Filtering

Recommends items based on the behavior or preferences of similar users.

Ex: If used A is similar to user B, and user B likes video 1, then it recommends A

2. Scoring

This model scores and ranks the candidates in order to select the set of items (on the order of 10) to display to the user. This is subset of Candidate generation

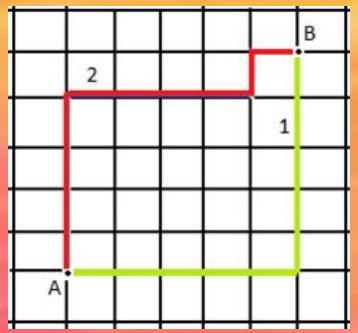
3. Re-ranking

This is final ranking which takes into account dislikes or likes of newer content. Re-ranking helps to ensure diversity, freshness and fairness.

To determine the degree of similarity, most recommendation systems rely on one or more of the following:

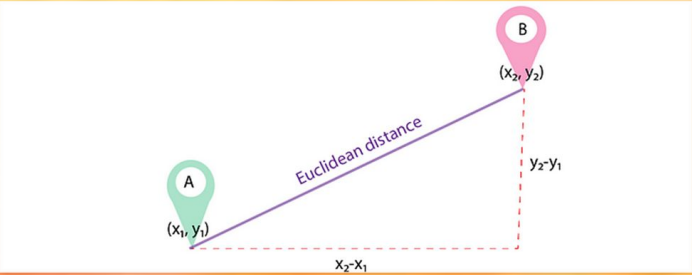
1. Manhattan

Also, know as city block distance, or taxicab geometry wherein the distance is measured between two data points in Grid-like path. As shown below. Mainly used where high dimensionality in the data.



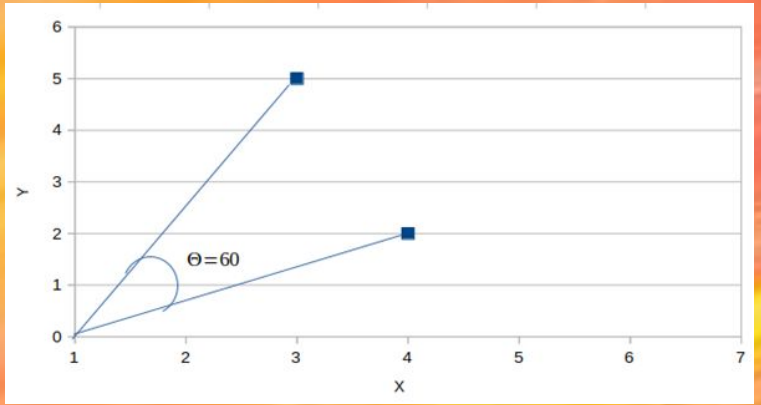
2. Euclidean

This is the shortest distance between 2 data points in the plane. Mainly used for the smaller dimensionality problems.



3. Cosine Distance

Also known as Cosine Similarity is used to find similarities between two data points. Cosine similarity is given by Cos θ, and cosine distance is 1- Cos θ. Mainly used in the Collaborative Filtering based recommender systems to offer future recommendations



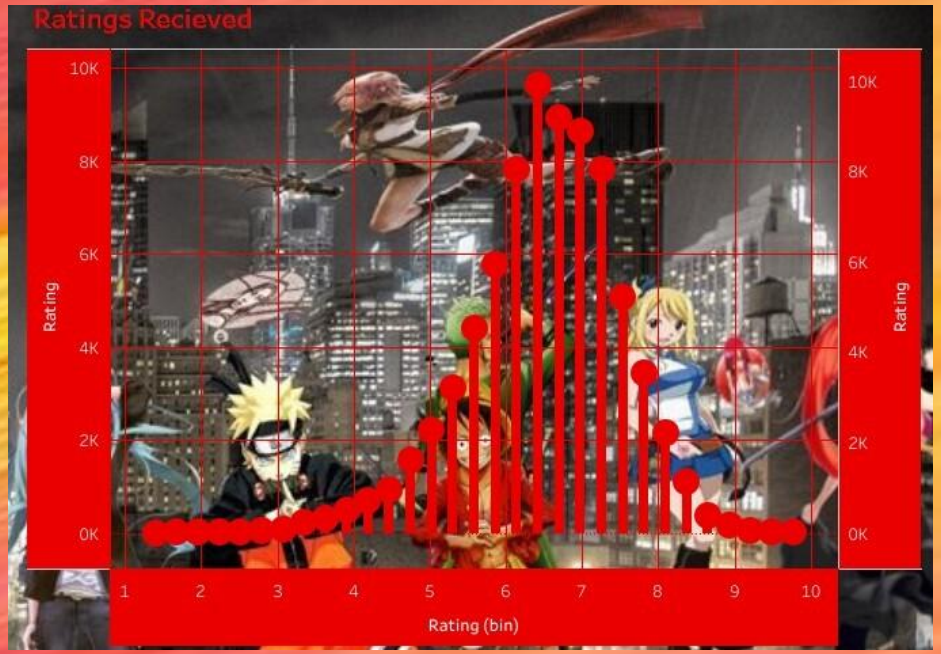
Imported Modules

Import pandas
Import numpy
Import pickle

Pre-Processing

from sklearn.preprocessing import StandardScaler,
OneHotEncoder, OrdinalEncoder
from sklearn.pipeline import Pipeline
from sklearn.compose import ColumnTransformer
from sklearn.impute import SimpleImputer
from sklearn.neighbors import NearestNeighbors

Visualizations Examples



Anime Search

Name	Avg. Rating	Episodes	Members
Oyako Club	6	1,818	160
Doraemon (1979)	8	1,787	14,233
Kirin Monoshiri Yakata	6	1,565	116
Manga Nippon Mukashiba..	6	1,471	406
Hoka Hoka Kazoku	6	1,428	194
Kirin Ashita no Calendar	6	1,306	59
Monoshiri Daigaku: Ashit..	7	1,274	112
Sekai Monoshiri Ryoko	6	1,006	153
Kotowaza House	6	773	110
Shima Shima Tora no Shi..	6	726	237
Ninja Hattori-kun	7	694	2,116
Perman (1983)	6	526	447
Obake no Q-tarou (1985)	7	510	161
Ninja-tai Gatchaman ZIP!	6	475	146
Kochira Katsushikaku Ka..	8	373	4,734
Bleach	8	366	624,055
Manga Jinbutsushi	7	365	71
Charady no Joke na Maini..	5	365	1,612
Keroro Gunsou	8	358	31,632
Kiteretsu Daihyakka	7	331	1,083


Live Demo
https://aidecisions.pythonanywhere.com/

Conclusions

With more than two thirds of the population of the United States enjoying some kind of anime, a recommender may be a useful tool for finding new shows and movies to watch amid the wide variety available. Hopefully a project like this can take some of the guesswork out of finding new entertainment.

Limitations/Bias

- The Dataset was last updated 8 years ago, meaning it contained no recent animes
- The companion rating dataset was too large to use for the scope of the project
- Fewer categories (ie: year, studio) than other similar datasets
- You must know at least one anime for it to work

Future Work

- Further refining of the nearest neighbors to yield stronger results
- Added search functionality
- Added image results to show cover art of recommended animes
- Optimize the recommender for people who have no experience with anime

Work Cited

worldpopulationreview.com/country-rankings/anime-popularity-by-country

- 2024 Anime statistics

public.tableau.com/app/profile/sakib.mahmud1560/viz/AnimeAnalyticsthroughStudios/AnimeAnalyticsDashboard

- Visualization inspiration

www.kaggle.com/datasets/CooperUnion/anime-recommendations-database/data

- Dataset

git.bootcampcontent.com/boot-camp-consortium-east-coast/DATA-PT-EAST-APRIL-041524/-/tree/main/01-Lesson-Plans/23-Project-4-Week-1/3/BOOOTH_RECOMMENDER_EXAMPLE?ref_type=heads

- Instructor project example

select2.org/data-sources/ajax

makitweb.com/loading-data-remotely-in-select2-with-ajax/

- Referenced for drop down elements in the HTML