MAL Data Analysis Project

Data Collection:

First version:

The first version of this project was done without using the jikan API and I didn't bother getting a MAL API key so I was going to try scraping the pages. The process for this was done by sending a request to the web page, storing the webpage in memory, parsing the data I want out of it then going onto the next page. This would be against MAL's TOS as well as being less convenient for me since manually writing out scripts to find all of the data I wanted takes a lot more time than reading from the JSON file that jikan or even the MAL API provides.

This never got to the stage where I started collecting data, just tested on a few pages.

Final Version:

First I started by finding out how to request anime data from the API by ID, then I wrote a script that iterates over those ID's, making requests, if the request comes back with response 200, then I load the json data, rip the title, create a local .json file with all of the information, update a text file to save what the last anime processed was and repeat until we have all of the anime.

The reason we save where we are to a text file is so if MAL goes down for maintenance, my PC turns off, or I get an incorrect response, we know where we are to pick back up and we know what ID's aren't giving correct responses.

After a few crashes of the program due to ID's not giving correct responses, I updated the code such that ID's that aren't giving responses get logged to a file and the program simply continues.

The speed of this with 1 instance is 0.9 anime per second, I at some point ran 2 instances which increased the speed of each to 1.5 anime per second for some reason, though without knowing the rate limit was 3 per second, there was another limit of 60 per minute, meaning that I had missed a lot of entries due to rate limits and needed to recollect the data.

Data Collected:

The data collected per anime is a JSON file with more than 50 things per anime, I will add more detail on the data to this section later but continuing on I want to talk about what I've done and plan to do with this data.

Data Processing:

As for how I am processing data, I will do this in two stages; the first stage I am playing around with the data in it's JSON format, this is obviously slow and inconvenient so for the second stage I will be loading everything into a database.

That being said, what I've done with the data so far is that I've checked the average rating of every genre and theme on MAL that has a rating and it goes as follows from low to high:

Avant Garde: 5.258 Music: 6.033 Erotica: 6.105 Hentai: 6.114 Horror: 6.168 Parody: 6.313 Strategy Game: 6.318 Educational: 6.322 Slice of Life: 6.425 Ecchi: 6.431 Racing: 6.435 Pets: 6.45 Comedy: 6.499 Mecha: 6.501 Idols (Male): 6.51 Boys Love: 6.531 Sci-Fi: 6.545 Space: 6.55 Crossdressing: 6.554 Video Game: 6.598 Reverse Harem: 6.611 Girls Love: 6.621 Gourmet: 6.627 Combat Sports: 6.636 Fantasy: 6.642 Idols (Female): 6.651 Anthropomorphic: 6.688 Sports: 6.688 Samurai: 6.689 Adventure: 6.69 Magical Sex Shift: 6.697	Action: 6.71 Super Power: 6.71 Delinquents: 6.725 Historical: 6.731 Mythology: 6.732 Martial Arts: 6.733 Harem: 6.738 Supernatural: 6.749 Psychological: 6.802 Military: 6.822 Romance: 6.825 Vampire: 6.828 Mahou Shoujo: 6.831 Drama: 6.841 School: 6.845 Performing Arts: 6.864 Detective: 6.881 Team Sports: 6.884 Gore: 6.884 Visual Arts: 6.898 Workplace: 6.936 Isekai: 6.937 High Stakes Game: 6.946 Mystery: 6.983 Medical: 6.987 Time Travel: 7.003 Suspense: 7.037 Survival: 7.041 Showbiz: 7.104 Reincarnation: 7.129 CGDCT: 7.154	Otaku Culture: 7.171 Gag Humor: 7.222 Adult Cast: 7.239 Organized Crime: 7.292 Award Winning: 7.298 Love Polygon: 7.33 Romantic Subtext: 7.338 Childcare: 7.342 Iyashikei: 7.493

Some ideas I have for further analysis are:

1. Query anime by genre or genre combinations.
2. Set minimum ratings or popularity for queries.
3. ML model to predict what anime you will like based rankings provided by the user.