

MyAnimeList (MAL) – Top Anime Scraper

CONTEXT

La pàgina web **MyAnimeList**¹, molts cops abreviada com MAL, és una de les principals comunitats virtuals de catalogació i xarxa social d'**anime**² i **manga**³. Aquesta web proporciona als usuaris un sistema basat en llistes per poder **organitzar** i **puntuar** l'anime i manga que s'hagi consumit.

Darrere té una gran base de dades que cataloga tant anime com manga, amb una gran quantitat d'informació com per exemple títol, sinopsi, data de sortida i de finalització, nota dels usuaris o gènere al que pertanyen.

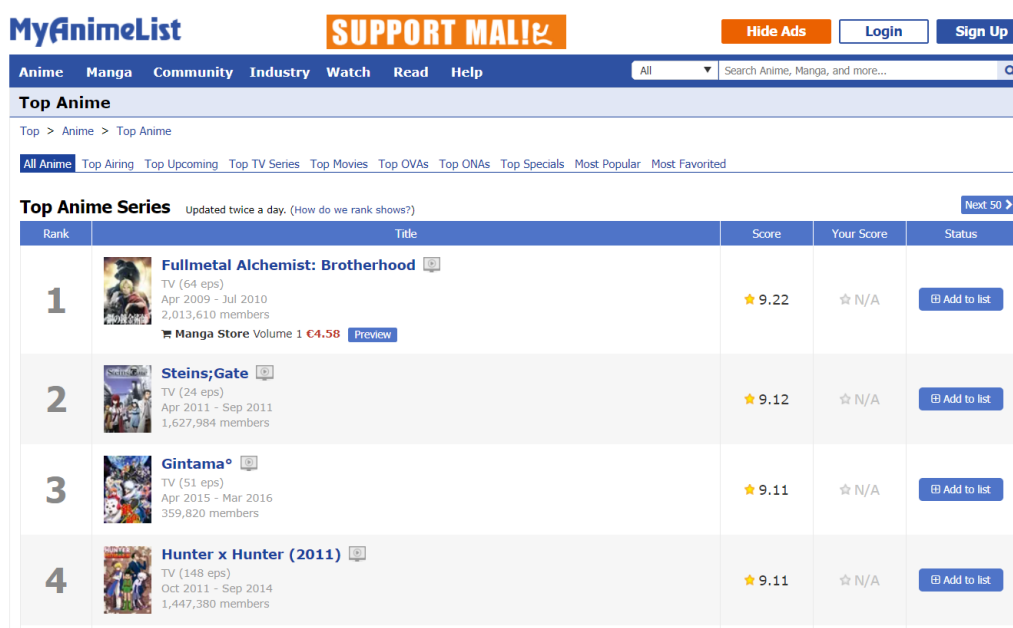
DESCRIPCIÓ

Aquest dataset conté informació sobre 500 sèries anime diferents, ordenades per la puntuació dels usuaris. Aquest Top 500 va canviant constantment, així que les dades sempre van associades a la data en que es realitza la seva extracció. En específic, les dades extretes per aquest dataset són del dia **7 de novembre de 2020**.

De cadascun dels anime s'extreu la seva posició en el top, el títol, la sinopsi, el nombre d'episodis, la nota mitjana dels usuaris, la data en que es va començar a emetre i la data en que va finalitzar, l'estat de la sèrie (si està acabada o en emissió) i els gèneres al que pertany. A tota aquesta informació també s'afegeix un *timestamp* del moment en que aquestes dades es van recopilar, ja que com són dades que varien ràpidament, és important mantenir-les en el seu context temporal.

REPRESENTACIÓ GRÀFICA

A la següent imatge es pot observar la pàgina web MyAnimeList, justament a la pestanya on es mostra el top. El disseny de la pàgina és àmpliament conegut pels usuaris, així que amb aquesta imatge a simple vista es pot comprendre què conté el dataset.



The screenshot shows the MyAnimeList website interface. At the top, there's a navigation bar with links like Anime, Manga, Community, Industry, Watch, Read, and Help. Below this, the 'Top Anime' section is visible, with a sub-header 'Top Anime Series' and a note 'Updated twice a day. (How do we rank shows?)'. The main content is a table listing the top anime series.

Rank	Title	Score	Your Score	Status
1	Fullmetal Alchemist: Brotherhood TV (64 eps) Apr 2009 - Jul 2010 2,013,610 members Manga Store Volume 1 €4.58 Preview	★ 9.22	☆ N/A	Add to list
2	Steins;Gate TV (24 eps) Apr 2011 - Sep 2011 1,627,984 members	★ 9.12	☆ N/A	Add to list
3	Gintama° TV (51 eps) Apr 2015 - Mar 2016 359,820 members	★ 9.11	☆ N/A	Add to list
4	Hunter x Hunter (2011) TV (148 eps) Oct 2011 - Sep 2014 1,447,380 members	★ 9.11	☆ N/A	Add to list

¹ <https://myanimelist.net/>

² <https://ca.wikipedia.org/wiki/Anime>

³ <https://ca.wikipedia.org/wiki/Manga>

CONTINGUT

Les dades extretes es troben emmagatzemades en un fitxer CSV “**TopAnime_2020-11-07.csv**”. El nom del fitxer variarà depenent de la data en que s’extreguin les dades.

El separador utilitzat pels valors del fitxer és “|”. Aquesta decisió està basada en el fet de que el text dins de les dades conté tant comes (,) com punts i comes (;), així que utilitzar un separador diferent a aquests dos caràcters ens assegura un fitxer de dades més robust.

El fitxer amb les dades conté els següents camps:

- **ranking_position:** Nombre sencer que indica la posició en la que es troba l’anime en el Top, segons la nota donada pels usuaris.
- **title:** Títol de l’anime.
- **score:** Nombre amb dos decimals que indica la nota mitjana dels usuaris.
- **n_episodes:** Nombre d’episodis. El camp pot estar buit, fet que significa que encara no se sap el nombre de capítols que tindrà un anime que està en emissió.
- **start_aired_date:** Data en que es va publicar el primer episodi, o si és una pel·lícula la data en que aquesta es va publicar.
- **end_aired_date:** Data en que es va publicar l’últim episodi. Tant si l’anime encara no està acabat com si és una pel·lícula, aquest camp es trobarà buit.
- **status:** Estat d’emissió de l’anime, que pot ser “Finished Airing” o “Currently Airing”.
- **sinopsis:** Text que descriu la sinopsi de l’anime.
- **genres:** Llista de gèneres als que pertany l’anime.
- **scrape_timestamp:** Timestamp del moment en que es va *scrappejar* l’anime.

Aquestes dades s’han recollit utilitzant el codi present en aquest repositori, completament realitzat amb **Python**.

AGRAÏMENTS

Les dades han sigut recollides de la pàgina web **MyAnimeList**. Per tant, els agraïments van dirigits tant a la pàgina web com als seus usuaris, ja que sense ells no hauria sigut possible obtenir un top d’anime per popularitat.

INSPIRACIÓ

Aquest conjunt de dades pot ser utilitzat amb diferents fins.

La principal utilitzat d’aquest dataset és poder realitzar un anàlisi descriptiu de les tendències en sèries anime. Algunes preguntes que es podrien contestar són:

- Quins gèneres són més populars?
- Són més populars els animes llargs o els curts?
- Són més populars els animes antics o els moderns?
- Com varia el Top d’anime amb el pas del temps?

Per altra banda, ja que s’estan extraient dades de les sinopsis, que són textos llargs, aquestes dades també es podrien utilitzar per entrenar un model de Machine Learning que necessiti textos extensos.

Per últim, ja que aquesta pàgina web encara no disposa d’una aplicació mòbil, si s’escalés aquest *scrapping* a tota la web, podríem obtenir la informació necessària per desenvolupar una aplicació mòbil amb les mateixes funcionalitats.

LLICÈNCIA

La llicència escollida per aquest dataset és **Creative Commons Attribution 4.0 International**. Aquesta llicència permet copiar i redistribuir el material en qualsevol medi o format, i també permet adaptar i construir noves solucions a partir d'aquest, fins i tot amb propòsit comercial.

Si un usuari vol utilitzar aquestes dades haurà de donar crèdit de forma adequada i indicar si s'han indicat canvis.

Aquesta llicència és la més adequada per que aquestes dades puguin arribar al major nombre d'usuaris i que puguin ser d'utilitat en diferents aplicacions.

CODI

A la carpeta `src` podem trobar el codi Python amb que s'han extret aquestes dades. A continuació es mostra l'estructura i la funció de cadascun dels fitxers:

- `/entities`
 - **Anime.py**: fitxer que conté la classe **Anime**, que s'utilitza per guardar la informació extreta d'un anime amb l'*scraper*. Té una funció per imprimir per pantalla la informació guardada en un objecte d'aquest tipus.
- **anime_scrapper.py**: fitxer que conté la classe **AnimeScrapper**. Conté tots els mètodes necessaris per obtenir tots els camps d'informació d'un anime (títol, nota, posició en el top ...).
- **top_anime_scrapper.py**: fitxer que conté la classe **TopAnimeScrapper**. Conté els mètodes necessaris per *scrappejar* un nombre N d'animes del top d'anime i retornar les URLs dels diferents anime.
- **main.py**: fitxer des del qual s'executa l'aplicació. Es crea un objecte **TopAnimeScrapper**, i un cop hem obtingut la llista d'animes a *scrappejar*, es creen els diferents objectes **AnimeScrapper** per obtenir la informació d'un anime en específic. Agrupa tota aquesta informació en un *pandas dataframe* i ho guarda en CSV. En aquest cas s'està extraient informació de 500 animes.

Per executar el codi, des de l'interior del directori `src` simplement s'ha d'executar el fitxer **main.py** i les dades extretes es guardaran a la carpeta `csv`, en un fitxer CSV amb el nom "TopAnime_YYYY-MM-DD.csv" amb la data en que s'han extret.

DOI DEL DATASET

El dataset s'ha publicat a Zenodo, com es pot veure al l'enllaç següent:

<https://zenodo.org/record/4256843#.X6b49mhKiUk>

El DOI assignat per aquest dataset és:

DOI 10.5281/zenodo.4256843

TAULA DE CONTRIBUCIONS

Contribucions	Signa
Recerca prèvia	Laura Planas Simón
Redacció de respostes	Laura Planas Simón
Desenvolupament del codi	Laura Planas Simón