Delin Mattéo, Timéo Margerand

SAÉ 3.VCOD.01 - Collecte automatisée de données web

Lien vers le repository : https://github.com/MatteoDelin/SAE-Web-Scraping

Introduction

Nous avons décidé de réaliser cette SAÉ sur l'animation japonaise, car ce domaine offre un grand volume de données à la fois qualitatives et quantitatives (nombre d'épisodes, studios d'animation, scores, popularité, genres) d'autant plus qu'un grand nombre de site sont des base de donn. Cela en fait un support idéal pour un projet de scraping et d'analyse de données.

Le but de cette analyse est de découvrir s' il existe des critères remarquables qui permettraient d'identifier de bons animés. Pour savoir si une œuvre est bonne nous avons décidé de nous baser sur sa note bien que l'on aurait pu prendre la popularité.

L'analyse cherchera donc à trouver dans qu'elle cas la note d'un animé est plus où moins élevée.

Liste des sites web choisis

Pour réaliser ce projet, nous nous sommes tournés vers MyAnimeList, site référence sur ce domaine qui offre des informations complètes sur plus de 25000 animes. De plus, c'est aussi le seul à autoriser le scraping de manière simple et sans limitation. Nous avons décidé de récupérer les 1535 premiers animés du site en fonction de leur score sur 10 afin d'avoir des données variées, sans avoir de données aberrantes.

Légalité et éthique

Le choix de MyAnimeList s'est fait dans le respect des règles de légalité et d'éthique liées à la collecte de données en ligne. Les informations extraites sont accessibles publiquement, sans nécessiter d'identification ou de contournement de protections techniques.

De plus, le fichier https://myanimelist.net/robots.txt ne nous empêche pas d'accéder à l'ongle Anime du site et nous avons fait en sorte de bien nous déclarer en tant que robots lors de la collecte de données.

Analyse de la récolte et du traitement

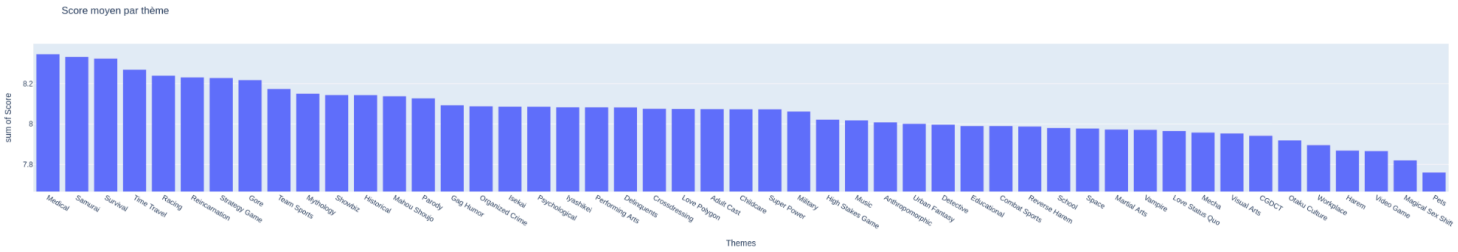
Le traitement est fait en récupérant le lien de chaque animé via le top anime puis dans chaque lien récupéré, nous nous sommes occupés de récupérer le bandeaux d'information sur le côté.

Après avoir récupéré les données sous plusieurs fichiers txt (une cinquantaine d'échantillons), j'ai d'abord dû regardé les différentes colonnes qui étaient exploitables. En effet, il y avait beaucoup de colonnes vides que j'ai donc décidé d'enlever car elles n'étaient pas pertinentes à l'analyse. Certaines colonnes nécessitaient aussi une modification de leurs valeurs comme les colonnes thème qui présentait chaque thème deux fois pour chaque animé, j'ai donc dû adapter mon code de formatage pour ne récupérer que le premier. Il y avait aussi les colonnes noms qui étaient composées de

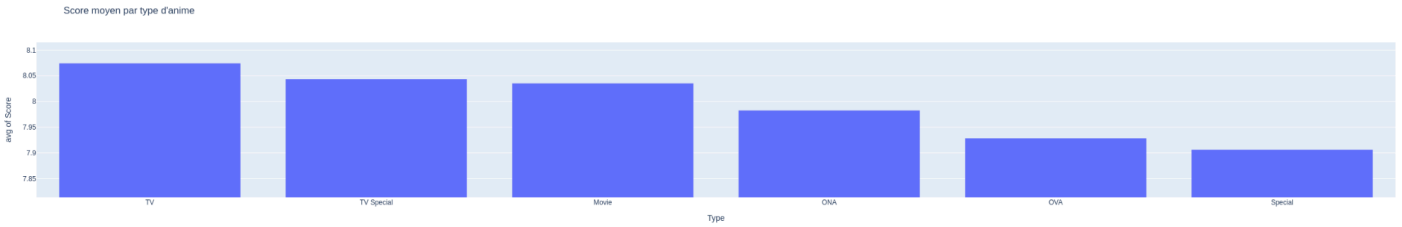
plusieurs langues, je n'ai donc gardé que la version anglaise (certains animés n'avaient pas de noms français). Enfin, nous avons pu tester le script final sur les 1500 animés totaux.

Conclusion suite à l'analyse

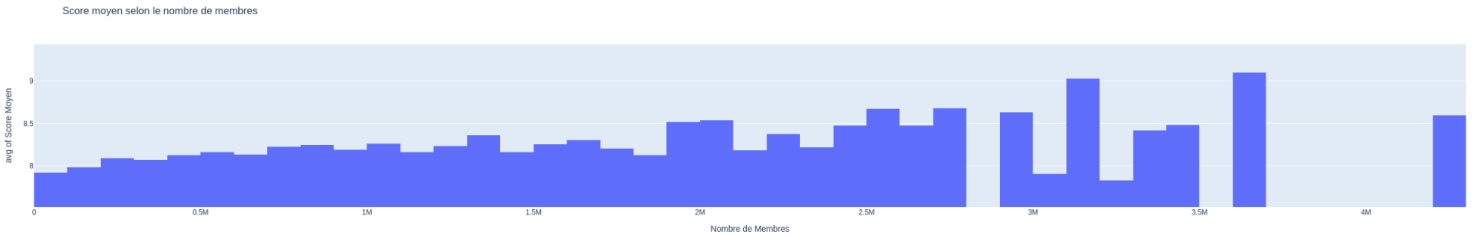
Ce graphique montre le score moyen en fonction du thème de l'animé. On peut voir que chaque



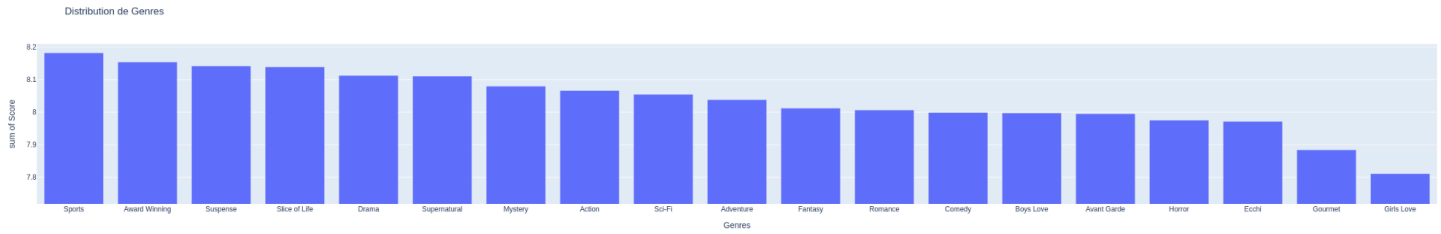
thème à un score autour de 8. Les thèmes les plus aimés sont Médical, Samurai et survie tandis que le moins aimé est le thème Pets (animaux). Cependant, il n'y a que 0.585 de score moyen de différences entre le thème le plus populaire et le moins populaire.



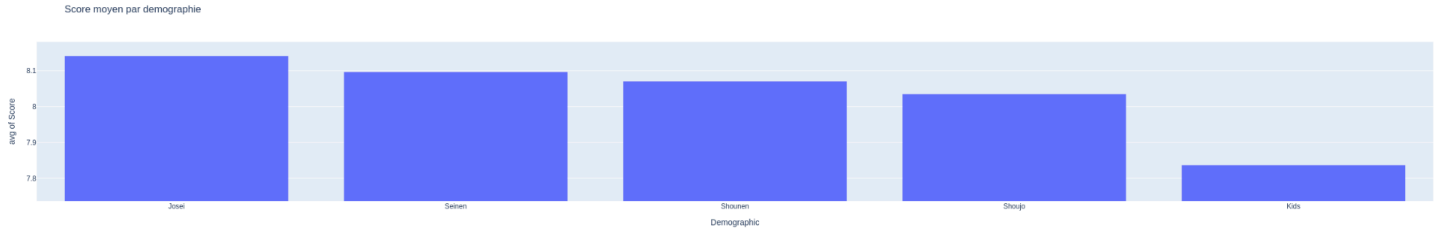
Ce graphique montre la note moyenne en fonction du type d'animation (TV, film, épisode spécial, ...). On peut voir que les notes influent peu en fonction du type (entre 7.8 et 8.2) même si les séries TV sont les plus appréciées.



Ce graphique montre le score moyen en fonction de la popularité, on peut voir que plus la popularité est élevée, plus le score l'est également dans une certaine mesure.



Cet histogramme montre la note moyenne en fonction du genre de l'animé, on peut voir que les catégories Sport et Awards winning sont les mieux notées (8.2 et 8.1 sur 10)



Ce graphique présente la note moyenne des animés par démographie (seinen, shounen, ...). La démographie des animés désigne le public cible principal (enfants, adolescents, hommes adultes, femmes adultes, etc.) d'une œuvre, ce qui influence directement ses thèmes, son contenu et son style. On peut notamment voir que la catégorie Kids est moins appréciée que les autres. On remarque aussi que plus la démographie de l'anime vise un public âgé, plus la note moyenne augmente.

Résultat et interprétation :

Les analyses menées sur les 1535 animes collectés sur MyAnimeList révèlent des schémas d'appréciation subtils au sein de l'échantillon d'œuvres les mieux notées. La faible dispersion des scores moyens, notamment entre les différents types d'animation (TV, Film) et thèmes, souligne une relative homogénéité qualitative perçue par la communauté. On observe que l'appréciation est fortement corrélée à la popularité, suggérant un effet de masse où les titres les plus largement vus tendent à conforter leur statut de favoris. De plus, les préférences des utilisateurs semblent s'orienter vers des genres spécifiques comme le Sport ou les œuvres primées ("Awards winning"), qui affichent les moyennes les plus élevées. Enfin, l'influence de la démographie cible est notable : les animes visant un public plus mature (Seinen, Josei) bénéficient d'une meilleure notation, potentiellement en raison d'une complexité thématique ou narrative que les utilisateurs valorisent davantage par rapport aux œuvres destinées aux enfants (Kids).

De plus, d'autres graphiques sont disponibles sur la WebApp mais n'étant pas jugé assez intéressant il n'ont pas été inséré dans ce rapport.

Pour conclure, il n'existe pas d'indicateur permettant d'être sûre de la qualité d'un animé dû à une forte homogénéité dans les données.

Partie individuelle et responsabilité :

Cette section détaille la contribution de chaque membre de l'équipe aux différentes phases du projet, en soulignant l'implication spécifique et la prise de recul critique apportée à chaque étape.

Analyse et Extraction des Données

- **Mattéo Delin (90%) :** J'ai principalement développé le code de *scraping* (extraction des informations des pages web) en Python. La prise de recul à cette étape s'est traduite par une réflexion constante sur l'optimisation des sélecteurs CSS pour garantir la robustesse du code face à d'éventuelles modifications de la structure du site, et pour minimiser le temps d'extraction.
- **Timéo Margerand (10%) :** J'ai analysé la structure des pages de MyAnimeList en amont afin d'identifier les balises cibles. Ma prise de recul a consisté à valider la complétude des données brutes initialement extraites par Mattéo, assurant que toutes les variables nécessaires à l'analyse future étaient bien capturées.

Traitement et Export

- **Timéo Margerand (85%) :** J'ai pris en charge le traitement des données brutes, incluant le nettoyage (gestion des valeurs manquantes et aberrantes), la transformation (conversion des types, standardisation des genres et thèmes) et le regroupement dans un fichier structuré au format CSV. Ma prise de recul a porté sur la garantie de l'intégrité des données après nettoyage et le choix du format d'export le plus pertinent pour faciliter l'analyse exploratoire et la visualisation.
- **Mattéo Delin (15%) :** J'ai optimisé la structure du code de traitement en adoptant une approche modulaire, ce qui a rendu l'exécution plus simple et flexible. Ma prise de recul a permis de s'assurer que le code de traitement pourrait être facilement adapté ou réutilisé si l'on décidait de *scraper* de nouvelles variables.

Visualisation et Interprétation

- **Mattéo Delin (55%) :** J'ai sélectionné les graphiques les plus pertinents pour illustrer les corrélations majeures (score vs. popularité, score par genre) et j'ai développé l'application Web (Web App) pour rendre ces visualisations interactives. La prise de recul s'est concentrée sur le choix des types de graphiques (nuages de points, histogrammes, heatmap) pour s'assurer qu'ils traduisent le plus clairement possible les schémas identifiés.
- **Timéo Margerand (45%) :** J'ai analysé en profondeur les résultats produits par les visualisations et j'ai rédigé la synthèse finale du rapport. Ma prise de recul a été essentielle pour éviter de tirer des conclusions hâtives et pour m'assurer que l'interprétation des faibles écarts de score moyen entre les thèmes ou les types d'animation soit rigoureuse et reflète fidèlement l'homogénéité de l'échantillon des animes les mieux notés.