ATHTHANAYAKA MUDALIGE Mahen – B3 DevOps

PROJET WEB SCRAPING PYTHON

Objectifs

L'objectif principal était d'extraire les données (web scraping) d'un site, les structurer et les nettoyer afin de pouvoir réaliser des analyses et visuels graphiques dessus. Ici nous avons utilisé un site web répertoriant un classement des animés : https://myanimelist.net/topanime.php

Méthodes

1. Récupération des données
   - Nous avons tout d'abord récupéré la page grâce à la bibliothèque requests, la bibliothèque BeautifulSoup pour analyser le contenu HTML de la page.

2. Traitement des données
   - Les titres des animes et leurs scores ont été extraits de la page web, organisés dans un dictionnaire, stocker dans un json file, puis convertis en un DataFrame* à l'aide de la bibliothèque pandas.

3. Visualisation des données
   - La bibliothèque matplotlib a été utilisée pour créer un tableau affichant le classement avec les titres des animes et leurs scores. La taille du tableau a été ajustée en fonction de la résolution de l'écran. Les données ont été triées en fonction des scores pour que ça soit plus cohérant à la visualisation.

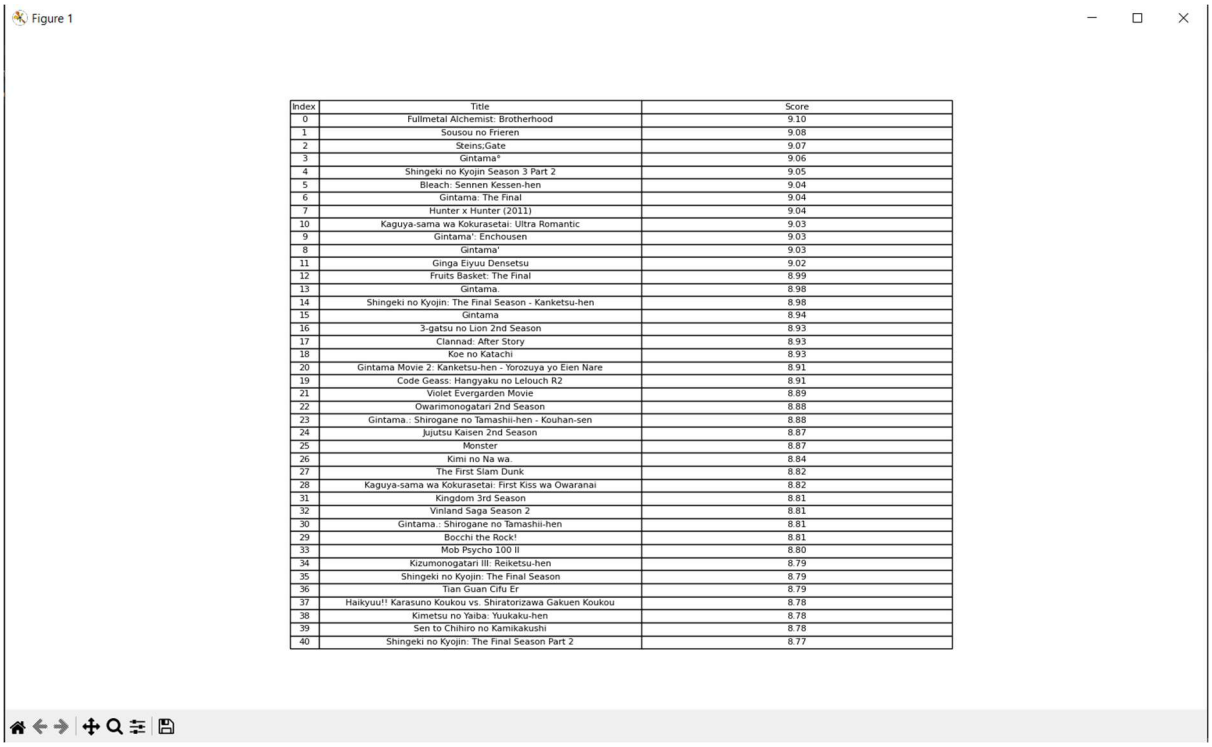
Défis Rencontrés

1. Difficultés en Visualisation Graphique
   - La manipulation des données afin d'avoir un bon affichage était un défi dû aux lignes de code complexe demandant beaucoup de temps à comprendre. Cependant, le fait que les bibliothèques utilisées soient bien documentées ont était d'une bonne aide afin d'avancer sur le projet.

2. Diversité des Structures HTML
   - Les sites de classement d'anime ont des structures HTML très différentes, ce qui peut rendre complexe le fait de faire un script global qui marcherait sur la plupart des sites. Cependant, le script développé ici est conçu de manière simple ce qui entraîne par conséquent le fait que le scraping utilise peu de données diverses et ne donne pas énormément d'information, d'un autre côté, c'est un script qui est facilement adaptable à d'autres sites web de classement avec des structures différentes.

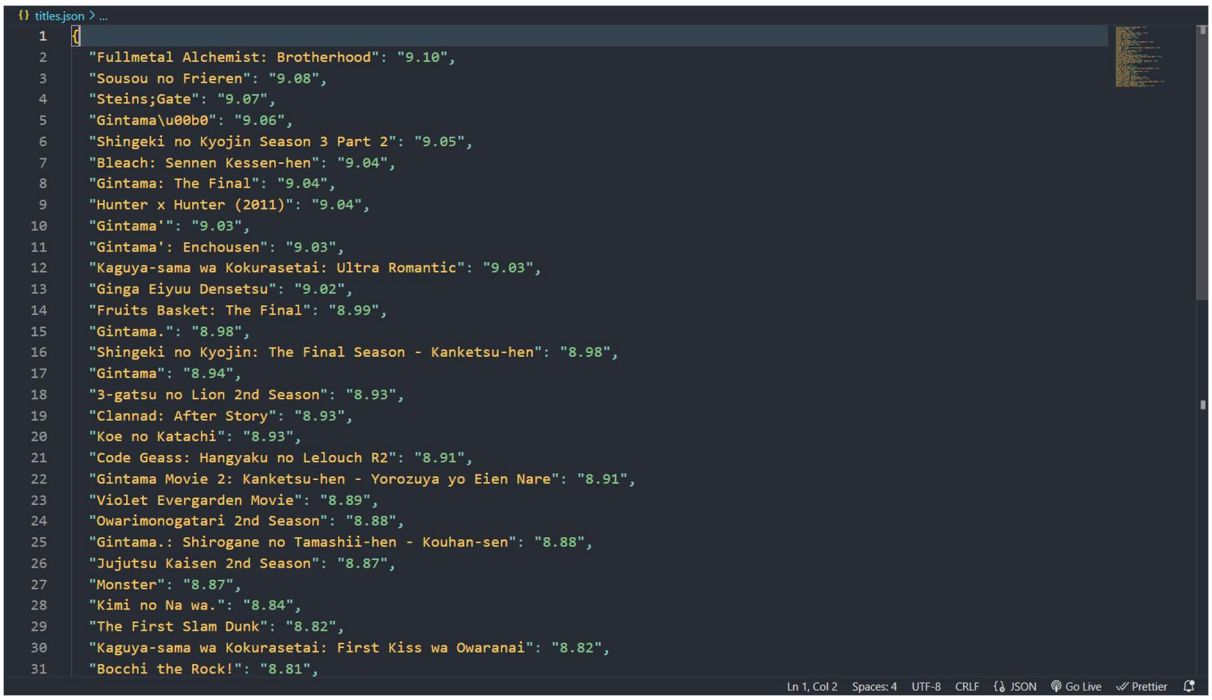
Résultats

Le script a réussi à récupérer avec succès les titres et scores des animes depuis MyAnimeList, les a organisées et les a affichées dans un tableau.



Index	Title	Score
0	Fullmetal Alchemist: Brotherhood	9.10
1	Sousou no Frieren	9.08
2	Steins;Gate	9.07
3	Gintama°	9.06
4	Shingeki no Kyojin Season 3 Part 2	9.05
5	Bleach: Sennen Kessen-hen	9.04
6	Gintama: The Final	9.04
7	Hunter x Hunter (2011)	9.04
10	Kaguya-sama wa Kokurasetai: Ultra Romantic	9.03
9	Gintama': Enchousen	9.03
8	Gintama'	9.03
11	Ginga Eiyuu Densetsu	9.02
12	Fruits Basket: The Final	8.99
13	Gintama.	8.98
14	Shingeki no Kyojin: The Final Season - Kanketsu-hen	8.98
15	Gintama	8.94
16	3-gatsu no Lion 2nd Season	8.93
17	Clannad: After Story	8.93
18	Koe no Katachi	8.93
20	Gintama Movie 2: Kanketsu-hen - Yorozuya yo Eien Nare	8.91
19	Code Geass: Hangyaku no Lelouch R2	8.91
21	Violet Evergarden Movie	8.89
22	Owarimonogatari 2nd Season	8.88
23	Gintama.: Shirogane no Tamashii-hen - Kouhan-sen	8.88
24	Jujutsu Kaisen 2nd Season	8.87
25	Monster	8.87
26	Kimi no Na wa.	8.84
27	The First Slam Dunk	8.82
28	Kaguya-sama wa Kokurasetai: First Kiss wa Owaranai	8.82
31	Kingdom 3rd Season	8.81
32	Vinland Saga Season 2	8.81
30	Gintama.: Shirogane no Tamashii-hen	8.81
29	Bocchi the Rock!	8.81
33	Mob Psycho 100 II	8.80
34	Kizumonogatari III: Reiketsu-hen	8.79
35	Shingeki no Kyojin: The Final Season	8.79
36	Tian Guan Cifu Er	8.79
37	Haikyuu!! Karasuno Koukou vs. Shiratorizawa Gakuen Koukou	8.78
38	Kimetsu no Yaiba: Yuukaku-hen	8.78
39	Sen to Chihiro no Kamikakushi	8.78
40	Shingeki no Kyojin: The Final Season Part 2	8.77

Nous avons également le json file avec les données récupérées.



Annexes

*DataFrame : Un DataFrame est une structure de données tabulaire bidimensionnelle organisée en lignes et colonnes, utilisée pour stocker et manipuler des données dans le langage de programmation Python avec la bibliothèque pandas. Illustration ci-dessous.