Pràctica 1: Web Scrapping

Stefany Chóez Bolaños i Daniel Panadero Espinosa

Context

Durant els últims anys hi ha hagut un augment de la popularitat de l'anime a fora Japó, això ha fet que plataformes com Netflix, HBO o Amazon afegeixin diferents animes al seu catàleg. Per poder decidir quins animes afegir pot ser interessant per les plataformes estudiar els animes més populars, és per això, que s'ha plantejat en generar un dataset dels animes més populars en l'actualitat. Tota la informació la recollirem de la web https://myanimelist.net/, que és la web d'anime més gran i popular que existeix en aquest moment, on podem trobar animes actuals, antics i fins i tot d'altres que no s'han estrenat encara.

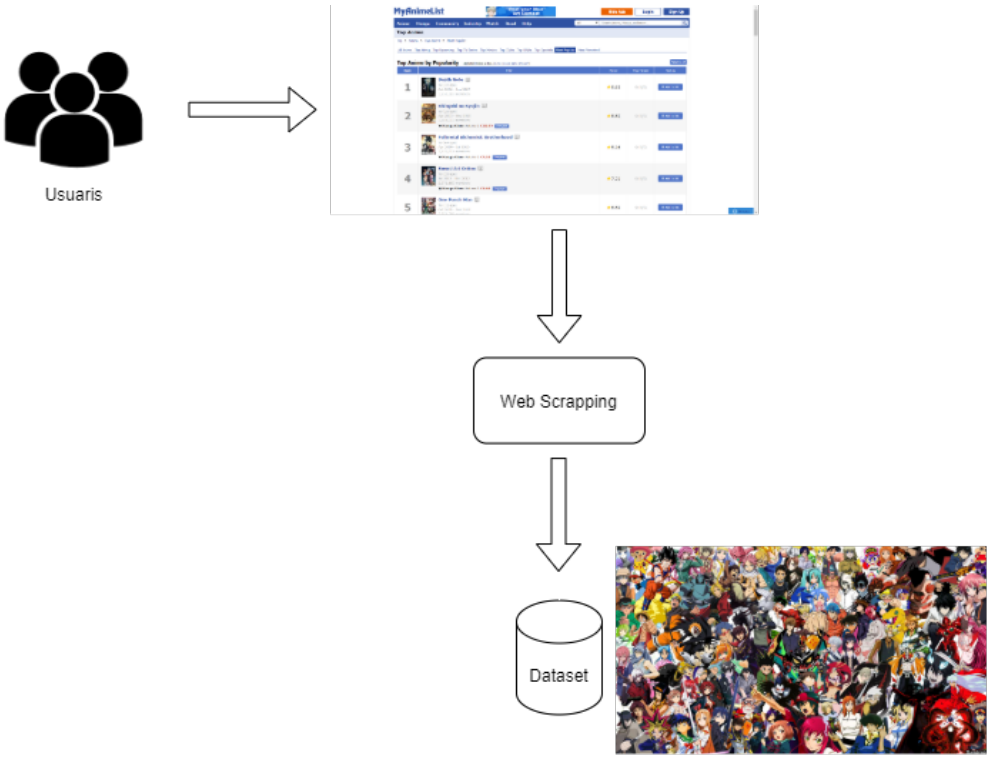
Títol

Top animes més populars

Descripció

El conjunt de dades generat està format per diferents característiques dels animes més populars en la pàgina "https://myanimelist.net/". Els animes més populars són els que tenen més seguidors a la pàgina. S'han agafat les característiques més rellevants de tota la informació de la pàgina.

Representació



Contingut

Cada fila del dataset corespon a un anime.

- Nom: títol del anime.
- Nota: Nota mitja de les valoracions dels usuaris.
- Ranking: Posició en el ranking de més populars.
- Seguidors: Nombre de seguidors en la pagina.
- Tipus: tipus del anime, el qual pot ser "TV" (serie de TV), "Movie" (pelicula), "OVA" i "Special" (episodis espacials) i "ONA" (serie online).
- Episodis: Número d'episodis totals.
- Llicència: Llista de distribuïdors.
- Estudi: Llista de productors.
- Gènere: temes relacionats amb aquell anime (exemple: Misteri, Suspens, etc.).
- Demografia: categoria japonesa per saber a qui va dirigit l'anime, per exemples: Shounen(adolecent masculí), Shojo(adolecent femení), Ecchi (adults), etc.
- Edat: Rang d'edat per visualitzar l'anime.

Agraïments

Les dades les hem capturat de la web https://myanimelist.net/, el propietari de la pàgina web es MyAnimeList Co., Ltd . Un treball similars seria el de Anime Recommendations Database (https://www.kaggle.com/CooperUnion/anime-recommendations-database), que recull informació d'aquesta mateixa pàgina web per tenir una base de dades per implementar un sistema de recomenacions.

Inspiració

Tal com s'ha explicat al context, les plataformes que existeixen per visualitzar series i pel · lícules cada vegada més estan afegint aquest contingut, és per això, tenir tota aquesta informació podria ser d'utilitat per millorar el sistema de recomanacions de les plataformes com Netflix, Amazon, etc. Amb aquesta informació es podrien respondre preguntes com: quins són els animes més famosos, quins animes tenen la millor nota, quins són els animes amb més episodis, quina distribuïdora té més animes, estudis amb més animes, generes més recurrents, demografia més vista, quants animes hi ha per cada edat, etc.

Llicència

La llicència triada per a la publicació d'aquest conjunt de dades ha estat CC BY-SA 4.0 License. Els motius per escollir aquesta llicència són:

- S'ha de proveir el nom del creador del conjunt de dades generat, indicant els canvis que s'han realitzat. D'aquesta manera, es reconeix el treball de tercers i en quina mesura s'han realitzat aportacions en relació amb el treball original.
- Es permet un ús comercial. Això faria que incrementin les probabilitats que una empresa utilitzi les dades generades i facin treballs de qualitat que aporti cert reconeixement a l'autor original.
- Les contribucions realitzades a posteriori sobre el treball publicat sota aquesta llicència hauran de distribuir-se sota aquesta. Això fa que el treball de l'autor original continuï distribuint-se sota els termes que ell mateix va plantejar.

Codi

S'ha escollit fer la pràctica amb Python com a llenguatge principalment per les llibreries per a realitzar web scraping, en el nostre cas Requests i BeatifulSoap. Els nostres fitxers són:

- src: carpeta on es troba tot el codi, en el qual trobem:
  - main.py: programa principal
  - Web_Scrapping.py: classe amb les funcions per fer web scrapping

- csv: carpeta amb el fitxer resultat del dataset
- informe: carpeta amb el document de la pràctica.

URL GitHub: https://github.com/schoezb/Practica1

Dataset

URL Zenodo: https://doi.org/10.5281/zenodo.5650093

DOI: 10.5281/zenodo.5650093

Contribucions

Contribucions	Signatura
Investigació prèvia	DPE, SCB
Redacció de les respostes	DPE, SCB
Desenvolupament del codi	DPE, SCB