PR1_web_scraping

October 18, 2024

1 Web Scraping PR1

Table of contents

1.1 Propuesta Inicial

La propuesta inicial era buscar un tema algo lúdico, por lo que esta propuesta girara en torno a sitios webs relacionados con lo lúdico (juegos de mesa, video juegos, etc.)¿Sobre que tema podríamos enfocar la PR? pues con esta premisa en mente, he recopilado información para empezar a trabajar.

A continuación listo algunos sitios webs que podrían ser interesantes:

1. SteamDB Base de datos de la plataforma Steam, en inglés. Los datos que se pueden recopilar fuera de la API estarían limitados.
2. BoardGameGeek fuente de información sobre juegos de mesa, con datos sobre la puntuación, la complejidad, la mecánica, etc.
3. Metacritic Video Games Recopila reseñas de videojuegos y les asigna una puntuación media. Información sobre el nombre del juego, la plataforma, la fecha de lanzamiento, el género, la puntuación de la crítica, la puntuación de los usuarios, etc.
4. HowlongBeat Tienen formación sobre la duración de los videojuegos, tanto para la historia principal como para completarlos al 100%. Contiene datos sobre el nombre del juego, la plataforma, el tiempo estimado para completarlo, etc.
5. vgchartz Ofrece información sobre las ventas de videojuegos, incluyendo el número de unidades vendidas por plataforma y región. Tiene datos sobre el nombre del juego, la plataforma, las ventas totales, etc.
6. TCGPlayer Se especializa en la venta de cartas coleccionables, incluyendo Magic: The Gathering, Pokémon y Yu-Gi-Oh!. Contiene datos sobre el nombre de la carta, el precio, la rareza, la edición, etc.
7. Cardmarket Parecida a TCGPlayer, esta web ofrece una amplia selección de cartas coleccionables y permitiría extraer datos sobre precios, disponibilidad, etc.
8. PriceCharting Rastrea el precio histórico de videojuegos, consolas, juegos de mesa y cartas coleccionables. Se podria extraer datos sobre la evolución del precio de un producto a lo largo del tiempo.
9. Camelcamelcamel Enfocada en Amazon. Rastrea el historial de precios de productos en Amazon, lo que permite analizar las fluctuaciones de precios y encontrar ofertas.

Destacar que todos los sitios web están en inglés.

A continuación he empezado por analizar los archivos Robots.txt:

1.1.1 Archivos Robots.txt

1. SteamdDB

En general, el archivo robots.txt de SteamDB está configurado para permitir el acceso a la información básica del sitio web, pero restringe el acceso a datos específicos y sensibles.

2. BoardGameGeek

El archivo robots.txt de BoardGameGeek es bastante restrictivo. Permite el acceso a la mayoría del contenido del sitio a todos los rastreadores web (*), pero con un retraso de 5 segundos entre cada solicitud (Crawl-delay: 5). Se Puede extraer información sobre juegos, como nombres, descripciones, mecánicas, puntuaciones, etc., siempre y cuando se respete el Crawl-delay. No podemos acceder a datos de usuarios, historiales, interacciones o precios de mercado, lo cual en todo caso carece de interés para el proyecto.

3. Metacritic Video Games

No se ha encontrado `robots.txt`, pero Metacritic forma parte de Fandom, por lo que el archivo robots.txt revisado es el de Fandom, no específicamente el de Metacritic. Esto significa que las reglas se aplican a todo el sitio de Fandom, incluyendo Metacritic.
Se pide a los motores de búsqueda que no indexen las siguientes secciones:
/d/u/
/f2/embed
/fandom?p=
/wp-content/uploads/
Disallow: Se prohíbe el acceso a los rastreadores a las mismas secciones mencionadas en “Noindex”.

Las reglas generales de Fandom permiten rastrear e indexar la información principal del sitio, incluyendo las páginas de reseñas de videojuegos.

4. HowlongBeat

El archivo robots.txt de HowLongToBeat es bastante sencillo.

- Permite el acceso a todos los rastreadores web (User-agent: *) a todo el sitio -(Disallow:) excepto a la sección /admin y /api.
- Bloquea completamente el acceso a tres rastreadores específicos: Exabot, PiplBot y GPTBot.

Ello supone que es posible rastrear la mayor parte del sitio web de HowLongToBeat, incluyendo las páginas de juegos, para extraer información sobre la duración de los juegos.

5. vgchartz

El archivo robots.txt de VGChartz es muy permisivo. En teoría, no hay restricciones explícitas para el web scraping en VGChartz según su robots.txt. Es posible rastrear cualquier sección del sitio web sin infringir sus reglas. La revision de los términos del servicio tampoco incluyen restricciones en el ámbito del raspado web.

[...]

Y he parado a la mitad por si te parece bien, completes por tu parte el resto de sitios web listados o propongas algunos nuevos.