Daniel Fernández Martínez

Práctica 1

Contexto

La información generada corresponde con el top de juegos más jugados del día 29 de octubre de 2021. En concreto, estos datos se han obtenido a través de las estadísticas públicas que proporciona la plataforma Steam creada por la empresa Valve, la cual ofrece uno de los servicios de distribución digital de videojuegos más reconocidos a nivel mundial.

Se ha seleccionado esta plataforma en particular dada su popularidad y su antigüedad, al existir desde el año 2003 y ofrecer un gran volumen de videojuegos en formato digital para ordenador. Además de esto, los datos vienen directamente de la fuente, al ser la propia empresa Valve quien tiene acceso a la información de juego de sus usuarios.

Dataset: Top 100 juegos más jugados de Steam

En esta sección se describe el conjunto de datos resultante y se menciona su contenido.

Figura 1: Logotipo y catálogo de Steam.

Como se ha comentado anteriormente, esta información se ha obtenido a través de los datos públicos de Steam a día 29 de octubre del 2021. En concreto, se ha tenido en cuenta el top 100 de juegos, que son el máximo que la plataforma permite consultar, ordenados según el número de jugadores activos en el momento de la consulta.

Los datos se han obtenido mediante un script en Python de web scraping que consulta la dirección web principal de estadísticas de Steam (https://store.steampowered.com/stats/Steam-Game-and-Player-Statistics) y los detalles resumidos de cada uno de los juegos obtenidos con un formato ligero (a través de https://store.steampowered.com/apphoverpublic/730, por ejemplo) para evitar realizar una petición a la entrada de cada producto en la propia tienda, la cual ralentizaría el proceso de extracción al cargar muchos más datos de los necesarios que acabarían siendo descartados.

Contenido

Los campos recogidos de cada juego son los siguientes:

- **Steam id**: identificador único del juego en la plataforma de Steam.
- **Game**: nombre del juego
- **Current players**: jugadores en el momento de la ejecución del programa.
- **Peak players today**: jugadores máximos en el día de la ejecución del programa.
- **Release date**: fecha en la que salió el juego.
- **Review summary**: texto resumen de las reviews de los usuarios.
- **Total reviews**: reviews de usuarios totales para el juego.
- **Tags**: conjunto de etiquetas que describen el juego.

Agradecimientos

Los datos han sido obtenidos de forma directa a través de la plataforma Steam, perteneciente a la empresa estadounidense Valve Corporation.

Como paso previo al desarrollo del proyecto y a la extracción de los datos, se ha realizado un análisis de la política de Valve en relación al web scraping, y además se han identificado las rutas no permitidas mediante una lectura del fichero "robots.txt" de la plataforma, el cual no bloquea ninguna de las rutas usadas por el programa desarrollado. En concreto, el contenido de dicho fichero en el momento de la redacción de este documento es el siguiente:

```
Host: store.steampowered.com
User-Agent: *
Disallow: /share/
Disallow: /news/externalpost/
Disallow: /account/emailoptout/?*token=
Disallow: /login/?*guestpasskey=
Disallow: /join/?*redir=
Disallow: /account/ackgift/
Disallow: /email/
Disallow: /widget/
```

Figura 2: Contenido del fichero "robots.txt" de la web store.steampowered.com.

En cuanto a la política de Valve en lo que a web scraping para obtener datos de Steam respecta, podemos decir con bastante seguridad que esto no supone un problema, siempre y cuando no se accedan a las direcciones de la figura 2 ni se realice un número desproporcionado de peticiones por segundo de forma malintencionada. Multitud de proyectos basados en Steam siguen siendo mantenidos a día de hoy por usuarios de todo tipo, y Steam no ha bloqueado el acceso ni opuesto resistencia legal a estos proyectos, siendo algunos de ellos bastante populares entre los consumidores que utilizan la plataforma.

Proyectos de web scraping sobre Steam

Otros proyectos con distintos enfoques han surgido a lo largo del tiempo para proporcionar información detallada sobre el uso de esta plataforma, permitiendo a los consumidores de este tipo de contenido tener un mejor conocimiento de la popularidad de cada producto antes de comprarlo.

Dos de los proyectos más utilizados y que se siguen actualizando a día de hoy son:

- Steamcharts: enfocado a almacenar un histórico de datos de los usuarios de cada juego, permitiendo realizar búsquedas por fecha de cada uno por individual y ver las diferencias mensuales, entre otros.
- SteamDB: de propósito más general, SteamDB contiene estadísticas resumidas y visuales de todos los aspectos de Steam, desde juegos populares actualmente hasta datos sobre las ofertas y rebajas de cada producto.

Inspiración

En muchas ocasiones, las búsquedas de videojuegos y de contenido audiovisual similar pueden ser complejas, dada la gran variedad de desarrolladores y editores que compiten por mostrar a los usuarios todo tipo de anuncios maliciosos y artículos muy influenciados según distintos tipos de intereses. Tener información objetiva sobre los datos de uso de estos productos pueden ayudar al consumidor a tomar una decisión final sin dejarse influenciar por este tipo de prácticas que llevan a cabo las industrias de distintos tipos de contenido audiovisual.

A diferencia de las herramientas mencionadas en el apartado anterior, las cuales están más orientadas al almacenamiento de datos a lo largo del tiempo y a la visualización de ofertas, los datasets como el generado en este proyecto permiten obtener una imagen limpia y rápida de los juegos más populares actualmente, incluyendo además elementos adicionales de filtrado como las etiquetas o el resumen y el volumen de las reseñas de los usuarios.

Licencia

La licencia elegida para este dataset es Creative Commons Attribution 4.0 International. Esta permite a cualquiera que lo desee copiar, modificar, distribuir e interpretar el dataset y solicita, a cambio, el reconocimiento o atribución del trabajo al autor (como se indica en el tercer apartado de este enlace). Se ha tomado esta decisión al tratarse de un proyecto creado con el objetivo de aprender conceptos básicos del web scraping, para el cual no se considera necesario el uso de una licencia con más restricciones que estas.

Código y dataset

Se puede acceder al código consultando este mismo repositorio, y el enlace del DOI en Zenodo es el siguiente:

https://zenodo.org/record/5655273#.YYlzIGDMKUk