videogamesScraper: Compilación de videojuegos Retro desde 1976 hasta 2018

Ricardo Garcia Ruiz

08 de Abril de 2018

Contents

1 Características de la práctica

1.1 Presentación

En esta práctica se elabora un caso práctico orientado a aprender a identificar los datos relevantes por un proyecto analítico y usar las herramientas de extracción de datos. Para hacer esta práctica tendréis que trabajar en grupos de 3 o 2 personas, o si preferís, también podéis hacerlo de manera individual. Tendréis que entregar un solo fichero con el enlace Github (https://github.com) donde haya las soluciones incluyendo los nombres de los componentes del equipo. Podéis utilizar la Wiki de Github para describir vuestro equipo y los diferentes archivos de vuestra entrega. Cada miembro del equipo tendrá que contribuir con su usuario Github. Podéis mirar estos ejemplos como guía:

- Ejemplo: https://github.com/rafoelhonrado/foodPriceScraper
- Ejemplo complejo: https://github.com/tteguayco/Web-scraping

1.2 Objetivos

Los objetivos concretos de esta práctica son:

- Aprender a aplicar los conocimientos adquiridos y su capacidad de resolución de problemas en entornos nuevos o poco conocidos dentro de contextos más amplios o multidisciplinarios.
- Saber identificar los datos relevantes que su tratamiento aportan valor a una empresa y la identificación de nuevos proyectos analíticos.
- Saber identificar los datos relevantes para llevar a cabo un proyecto analítico.

- Capturar datos de diferentes fuentes de datos (tales como redes sociales, web de datos o repositorios) y mediante diferentes mecanismos (tales como queries, API y scraping).
- Actuar con los principios éticos y legales relacionados con la manipulación de datos en función del ámbito de aplicación.
- Desarrollar la capacidad de búsqueda, gestión y uso de información y recursos en el ámbito de la ciencia de datos.

1.3 Descripción de la Práctica a realizar

El objetivo de esta actividad será la creación de un dataset a partir de los datos contenidos al web. Tenéis que indicar las siguientes características del dataset general:

1. Título del dataset. Poned un título que sea descriptivo.
2. Subtítulo del dataset. Agregad una descripción ágil de vuestro conjunto de datos por vuestro subtítulo.
3. Imagen. Agregad una imagen que identifique vuestro dataset visualmente
4. Contexto. ¿Cuál es la materia del conjunto de datos?
5. Contenido. ¿Qué campos incluye? ¿Cuál es el periodo de tiempo de los datos y cómo se ha recogido?
6. Agradecimientos. ¿Quién es propietario del conjunto de datos? Incluid citas de investigación o análisis anteriores.
7. Inspiración. ¿Por qué es interesante este conjunto de datos? ¿Qué preguntas le gustaría responder la comunidad?
8. Licencia. Seleccionad una de estas licencias y decid porqué la habéis seleccionado:

- Released Under CC0: Public Domain License
- Released Under CC BY-NC-SA 4.0 License
- Released Under CC BY-SA 4.0 License
- Database released under Open Database License, individual contents under Database Contents License
- Other (specified above)
- Unknown License

9. Código: Hay que adjuntar el código con el que habéis generado el dataset, preferiblemente con R o Python, que os ha ayudado a generar el dataset
10. Dataset: Dataset en formato CSV

2 Realización de la práctica

2.1 Título del dataset: Base de datos general de videojuegos retro

Para construir nuestro dataset **Base de datos general de videojuegos retro**, el conjunto de datos escogido para esta práctica ha sido el de la web http://www.retrocollect.com/. Se pretende compilar una base de datos de videojuegos retro, y despues de un exhaustivo análisis de las diversas compilaciones de videojuegos retro existentes en Internet, se ha escogido la que ofrece RetroCollect, que corresponde con videojuegos de los denominados **'retro'**, y que ha resultado ser amplia y diversificada.

Logotipo de RetroCollect extraído de http://www.retrocollect.com/

El periodo de datos que abarca la base de datos está entre 1976 y 2018, existiendo algunos videojuegos que no están correctamente datados en la própia página web.

Las principales variables de este conjunto corresponden con la **fecha**, el **nombre** del videojuego y la **plataforma** de juegos.

2.2 Imagen identificativa

Protagonistas del videojuego iconico por excelencia: Mario Bros.

Cortesía de Alexas, https://pixabay.com/es/mario-luigi-yoschi-cifras-gracioso-1557240/

2.3 Contexto

El conjunto de datos se corresponde con videojuegos de tipo *'retro'* que se hayan poblicado en un periodó que abarca 1976 al 2018. Aunque existen videojuegos que abarcan desde los años 50, estos son considerados como *arcanos*, y en algunos casos no son mas que prototipos de los juegos que verdaderamente empezaron a comercializarse en la década de los 70 del siglo pasado.

Entre ellos pueden encontrarse todo tipo de géneros:

- Action
- Adventure
- Compilation
- DLC
- Add-on
- Educational
- Puzzle
- Racing
- Driving
- Role-Playing (RPG)
- Simulation
- Special Edition
- Sports
- Strategy/Tactics

2.4 Contenido

Para cada videojuego, el cual se corresponde con un registro en el conjunto de datos, se recogen las siguientes características:

1. **Original System**: .
2. **Title**: Titulo original del videojuego (en ingles).
3. **Year**: año de publicación del videojuego.
4. **Publisher**: compañía que lo lanzó al mercado comercial en el año citado.
5. **Developer**: compañía desarrolladora del videojuego. Puede coincidir en algunos casos con el *'Publisher'* pero es importante, ya que en los 70, 80 y 90 del siglo pasado había muchas compañías que desarrollaban los juegos por encargo.
6. **Países de lanzamiento**: Se agrupan en tres categorías (Europe, US y Japan). Cada registro puede estar vacío o con una **'x'**. La **'x'** indica que en ese país o región no fue comercializado:
    1. **Europe**: cualquier país de Europa.
    2. **US**: en Estados Unidos.
    3. **Japan**: en Japón.

RetroCollect solo almacena datos de juegos que se consideran **'retro'**, de forma que aunque algunos juegos puedan ser sobradamente conocidos, deben alcanzar la categoría de **'retro'** para poder estar incluídos en esta web.

Existen otras web que compilan datos sobre videojuegos, pero la extracción de datos únicamente de videojuegos *retro* requería de un tratamiento posterior que no es el objetivo de esta práctica, sino el adquirir los datos directamente de la web.

2.4.1 Ficheros del código fuente

1. **src/videogamesScraper**: Es el código de entrada al scraping y contiene el código principal utilizado para gestionar el trabajo de compilación de toda la base de datos retro de videojuegos de la web **RetroCollect**.
2. **src/getPlatformDB**: Contiene el código fuente de la función **getPlatformDB()**. Esta función accede a la web de RetroCollet y obtiene un data frame con los códigos numerícos y sus equivalencias en texto de los nombres de las Plataformas disponibles en RetroCollect. Con esta función se puede realizar un filtro por tipo de plataforma, o bien toda la base de datos de videojuegos (por defecto).
3. **src/searchPaginationDB**: Contiene el código fuente de la función **searchPaginationDB()**. La función realiza una búsqueda en la web localizando la página web última en la que se deben buscar los datos de scraping, devolviendo un valor numérico con la última página que se debe acceder. Los paramétros son los siguientes:
   - **url_base**: La dirección web generalde acceso a RetroCollect
   - **listview**: Sistema de visualización, por defecto *'list'*
   - **modeview**: Por defecto se buscan *'games'*
   - **plataforma**: La plataofrma de filtro, por defecto = 0, todas sin excepcion
   - **sort**: Esquema de ordenación, puede tomar 4 parámetros:
     - *'title'*, es el defectivo y es igual a **NA**
     - *'system'*, organiza por S.O. y es igual a *"platform"*
     - *'publisher'*, organiza por cia. de publicación y es igual a *"publisher"*
     - *'year'*, organiza por año de publicación y es igual a *"year"*
   - **filas**: Indica el numero de filas de visualización por página, defecto = 20
   - **verbose**: Indica si se desea o no información de progreso, defecto = *TRUE*
4. **src/accessVideoGameDatabase**: Contiene el código fuente de la función **accessVideoGameDatabase()**. La función realiza un web scrapin en RetroCollect, posibilitando un acceso dinamico a la misma y configurando algunos parametros de control en la llamada a la pagina web de RetroCollect indicando algunas variables de carga y control de visualización. Los paramétros son los siguientes:
   - **url_base**: La dirección web generalde acceso a RetroCollect
   - **listview**: Sistema de visualización, por defecto *'list'*
   - **modeview**: Por defecto se buscan *'games'*
   - **plataforma**: La plataofrma de filtro, por defecto = 0, todas sin excepcion
   - **sort**: Esquema de ordenación, puede tomar 4 parámetros:
     - *'title'*, es el defectivo y es igual a **NA**
     - *'system'*, organiza por S.O. y es igual a *"platform"*
     - *'publisher'*, organiza por cia. de publicación y es igual a *"publisher"*
     - *'year'*, organiza por año de publicación y es igual a *"year"*
   - **filas**: Indica el numero de filas de visualización por página, defecto = 20
   - **verbose**: Indica si se desea o no información de progreso, defecto = *TRUE*

2.5 Agradecimientos

Los datos han sido compilados desde la base de datos online RetroCollect. Se ha utilizado el lenguaje '**R**' y de técnicas de *Web Scraping* para extraer la información alojada en las múltiples páginas de gestión de esta base de datos online.

En la presentación de este trabajo se han utilizado referencias a los siguientes trabajos:

1. Simon Munzert, Christian Rubba, Peter Meißner, Dominic Nyhuis. (2015). *Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining.* John Wiley & Sons
2. Garcia Ruiz, Ricardo. (2014). *Estudio y caracterización de marcas de videojuegos mediante análisis de la producción de patentes y el desarrollo técnico de software para plataformas de videojuegos.* Universitat Internacional de Catalunya, DOI: 10.13140/2.1.5162.0166.

2.6 Inspiración

El conjunto de datos puede resultar muy útil para trabajos relacionados con la minería de datos relativa al comportamiento de los videojuegos (García Ruiz 2014).

Tambien sirve para verificar el comportamiento analítico sobre los desarrolladores y publicadores de videojuegos a lo largo del tiempo.

De igual manera, también puede utilizar verificar el comportamiento de aceptación de los diversos videojuegos a lo largo del tiempo en las distintas zonas geográficas, y su impacto relativo o absoluto.

2.7 Licencia

La licencia escogida para la publicación de este conjunto de datos ha sido **CC BY-NC-SA 4.0 ES**. Los motivos que han llevado a la elección de esta licencia tienen que ver con la idoneidad de las cláusulas que esta presenta en relación con el trabajo realizado:

Se permite con nuestro trabajo y la base de datos extraída de la web:

- *Compartir* — copiar y redistribuir el material en cualquier medio o formato
- *Adaptar* — remezclar, transformar y crear a partir del material

Por otro lado, la licencia activa las siguientes restricciones:

- **Reconocimiento**: Debe reconocer adecuadamente la autoría, proporcionar un enlace a la licencia e indicar si se han realizado cambios. Puede hacerlo de cualquier manera razonable, pero no de una manera que sugiera que tiene el apoyo del licenciador o lo recibe por el uso que hace.
- **NoComercial**: No puede utilizar el material para una finalidad comercial.
- **CompartirIgual**: Si remezcla, transforma o crea a partir del material, deberá difundir sus contribuciones bajo la misma licencia que el original.
- **No hay restricciones adicionales**: No puede aplicar términos legales o medidas tecnológicas que legalmente restrinjan realizar aquello que la licencia permite.

2.8 Código fuente y dataset

Tanto el código fuente escrito para la extracción de datos como el dataset generado pueden ser accedidos a través de este enlace.

Para el desarrollo de la práctica se han tomado en consideración las técnicas de Web Scraping de (Munzert et al. 2014).

Recursos

García Ruiz, Ricardo. 2014. "Estudio Y Caracterización de Marcas de Videojuegos Mediante análisis de La Produccion de Patentes Y El Desarrollo Técnico de Software Para Plataformas de Videojuegos." PhD thesis, Universitat Internacional de Catalunya.

Munzert, Simon, Christian Rubba, Peter Meißner, and Dominic Nyhuis. 2014. *Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining.* John Wiley & Sons.