PRÁCTICA 1

EQUIPO

Los integrantes de esta práctica somos:

- Ignacio Such Ballester (isuch@uoc.edu)
- Andrés Fonts Santana (afontss@uoc.edu)

WEB SCRAPING DE LA PÁGINA WEB MISHIGEEK.COM

1 CONTEXTO

El objetivo de esta práctica de Web Scraping es la búsqueda de información sobre las valoraciones que han recibido una variedad de juegos de mesa con la finalidad de realizar un estudio de mercado. Para ello, se ha decidido escoger la página web www.mishigeek.com la cual recopila una gran variedad de juegos de mesa. Asimismo, contiene diversos tipos de valoraciones como "Originalidad" y "Mecánicas" los cuales ayudarán a obtener una buena información sobre lo que más interesan a los jugadores.

2 TÍTULO

2. El título empleado para el dataset obtenido es Boardgames Ranking.

3 DESCRIPCIÓN DEL DATASET

3. Mediante el uso de librerías de web Scrapping, lo que se ha obtenido de la web mencionada en el apartado anterior son las valoraciones de todas las reseñas de los juegos de mesa que hay en la propia web.

4 REPRESENTACIÓN GRÁFICA

4. XXXXXXXXXXX

5 CONTENIDO

5. x

A la hora de recopilar los datos de la web mencionada anteriormente, se han empleado técnicas de buena práxis para el web scraping. Para este proyecto, se ha empleado el uso de la librería *Beautiful Soup (bs4)*.Esta librería transforma en texto plano la web y mediante el uso de funciones de la misma librería y expresiones regulares, se ha logrado obtener los datos necesarios para la creación del dataset. Para lograr esto, se ha creado un script que extraiga del archivo *sitemap.xml* todos los links que ponga la regex *-resena*. Una vez extraídos los links, se ha creado una función que por cada web, extraiga las valoraciones de ese juego y categorías de dichas valoraciones, y las almacene en un diccionario. Para hacer este proceso, se ha de implementar un delay de tiempo para que la página web no bloquee los scripts aún teniendo el permiso del autor. Es por ello que por cada extracción de datos, se dará un margen de unos 30 segundos para que el servidor web entienda que no estamos saturando la web.

AÑADIR : Obtener como se han recogido Responder : QUÉ, DÓNDE Y CUÁNDO

6 AGRADECIMIENTOS

6. En primer lugar, agradecemos la colaboración de Javier Rodríguez Menéndez, autor del blog de mishigeek.com el cual nos dio permiso para tratar sus datos de manera privada y no distribuirlos al final del proceso. CABECERA, DELAY Y PERMISO DEL AUTOR.

7 INSPIRACIÓN

7.

• Identificar los mejores juegos publicados cada año según mishigeek.com • Saber qué editoriales de juegos de mesa tienen más éxito • Estudiar los tipos de juego de mesa según la temática/mecánica de juego por editorial, diseñador... • Comparar precios entre editoriales • Buscar un juego de mesa por puntuación y según diferentes criterios

8 LICENCIA

8. La licencia empleada es la licencia *CC BY-NC-SA*, ya que el dueño de la web nos indicó que no se deseaba que se hiciera pública la información por si alguien o empresa la empleaban con uso comerciales. Es por ello que esta licencia se acopla a las necesidades del autor. Esta licencia ha de reconocer el auténtico autor de los datos, que en este caso es Javier Rodríguez Menéndez, en caso de divulgación de la ingformación obtenida en el dataset resultante. Asimismo, si remezcla, transforma o crea a partir del material, deberá difundir sus contribuciones bajo la misma licencia que el original.

9 CÓDIGO

9. XXXXXXXX

Esto es un cambio

10 DATASET

11 VÍDEO