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 contexto que hemos escogido en esta práctica de Web Scraping la recopilación de valoraciones de juegos de mesa. Hemos escogido la página web www.mishigeek.com para realizar esta tarea.

MishiGeek.com contiene centenares de reseñas. En cada una se publican diveros tipos de valoraciones como "Originalidad" y "Mecánicas". También se asigna una valoración total y una clasificación cualitativa como "Suspenso", "Recomendado" o "Juegaco".

2 TÍTULO

El título del dataset es **Boardgames Ranking**.

3 DESCRIPCIÓN DEL DATASET

El dataset Boardgames Ranking contiene N registros y NxM datos. Cada registro corresponde a un juego de mesa.

Hemos recopilado N reseñas publicadas en mishigeek y para cada una hemos identificado M atributos de interés. El formato del dataset obtenido es un fichero CSV.

4 REPRESENTACIÓN GRÁFICA

5 CONTENIDO

Cada registro de Boardgames Ranking cuenta con los siguientes campos:

Campo	Tipo	Descripción
nombre	string	Nombre del juego de mesa
n_jug	string	Número de jugadores
duracion	string	Duración de una partida, en minutos
dureza	string	Preguntar a Mishieek
edad	string	Edad mínima recomendada o rango de edad
precio	float	Precio del juego en euros

Campo	Tipo	Descripción
genero	string	Género al que pertenece el juego
editorial	string	Editorial que publica el juego
diseño	string	Diseñadores del juego
val_asp	float	Valoración en cuanto Aspecto / Componentes
val_inter	float	Valoración en cuanto a nivel de Interacción
val_var	float	Valoración en cuanto Variabilidad
val_org	float	Valoración en cuanto a Originalidad
val_mec	float	Valoración en cuanto Mecánicas
val_glob	float	Valoración global del juego
val_cual	string	Valoración Cualitativa

Los datos de nuestro dataset se han recogido de publicaciones en mishigeek.com comprendidas entre las fechas día/mes/año y día/mes/año.

Hemos estructurado la solución en 3 partes:

1. Script de python que analiza las páginas de reseñas y obtiene las direcciones url de cada reseña publicada en mishigeek.com. La lista se escribe en un fichero CSV.
2. Función de python que dada una reseña obtiene cada uno de los campos
3. Script de python queitera sobre la lista de enlaces obtenida en el punto 1. y obtiene de cada enlace los campos llamando a la función definida en 2.

Para recopilar los datos 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.

6 AGRADECIMIENTOS

Para poder realizar la tarea de *web scraping* hemos necesitado tomar tres medidas básicas:

- Solicitar permiso a mishigeek.com para realizar la actividad.
- Consulatar *robots.txt* de mishigeed.com y modificar la cabecera de la petición HTTP para evitar el bloqueo de mishigeek.com
- Prevenir la saturación del servidor debido a nuestras conexiones.
    - Hemos calculamos un tiempo entre peticiones de *n* milisegundos.

    - El código se estructura en dos archivos que se ejecutan de forma secuencial. En primer lugar se realiza una búsqueda de enlaces con reseñas y en segundo lugar se analizan los enlaces con reseñas para obtener el *dataset*. Así prevenimos también realizar centenares de peticiones al servidor en un espacio corto de tiempo.

Los datos recogidos son propiedad de Javier Rodríguez Menéndez, autor del blog mishigeek.com. Agradecemos la colaboración de Javier al consentir explícitamente la labor de *web scrpaing* en su dirección web, con la condición de tratar sus datos de manera privada y no distribuirlos de forma pública.

Añadir citas de análisis anteriores. Justificar búsqueda con análisis similares (pendiente Andrés Fonts).

7 INSPIRACIÓN

- 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

a 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

10 DATASET

11 VÍDEO