Guía inicial de Web Scraping

Importante: Para la implementación de las actividades de la presente guía se deben tener instalados en el sistema los módulos de terceros Requests y Beautiful Soup.

Ejercicio 1

Obteniendo HTML

Implementar un script que utilice el módulo pypercLip para obtener la url de un sitio web luego se debe utilizar requests para obtener el contenido HTML de dicha web para finalmente guardar el contenido de la página en un archivo de texto de manera persistente en el sistema.

Adicional: Es importante que cada vez que se realice un ctrl+C el script determine si la url es válida antes de comenzar con el procedimiento de descarga. Finalmente los nombres de los archivos alojados en el sistema deben comenzar con un número asociado al número de descargas realizadas hasta el momento.

Ejercicio 2

Obteniendo párrafos

Implementar un script que utilice el módulo pypercLip para obtener la url de un sitio web luego se debe utilizar requests para obtener todos los párrafos de la página alojandose en una lista con el orden asociado a la aparición en la respectiva web.

Finalmente el script deberá imprimir la lista construida a través de la consola del sistema.

Adicional: Es importante que cada vez que se realice un ctrl+C el script determine si la url es válida antes de comenzar con el procedimiento de descarga.

Ejercicio 3

Atravesando la paginación

Implementar un script que utilice el módulo pypercLip para obtener la url de un sitio web luego se debe utilizar requests de manera iterativa para ir obteniendo el contenido de artículos que se encuentren dentro de una paginación. Dichos contenidos deben ser alojados en una lista y posteriormente el contenido de la lista debe ser mostrado utilizando la consola del sistema.

Adicional: Es importante que cada vez que se realice un ctrl+C el script determine si la url es válida antes de comenzar con el procedimiento de descarga. Adicionalmente si la paginación es muy grande se puede establecer un límite utilizando un parámetro adicional n.

Ejercicio 4

Descargando PDF's

Implementar un script que utilice el módulo pypercLip para obtener la url de un sitio web luego para luego obtener y descargar (en caso de que hubiere) todos los documentos en PDF del sitio en la memoria persistente del sistema.

Adicional: Es importante que cada vez que se realice un ctrl+C el script determine si la url es válida antes de comenzar con el procedimiento de descarga. Adicionalmente los PDF's deben alojarse en un subdirectorio con el mismo nombre.

Ejercicio 5

Obteniendo promedio de análisis

Implementar un script que utilice el módulo pypercLip para obtener la url de un sitio web de venta de productos (Por ej. ML) enfocado en un producto en particular (dicho producto debe ser solicitado por consola al usuario), luego se deben obtener las valoraciones de los usuarios para dicho producto y obtener un promedio de las mismas así como también una recolección de las palabras que ocurran con más frecuencia entre las devoluciones realizadas por los usuarios de la plataforma. Finalmente se solicita cargar la siguiente información en un documento en el sistema (que se va a ir escribiendo a medida que se busquen productos):

- Nombre del producto
- Sitio de ventas
- Promedio de valoraciones
- Palabras clave más frecuentes

Adicional: Es importante que cada vez que se realice un ctrl+C el script determine si la url es válida antes de comenzar con el procedimiento de descarga. Adicionalmente los PDF's deben alojarse en un subdirectorio con el mismo nombre.

Ejercicio 6

El precio más bajo

Aplicando lo trabajado en ejercicios anteriores se solicita implementar un script que solicite al usuario el nombre de un producto alimenticio con su respectiva marca. Luego se solicita realizar una búsqueda del mismo en las webs de los principales supermercados de la zona.

El objetivo es determinar el supermercado que tenga stock del producto al menor precio posible al momento de realizar la consulta del mismo.