



Contenido

Manual de Usuario Practica Final – Web Scraping.....	1
Nota del autor.	3
Introducimos la web que queremos scrapear, con el formato 'ideal.es'.	4
En la pestaña General	5
En la pestaña Palabras Clave	6
En la pestaña de Multimedia.....	7
En la pestaña de MetaDescripción.....	8

Nota del autor.

Ha sido un proyecto que me ha llevado prácticamente un mes llevar a cabo y complicarme la cabeza, no ha sido para nada fácil, pero lo curioso es que lo veo ahora y digo: “pues oye no ha sido tan complicado.”

Lo malo de scrapear las páginas webs (obtener todo su código, texto y lo que quieras), es que es complicado, porque hay algunas reglas por las que nos podemos regir para poder analizarlas como por ejemplo que los encabezados son h1, h2, h3... pero cuando queremos scrapear la “metadescripcion” el atributo cambia, dependiendo de la web, y no es nada fácil lidiar con eso.

Hay muchos métodos diferentes que he usado, muchos normales, es decir, que hemos visto. Pero hay muchos otros que no hemos visto y que he implementado aquí. HashMap, listas iterator, exportar en archivos, una librería JSoup(que es la que utilizo para sacar los datos de las páginas web).

En definitiva, ha sido algo movidito pero que lo he disfrutado mucho, a continuación, te dejo el manual para que puedas ver todo lo que tiene y lo que puedes hacer con el programa.

Atentamente, Javier Heredia.

Introducimos la web que queremos scrapear, con el formato 'ideal.es'.

Website:

Analizar Web

General

Palabras Clave

Multimedia

MetaDescripción

KeyWords

Buscar KeyWords

Buscar keywords

1 palabra

Tabla de palabras

Exportar a CSV

Veces	Palabras	Densidad KeyWord
-------	----------	------------------

Encabezado	Longitud	Contenido
------------	----------	-----------

En la pestaña General

Cuando pulsamos el botón “Cuenta las palabras” en la parte derecha aparecen la cantidad javascript y encabezados que hay, además de las palabras totales.

Abajo están la media de palabras que hay en los párrafos y en los encabezados.

Website
Website:

General Palabras Clave Multimedia MetaDescripcion

General

Palabras totales: 0

JavaScript: 0

H1: 0 H2: 0 H3: 0 H4: 0

Media de palabras
párrafos: 0
h1 0
h2 0
h3 0
h4 0

En la pestaña Palabras Clave

Podemos pulsar el botón Buscar KeyWords para buscar las palabras que hay, cuantas veces aparece y la densidad de palabras claves((veces que aparece*100)/palabras totales).

En el botón de Exportar a csv, tenemos la opción de exportar la primera o la segunda tabla como nosotros queramos en la dirección: output/(nombreweb) palabras.csv

Website:

General Palabras Clave Multimedia MetaDescripcion

KeyWords (1) 1 palabra ▼

Buscar KeyWords (2)

Veces que aparece	Palabra	Densidad de KeyWord
194	de	6 %
116	la	4 %
82	el	3 %
74	en	2 %
60	granada	2 %
46	a	1 %
41	y	1 %
41	del	1 %
37	los	1 %
34	para	1 %
34	que	1 %
31	las	1 %
28	/	1 %
28	un	1 %
21	con	1 %

Encabezado	Longitud	Contenido
h2	1	Granada
h2	1	Provincia
h2	1	Andalucía
h2	2	Granada CF
h2	1	Deportes
h2	1	España
h2	1	Mundo
h2	2	Compartiendo conocimiento
h2	1	Videos
h2	1	Culturas
h2	1	Butaca
h2	1	VIVIR
h2	3	Salud y ciencia
h2	4	A pie de calle
h2	1	Gente&Estilo

En la pestaña de Multimedia

El botón de obtener imágenes nos visualizará todas las imágenes que tiene la página web.

En la lista de la derecha nos aparece el texto que tiene el atributo de la imagen ALT:

```
 </img>
```

Después nos aparece la cantidad de imágenes totales.

(hay errores si encuentra imágenes con el archivo .svg)

En el cuadro de debajo “Videos” si pulsamos el botón a Obtener Videos nos abrirá en el navegador todos los videos que tiene la web.

Website

Website: Analizar Web

General

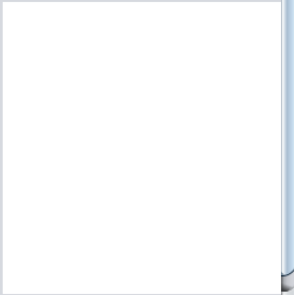
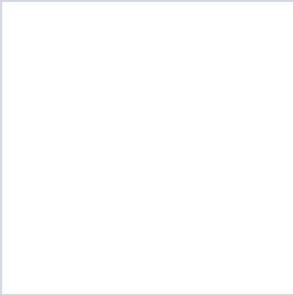

Palabras Clave

Multimedia

MetaDescripcion

Imágenes

Obtener imágenes



La Claqueta
WC luz nocturna, Adoric L
mDesign Organizador de
TTMOW Cubos de Basura
La Croqueta - Color Ver
Healthy Clubs Limpiador
Tapas de silicona elástica
Dispositivo de ahorro de a
Metaltex Clean-Tex Bande
Seguridad para el Hogar
Climatización Catálogo de
Electrodomésticos Amplio
Iluminación Amplio catálo
Muebles Los mejores cat
Jardín Los mejores catálo
Baño Los mejores catálo
Comedor Los mejores ca
Cocina Los mejores catál
Hogar Descubre product
Descubre Descubre de

Cantidad: 52

Videos

Obtener videos

En la pestaña de MetaDescripción.

Aparece el título de la página web, la metadescripción de la web, el lenguaje y si la web tiene el archivo robots.txt

Website

Website:

laclaqueta.net

Analizar Web

General

Palabras Clave

Multimedia

MetaDescripcion

MetaDescripcion

Titulo:

La Claqueta Artículos para el Hogar y Trucos para la Vida Diaria

MetaDescripcion:

La Claqueta - Aquí donde encontrarás gran cantidad de opiniones y reseñas de artículos para el hogar más populares y los mejores trucos para la vida diaria.

Lenguaje:

es

Robots:

User-agent: * Disallow: /wp-login Disallow: /wp-admin Disallow: //wp-includes/ Disallow: /*/trackback/ Disallow: /*/attachment/ Disallow: /author/ Disallow: /*/page/ Disallow: /tag/*