**Estructura del Site**

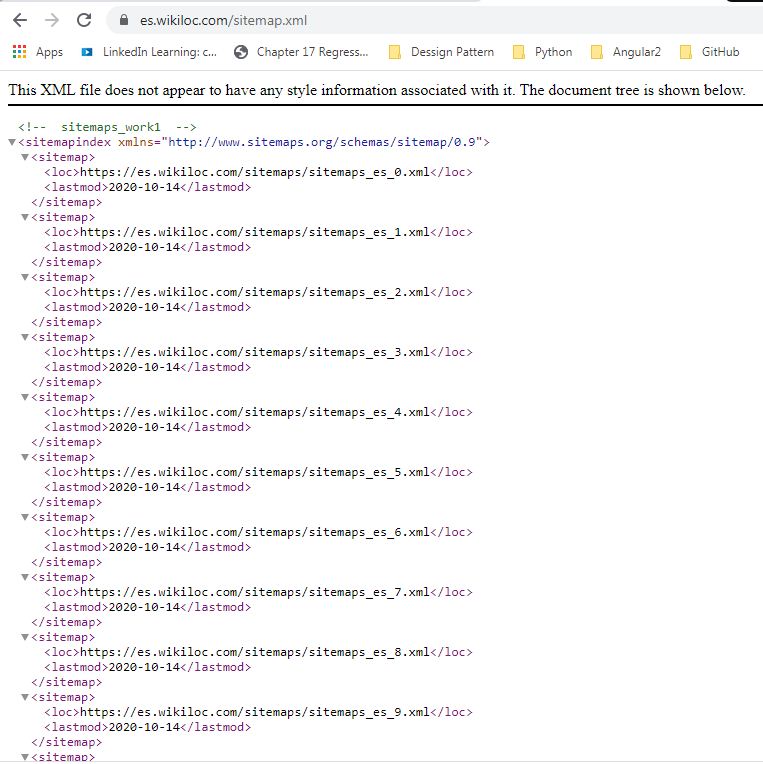
Es.wikiloc.com/rutas

En este portal podemos encontrar 2 tipos de ficheros que nos pueden ayudar en gran medida a desarrollar una actuación de Web Scrapping el máximo de ética posible, por una parte tenemos el fichero [robots.txt] , del cual mostramos su contenido:



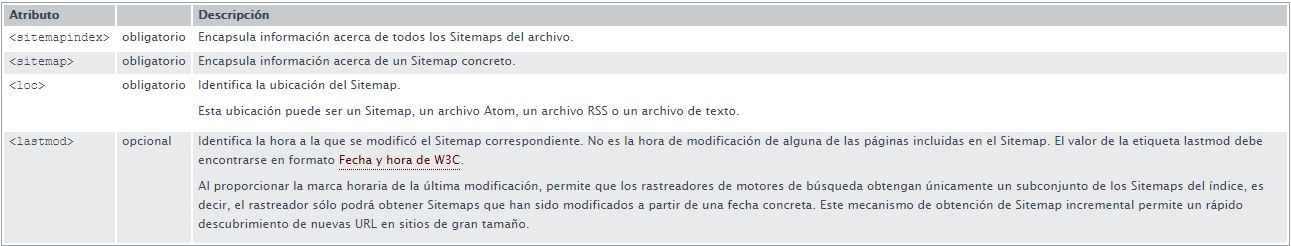
Que observamos en este fichero?

* Todas las reglas son aplicables a cualquier “user.agent”
* Encontramos directivas “allow” y “disallow” las cuales tienen significados opuestos y además tienen interpretaciones también diferentes:
  + Standard Implementation: la 1º directiva que coincida con la url por la que se pregunta se aplicará.
  + Google Implementation: si el patrón indicado en la directiva “allow” tiene una longitud igual o superior a la de la directiva “disallow” entonces se aplicará la directiva “allow”.
  + Bing Implementation: si la url preguntada coincide con más de una directiva, independientemente de que sea “allow” o “dissallow” se aplicará la que tenga una longitud mayor.
* Se bloquean algunas páginas con extensión “.do”
* Se bloquean las peticiones a url’s que empiecen con “/enterwikiloc/worlmap”
* Se bloquean las peticiones a url’s que acaben en “\_full.jpg”
* Se bloquean las peticiones url’s a la carpeta “/cdn-cgi/”
* En la última línea se encuentra una referencia al fichero SiteMaps, del cual mostramos una parte de su contenido



Este fichero no es exactamente un fichero que se ciña al formato del protocolo SiteMap ya que no contiene algunos de los tags necesarios especificados en este protocolo, como pueden ser <urlset> o <url>.

En la siguiente imagen se puede observar el formato de este tipo de ficheros:



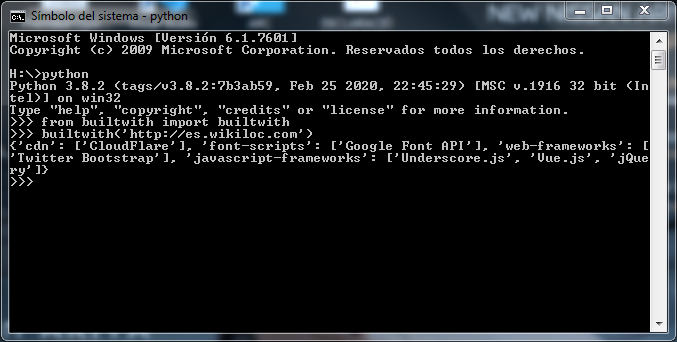
En su lugar nos encontramos con un archivo índice de SiteMap que nos permite agrupar varios archivos SiteMap. Este tipo de archivos a los que se hace referencia dentro del fichero índice de SiteMap deben cumplir unas especificaciones:

* Deben contener un máximo de 50.000 direcciones URL
* No deben superar los 50 Mb

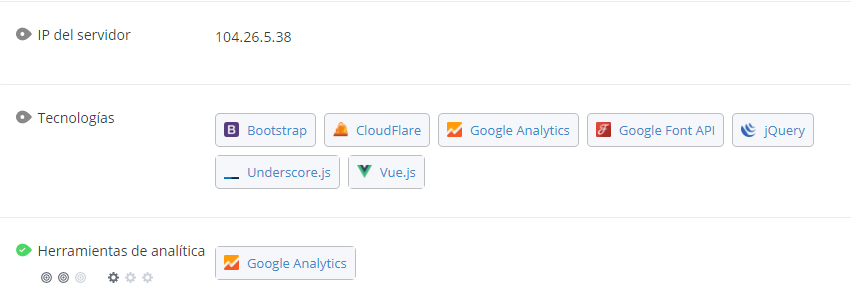
En este caso particular el fichero índice de SiteMap hace referencia a un total de 160 ficheros SiteMap, con esta cantidad de información ya podemos ver que se trata de un site con muchísimas url’s y que por lo tanto deberemos ser muy cuidadosos a la hora de realizar acciones.

Tecnologia utilizada en el Site

Para ello utilizaremos el package [builtwith] el cual nos dará información al respecto.

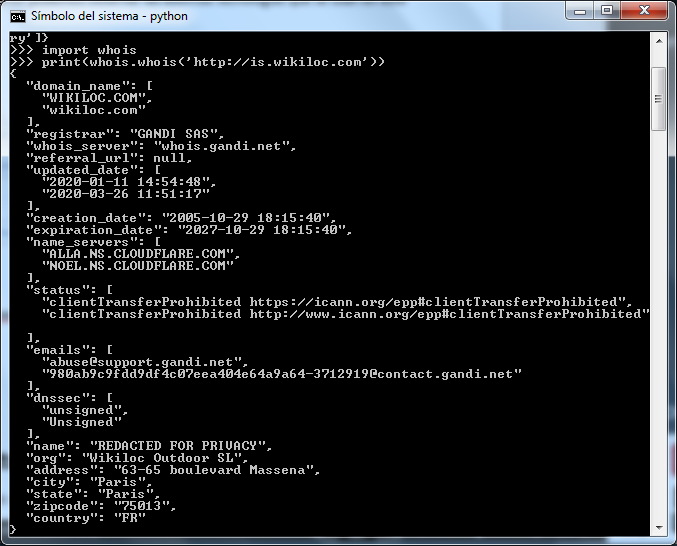


Una vez ejecutado en la imagen superior podemos observar las distintas tecnologías que se usan en este Site



Propietari del Site

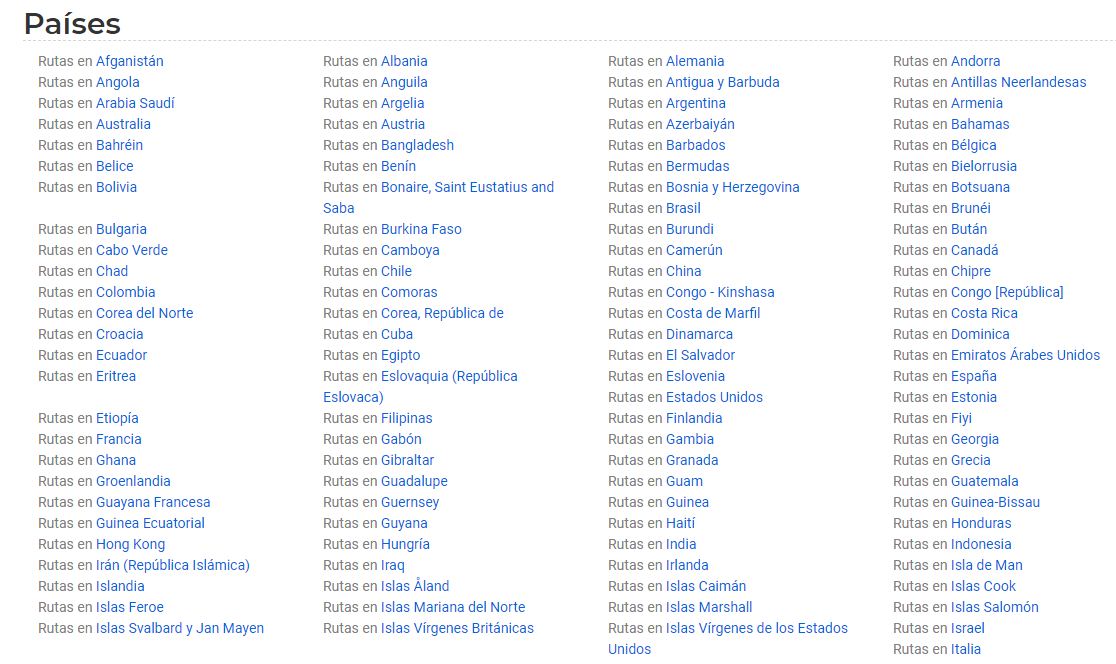
Per aconseguir informació del propietari del site utilitzarem el package [whois], el qual ens tornarà la següent informació:



**Estudi de l’ Estructura del Site**

El site al qual volem accedir és el: **https://es.wikiloc.com**.

Si accedim a l’url **“https://es.wikiloc.com/rutas”** se’ns renderitza una pàgina HTML que ens agrupa la información en 2 grups ben diferenciats, tenim les “Activitats” i els “Paisos”. La nostra feina es foculitzarà en l’opció del “Paisos” com es pot apreciar a la següent imatge:

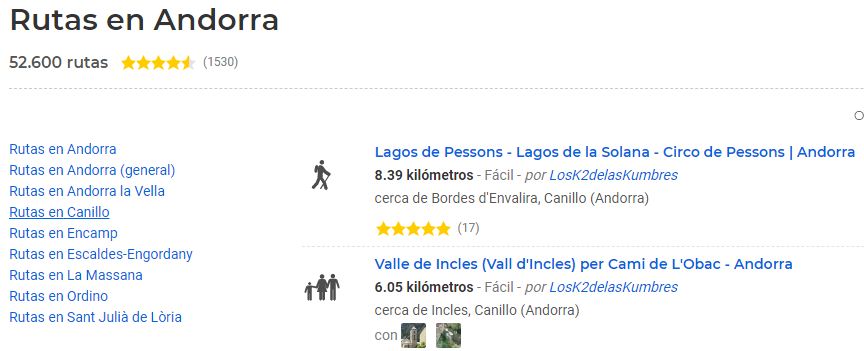


Si procedemos a observar la estructura de esta página apreciamos el uso de:

* <ul>: muestra una lista donde no se muestra un nº ordinal de referencia.
* <li>: muestra cada uno de los ítems de la lista.
* <a>: link a la url indicada



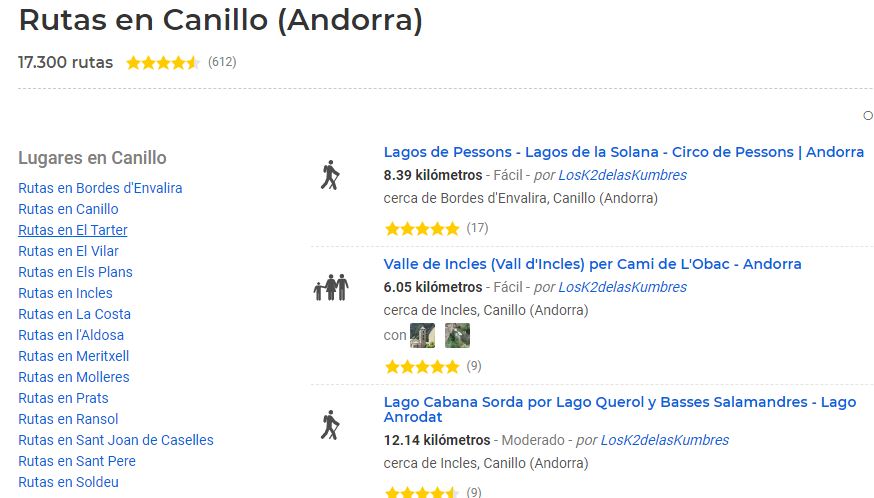
Si realizamos un click sobre “Andorra” se nos dirige a otra url “**es.wikiloc.com/rutas/outdoor/andorra**” donde observamos



Si obtenemos la estructura de página, observamos que sigue el mismo formato que anteriormente



Realizamos un click sobre “Rutas en el Canillo” y se nos dirige a la url “**es.wikiloc.com/rutas/outdoor/andorra/canillo**”



Volvemos a observar la estructura y sigue el mismo patrón que las vistas anteriormente.



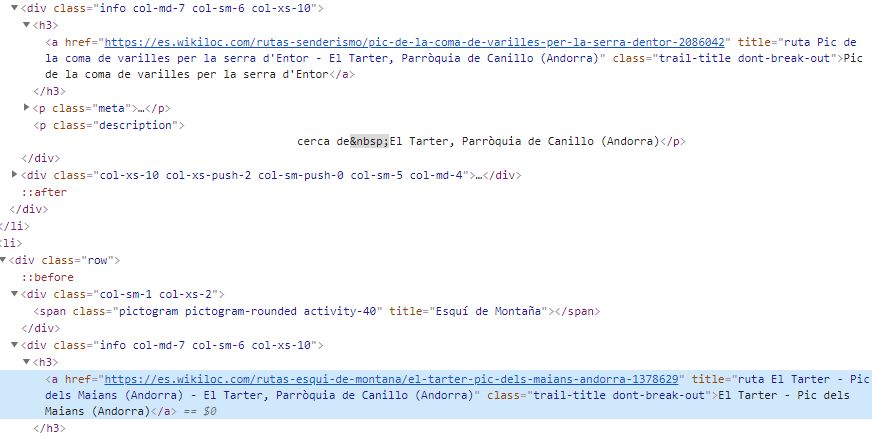
Realizamos un click sobre “Rutas en el Tarter”, que nos llevará a la url “es.wikiloc.com/rutas/outdoor/andorra/canillo/el-tarter”



Donde ya nos aparecen todas las rutas y el tipo de ruta que podemos realizar en la zona seleccionada.

La estructura de esta página difiere un poco de las anteriores:

* <h3>: heading de tipo 3
* <a>: link a la url indicada.



Y ahora si que ya sólo nos queda seleccionar la ruta deseada



Donde ya podremos obtener la información específica de la ruta:

* Tipo: Senderismo
* Nombre: Pic de la Coma de Varilles
* Distancia: 12,65 km
* Desnivel acumulado subiendo: 1076 m
* Desnivel acumulado bajando: 1027 m
* Dificultad Técnica: moderado
* Ruta circular: si
* Altitud máxima: 2761 m
* Altitud mínima: 1923 m
* Tiempo: 7h 31 minutos
* Fecha realización: Octubre 2011

Informacional adicional:

* Valoración: de 1 a 5 estrellas
* Información: de 1 a 5 estrellas
* Fácil de seguir: de 1 a 5 estrellas
* Entorno: de 1 a 5 estrellas
* Características de la ruta:
  + Información
  + Intersección
  + Fuente
  + Panoramica
  + Parquing

Referencias

<https://www.sitemaps.org/es/protocol.html>

<https://www.lawebera.es/como-hacer/ejemplos-html-xhml/protocolo-sitemap.php>

<https://www.tutorialspoint.com/python3/python_overview.htm>

<https://tedboy.github.io/bs4_doc/generated/bs4.html#classes>

<https://beautiful-soup-4.readthedocs.io/en/latest/#navigating-the-tree>

<https://www.crummy.com/software/BeautifulSoup/bs4/doc/#calling-a-tag-is-like-calling-find-all>

<https://entrenamiento-python-basico.readthedocs.io/es/latest/index.html>

<https://j2logo.com/python/web-scraping-con-python-guia-inicio-beautifulsoup/>

<https://www.geeksforgeeks.org/queue-in-python/>

<https://www.geeksforgeeks.org/different-ways-to-clear-a-list-in-python/>