**Pràctica 1 (35% nota final)**

# Descripció de la Pràctica a realitzar

L'objectiu d'aquesta activitat serà la creació d'un dataset a partir de les dades contingudes en una web. Per a la seva realització, s'han de complir els següents punts:

1. Context. Explicar en quin context s'ha recol·lectat la informació. Explicar per què el lloc web triat proporciona aquesta informació.

Hem pensant en la necessitat que tenim els que ens agrada anar a fer passejades per la natura d’ aconseguir informació de rutes fetes per altres persones anónimes.

Generar aquest dataset ens donarà la possibilitat de poder consultar una gran quantitat de rutes sense haver d’ anar navegant d’ una a una altre.

L’ objectiu d’ aquest site es poder compartir rutes a l’ aire lliure entre els diferents usuaris tant si están enregistrats com si no. Els usuaris enregistrats fan rutes i posteriorment les penjen al site, d’ aquesta manera la resta pot veure el tipus de ruta i decidir si li agradaría fer-la o no.

1. Definir un títol pel dataset. Triar un títol que sigui descriptiu.

El títol que hem escollit pel dataset és: Trail & Track DataSet.

1. Descripció del dataset. Desenvolupar una descripció breu del conjunt de dades que s'ha extret (és necessari que aquesta descripció tingui sentit amb el títol triat).

El dataset resultant contindrà informació de totes les rutes que han estat valorades per algún altre usuari, en funció d’ un país, regió1 i regió2 específica, o bé en funció d’ una activitat i una regió.

Les dades que hi podrem trobar són: tipus, puntuació de la ruta, puntuació del tracker, distancia, alçada, desnivell, circular, votació dels usuaris, etc…

1. Representació gràfica. Presentar una imatge o esquema que identifiqui el dataset visualment



1. Contingut. Explicar els camps que inclou el dataset, el període de temps de les dades i com s'ha recollit.

El dataset inclou els següents camps:

* TypeTrack: tipus de ruta
* TrailRank: ranking de la pista
* UserRank: ranking de l’ usuari que ha pujat la informació.
* Distància: longitud de la ruta en km.
* Loop: indica Si la ruta és circular o No.
* Desnivell acumulat pujant en metres.
* Altitud màxima en metres.
* Desnivell acumulat baixant en metres.
* Altitud mínima en metres.
* Dificultat Tècnica : Dificil, Moderat o Fàcil.
* Data en la que es va fer.
* Valoració sobre 5 punts.
* Vots (quantitat)

SI trobem o NO els següents elements a la ruta:

* + Lake: SI/NO.
  + Fountain: SI/NO.
  + Panoramic: SI/NO.
  + River: SI/NO.
  + Parquing: SI/NO.
  + Bridge: SI/NO
  + WaterFall: SI/NO.
  + Refuge: SI/NO.
  + Shelter: SI/NO.
  + Url: l’ adreça de la página on trobem la informació.

En quant a la durada de les dades hem de dir que aquestes poden arribar a estar emmagatzemades durants uns quants anys, ja que ens hem trobat amb rutes de l’ any 2006. De totes maneres l’ antiguitat de la informació en el nostre cas no és gaire important degut a que possiblement el contingut i l’ explicació de la ruta poc variarà encara que passin uns quants anys.

Aquest site està estructurat en tres nivells tal i com s’ observa en el path de l’ url quan volem arribar a consultar una ruta, per exemple una ruta que es trobi a Andorra, Canillo, el Tarter tindrà un path com <https://ca.wikiloc.com/rutes/outdoor/andorra/canillo/el-tarter> on podem observar que les diferents parts del path indiquen les zones geogràfiques que acotem. Un cop ja estem situats a la zona desitjada ja podem escollir la ruta i pasar a obtener la informació sobre la mateixa. Es un cop ja dins d’ aquest últim nivell quan començarem a fer Web Scrapping i a obtenir totes les dades que s’ han comentat amb anterioritat.

1. Agraïments. Presentar el propietari del conjunt de dades. És necessari incloure cites de recerca o anàlisis anteriors (si n'hi ha).
2. Inspiració. Explicar per què és interessant aquest conjunt de dades i quines preguntes es pretenen respondre.

A part de l’ utilitat funcional que ja he comentat prèviament també li podem trobar una part d’ interés de cara a l’ estudi, per exemple podríes arribar a saber:

* Quin tipus de rutes són les més valorades
* Si les rutes que pujen els usuaris són circulars normalment
* Quines són les zones amb més rutes
* Quines són les zones que tenen rutes específiques en una modalitat
* Quins són els usuaris més fiables

1. Llicència. Seleccionar una d'aquestes llicències pel dataset resultant i explicar el motiu de la seva selecció:

○ Released Under CC0: Public Domain License

○ Released Under CC BY-NC-SA 4.0 License

○ Released Under CC BY-SA 4.0 License

○ Database released under Open Database License, individual contents under Database Contents License

○ Other (specified above)

○ Unknown License

1. Codi. Adjuntar el codi amb el qual s'ha generat el dataset, preferiblement en Python o, alternativament, en R.

Aprofitem aquesta secció per explicar una mica tot l’ estudi previ que hem fet del site i com hem arribat a extreure informació de les url’s desitjades així com informació propia del site.

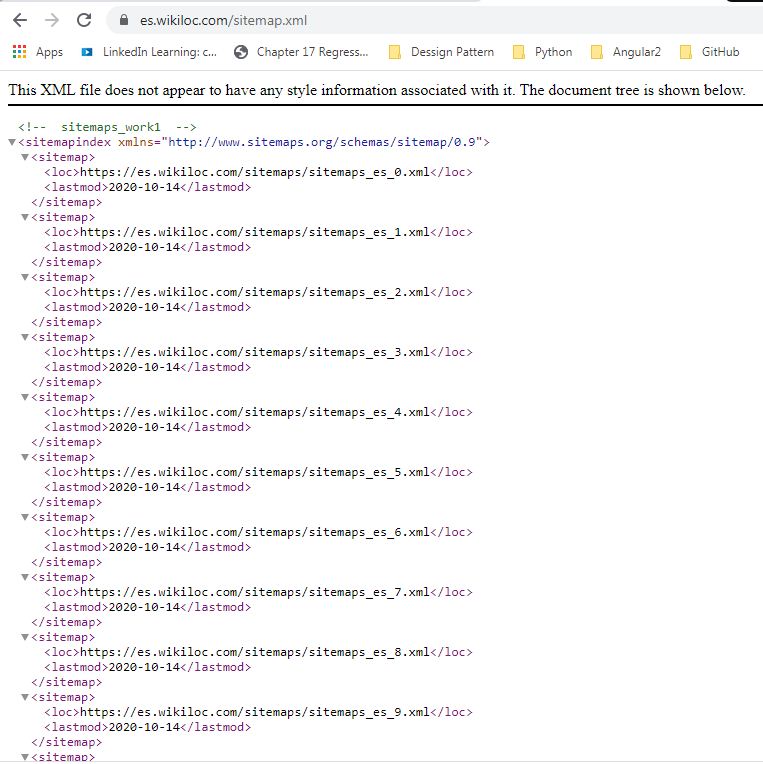
**Estructura del Site**

En aquest portal podem trovar 2 tipus de fitxers que ens poden ajudar en gran messura a desenvolupar una actuació de Web Scrapping el màxim d’ ética possible, per una parte tenim el fitxer [robots.txt] , del qual mostrem el seu contingut:



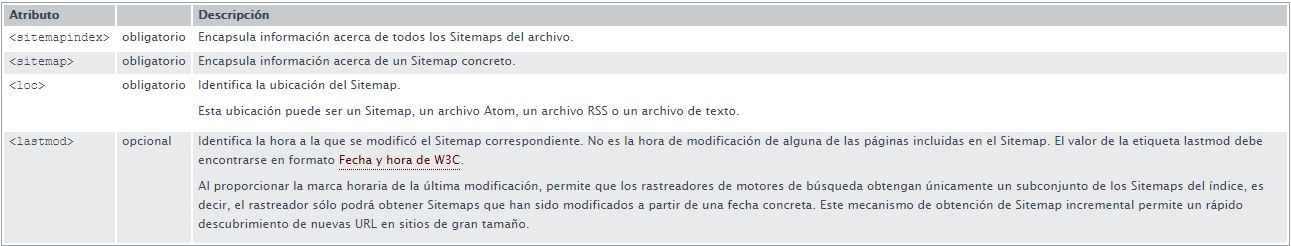
Que observem en aquest fitxer?

* Totes les rules són aplicables a qualsevol “user.agent”
* Trobem directives “allow” i “disallow” les quals tenen significats diferents i a més tenen interpretacions diferents:
  + Standard Implementation: la 1º directiva que coincideixi amb la url per la qual es pregunta s’ aplicarà.
  + Google Implementation: si el patró indicat en la directiva “allow” té una longitud igual o superior a la de la directiva “disallow” llavors s’ aplicarà la directiva “allow”.
  + Bing Implementation: si la url preguntada coincideix amb més d’ una directiva, independentement de que sigui “allow” o “dissallow” s’ aplicarà la que tingui una longitud major.
* Es recomana no accedir a algunes pàgines amb extensió “.do”
* Es recomana no accedir a url’s que comencin amb “/enterwikiloc/worlmap”
* Es recomana no accedir a url’s que acabin amb “\_full.jpg”
* Es recomana no accedir a la carpeta “/cdn-cgi/”
* A la última línea es troba una referència al fitxer SiteMaps, del qual mostrem una part del seu contingut.



Aquest fitxer no és exactament un fitxer que segueixi el format del protocol SiteMap ja que no conté algun dels tags necessaris especificats en aquest protocol, com poden ser <urlset> o <url>.

En la següent imatge es pot observar el format d’ aquest tipus de fitxers:



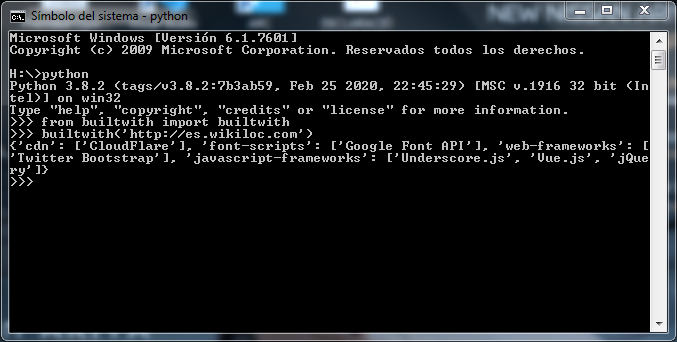
En el seu lloc ens trobem amb un arxiu índex de SiteMap, que ens permet agrupar varis fitxers SiteMap. Aquest tipus de fitxers als que es fa referencia dins de l’ arxiu índex de SiteMap han de cumplir unes especificacions:

* El nombre màxim d’ urls no ha de superar les 50.000.
* Els fitxers no han de superar un size de 50 Mb

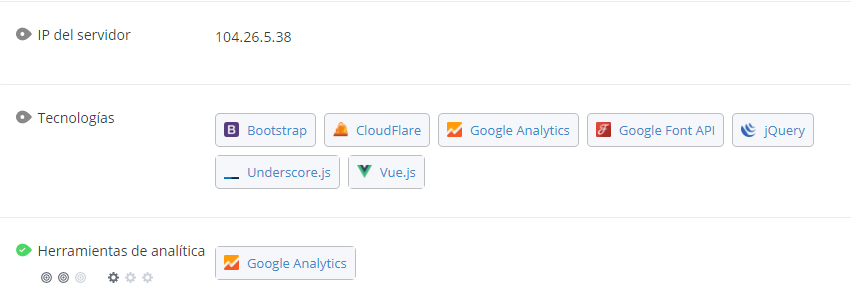
En aquest cas particular el fitxer índex de SiteMap fa referencia a un total de 160 fitxers SiteMap, amb aquest quantitat d’ informació ja podem veure que es tracta d’ un site amb moltísimes urls i que per tant hem de ser molt cuidadosos a l’ hora de realizar accions.

Tecnologia utilitzada en el Site

Per aconseguir aquesta informació utilitzarem el package [builtwith].

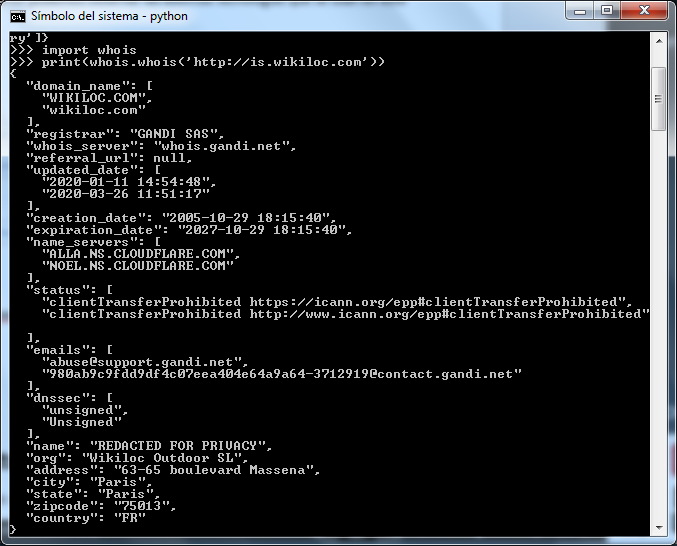


Un cop executat, a l’ imatge superior podem observar les diferents tecnologies que s’ utilitzen en aquest Site.



Propietari del Site

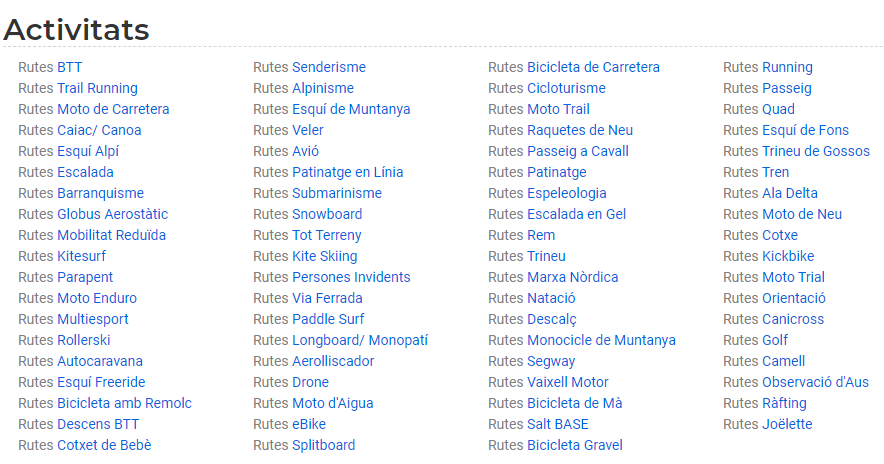
Per aconseguir informació del propietari del site utilitzarem el package [whois], el qual ens tornarà la següent informació:



**Estudi de l’ Estructura del Site**

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

Si accedim a l’url **“https://ca.wikiloc.com/rutes”** se’ns renderitza una pàgina HTML que ens agrupa la informació en 2 grups ben diferenciats, tenim les “Activitats” i els “Paisos”.



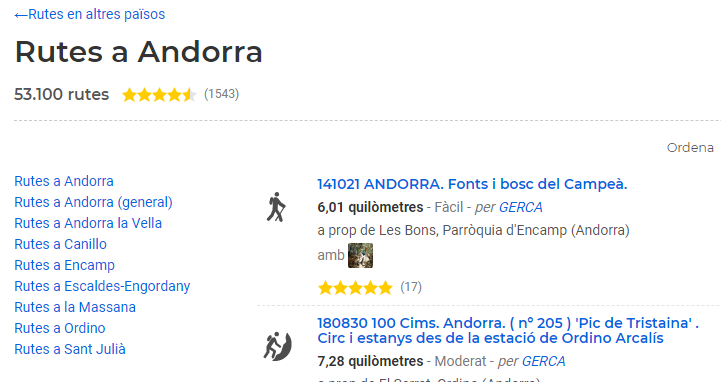


Si procedim a observar l’ estructura d’ aquesta página observem l’ ús de:

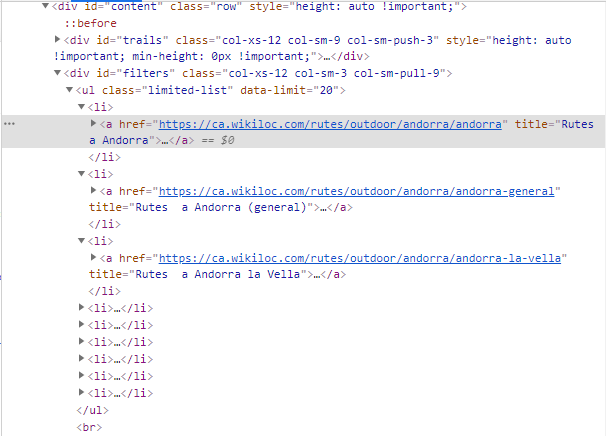
* <ul>: mostra una llista sense ensenyar l’ ordinal de la llista.
* <li>: mostra cadascun dels ítems de la llista.
* <a>: link a l’ url indicada.



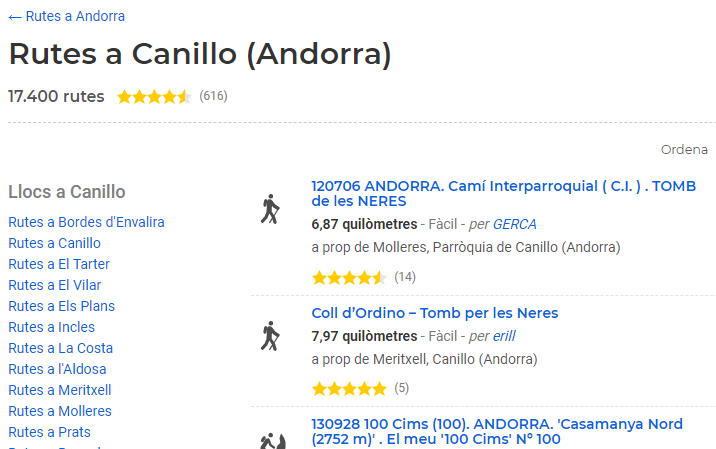
Si realitzem un click a sobre “Andorra” se’ns dirigeix a una altre url “**ca.wikiloc.com/rutes/outdoor/andorra**” on observem:



Si obtenim l’ estructura de la página, observem que segueix el mateix format que anteriorment



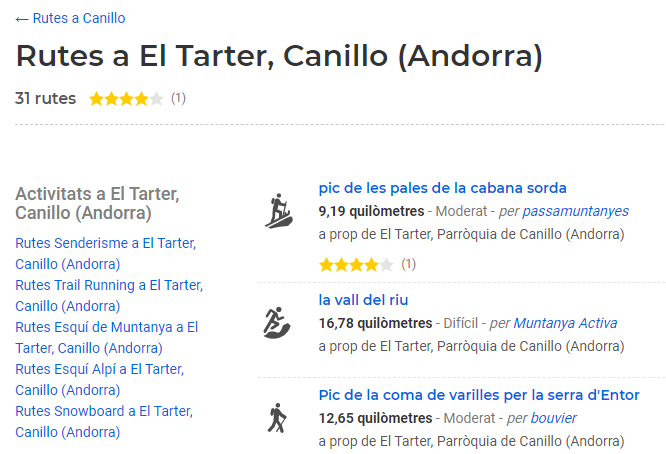
Realitzem un click sobre “Rutes en Canillo” i se’ns dirigeix a l’ url “**ca.wikiloc.com/rutes/outdoor/andorra/canillo**”



Tornem a observar l’ estructura i segueix el mateix patró que les vistes anteriorment.



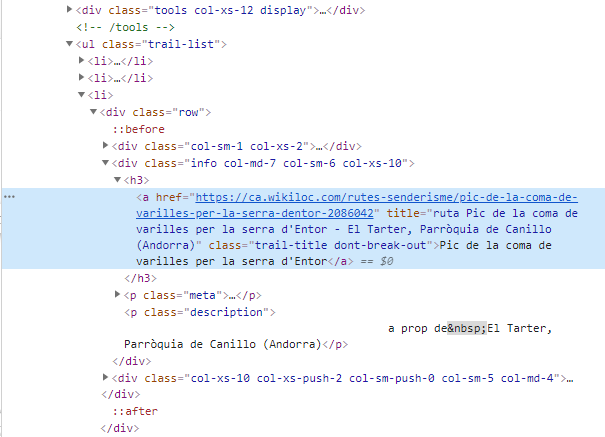
Realitzem un click sobre “Rutes a el Tarter”, que ens portarà a l’ url “ca.wikiloc.com/rutes/outdoor/andorra/canillo/el-tarter”



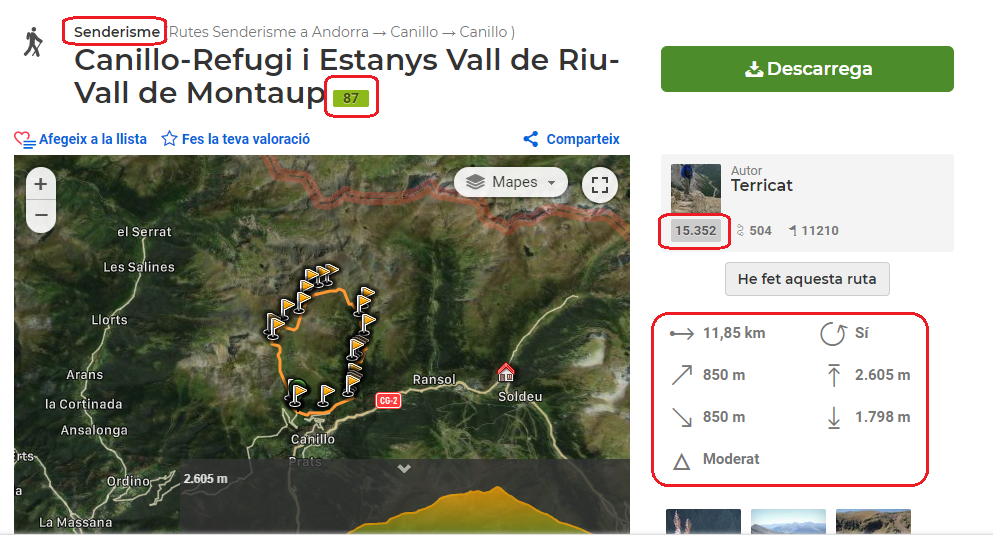
On ja ens apareixeran totes les rutes i el tipus de ruta que podem realizar a la zona seleccionada.

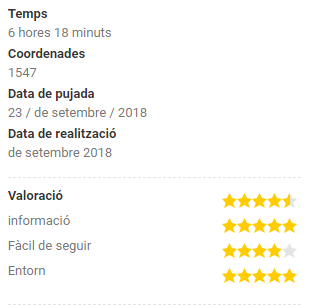
L’ estructura d’ aquesta página és diferent a les anteriors:

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



I ara si que ja nomès ens queda seleccionar la ruta destijada.





On ja podrem obtenir la informació especifica de la ruta:

* Tipus: Senderisme
* Distància: 11,85 km
* Desnivell acumulat pujant: 850 m
* Desnivell acumulat baixant: 850 m
* Dificultat Tècnica: moderat
* Ruta circular: si
* Altitud màxima: 2605 m
* Altitud mínima: 1798 m

Informació adicional:

* Valoració: de 1 a 5 estrellas
* Vots: 2 vots
* Data realització: Setembre 2018
* Característiques de la ruta, indicant si existeix algún d’ aquests elements:
  + Llac
  + Font
  + Riu
  + Panoramica
  + Parquing
  + Pont
  + Catarata
  + Refugi
  + Refugi Lliure

Un cop vist l’ estudi previ ara el que farem será analitzar breument el contingut del fitxer .log que genera la nostre aplicació.



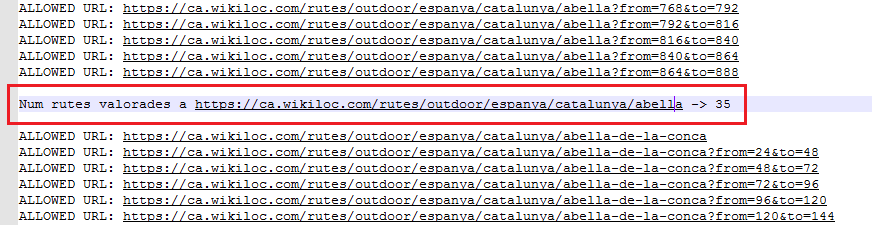
Les primeres línies ens mostren els paràmetres que volem utilitzar a l’ aplicación, en aquest cas volem obtener totes les rutes del país [Espanya] i regió1 [Catalunya].

A continuació ens mostra que l’url resultant està permesa segons les directives del fitxer [Robots.txt] del site i tot seguit totes les urls descendents de la 1ª.

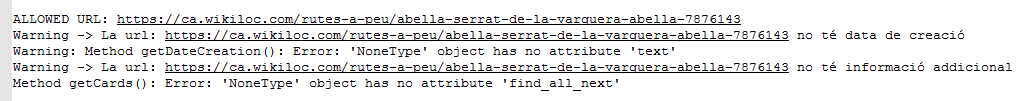


En funció de les rutes que hi hagin emmagatzemades en cada url haurem de continuar fent “gets” d’ aquesta zona o no. Cada “get” ens retorna un total de 24 rutes. Hi han moltes url’s que arriben a tenir fins a 10000 rutes per visitar.

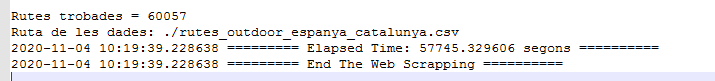
Cada “allowed” associat a una url ens indica que aquesta està permessa segons les directives del fitxer “Robots.txt”.



Un cop hem visitat totes les rutes disponibles en aquella url indiquem les rutes valorades que hi hem trobat.



Com ja hem comentat anteriorment hi ha un conjunt de dades que no sempre existeixen al DOM de les url’s analitzades, quan ens trobem amb aquest casos treuren missatges informatius com es pot veure a la imatge anterior.



Un cop finalitzi ens indicarà quantes rutes ha analitzat, el nom del fitxer on emmagatzemarà les dades i el temps que ha dedicat per obtenir tota la informació, que en aquest cas s’ aproxima a les 16h de treball.

1. Dataset. Publicar el dataset en format CSV a Zenodo (obtenció del DOI) amb una breu descripció.

# Format i data de lliurament

Durant la setmana del 26 al 30 d’octubre, el grup podrà lliurar al professor un lliurament parcial opcional. Aquest lliurament parcial és molt recomanable per rebre assessorament sobre la pràctica i verificar que la direcció presa és la correcta. Es lliuraran comentaris als estudiants que hagin efectuat el lliurament parcial però no comptaran per a la nota de la pràctica. En el lliurament parcial els estudiants hauran de lliurar per correu electrònic, al profesor encarregat de l’aula, l'enllaç al repositori Github amb el que hagin avançat.

En referent al lliurament final, cal lliurar un únic fitxer que contingui l'enllaç a Github on hi hagi:

1. Una Wiki on estiguin els noms dels components del grup i una descripció dels fitxers.
2. Un document PDF amb les respostes a les preguntes i els noms dels components del grup. A més, al final del document, ha d'aparèixer la següent taula de contribucions al treball, la qual ha de signar cada integrant del grup amb les seves inicials. Les inicials representen la confirmació per part del grup que l'integrant ha participat en aquest apartat. Tots els integrants han de participar a cada apartat, per la qual cosa, idealment, els apartats haurien d'estar signats per tots els integrants.

|  |  |
| --- | --- |
| **Contribucions** | **Signa** |
| Recerca prèvia | Integrant 1, Integrant 2, ... |
| Redacció de les respostes | Integrant 1, Integrant 2, ... |
| Desenvolupament codi | Integrant 1, Integrant 2, ... |

1. Una carpeta amb el codi Python o R generat per obtenir les dades.
2. El DOI a les dades.

Aquest document del lliurament final s'ha de lliurar a l'espai de Lliurament i Registre d'AC de l'aula abans de les **23:59 del dia 9 de novemebre**. No s'acceptaran lliuraments fora de termini.