



TIPOLOGIA I CICLE DE VIDA DE LES DADES

PRÀCTICA 1

L'objectiu d'aquesta activitat serà la creació d'un dataset a partir de les dades contingudes a wikiloc.

Aitor Hernández i Anna Mayoral

Contenido

1.	Context.	2
2.	Títol.	2
3.	Descripció del dataset.	2
4.	Representació gràfica.	3
5.	Contingut.	3
6.	Agraïments.	3
7.	Inspiració.	5
8.	Llicència.	5
9.	Codi.	5
10.	Dataset.	5
11.	Vídeo.	5
12.	Annex: Valoració inicial.	6

1. Context.

Per a la realització de la pràctica vam estar analitzant diferents temes i pàgines web on fer la pràctica de Web Scrapping. Finalment ens vam decantar tots dos pel senderisme i van triar la pàgina <https://ca.wikiloc.com/>, que et permet crear manualment la teva ruta o track, així com compartir rutes o lloc d'interès. Actualment ja son més de 9M de membres explorant i compartint 32M de rutes i més de 56M fotos a l'aire lliure. Per ser més específics, la pràctica és centra en concret en rutes per Catalunya:

<https://ca.wikiloc.com/rutes/senderisme/espanya/catalunya>

L'objectiu de la pràctica és aconseguir les dades de les millors rutes de senderisme de Catalunya.

2. Títol.


Les millors rutes de senderisme a Catalunya

3. Descripció del dataset.

El dataset obtingut correspon a les millors rutes de senderisme a Catalunya definits per el usuaris de Wikiloc. Les dades que obtindrem serà el nom de la ruta, la localització, la distància, el desnivell i la puntuació de la ruta. A més obtindrem els links de les imatges que els usuaris han compartit.

També volem puntualitzar que la descripció, hem decidit no extraure-la ja que no li veiem gaire utilitat d'anàlisis.



A continuació mostrem un exemple del que extraurem:




 Senderisme

Vallter - Pic de la Dona - Pic de Prat de Bacivers - Pic de Bastiments - Vallter

a prop de Setcases, Catalunya (Espanya)

Distància	Desnivell +	TrailRank
12,01km	937m	99 ★ 4.85

 **JEP BOIX**
amb 



Després de sis anys i amb el record del bon regust, he tornat a fer aquesta magnífica ruta:
<http://ca.wikiloc.com/wikiloc/view.do?id=443945> en companyia del meu germà i la meva neboda...

★★★★★
«Moltes gràcies per compartir la ruta.
La vam fer la setmana passada, però...»
91lauramarti

★★★★★
«Ruta fantàstica completada aquest
estiu passat. Indispensable!»
PedraToska

★★★★★
«Com sempre molt ben explicada.
Aquesta vegada, però, no ens va caldr...»
mussola

[Veure ruta →](#)

4. Representació gràfica.



5. Contingut.

El conjunt de dades conté:

- **Nom de la ruta:** Títol de la ruta.
- **Tipus d'activitat:** Tipologia de l'activitat.
- **Punt d'inici:** Localització d'inici de la ruta.
- **Distància:** Longitud total de recorregut en kilòmetres.
- **Desnivell:** Diferència de nivell entre el punt de sortida i d'arribada en metres.
- **Ranking1:** Valoració de la ruta. Puntuació sobre 100. Aquesta puntuació la calcula directament Wikiloc en funció si la ruta té molts seguidors, fotografies i valoracions positives.
- **Ranking2:** Valoració de la ruta. Puntuació sobre 5. Aquesta valoració és calculada directament amb les valoracions dels usuaris de Wikiloc.
- **ImageURL1:** URL1 on es guarda la primera imatge.
- **ImageURL2:** URL2 on es guarda la segona imatge.
- **ImageURL3:** URL3 on es guarda la tercera imatge.

6. Agraïments.

En primer lloc volem agrair a Jordi Ramot el fundador de wikiloc per la realització d'aquesta web (i també APP) ja que fomenta no només la compartició d'informació sinó també promou un estil de vida saludable amb rutes de tot nivell, dificultat i amb activitats per fer amb família o amics de forma totalment segura.

També volem comentar que hem fet una recerca en Kaggle i no hem trobat cap dataset provinent del Wikiloc.

Datasets

[+ New Dataset](#)[Filters](#)[Computer Science](#)[Education](#)[Classification](#)[Computer Vision](#)[NLP](#)[Data Visualization](#)[Pre-Trained Model](#)[Datasets](#)

Hotness ▾



No Datasets Found

We couldn't find any relevant datasets on Kaggle.

[Try Google Dataset Search](#)

En la web de Zenodo, l'únic que hem trobat ha estat un dataset però d'imatges que provenen del Wikiloc, res a veure amb el nostre objectiu de la pràctica.



June 21, 2021

[Dataset](#)[Open Access](#)

Social media images from Peneda-Gerês National Park, Northern Portugal

[Cardoso, Ana Sofia](#); [Moreno-Llorca, Ricardo](#); [Alcaraz-Segura, Domingo](#); [Vaz, Ana Sofia](#)

This dataset contains images from the Peneda-Gerês National Park, Northern Portugal. The images were collected from the Flickr and **Wikiloc** platforms considering a time period from 2003 to 2017. In respect to the General Data Protection Regulation 2016/679, social media data protected by users' rights was not downloaded nor analysed. Public data that would potentially contain personal information from social media users was kept anonymous through the study. Data was retrieved through the use of the freely available Flickr's Application Programming Interface (API), indicating a time window and a bounding box with a pair of coordinates (in our case: minimum latitude: 41.653104; maximum lat.: 42.083595; min. longitude: -8.426270; max. lon.: -7.754076) around Peneda-Gerês. This information was then saved as an excel file with the following attributes: user-id, date taken, latitude, longitude, picture uniform resource locator (url).

A first annotation, in the context of cultural ecosystem services (CES), was performed by dividing the photographs of the dataset into "Indoor" and "Outdoor" classes. Only the "Outdoor" pictures were included in this study, since CES are directly connected to nature and environment, which in turn are related to the outside/outdoor. The "Outdoor" images were further divided into two main classes, "Natural" and "Human", depending on whether the image was dominated by natural or man-made elements. Lastly, a finer annotation for outdoor images was also provided, which encompasses the following six classes: "Species", "Landscape", "Nature", "Human activities", "Human structures" and "Posing". "Species" pictures respectively pertained to close-up shots of animals or plants in the wild, translating CES pertaining to biodiversity appreciation. "Landscape" pictures show wide-open shots of nature in the wild, often with a visible horizon most often representing people's enjoyment of landscape aesthetics. "Human activities" include pictures where people engage in by recreational activities, for instance related to sports such as ski or cycling. "Human structures" include those pictures where man-made structures dominate in the wild, e.g., historical monuments and churches, capturing situations of cultural heritage and spiritual enrichment. "Posing" refers to pictures with people looking at the camera, with recognizable faces, testifying social enjoyment and sense of identity. Finally, "Nature" pictures capture natural elements with no particular feature (such as species) but with an intermediate shot (differing from wide-open shots attributed to landscapes), expressing the appreciation of nature by people.

Adicionalment també hem buscat en els repositoris públic del Github i en un primer moment no vam trobar cap repositori que fes el que nosaltres hem plantejat. Si que és cert que a posteriori hem vist que hi ha una activitat de web Scraping sobre Wikiloc però totalment diferent a la que hem fet nosaltres: <https://github.com/polmoya/rutesWikiloc>

7. Inspiració.

Aquest conjunt de dades és interessant per totes aquelles persones que els hi agradi la naturalesa i fer rutes per Catalunya. Abans d'anar a la muntanya és important conèixer quin desnivell, distància i dificultat té una ruta per anar ben preparats. Cada cop és més important gaudir de la natura però de forma segura. A més, aquest data set et dona recomanacions de les rutes més valorades i et permet conèixer Catalunya més profundament. Com hem comentat abans no hem trobat cap projecte similar al nostre i per tant creiem que és molt interessant i necessari.

En l'àmbit de la ciència de dades, el data set permet extraure perfils de rutes basats en mètriques de distància i desnivell, i comprar-ho amb les puntuacions dels usuaris. Geogràficament, permetria també extreure conclusions a nivells de qualitat / duresa de les rutes en funció dels municipis o comarques.

8. Llicència.

La llicència escollida per a aquest projecte ha estat la llicència CC BY-SA 4. Els motius per fer-ho son:

1. Es permet la lliure distribució o còpia del material que es proveeix per qualsevol motiu, inclosos motius comercials.
2. S'ha d'acreditar el nom del creador de les dades, indicant si s'han fet canvis al conjunt de dades.
3. Si es modifica o es transformen les dades del dataset, s'han de distribuir les modificacions sota la mateixa llicència plantejada.

9. Codi.

El repositori Git de la pràctica és: https://github.com/aitorhdez/Tipologia_PRA1

10. Dataset.

El data set es pot trobar a Zenodo a la url següent:

<https://zenodo.org/record/6421086#.Yk7GHsjP2Uk>

11. Vídeo.

12. Annex: Valoració inicial.

També volem fer una ràpida menció a la valoració inicial de la pàgina web amb la que hem treballat el Web Scrapping:

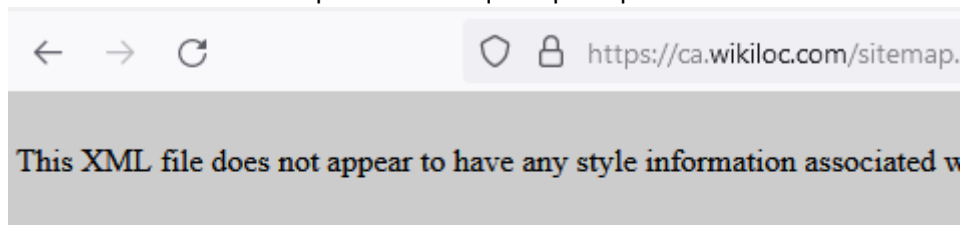
- 1) **L'arxiu robots.txt:** Els propietaris de les pàgines web utilitzen el fitxer /robots.txt per donar instruccions sobre el seu lloc web.

En el nostre cas observem, com es permet que tots els robots rastregin tot el web però hi han pàgines o directoris puntuals que no es permet l'accés a un bot (*_full.jpg\$, .do etc)



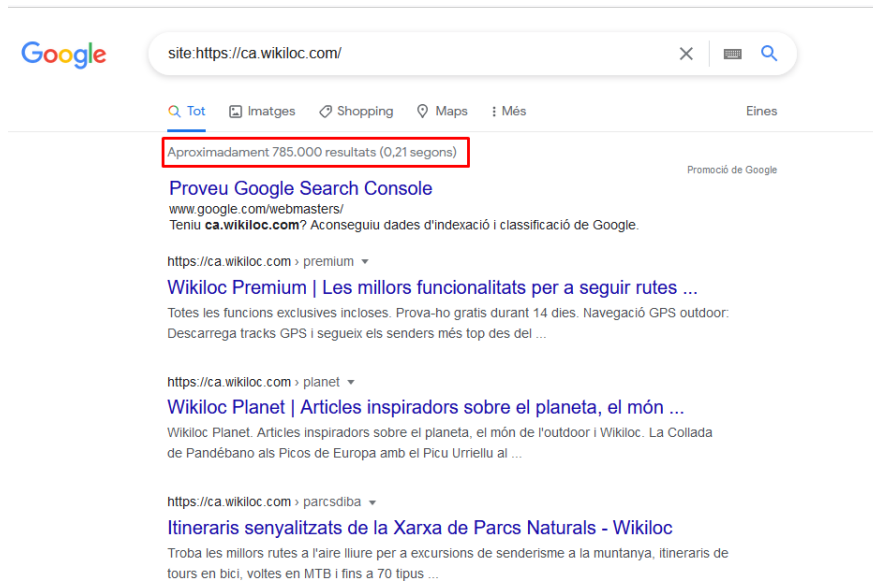
```
User-Agent: *
Allow: /
Disallow: /*_full.jpg$
Disallow: /wikiloc/companionRequest.do
Disallow: /wikiloc/tr.do
Disallow: /wikiloc/map.do
Disallow: /wikiloc/geocode.do
Disallow: /cdn-cgi/
Sitemap: https://ca.wikiloc.com/sitemap.xml
```

- 2) **El mapa del lloc web:** Observem com el mapa web no comença amb etiqueta d'obertura <urlset> i ni acaba amb una etiqueta de tancament </urlset>. El que si que inclou és una entrada secundària <loc> per a cada etiqueta principal.

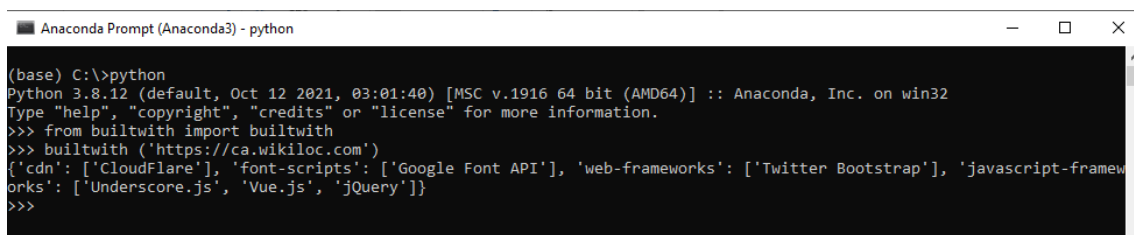


```
<!--<sitemapindex>
  <!--<sitemap>
    <loc>https://ca.wikiloc.com/planet/sitemap.xml</loc>
  <!--</sitemap>
  <!--<sitemap>
    <loc>https://ca.wikiloc.com/sitemaps/sitemaps_ca_0.xml</loc>
  <!--</sitemap>
  <!--<sitemap>
    <loc>https://ca.wikiloc.com/sitemaps/sitemaps_ca_1.xml</loc>
  <!--</sitemap>
  <!--<sitemap>
    <loc>https://ca.wikiloc.com/sitemaps/sitemaps_ca_2.xml</loc>
  <!--</sitemap>
  <!--<sitemap>
    <loc>https://ca.wikiloc.com/sitemaps/sitemaps_ca_3.xml</loc>
  <!--</sitemap>
  <!--</sitemapindex>
```

3) La seva grandària: Podem observar com Google estima 785k pàgines



4) La tecnologia emprada: A continuació podem observar les tecnologies utilitzades:



5) El propietari del lloc web: Finalment observem el propietari:

```
>>> import whois
>>> print(whois.whois('https://ca.wikiloc.com'))
...
{
  "domain_name": [
    "WIKILOC.COM",
    "wikiloc.com"
  ],
  "registrar": "GANDI SAS",
  "whois_server": "whois.gandi.net",
  "referral_url": null,
  "updated_date": "2022-01-08 15:01:26",
  "creation_date": [
    "2005-10-29 18:15:40",
    "2005-10-29 16:15:40"
  ],
  "expiration_date": "2027-10-29 18:15:40",
  "name_servers": [
    "ALLA.NS.CLOUDFLARE.COM",
    "NOEL.NS.CLOUDFLARE.COM"
  ],
  "status": [
    "clientTransferProhibited https://icann.org/epp#clientTransferProhibited",
    "clientTransferProhibited http://www.icann.org/epp#clientTransferProhibited"
  ],
  "emails": [
    "abuse@support.gandi.net",
    "980ab9c9fdd9df4c07eea404e64a9a64-3712919@contact.gandi.net"
  ],
  "dnssec": [
    "unsigned",
    "Unsigned"
  ],
  "name": "REDACTED FOR PRIVACY",
  "org": "Wikiloc Outdoor SL",
  "address": "63-65 boulevard Massena",
  "city": "Paris",
  "state": "Paris",
  "zipcode": "75013",
  "country": "FR"
}
```