Dataset: Refugis de muntanya i llocs d'acampada a Espanya, Andorra i sud de França

Tipologia i cicle de vida de les dades
Novembre 2022

Joan Peracaula Prat
Sara Jose Roig

1. Context

En aquesta pràctica hem escollit crear un dataset de llocs on dormir a la muntanya, siguin refugis (lliures o guardats) o bé espais d'acampada habilitats, en les regions d'Espanya, Andorra i sud de França.

El lloc web d'on s'ha extret aquesta informació és https://www.walkaholic.me/. Aquesta es tracta d'una plataforma encarada als amants del senderisme, amb rutes i allotjaments (refugis i llocs d'acampada) a Espanya, Andorra i França.

Més concretament, hem extret la informació de les següents landings:

- Refugis (https://www.walkaholic.me/shelter)
- Llocs d'acampada (https://www.walkaholic.me/campsite)

Tal com es pot comprovar, aquestes landings estan formades per una llista extensa d'enllaços, on cada un redirigeix a una pàgina amb dades de l'allotjament en qüestió.

2. Títol

Un títol descriptiu per aquest dataset seria: "*Refugis de muntanya i llocs d'acampada a Espanya, Andorra i al sud de França*".

En anglès: "Mountain shelters and campsites in Spain, Andorra and South of France".

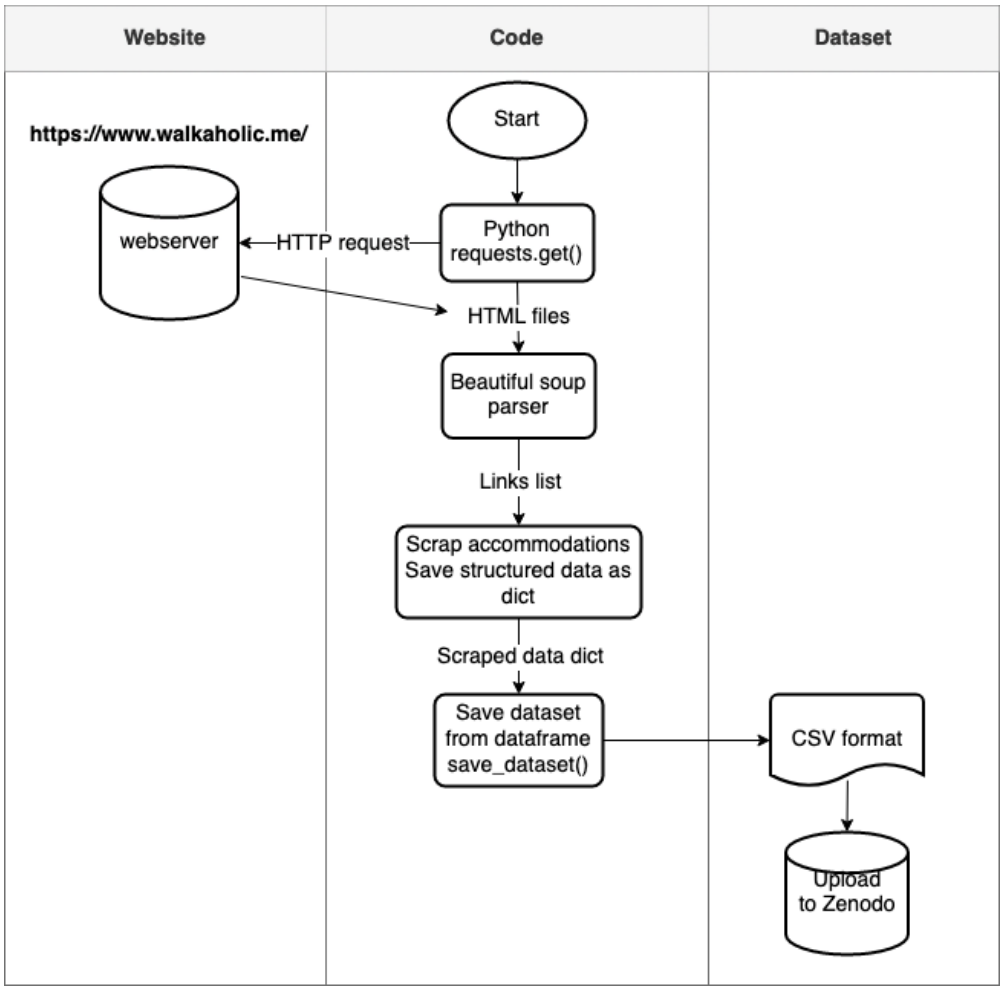
3. Descripció del dataset

Tal com s'ha comentat anteriorment, aquest dataset consisteix en un fitxer de tipus CSV amb registres de llocs on dormir a la muntanya en les zones d'Espanya, Andorra i sud de França. Hi ha dos tipus principals de registres: refugis i llocs d'acampada.

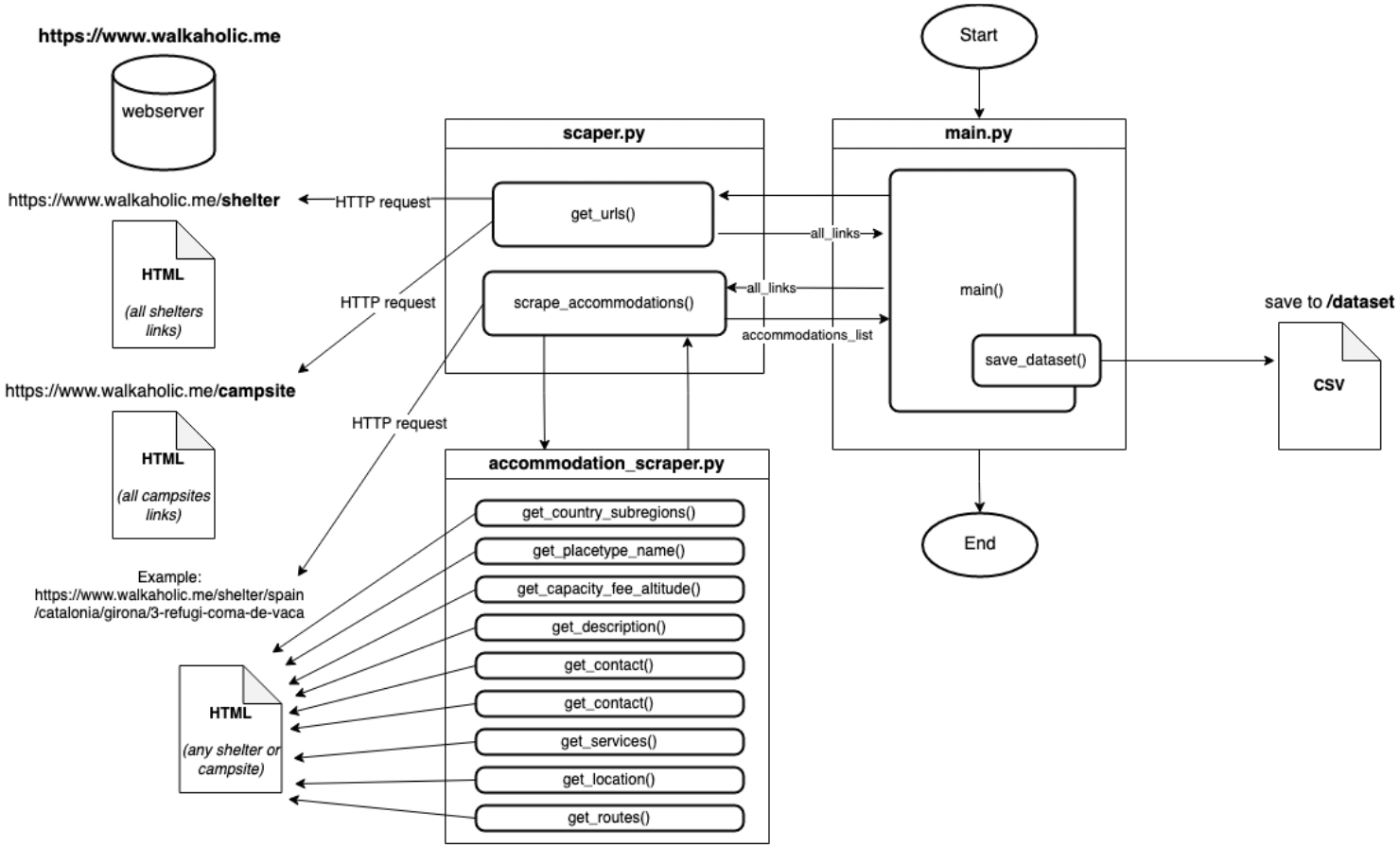
En general, les dades tenen característiques comunes que són presents en qualsevol registre, com ara el tipus, el nom i la localització. Tot i això, hi ha una varietat d'atributs diferents què podem ser coneguts o no, segons la informació disponible de cada allotjament real. Per exemple, hi ha refugis o llocs d'acampada que tenen informació detallada respecte a les formes de contacte o serveis disponibles, però n'hi ha d'altres que no en tenen o es desconeix. Aquestes variables desconegudes (per la no presència), estan etiquetades amb el caràcter '?'.

4. Representació gràfica

A la figura que hi ha a continuació es representa el procés general de les dades des del *web scraping* fins que són guardades a un fitxer CSV, mostrant les tecnologies principals utilitzades (Python Requests, BeautifulSoup i pandas):



Per últim, es mostra un diagrama més detallat del codi i l'estructura del projecte:



5. Contingut

El dataset inclou els següents atributs:

Atribut	**Descripció**
Place type	Especifica quin tipus de lloc es tracta, refugi o espai d'acampada.
Name	Nom de l'allotjament.
Place list	Regió geogràfica, en forma de llista de més a menys àrea. Segons el país, canvia l'estructura de la llista: - Espanya: país, comunitat autònoma, província - França: país, regió, departament - Andorra: país, parròquia

Capacity	Nombre de llits totals. (*Pot ser desconegut*)
Fee	Indica si és gratuït o de pagament.
Altitude	Altitud en metres sobre el nivell del mar en què es troba l'allotjament. (*Pot ser desconegut*)
Description	Descripció variada de l'allotjament en format text. (*Pot ser buida*)
Telephone	Telèfon de contacte. (*Pot ser desconegut o inexistent*)
Website	Pàgina web de l'allotjament. (*Pot ser desconeguda o inexistent*)
Email	Correu electrònic de contacte. (*Pot ser desconegut o inexistent*)
Hiking association	Associació senderista a càrrec. (*Pot ser desconeguda o inexistent*)
Guard name(s)	Nom del guàrdia/ guàrdies. (*Pot ser desconegut o inexistent*)
Services	Serveis que ofereix l'allotjament en forma de llista. Inclouen: bany, calefacció, dutxa, internet, llar de foc, mantes, menjar i begudes, ràdio d'emergència… (*Poden ser desconeguts*)
Coordinates	Coordenades de latitud i longitud de la localització de l'allotjament.
Access	Instruccions de com arribar a l'allotjament, en forma de llista segons el tipus d'accés: a peu, en cotxe… (*Poden ser desconegudes*)
Zones	Zona/es dins la província on es troba l'allotjament, a vegades també inclou la comarca. (*Poden ser desconegudes*)
Emplacement	Instruccions més específiques dintre de la zona per ajudar a situar l'allotjament, en format llista. (*Poden ser desconegudes*)
Nearby routes	Llista de noms de rutes de senderisme que es troben a prop de l'allotjament. (*Poden ser desconegudes*)

Tot i que alguns camps són de caràcter numèric (com ara la capacitat, l'altitud o les coordenades), **tots els camps del dataset estan en format text (*string*) o llista de *strings***, ja que s'ha construït el dataset amb les dades tal com s'han extret, sense netejar-les ni preprocessar-les.

Respecte al període de les dades, **la data d'extracció de les dades del dataset final és el 18/11/2022**.

La pàgina web no indica la data d'addició o modificació de les dades. No obstant això, sí que comuniquen que l'any 2017 i el 2018 van rebre una subvenció amb el finançament de la Generalitat i la Unió Europea. A més a més, totes les pàgines es van indexar per primera

vegada a Google l'any 2018. Per tant, podem assumir que la informació és del 2018 o més recent.

Cal esmentar que aquest tipus de dades, en general, són bastant “estàtiques” en el temps, en el sentit que és complicat que gran part de la informació d’aquestes dades canviï en els anys (sobretot informació més rellevant com el nom, la localització o l’existència en si) per la pròpia naturalesa del tipus d’entitat real que representen. Alguna informació més secundària com algun dels serveis o forma de contacte, sí que pot canviar amb més freqüència. Així doncs, podríem concloure que aquestes dades tindrien un temps de vida d’uns quants anys (entre 5 i 10 per posar un número).

6. Propietari

Les dades d’aquest dataset que presentem han estat extretes exclusivament de la plataforma web **Walkaholic**. És per aquest motiu, que el propietari del conjunt de dades és **Carlos Alberto Martínez Gadea**, creador i titular de la plataforma Walkaholic, tal com es pot veure en la pàgina de Termes Legals.

Abans d’escollir aquesta pàgina web per fer web scraping, n’havíem considerat d’altres, però que finalment vam descartar. A continuació s’esmenten les pàgines web analitzades anteriorment amb un una petita explicació de per què no s’han acabat escollint:

- **Refugios Libres**: web amb informació de refugis lliures, principalment d’Espanya, creada i actualitzada pels usuaris. El principal inconvenient és que es tracta d’una pàgina molt senzilla d’accés sota registre d’usuari i que no té publicats (almenys visiblement) polítiques i termes legals, i, doncs, no ens semblava una bona elecció si no podíem comprovar la prohibició o no d’extracció de dades. Per altra banda, el nombre de dades disponibles és força petit.
- **Entre Montañas**: té una landing amb molts refugis de muntanya a Espanya i Andorra, però hi ha poca informació de cada allotjament i a part de refugis s’inclouen hotels/hostals de muntanya, que creiem que s’escapa una mica de l’objectiu inicial del dataset.
- **EUMA (European Mountaineers):** té una pàgina web amb un mapa que localitza una gran quantitat de refugis de diversos països d’Europa, d’associacions d’excursionisme pertanyents a l’associació EUMA. Els principals inconvenients són la poca informació de cada refugi i la dificultat per scrapejar la informació únicament present en elements del mapa. Com a curiositat, aquesta pàgina té refugis de tot Espanya menys de Catalunya, ja que la FEEC (Federació d'Entitats Excursionistes de Catalunya) no forma part de l’associació.
- **FEEC (Federació d'Entitats Excursionistes de Catalunya):** té una pàgina web amb refugis de Catalunya pertanyents a la federació. Els refugis estan molt ben documentats,

però en són molt pocs. De manera que en sortiria un dataset molt petit en nombre de dades.

Per últim, havíem considerat extraure les dades de diverses d'aquestes pàgines web, però el principal problema era la inconsistència en atributs de les dades segons la pàgina web de provinença. La plataforma **Walkaholic**, finalment escollida, ofereix una gran quantitat de refugis, tant a Espanya (inclosa Catalunya) com a Andorra i el sud de França. A més, conté dades de llocs d'acampada a part de refugis, cosa que permet ampliar el tipus de dades i la quantitat d'aquestes de cara a obtenir un dataset més ric en dades i igual de coherent.

Per seguir els principis ètics i legals del webscraping hem seguit els següents passos. Ens hem assegurat que estàvem complint els termes i condicions de la pàgina utilitzada. El titular de la pàgina és Carlos Alberto Martínez Gadea i les normes d'ús es regeixen segons la legislació espanyola. Hem complert amb els termes i condicions de la pàgina, on diu que les dades poden ser utilitzades per tercers, però que el titular no n'és responsable del seu mal ús.

A més a més, en aquest cas les dades es consideren ús legítim dintre de la legislació legal de l'ús raonable, ja que el projecte està fet en fins acadèmics.

També hem comprovat el fitxer robots.txt de la pàgina, el conté el següent contingut:

```
Sitemap: https://www.walkaholic.me/sitemap.xml
User-agent: *
Disallow: /admin/
```

Segons aquest fitxer, tots els *user-agents* tenen permès fer crawling d'aquesta pàgina web i l'única pàgina deshabilitada és la d'admin. Per tant, estem seguint les recomanacions exposades.

Per seguir els principis de bones pràctiques, hem desenvolupat el codi només rastrejant una petita mostra d'informació de tota la disponible, de forma aleatòria per obtenir varietat d'execucions però sobretot per evitar accessos repetitius i en un mateix ordre, intentant prevenir ser detectats pel servidor. També hem definit uns *headers* amb *user-agents* variats per evitar ser detectats com un script i ser bloquejats.

7. Inspiració

Aquest joc de dades és interessant per diversos motius. En primer lloc, els refugis i camps d'acampada són allotjaments d'interès pels excursionistes, perquè, entre altres motius, molts d'ells són els únics allotjaments disponibles en rutes de muntanya. Per tat, per començar, el tipus de dades és d'interès per l'usuari final, independentment de qui i com utilitzi el dataset.

En la mateixa línia, gràcies que aquestes dades tenen un valor per uns perfils d'usuari concrets, aquest dataset pot servir com a font de dades per a diferents aplicacions, siguin comercials o no. Per exemple, es poden emprar en aplicacions (web o mòbil) del sector de l'excursionisme, per crear cartografia específica, per tenir un control d'allotjaments de muntanya del territori (per exemple, per motius de manteniment o rescat).

Des del punt de vista de la ciència de dades, aquest dataset en si sol semblaria que no té una aplicació molt clara amb els tipus d'algorismes habituals, com classificació o regressió. No obstant això, se'n poden fer anàlisis estadístiques a partir dels atributs i, per altra banda, amb combinació d'altres dades sí que guanya molt d'interès. Per exemple, si en un futur es poguessin recollir dades temporals de persones que passen i s'allotgen en aquests llocs, es podrien aplicar algorismes de predicció de capacitat/demanda dels allotjaments en següents temporades.

Comparant-ho amb les anàlisis anteriors, com que la majoria de llocs web tenien poca informació extra dels allotjaments o molt justa (la capacitat és una de les característiques que no apareix en la majoria), aquest tipus d'anàlisis hipotètiques no serien possibles. Un segon problema de les pàgines web analitzades anteriorment és que moltes d'elles tenien dades de zones molt particulars (només Catalunya, o països d'Europa amb Espanya però sense Catalunya, o bé refugis i hotels només d'Espanya), provocant que aquestes possibles anàlisis no siguin tant d'utilitat per aquestes incoherències o falta de dades representatives d'una zona.

Un últim comentari és que aquest joc de dades és interessant perquè (sembla que) no existeixen jocs de dades similars públics. Aquesta originalitat li dona un valor afegit, donant peu a possibles futurs estudis o anàlisis inexistents fins al moment.

8. Llicència

Per aquest projecte hem seleccionat la llicència MIT, ja que és una llicència *open source* que permet conservar els drets d'autor, però a la vegada és força permissiva. Això vol dir que terceres persones poden modificar i fer ús del projecte, sigui per ús comercial, privat o distribució.

També, a diferència de les llicències *copyleft*, com per exemple, la *GNU (General Public License)* o la *LGP*, els projectes que facin ús del nostre codi o dataset no necessiten portar el mateix tipus de llicència. A més a més, en aquest cas és preferible una llicència MIT en lloc d'una de *Public* ja que aquesta darrera pot presentar àrees legals grises que una llicència permissiva com la MIT soluciona.

Per aquesta raó la llicència MIT es força escaient. Altres llicències que serien adients per aquest tipus de codi serien *CC BY-NC-SA 4.0 License*, la llicència *Apache* o la *BSD,* que tenen nivells similars de permissivitat a l'escollida.

9. Codi

Enllaç al repositori de Github del projecte:
https://github.com/jperacaula/mountain-shelters-campsites-scraping.git

Hem desenvolupat el projecte en Python, utilitzant les següents llibreries principals:

- **requests (v 2.28.1)**: per a realitzar requests HTTP de les pàgines a scrapejar.
- **beautifulsoup4 (v 4.11.1)**: per a parsejar i navegar pels documents HTML.
- **pandas (v 1.5.1)**: per a crear un dataframe amb el conjunt de dades i posteriorment guardar-les en un fitxer CSV.

En el directori **/source** del repositori es pot trobar el fitxer *requirements.txt* (generat mitjançant la comanda *pip freeze*) amb aquestes llibreries i les seves subdependències. És necessari instal·lar aquest conjunt de dependències per tal que el codi s'executi correctament.

El codi està format per tres fitxers python:

- **main.py**: fitxer principal i punt d'entrada al codi. Executa les funcions *get_urls()* i *scrape_accommodations()* del fitxer *scraper.py*. Finalment, s'encarrega de guardar el dataset en un fitxer amb format CSV.

- **scraper.py**: s'encarrega de fer les requests HTTP al servidor web i parsejar els documents HTML. Mitjançant les diverses funcions del fitxer *accommodation_scraper.py*, genera una llista dels allotjaments en format diccionari de python.

- **accommodation_scraper.py**: s'encarrega d'extreure els atributs de cada apartat d'un allotjament a partir del document HTML parsejat com a objecte *soup*.

El flux del codi, doncs, és el següent:

1. Iniciar el programa (*main.py*).
2. Scrapejar les landings de refugis i camps d'acampada, per obtenir una llista de tots els enllaços a allotjaments (*scraper.py*).
3. Reordenar aleatòriament tots els enllaços, amb l'objectiu de no seguir un ordre lògic d'accessos de cara al servidor web (*scraper.py*).
4. Recórrer la llista d'enllaços. Per cada enllaç, extreure'n el contingut HTML i parsejar-lo (*scraper.py*).
5. Scrapejar totes les seccions del document d'informació de l'allotjament, retornant-ne la informació valuosa com a atributs (*accommodation_scraper.py*).
6. Construir un objecte de tipus diccionari amb tots els atributs de l'allotjament (*scraper.py*).
7. Transformar les dades a *dataframe* de *pandas* i guardar-les en un fitxer CSV dins el directori /dataset (*main.py*).

Finalment, pel que fa a les dificultats que presenta el lloc web triat, podem destacar-ne les següents:

En primer lloc, cada entitat de dades (un allotjament: refugi o camp d'acampada), conté la informació en una pàgina diferent. Per tant, per extreure'n la informació calia fer moltes requests HTTP al servidor web. Per tal de no ser bloquejats pel servidor, vam utilitzar les següents estratègies:

- Utilitzar *headers* que simulin un accés des de navegador web.
- Reordenar aleatòriament la llista d'enllaços a accedir en cada execució.
- Desenvolupar el codi realitzant pocs accessos seguits en les execucions de prova.

La segona dificultat trobada té a veure amb la informació disponible de cada allotjament. Vam detectar que hi havia allotjaments amb molta informació i d'altres amb poca informació (amb alguns apartats d'informació que no apareixien en l'estructura HTML del document). Per tal que el codi fos el més robust i modular possible, vam desenvolupar-lo fent que les funcions d'extracció d'informació dels apartats siguin funcionals independents segons el tipus d'allotjament i que controlin l'aparició o no de l'atribut en l'estructura de la pàgina.

10. Dataset

El dataset obtingut de la pràctica es troba en el directori **/dataset** de l'arrel del projecte i està publicat en la plataforma Zenodo.

DOI: 10.5281/zenodo.7338336
Pàgina web: https://zenodo.org/record/7338336#.Y3lNhXaZO3A

11. Vídeo

Link del vídeo explicatiu de la pràctica:
https://drive.google.com/file/d/1sKYpg2N_2lM94gUY2Nd5RZDVJVKTD2uR/view?usp=sharing

12. Contribucions

Contribucions	**Signatura**
Investigació prèvia	SJR, JPP
Redacció de les respostes	SJR, JPP
Desenvolupament del codi	SJR, JPP
Participació al vídeo	SJR, JPP