

Tipologia i cicle de vida de les dades

Pràctica 1

**Web Scraping de les escales de vaixells al Port de Tarragona
durant l'any 2019**

Arnau Giribet March

Oscar Buisan Viniegra

1. Títol

Web Scraping de les escales de vaixells al Port de Tarragona durant l'any 2019.

2. Context

Les dades recollides pel scraper eren necessàries per l'estudi previ a la creació de l'impost d'emissions dels grans vaixells que actualment s'està desenvolupant des de la conselleria d'Acció Climàtica i la conselleria d'Economia de la Generalitat de Catalunya.

3. Descripció del dataset

El dataset conté un registre per a cada escala de cada vaixell realitzada al Port de Tarragona.

Cada escala està caracteritzada pel vaixell en qüestió i les seves propietats, el temps que va durar, les empreses implicades, el moll on es va realitzar i el tipus d'operació o operacions que es van dur a terme.

4. Representació Gràfica

gisweb

estat actual moviments **històrics** diari explotació previsió

VAIXELLS PORT

atracades del vaixell entre 01/01/2019 00:00 i 31/12/2019 23:59

Vaixell

Arribada
01/01/2019 00:00

Sortida
31/12/2019 23:59

VEURE

		operacions								
Escala	Vaixell	Moll	Entrada	Sortida	Consignatari	Eslora	Tipus	Mercaderia	Tn	Estibador
- tarragona										
2.870	NEPTUNE ODYSSEY	25 CANTABRIA (del 0 al 11)	31/12/2019 05:45	31/12/2019 14:53	BERGE MARITIMA, S.L.	170	Desembarque	Vehiculos	500	BERGE MARITIMA SL
2.872	FALESIA	15 QUIMICA 2	31/12/2019 03:30	31/12/2019 19:08	ERSHIP, S.A.	100	Desembarque	Gases energéticos del petróleo	3500	TERQUIMSA
2.836	EMERALD LEADER	24 GALICIA (del 3 al 14)	30/12/2019 23:50	31/12/2019 15:49	COMBALIA AGENCIA MARITIMA S.A.	176	Embarque	Vehículos	800	NOATUM TERMINAL TARRAGONA, S.A.
2.823	LISA ESSBERGER	17 QUIMICA 4	30/12/2019 19:57	31/12/2019 14:11	NEXT MARITIME, S.L.	115	Desembarque	Productos químicos	3000	TERQUIMSA
2.871	ESHIPS DANA	34 FONDEIG ZONA II	30/12/2019 18:54	31/12/2019 02:00	MARITIMA DEL MEDITERRANEO S.A.	116				
2.128	ZIM YOKOHAMA	23 ANDALUSIA (del 19 al 34)	30/12/2019 18:15	31/12/2019 02:29	A. PEREZ Y CIA, S.L.	261	Desembarque Trasbordo Embarque	Contenedores llenos Contenedores llenos Contenedores llenos	50 800 100	DP WORLD TARRAGONA DP WORLD TARRAGONA DP WORLD TARRAGONA
2.873	BOMAR SEDNA	29 PANTALA REPSOL 35-T	30/12/2019 17:00	31/12/2019 15:29	NEXT MARITIME, S.L.	101	Embarque Residuos	Gasolina Basuras	5000 0	REPSOL PETROLEO
2.864	SIDER RODI	03 ALCUDIA (del 0 al 6)	30/12/2019 13:15	31/12/2019 15:12	EUROPORTS IBERICA TPS S.L.	108	Embarque	Carbones y coque de petroleo	7000	EUROPORTS IBERICA TPS, S.L.
2.811	SYMPHONY SPIRIT	26 CANTABRIA PROLONGACIO (del 20 al 28)	30/12/2019 10:05	31/12/2019 14:28	EUROPORTS IBERICA TPS S.L.	123	Desembarque	Papel y pasta	3800	EUROPORTS IBERICA TPS, S.L.
2.818	RUSICH-7	05 ARAGO (del 24 al 32)	30/12/2019 07:30	30/12/2019 16:18	ALFASHIP MARITIME SERVICES, S.L.	128				

16 144

Pàgina 1 de 367

10

Mostrant 1 - 10 de 3.664

Aquesta és una captura de pantalla de l'aplicació amb la qual s'interactua.

Com es pot observar, les dades que hem d'obtenir estan distribuïdes en 367 pàgines i taules de 10 files que el scraper haurà de navegar.

Com que podem augmentar fins a 30 el número de registres per pàgina, el nostre programa només haurà d'iterar 123 pàgines.

The screenshot shows the 'gisweb' interface for 'Port Tarragona'. It includes a map on the left, a table of ship movements in the center, and a browser's developer console on the right. The table has columns: Escala, Vaixell, Moll, Entrada, Sortida, Consignatari, Eslora, and Tipus. The developer console shows HTML elements and CSS styles, with a search for 'historicos' in the header.

	Escala	Vaixell	Moll	Entrada	Sortida	Consignatari	Eslora	Tipus
2.401	WILHELMINE 17 QUIMICA 4	19/11/2022				NEXT MARITIME, SL	115	Disent
2.313	AL FAROUK 26 CAMPABRIA PROLONGADO (del 3 al 10)	19/11/2022		13:00		TRANSMAR SHIP AND FORWARDING AGENCY, S.L.	92	Embarc
2.309	TRONA 12 CASTELLA SUD (del 14 al 23)	18/11/2022		20:51		EUROPORTS BERICA TPS S.L.	140	Disent
2.412	MARLEY 34 FONDIGO ZONA II	18/11/2022		15:55		EUROFRANSTOZE Y CONSIGNACIONES MARITIMAS EUROPAEMCO	88	Disent
2.332	ARKLOW SPRAY 04 NUBARRA (del 5 al 11)	18/11/2022		15:50		ERDAP, S.A.	182	Disent
2.404	LILLY BOLTON 02 CATALUNYA (del 12 al 22)	18/11/2022		10:10		EUROPORTS BERICA TPS S.L.	179	Embarc

Per identificar cada lloc on el scraper ha de clicar o obtenir-ne dades s'utilitza l'opció *Inspecciona* del cercador.

```
#clica a històrics quan la pàgina s'hagi carregat
WebDriverWait(driver, 20).until(EC.element_to_be_clickable(
    (By.XPATH, "//span[@id='historicos']"))).click()
```

I aquí es pot veure la instrucció que li donem al scraper per a que detecti, en aquest cas, la pestanya històrics. Més endavant analitzarem el codi amb detall.

5. Contingut

S'han recollit totes les escales realitzades durant l'any 2019 (al Port de Tarragona).

De cada escala se n'enregistra el **número d'escala**, **nom del vaixell**, a quin **moll** s'ha realitzat, **data i hora d'entrada i de sortida**, el **consignatari**, l'**eslora** del vaixell, el **tipus d'operació**, les **mercaderies** transportades, el **tonatge** i l'**estibador**. Cal dir que aquestes últimes quatre dades – operació, mercaderia, tonatge i estibador - poden ser múltiples dins de cada escala (diverses operacions per una sola escala), de manera que en un posterior tractament de les dades caldrà tenir-ho en compte.

6. Propietari

Les dades són propietat del Port de Tarragona.

Per poder fer-les accessibles van crear l'eina GISWEB, que és la plataforma amb la qual nosaltres interactuem amb el scraper. La seva intenció és proporcionar-les obertament a la ciutadania però no acaben de facilitar-ne la descàrrega directa i massiva.

Es pot trobar més informació en el següent enllaç:

<https://www.porttarragona.cat/ca/situacio-de-vaixells-link>

- **Accessibilitat a través de scraper**

L'eina GISWEB (amb enllaç <https://tarragona.posidoniaport.com/>) no disposa de fitxer robots.txt. Es pot constatar seguint l'adreça <https://tarragona.posidoniaport.com/robots.txt>, que no condueix enlloc. Per tant, assumim que navegar-la amb un scraper no serà problemàtic. La pàgina web del Port de Tarragona sí que té fitxer robots.txt i les úniques adreces que inhabilita són les següents, sense fer esment de GISWEB:

```
User-agent: *
Allow: /*.js*
Allow: /*.css*
Allow: /*.png*
Allow: /*.jpg*
Allow: /*.gif*

Disallow: /administrator/
Disallow: /bin/
Disallow: /cache/
Disallow: /cli/

Disallow: /includes/
Disallow: /installation/
Disallow: /language/
Disallow: /layouts/
Disallow: /libraries/
Disallow: /logs/

Disallow: /tmp/
```

7. Inspiració

Aquestes dades són interessants ja que permeten observar amb tota transparència els moviments que es duen a terme al Port de Tarragona.

A més a més, estan resultant pràctiques en l'elaboració de l'impost a les emissions dels grans vaixells que s'està desenvolupant conjuntament a les conselleries d'Acció Climàtica i la d'Economia. A partir del temps al port, tonatge i tipologia de vaixell es pot fer un estudi de les emissions contaminants habituals al llarg de l'any i a partir d'aquí gravar tan justament com es pugui els grans vaixells.

8. Llicència

La llicència escollida és Open Database License (ODbL-1.0).*

El motiu és que, en tractar-se de dades obertes que facilita el Port de Tarragona, sembla coherent mantenir el dataset resultant també obert.

*Tot i creure que aquesta és la llicència més adient per la base de dades, en la publicació a Zenodo aquesta no era una de les opcions possibles, i s'ha acabat aplicant-hi la llicència *Creative Commons Attribution 4.0 International*.

9. Codi

El codi íntegre del scraper es pot trobar a:

<https://github.com/arnaugiribet/scrapings/blob/main/source/Escales%20Vaixells.py>

Una explicació detallada del funcionament del codi es pot trobar a:

<https://github.com/arnaugiribet/scrapings/blob/main/Explicaci%C3%B3%20codi.pdf>

10. Dataset

El dataset s'ha publicat a Zenodo amb DOI 10.5281/zenodo.7337235.

<https://doi.org/10.5281/zenodo.7337235>

11. Vídeo

El vídeo de presentació de la pràctica es pot trobar a:

[Enllaç](#)