



MÀSTER UNIVERSITARI EN CIÈNCIA DE DADES

## Pràctica 1: Web Scraping

TIPOLOGIA I CICLE DE VIDA DE LES DADES

ANTONI VANRELL VALLS

GUILLEM MIR FUSTER

6 d'abril de 2021

## 1. Context.

L'actual pandèmia provocada per la Covid-19 ha normalitzat l'ús de material higiènic i sanitari de tot tipus a les llars del territori espanyol. Si fa un any la compra d'aquest tipus de material es trobava majoritàriament enfocada a farmacèutiques o negocis especialitzats, avui en dia és una necessitat present al dia a dia de la ciutadania. A més, les restriccions comercials i de mobilitat han intensificat una tendència ja present, les compres online. En aquest aspecte, la pàgina web d'Amazon s'ha erigit com a líder indiscutible en el mercat online. És per això que considerem el seu apartat de novetats en salut i cura personal, una font d'informació especialment sensible per conèixer les tendències canviants en aquest mercat.

## 2. Definir un títol pel dataset.

Últimes novetats en salut i cura personal d'Amazon Espanya.

## 3. Descripció del dataset.

El dataset generat recull la informació bàsica de les últimes novetats aparegudes a la pàgina de salut i cura personal d'Amazon Espanya. En concret, es recull el nom del producte, les valoracions i el preu, entre d'altres.

## 4. Representació gràfica.

A continuació anem a veure com estan disposats els elements a la pàgina web.

Si ens fixem amb el primer, observem que l'identificador està a dalt a l'esquerra, una fotografia d'aquest enmig, seguit del nom i la puntuació entre 0 i 5 estrelles. Al costat de la puntuació està el nombre de ressenyes i finalment l'interval de preu.







<p>#1</p>  <p>Mascarillas Quirúrgicas, higiénicas, desechables, Tipo IIR, en diferentes colores, filtración (BFE)...</p> <p>★★★★☆ 5</p> <p>9,99 € - 62,00 €</p>	<p>#2</p>  <p>Mascarilla FFP2 CE 1463 WottoCare, 20 uds FFP2 Protección Personal. Bolsa Individual 5 capas....</p> <p>★★★★☆ 371</p> <p>12,71 €</p>	<p>#3</p>  <p>KDOD Mascarillas FFP2 Homologadas, 20 Unidades, CE2163 EN 149:2001 + A1:2009 FFP2...</p> <p>★★★★☆ 26</p> <p>12,99 € - 13,00 €</p>
<p>#4</p>  <p>CRAZYCHIC - Mascarilla FFP2 Negra Homologada Certificada CE EN149 - Mascarilla de Protección...</p> <p>★★★★☆ 546</p> <p>9,99 €</p>	<p>#5</p>  <p>Oxímetro de Pulso, Pulsioxímetro de Dedo Profesional, saturación oxígeno oxímetro con...</p> <p>★★★★☆ 8.800</p> <p>16,99 €</p>	<p>#6</p>  <p>Protección con Elástico para Los Oídos Pack 100 Unidades</p> <p>★★★★☆ 73</p> <p>5,99 €</p>

Figura 1: Esquema dels 6 primers elements de la pàgina web d'Amazon: *Últimas novedades en Salud y cuidado personal* a dia 29/03/2021

## 5. Contingut.

El dataset obtingut consta d'un total de 100 entrades amb 7 camps o columnes:

- **ID:** posició del producte en les darreres novetats en salut i cuidat personal.
- **Nom:** nom complet del producte.
- **Link\_Productes:** link del producte per accedir a comprar-lo.
- **Preu:** interval de preu del producte.
- **Puntuació:** puntuació entre el 0 i el 5 mitjana de diferents ressenyes.
- **Ressenyes:** total de ressenyes del producte.
- **Link\_Ressenyes:** link amb les ressenyes dels clients.

El dataset recull les últimes novetats en productes a la venta en Amazon en salut i cura personal del dia 06/04/2021.

La pàgina web recull els productes nous amb majors ventes sobre salut i cura personal actualitzant cada hora.

Les dades han estat recollides utilitzant el llenguatge Python per fer Web Scraping per extreure la informació de la pàgina HTML.

Més concretament, el que hem fet és usar el BeautifulSoup per transformar la pàgina HTML en un arbre complex d'objectes (Tag, NavigableString, BeautifulSoup i Comment) per poder navegar més còmodament per l'estructura imbricada resultant.

Per extreure tots els registres dels 7 camps hem creat una funció que busca a través de l'estructura (BeautifulSoup) cada un dels 7 camps de cada registre i els emmagatzema cada un en la llista corresponent. Al haver-hi 100 entrades separades entre dues pàgines (50 i 50) hem cridat la funció dues vegades, un cop per pàgina, per obtenir així els 100 registres totals

Un cop hem omplert les 7 llistes hem creat i extret el fitxer csv resultant que podeu trobar penjat a Zenodo.

	A	B	C	D	E	F	G
1	ID	Nom	Link_Productes	Preu	Puntuació	Ressenyes	Link_Ressenyes
2	#1	Mascarillas Quir	<a href="https://www.amazon.es/product-reviews/B08WHN6NV">https://www.amazon.es/product-reviews/B08WHN6NV</a>	9,99 € - 62,00 €	3,5 de un máxim	5	<a href="https://www.amazon.es/product-reviews/B08WHN6NV">https://www.amazon.es/product-reviews/B08WHN6NV</a>
3	#2	Mascarilla FFP2	<a href="https://www.amazon.es/product-reviews/B08VD6655G">https://www.amazon.es/product-reviews/B08VD6655G</a>	16,95 €	4,1 de un máxim	378	<a href="https://www.amazon.es/product-reviews/B08VD6655G">https://www.amazon.es/product-reviews/B08VD6655G</a>
4	#3	KDOD Mascarilla	<a href="https://www.amazon.es/product-reviews/B08WPPYQLD">https://www.amazon.es/product-reviews/B08WPPYQLD</a>	12,99 € - 13,00 €	4,2 de un máxim	28	<a href="https://www.amazon.es/product-reviews/B08WPPYQLD">https://www.amazon.es/product-reviews/B08WPPYQLD</a>
5	#4	Oxímetro de Pul	<a href="https://www.amazon.es/product-reviews/B0914QPTHY">https://www.amazon.es/product-reviews/B0914QPTHY</a>	16,99 €	4,6 de un máxim	8.78	<a href="https://www.amazon.es/product-reviews/B0914QPTHY">https://www.amazon.es/product-reviews/B0914QPTHY</a>
6	#5	CRAZYCHIC - M	<a href="https://www.amazon.es/product-reviews/B08R6CTHNK">https://www.amazon.es/product-reviews/B08R6CTHNK</a>	9,99 €	4,5 de un máxim	547	<a href="https://www.amazon.es/product-reviews/B08R6CTHNK">https://www.amazon.es/product-reviews/B08R6CTHNK</a>
7	#6	Protección con E	<a href="https://www.amazon.es/product-reviews/B08RZBTGQ6">https://www.amazon.es/product-reviews/B08RZBTGQ6</a>	5,99 €	4,2 de un máxim	73	<a href="https://www.amazon.es/product-reviews/B08RZBTGQ6">https://www.amazon.es/product-reviews/B08RZBTGQ6</a>
8	#7	WottoCare Masc	<a href="https://www.amazon.es/product-reviews/B08R683X3V">https://www.amazon.es/product-reviews/B08R683X3V</a>	12,60 €	4,5 de un máxim	63	<a href="https://www.amazon.es/product-reviews/B08R683X3V">https://www.amazon.es/product-reviews/B08R683X3V</a>
9	#8	konjac Mascarilla	<a href="https://www.amazon.es/product-reviews/B08ZKSHS7X">https://www.amazon.es/product-reviews/B08ZKSHS7X</a>	20,99 €	4,9 de un máxim	589	<a href="https://www.amazon.es/product-reviews/B08ZKSHS7X">https://www.amazon.es/product-reviews/B08ZKSHS7X</a>
10	#9	10 Piezas Niños	<a href="https://www.amazon.es/product-reviews/B08VJHJT56">https://www.amazon.es/product-reviews/B08VJHJT56</a>	4,19 € - 4,99 €	3,4 de un máxim	31	<a href="https://www.amazon.es/product-reviews/B08VJHJT56">https://www.amazon.es/product-reviews/B08VJHJT56</a>
11	#10	MAYJAM Lemon	<a href="https://www.amazon.es/product-reviews/B08S6RRQ3S">https://www.amazon.es/product-reviews/B08S6RRQ3S</a>	11,89 € - 19,99 €	4,2 de un máxim	405	<a href="https://www.amazon.es/product-reviews/B08S6RRQ3S">https://www.amazon.es/product-reviews/B08S6RRQ3S</a>
12	#11	KKmier Mascaril	<a href="https://www.amazon.es/product-reviews/B08T5R6RG1">https://www.amazon.es/product-reviews/B08T5R6RG1</a>	19,99 €	4,8 de un máxim	420	<a href="https://www.amazon.es/product-reviews/B08T5R6RG1">https://www.amazon.es/product-reviews/B08T5R6RG1</a>
13	#12	N / P 50 Unidad	<a href="https://www.amazon.es/product-reviews/B08V8MQRZY">https://www.amazon.es/product-reviews/B08V8MQRZY</a>	5,78 € - 6,89 €	3,7 de un máxim	38	<a href="https://www.amazon.es/product-reviews/B08V8MQRZY">https://www.amazon.es/product-reviews/B08V8MQRZY</a>
14	#13	Mobiclinic, Alm	<a href="https://www.amazon.es/product-reviews/B08TMHB8GK">https://www.amazon.es/product-reviews/B08TMHB8GK</a>	19,99 €	3,5 de un máxim	3	<a href="https://www.amazon.es/product-reviews/B08TMHB8GK">https://www.amazon.es/product-reviews/B08TMHB8GK</a>
15	#14	PACK 50 Masca	<a href="https://www.amazon.es/product-reviews/B08S49HD77">https://www.amazon.es/product-reviews/B08S49HD77</a>	9,99 €	4,2 de un máxim	56	<a href="https://www.amazon.es/product-reviews/B08S49HD77">https://www.amazon.es/product-reviews/B08S49HD77</a>
16	#15	BIODEFENCE M	<a href="https://www.amazon.es/product-reviews/B08VRZMJRH">https://www.amazon.es/product-reviews/B08VRZMJRH</a>	23,20 €	4,6 de un máxim	140	<a href="https://www.amazon.es/product-reviews/B08VRZMJRH">https://www.amazon.es/product-reviews/B08VRZMJRH</a>
17	#16	IDOIT Mascarilla	<a href="https://www.amazon.es/product-reviews/B08T7GZPQN">https://www.amazon.es/product-reviews/B08T7GZPQN</a>	14,99 €	4,7 de un máxim	422	<a href="https://www.amazon.es/product-reviews/B08T7GZPQN">https://www.amazon.es/product-reviews/B08T7GZPQN</a>
18	#17	Báscula de Bañ	<a href="https://www.amazon.es/product-reviews/B08T9Q6N2N">https://www.amazon.es/product-reviews/B08T9Q6N2N</a>	15,29 €	4,5 de un máxim	462	<a href="https://www.amazon.es/product-reviews/B08T9Q6N2N">https://www.amazon.es/product-reviews/B08T9Q6N2N</a>
19	#18	Clemars- Copa	<a href="https://www.amazon.es/product-reviews/B08TCC4S79">https://www.amazon.es/product-reviews/B08TCC4S79</a>	12,59 €	4,8 de un máxim	10	<a href="https://www.amazon.es/product-reviews/B08TCC4S79">https://www.amazon.es/product-reviews/B08TCC4S79</a>
20	#19	Arándano Rojo	<a href="https://www.amazon.es/product-reviews/B08RRZPPNP">https://www.amazon.es/product-reviews/B08RRZPPNP</a>	15,49 €	4,6 de un máxim	28	<a href="https://www.amazon.es/product-reviews/B08RRZPPNP">https://www.amazon.es/product-reviews/B08RRZPPNP</a>
21	#20	PUBLIMER Mas	<a href="https://www.amazon.es/product-reviews/B08VR22M9H">https://www.amazon.es/product-reviews/B08VR22M9H</a>	23,95 €	4,1 de un máxim	38	<a href="https://www.amazon.es/product-reviews/B08VR22M9H">https://www.amazon.es/product-reviews/B08VR22M9H</a>

Figura 2: Primers registres del dataset resultant per a visualitzar els diferents camps i els seus valors del dia 06/04/2021

## 6. Agraïments.

Agrair a **Amazon.com** el fet de deixar-nos accedir a la seva pàgina web (<https://www.amazon.es>) per poder extreure la informació que volíem. Hem revisat l'arxiu robots.text d'amazon Espanya i les limitacions fan referència sobre tot a les zones reservades per usuaris administradors i les zones d'usuari. Tanmateix sí que és cert que certs departaments (com els de vídeo) també tenen l'accés limitat. Pel que fa al material sanitari es permet la seva explotació sense més limitacions que les pròpies de qualsevol altra pàgina.

Pel que fa a altres anàlisis, Amazon és un objectiu prototípic dels scrapers. Al cap i a la fi, actualment és el major mercat online a nivell Europeu i tot una referència en aquest sector. De fet, existeixen nombroses pàgines web que ofereixen APIs de web scraping orientades a l'explotació de dades en Amazon com podem veure al següent article web: [2]. Resulta especialment rellevant l'orientació de molts anàlisis de cara a les mateixes empreses que ofereixen els seus productes a Amazon per tal que puguin fer seguiment dels seus productes o puguin realitzar anàlisis de mercat sobre la competència.

## 7. Inspiració.

Tal com hem explicat, en aquest context de crisi sanitària és especialment important la rellevància que han pres els productes especialitzats en salut en el dia a dia dels consumidors. Aquest conjunt de dades pot servir com una aproximació simplificada a l'oferta generada. A partir d'ell es pot respondre quines són les tendències en l'oferta de productes online de tipus sanitària, així com es valoren i quins preus tenen. D'aquesta manera es poden intuir oportunitats d'inversió, conèixer de manera simplificada quina és l'orientació productiva d'aquest mercat o el dinamisme del "branching" en aquest sector. El dataset generat per aquest codi pot respondre les següents preguntes:

- Quina és la tendència actual en productes sanitaris?
- Quins són els preus de referència oferits per una gran corporació com és Amazon en material sanitari?
- De quin tipus de productes sanitaris i de cura personal existeix major varietat a les novetats?
- Com afecta la pandèmia global al tipus de productes ofertats a Amazon Espanya?
- Les tendències d'oferta en material sanitari d'Amazon són reactives a les restriccions que s'apliquen en el territori?

## 8. Llicència.

La llicència escollida és Released Under CC BY-SA 4.0 License. Els motius que porten a aquesta selecció són diversos. Principalment perquè és una llicència que permet l'ús comercial de la feina feta i tenint en compte que les preguntes que respon el nostre dataset poden ser útils per empreses la fa idònia en aquest aspecte. A més, la llicència blindia la distribució posterior sota aquests termes i reconeix la feina de les persones col·laboradores, el que considerem bàsic per tal de reconèixer la feina aliena.

## 9. Codi.

S'adjunta el codi en python. L'enllaç de github on es troben tots aquest elements és el següent: <https://github.com/GuieMF/HealthCareAmazonNRScraping>

## 10. Dataset.

El DOI de Zenodo del dataset és: 10.5281/zenodo.4646807

<https://doi.org/10.5281/zenodo.4646807>

## Referències

- [1] LAIA SUBIRATS MATÉ I MIREIA CALVO GONZÁLEZ. *Web scraping*. PID\_00256968.
- [2] WEBSCRAPINGAPI. *Best 5 Web Scraping APIs For Amazon Product Prices, Reviews, and Market Analysis*. <https://medium.com/api-world/best-5-web-scraping-apis-for-amazon-product-prices-reviews-and-market-analysis-12bab19e3afe>

## Taula de contribucions al treball:

Contribucions	Signa
Recerca prèvia	GMF, AVV
Redacció de les respostes	GMF, AVV
Desenvolupament codi	GMF, AVV