TIPOLOGÍA Y CICLO DE VIDA DE LOS DATOS (PRA 1)
Entrega de trabajo

Fernando Cachadiña
fcachadinag@uoc.edu

Héctor Bastidas Vallejo
hbastidasv@uoc.edu

fcachadinag@uoc.edu hbastidasv@uoc.edu

ÍNDICE

1 CONTEXTO 3
1.1 Motivación y entorno 3
2 TÍTULO DE LOS DATASETS 4
3 DESCRIPCIÓN DE LOS DATASETS 5
4 REPRESENTACIÓN GRÁFICA 9
5 CONTENIDO 10
5.1 DESCRIPCIÓN DEL DATASET 1: evolución de victimas a nivel mundial de la pandemia covid 19. 10
5.2 DESCRIPCIÓN DEL DATASET 2: datos de equipos de protección FACIAL (MASCARILLAS) ofertados en amazon 11
5.3 RELACIÓN ENTRE AMBOS DATASETS. 12
6 AGRADECIMIENTOS 13
7 INSPIRACIÓN 14
8 LICENCIA 15
9 TABLAS DE CONTRIBUCIONES 16

fcachadinag@uoc.edu hbastidasv@uoc.edu

1 CONTEXTO

1.1 MOTIVACIÓN Y ENTORNO

"Cada cicatriz que tienes no es un recuerdo de que te hirieron, sino de que sobreviviste".

Michelle Obama

Al plantearnos este trabajo y su motivación o aportación de valor, ha sido difícil abstraerse del entorno social que afecta al planeta. Es por esto que teníamos que encontrar alguna aportación o vincular el trabajo con el entorno actual.

Por ello y tras analizar el entorno de la pandemia del COVID 19 nos encontramos con este tipo de noticias:

- La Vanguardia 24_3_2020: https://byzness.elperiodico.com/es/el-ultimo-dato/20200227/coronavirus-precio-mascarillas-aumenta-700-7866771
- El Periódico 20_2_2020: https://byzness.elperiodico.com/es/el-ultimo-dato/20200227/coronavirus-precio-mascarillas-aumenta-700-7866771

Nos hemos propuesto trazar el precio de elementos similares de protección personal, como mascarillas de características similares y su precio en los sitios web de los puntos de venta de Amazon en cada país. Es de conocimiento público que Amazon oferta millones de productos a través de la web en donde se detalla información de cada uno de estos, incluido el precio como campo relevante para nuestra investigación.

Por otro lado, se utilizará el sitio web Worldometer para la extracción de las estadísticas sobre las víctimas del COVID 19 ya que se trata de un facilitador global de datos sobre la pandemia para entidades gubernamentales, universidades, e instituciones alrededor del mundo. La fuente de datos que maneja Worldometer son los sitios web oficiales de los ministerios de salud o agencias del gobierno de cada país con lo cual se garantiza datos avalados por cada nación.

fcachadinag@uoc.edu hbastidasv@uoc.edu

2 TÍTULO DE LOS DATASETS

Preparamos dos data sets:

- El primer dataset contiene datos de los elementos de mascarillas en cada país en su página de Amazon, el cual incluye descripción del producto, país y precio local.

Título: Datos de equipos de protección facial (mascarillas) ofertados en Amazon

- El segundo dataset contiene los datos oficiales por país de los afectados por la pandemia: infectados, muertos, críticos. etc.

Título: Evolución de las víctimas a nivel mundial de la pandemia COVID 19

3 DESCRIPCIÓN DE LOS DATASETS

- ***Captura de los datos unificados de las víctimas en múltiples países.***

Se ha hecho seguimiento del número de casos, número de muertes, número de personas recuperadas, personas en estado crítico y víctimas de la pandemia en una serie de países asociados.

Obtendremos los valores desde el «webscrapping» de:

https://www.worldometers.info/coronavirus/

Comentar que esta web es especialmente dinámica y ha cambiado el formato y estructuración de los datos varias veces desde que comenzamos el proyecto.

Now Yesterday Search:

Country, Other	Total Cases	New Cases	Total Deaths	New Deaths	Total Recovered	Active Cases	Serious, Critical	Tot Cases/ 1M pop	Tot Deaths/ 1M pop
China	81,171	+78	3,277	+7	73,159	4,735	1,573	56	2
Italy	63,927		6,077		7,432	50,418	3,204	1,057	101
USA	48,778	+5,044	588	+35	295	47,895	1,175	147	2
Spain	39,676	+4,540	2,800	+489	3,794	33,082	2,355	849	60
Germany	31,370	+2,314	133	+10	749	30,488	23	374	2
Iran	24,811	+1,762	1,934	+122	8,913	13,964		295	23
France	19,856		860		2,200	16,796	2,082	304	13
Switzerland	9,117	+322	122	+2	131	8,864	141	1,053	14
S. Korea	9,037	+76	120	+9	3,507	5,410	59	176	2
UK	8,077	+1,427	422	+87	135	7,520	20	119	6
Netherlands	5,560	+811	276	+63	2	5,282	435	324	16
Austria	5,010	+536	28	+7	9	4,973	22	556	3
Belgium	4,269	+526	122	+34	461	3,686	381	368	11
Norway	2,766	+141	12	+2	6	2,748	44	510	2
Portugal	2,362	+302	30	+7	22	2,310	48	232	3
Sweden	2,286	+240	36	+9	16	2,234	136	226	4
Canada	2,176	+85	24		112	2,040	1	58	0.6
Australia	2,144	+257	8	+1	118	2,018	11	84	0.3
Brazil	1,980	+56	34		2	1,944	18	9	0.2

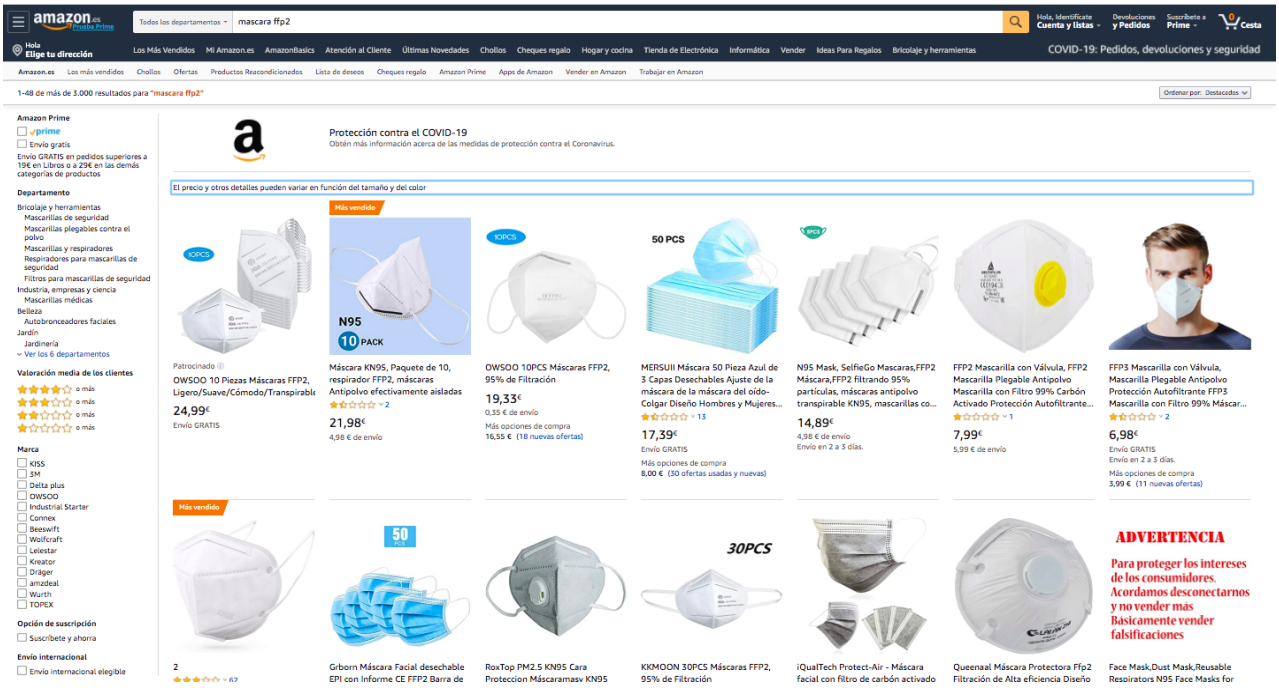
- ***Datos de precios de venta de múltiples productos de protección facial en Amazon:***

También vamos a hacer seguimiento del precio de mascarillas en los países asociados a las webs de venta online de AMAZON.

Por lo cual se han capturado las ofertas «frontales de primera página» de mascarillas en las webs de AMAZON de UK, USA, ES, AU, NL, DE. Se hace foco en las mascarillas tipo FFP2 y N95.

Tras valorar la forma de controlar el precio de mascarillas hemos tenido múltiples problemas.

1. ***Falta de marcas identificables.***
2. ***Hay miles de formas, marcas blancas (sin marca) y tipos de filtrado.***
3. ***Poca uniformidad entre países.***
4. ***Hemos decidido tomar el fabricante 3M como líder mundial como referencia.***
5. ***Hemos tomados los estándares de calidad FFP2 y N95***
6. ***En algún país no se ha encontrado nada equivalente disponible por lo que realizaremos un conjunto de medidas de varios productos.***



Las webs (con búsqueda incluida) implicadas son:

urlSpain =

'HTTPS://WWW.AMAZON.ES/S?K=MASCARILLAS+FFP2&__MK_ES_ES=%C3%85M%C3%85%C5%BD%C3%95%C3%91&REF=NB_SB_NOSS'

urlUSA =

HTTPS://WWW.AMAZON.COM/S?K=FFP2+MASK&REF=NB_SB_NOSS_2

urlFrance =

HTTPS://WWW.AMAZON.FR/S?K=MASQUE+FACIAL+FFP2&__MK_FR_FR=%C3%85M%C3%85%C5%BD%C3%95%C3%91&REF=NB_SB_NOSS_2

urlUK =
HTTPS://WWW.AMAZON.CO.UK/S?K=FFP2+MASK&REF=NB_SB_NOSS_2

urlGermany =
HTTPS://WWW.AMAZON.DE/S?K=GESICHTSMASKE+FFP2&__MK_DE_DE=%C3%85M%C3%85%C5%BD%C3%95%C3%91&REF=NB_SB_NOSS_2

urlItaly =

HTTPS://WWW.AMAZON.IT/S?K=MASCHERA+FFP2&__MK_IT_IT=%C3%85M%C3%85%C5%BD%C3%95%C3%91&REF=NB_SB_NOSS_2

urlNetherlands =

HTTPS://WWW.AMAZON.NL/S?K=GEZICHTSMASKER+FFP2&__MK_NL_NL=%C3%85M%C3%85%C5%BD%C3%95%C3%91&CRID=296W8O9PQ26GE&SPREFIX=GEZICHTSMASKER+%2CAPS%2C320&REF=NB_SB_SS_I_4_15

urlAustralia =

HTTPS://WWW.AMAZON.COM.AU/S?K=FFP2+MASK&REF=NB_SB_NOSS_2

Adicionalmente, se ha realizado la captura de:

- ***Datos de Free proxies activos para navegación a través de ellos.***

Sabemos que AMAZON tiene una estricta problemática y controles «anti-Scrapping» por ello las capturas de los datos de las paginas previstas, se harán contando con unos proxies de cada país. El código a generar buscará la lista de proxies libres activos en cada país y los rotará para evitar problemas de bloqueo de IP. Esto sería extrapolable a cualquier otra campaña mucho más agresiva de captación de precios.

Ha sido imposible recabar su autorización.
Se han capturado los datos de la web:

https://www.us-proxy.org/

No se generará un dataset para este fin, sin embargo, se usarán en la captura de datos posterior.

fcachadinag@uoc.edu hbastidasv@uoc.edu

Free Proxy List

FREE PROXY WEB PROXY SOCKS PROXY BUY PROXY COMPANY

US Proxy

US proxies that are just checked and updated every 10 minutes

Show 20 entries

Search all columns:

IP Address	Port	Code	Country	Anonymity	Google	Https	Last Checked
142.93.4.1	3128	US	United States	anonymous	no	no	13 seconds ago
50.226.151.3	80	US	United States	anonymous	no	no	13 seconds ago
140.82.23.25	8080	US	United States	elite proxy	no	yes	13 seconds ago
159.203.181.112	8080	US	United States	anonymous	no	no	13 seconds ago
3.235.93.102	8080	US	United States	anonymous	no	no	13 seconds ago
138.68.24.145	8080	US	United States	anonymous	no	no	10 minutes ago
52.42.107.107	8080	US	United States	anonymous	no	no	10 minutes ago
142.93.14.224	8080	US	United States	anonymous	no	no	10 minutes ago
13.65.195.36	8080	US	United States	anonymous	no	no	10 minutes ago
149.28.104.226	3128	US	United States	anonymous	no	no	10 minutes ago
45.55.27.15	8080	US	United States	anonymous	no	no	10 minutes ago
67.205.132.241	8080	US	United States	anonymous	no	no	10 minutes ago

4 REPRESENTACIÓN GRÁFICA

Dataset 1: Evolución de las víctimas a nivel mundial de la pandemia COVID 19

dateTime	country	total_cases	new_cases	total_deaths	new_deaths	total_recovered
2020-03-29T21:32:14.621067	USA	139745	16167	2448	228	4435
2020-03-29T21:32:14.621067	Italy	97689	5217	10779	756	13030
2020-03-29T21:32:14.621067	Spain	80031	6796	6802	820	14709
2020-03-29T21:32:14.621067	Germany	62095	4400	525	92	9211
2020-03-29T21:32:14.621067	France	40174	2599	2606	292	7202
2020-03-29T21:32:14.621067	Iran	38309	2901	2640	123	12391

Dataset 2: Datos de equipos de protección facial (mascarillas) ofertados en Amazon

asin	description	dateTime	date	country	currency	price
B0863LTD6C	Máscara KN95, Paquete de 10, respirador FFP2, máscaras Antipolvo efectivamente aisladas	2020-03-29T21:31:58.978378	2020-03-29	Spain	Euro	22.79
B085VV4VRV	FENBO EU Máscara FFP2, Paquete de 20, respirador N95, máscaras Antipolvo efectivamente aisladas	2020-03-29T21:31:58.978378	2020-03-29	Spain	Euro	29.0
B085TBDXWY	BENKS [5 Piezas] Anti P-M-2. 5 Pro -tec -ción Contra el Pol -vo y la Conta- mi- na-ción – RES M	2020-03-29T21:31:58.978378	2020-03-29	Spain	Euro	36.54

5 CONTENIDO

5.1 DESCRIPCIÓN DEL DATASET 1: EVOLUCIÓN DE VICTIMAS A NIVEL MUNDIAL DE LA PANDEMIA COVID 19.

Se ha generado un dataset con resumen de la evolución de los datos capturados de víctimas de la pandemia COVID 19.

- ***Nombre:***

  afectados_pandemia_pais_covid_19.csv

- ***Código DOI:***

  Código de DOI publicado en ZENODO.
  10.5281/zenodo.3747902

- ***Contenidos:***

  Datos de parámetros de evolución de los afectados por la pandemia COVID 19 por país.

- ***Periodo de tiempo:***

  Desde 29/03/2020 a 9/4/2020

- ***Hora de Extracción:***

  Las extracciones se han realizado a las 11:00 y 21:30, ambas son horas UTC.

- ***Detalle de campos:***

'fecha'	tipo STRING	"Fecha de la captura del datos
'pais',	tipo STRING	"País al que hacen referencia los datos, salvo un crucero"
'total_casos'	tipo INT	"Número total de casos que se han contabilizado en el país"
'nuevos_casos'	tipo INT	"Número total de casos incrementados desde la info anterior en el país"
'muertes_totales'	tipo INT	"Número total de fallidos en el país "
'nuevas_muertes'	tipo INT	"Número total de fallecidos incrementados desde la info anterior en el país"
'total_recuperados'	tipo INT	"Número total de recuperados en el país "
'casos_activos'	tipo INT	"Número total de casos en tratamiento en el país "
'criticos'	tipo INT	"Número total en tratamiento en UCI en el país "
'casos_por_millon'	tipo INT	"Ratio de casos por millón de habitantes en el país "
'muertes_por_millon'	tipo INT	"Ratio de muertes por millón de habitantes en el país "

- ***Comentarios:***

  Es posible que existan días sin contenidos por falta de capturas debido al cambio constante en la estructura de la página.

5.2 DESCRIPCIÓN DEL DATASET 2: DATOS DE EQUIPOS DE PROTECCIÓN FACIAL (MASCARILLAS) OFERTADOS EN AMAZON

Se ha generado una tabla con la captura de productos de protección facial por día y por país con precios unitarios.

- ***Nombre:***

  Evolucion_precios_EPP

  (nota: EPP = Elementos de protección personal)
- ***Código DOI:***

  Código de DOI publicado en ZENODO.
  10.5281/zenodo.3747888
- ***Contenidos:***

  Datos de los productos de Amazon de mascarillas con precios por elemento identificados.
- ***Periodo de tiempo:***

  Desde 29/03/2020 a 9/4/2020
- ***Hora de Extracción:***

  Las extracciones se han realizado a las 11:00 y 21:30, ambas son horas UTC.
- ***Detalle de campos:***

'asin'	tipo STRING	"Código de identifícación único de product de AMAZON"
'description'	tipo STRING	'Texto descriptivo del producto'
'dateTime'	tipo TIMESTAMP	'Cadena de carateres que contiene fecha y hora GMT'
'date'	tipo DATETIME	' Formato diferente de la misma fecha / hora de captura '
'country'	tipo STRING	'Pais al que pertenece a distribución del producto ' Valores posibles: 'Spain', 'USA', 'France', 'UK', 'Germany', 'Italy', 'Netherlands', 'Australia'
'currency'	tipo STRING	' Moneda aplicable al precio inferior ' Valores posibles: "Euro', 'Dólar', 'Libra Esterlina'
'price'	tipo FLOAT.64	' Precio de venta del producto en moneda local (de paquete, puede no ser un precio unitario)'

fcachadinag@uoc.edu hbastidasv@uoc.edu

5.3 RELACIÓN ENTRE AMBOS DATASETS.

Ambos dataset se refieren a hechos diferentes por lo que sólo tienen en común la ubicación geográfica.

En este caso los campos relacionales en cada dataset son:

- ***DATASET 1: 'pais'***
- ***DATASET 2: 'country'***

Estos campos al momento se encuentran normalizados por lo que pueden ser utilizados para relacionar ambas bases de datos.

En un desarrollo futuro se relacionarán en un dataset único mediante un campo calculado "country_id"

6 AGRADECIMIENTOS

- ***AMAZON***

Pese a que AMAZON en sus documentos de robots.txt prohíbe expresamente el «Webscraping», hemos solicitado la autorización vía correo electrónico sin respuesta al día de hoy, pero suponemos que no tendrá inconveniente en una investigación de carácter meramente divulgativo y sin uso comercial.

- ***WORLDOMETERS***

Agradecemos a Worldometer a través de sitio web https://www.worldometers.info/coronavirus/ ***por el trabajo diario de agregación de datos que efectúan a fin de facilitar información oficial sobre los afectados por la pandemia en cada país.***

- ***US-PROXY***

Agradecemos a https://www.us-proxy.org/ ***el trabajo de monitorización constante de «free proxies» de internet precisamente orientada a usos como el de este informe.***

fcachadinag@uoc.edu hbastidasv@uoc.edu

7 INSPIRACIÓN

- ***Futuro uso de los datos e Hipótesis de investigación futuras.***

Vamos a intentar observar si de promedio, un conjunto de ofertas de mascarillas de protección facial sube de precio en base a:

- El paso del tiempo.
- La evolución de los casos por país y sus variables asociadas: Infectados y datos asociados.

Para ello dejamos las bases puestas para analizar si hay alguna correlación entre el incremento de casos diario por país (o mediciones accesorias) y la variación de precios.

No tenemos claro que los titulares de prensa serán tan reales en el suministro «on-line» que es el único que queda activo en este momento.

fcachadinag@uoc.edu hbastidasv@uoc.edu

8 LICENCIA

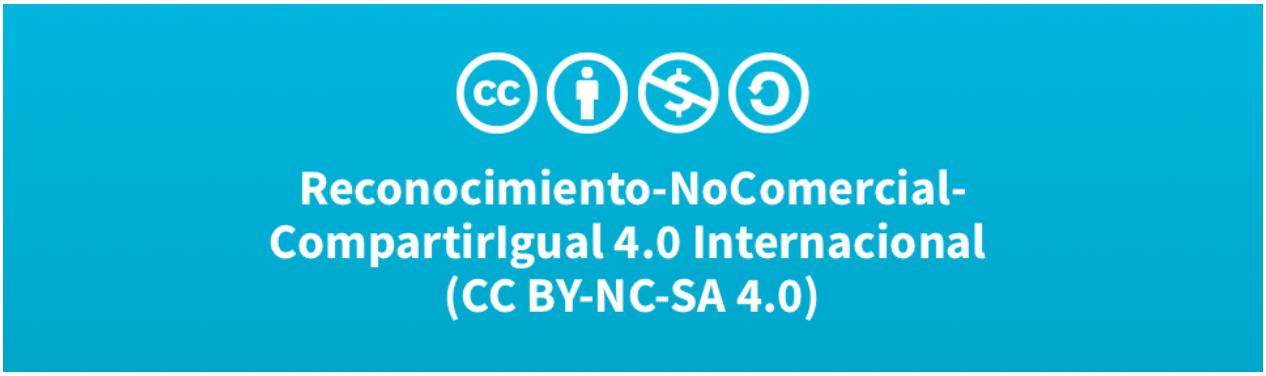
- ***LICENCIA DE USO DE LOS DATASETS.***

Usted es libre de:

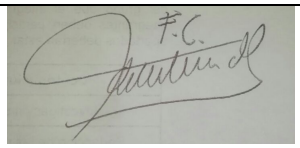
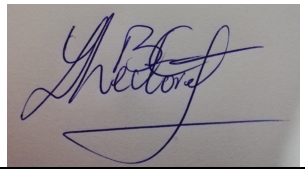
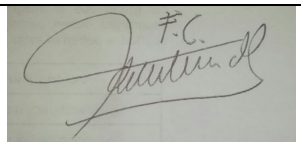
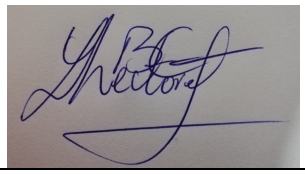
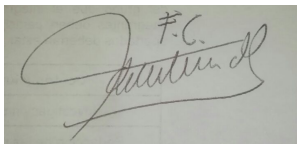
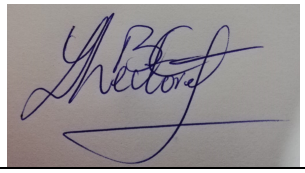
Compartir — copiar y redistribuir el material en cualquier medio o formato
Adaptar — remezclar, transformar y crear a partir del material

El licenciador no puede revocar estas libertades mientras cumpla con los términos de la licencia.

RECONOCIMIENTO -NOCOMERCIAL-COMPARTIR-IGUAL4.0 INTERNACIONAL
CC BY-NC-SA 3.0



9 TABLAS DE CONTRIBUCIONES

CONTRIBUCIONES	INTEGRANTE	FIRMA
INVESTIGACION PREVIA	FERNANDO CACHADIÑA	
	HÉCTOR BASTIDAS	
REDACCION DE LAS RESPUESTAS	FERNANDO CACHADIÑA	
	HÉCTOR BASTIDAS	
DESARROLLO DEL CODIGO	FERNANDO CACHADIÑA	
	HÉCTOR BASTIDAS