PRAC1:

Tipología y ciclo de vida de los datos.

Web scraping: COVID-19

Raúl Sánchez Campos

Pablo Santos Ramos

12 de abril de 2021

1. Contexto

Explicar en qué contexto se ha recolectado la información. Explique por qué el sitio web elegido proporciona dicha información.

La importancia de este proyecto radica en encontrar y recabar información sobre el COVID-19, una pandemia que ha generado una crisis a la que el mundo pocas veces se ha enfrentado. Esta pandemia está afectando a países de manera internacional, y conocer las estadísticas de cómo evoluciona esta enfermedad a nivel mundial nos sirve de baremo para saber qué países están haciendo mejor las cosas para controlar esta afección.

La página web es Cambio político, un sitio web (CPNEWS) que nació en virtud de la necesidad de contar con un espacio donde se pueda leer columnas de opinión, análisis, entrevistas y reportajes de un sector de pensamiento que usualmente no es muy divulgado, se constituye en un esfuerzo por fortalecer el análisis, la reflexión y el estudio de temas políticos, económicos y sociales de la realidad costarricense y latinoamericana, mediante el desarrollo de un portal electrónico que contenga artículos, análisis, documentos y columnas en esa línea.

Este sitio web ofrece la información en línea que publica la OMS de las estadísticas globales del COVID 19 a nivel mundial.

2. Definir un título.

Elegir un título que sea descriptivo.

Estadísticas del virus SARS-CoV-2 en diferentes países y a nivel mundial.

3. Descripción del data set.

Desarrollar una descripción breve del conjunto de datos que se ha extraído (es necesario que esta descripción tenga sentido con el título elegido).

El data set de estadísticas sobre el Covid-19 contiene datos sobre los casos actuales en distintos países, así como el número de fallecidos, recuperados, casos activos y casos críticos de cada país, actualizados cada día. Para los casos en los que no tenemos información damos un valor 0 de manera que no afecte al recuento total.

Podemos observar la cabecera de los datos para hacernos una idea de la información que poseemos.

	Pais	Total de casos	Casos nuevos	Total de muertes	Muertes nuevas	Recuperados	Casos activos	Casos criticos
0	USA	31560438.0	0.0	570260.0	0.0	24122221.0	6867957.0	8960.0
1	Brasil	13106058.0	0.0	337364.0	0.0	11558784.0	1209910.0	8318.0
2	India	12801785.0	2039.0	166208.0	0.0	11792135.0	843442.0	8944.0
3	Francia	4841308.0	0.0	97273.0	0.0	301299.0	4442736.0	5626.0
4	Rusia	4597868.0	0.0	101106.0	0.0	4220035.0	276727.0	2300.0

Imagen 1. Cabecera del data set Covid-19.

Tenemos en total 8 columnas de información, todas con tipos de datos *float64* (coma flotante de doble precisión) menos el nombre del país que es tipo *String*. De estas columnas podemos extraer datos como el promedio a nivel mundial, la desviación, los cuartiles y el máximo y el mínimo de cada columna. Tenemos:

- País. Indica el país al que se refieren los datos.
- Total de casos. – Suma de todos los casos en el país. Promedio 602036.4
- Casos nuevos. - Cantidad de casos nuevos por día. Promedio 156.57
- Total de muertes. – Suma de las muertes producidas por el covid-19. Promedio 13062.12
- Muertes nuevas. – Cantidad de fallecidos nuevos por día. Promedio 5.8
- Recuperados. – Número de personas recuperadas. Promedio 485421.9
- Casos activos. – Cantidad de casos activos en cada país. Promedio 103552.3
- Casos críticos. – Casos con riesgo mayor de fallecimiento. Promedio 450.53

La información la hemos obtenido directamente del análisis de los datos posterior al scraping. Podemos encontrar más información en la Imagen 2.

	Total de casos	Casos nuevos	Total de muertes	Muertes nuevas	Recuperados	Casos activos	Casos criticos
count	2.210000e+02	221.000000	221.000000	221.000000	2.210000e+02	2.210000e+02	221.000000
mean	6.020364e+05	156.565611	13062.117647	5.855204	4.854219e+05	1.035523e+05	450.524887
std	2.537686e+06	1140.785077	50557.002385	52.270858	2.035636e+06	5.687495e+05	1309.007398
min	1.000000e+00	0.000000	0.000000	0.000000	0.000000e+00	0.000000e+00	0.000000
25%	4.297000e+03	0.000000	63.000000	0.000000	2.825000e+03	2.480000e+02	0.000000
50%	3.696600e+04	0.000000	576.000000	0.000000	2.584100e+04	2.834000e+03	11.000000
75%	2.589170e+05	0.000000	4636.000000	0.000000	2.206450e+05	2.640600e+04	194.000000
max	3.156044e+07	15415.000000	570260.000000	603.000000	2.412222e+07	6.867957e+06	8960.000000

Imagen 2. Resumen de estadísticas del data set covid-19..

4. *Representación gráfica.*

Presentar esquema o diagrama que identifique el DATASET visualmente y el proyecto elegido.

Este proyecto fue realizado con SCRAPY que es una plataforma colaborativa de código libre que corre en Python para extraer datos de páginas web usado para una serie de aplicaciones como minería de datos, procesamiento de información o registro histórico.

SCRAPY tiene las siguientes características:

- Rápida y poderosa: Escribes las reglas para extraer los datos y SCRAPY hace el resto.
- Fácilmente extensible: Dada su configuración, puede generar nueva funcionalidad sin tener que modificar el código fuente.
- Portable y PYTHONICO: Escrito en Python y puede correr en Linux, Windows, Mac y BSD.

EL algoritmo consta de consta de las siguientes partes:

- Proceso de extracción: La primera parte consiste en extraer la información del WEBSITE cambio político el cual ofrece una información estructurada de las estadísticas a nivel mundial del COVID-19.

- Proceso de transformación: El proceso de transformación consiste en reformateo de datos, conversión de unidades y evaluación de datos faltantes.

- Proceso de carga: La última parte del proceso ETL es llevar todos los datos limpios a un fichero CSV llamado estadísticas_a_nivel_mundial_covid_19.csv

5. Contenido.

Explicar los campos que incluye el DATASET, el periodo de tiempo de los datos y cómo se ha recogido.

El set de datos consta de ocho columnas, estas ya se han explicado anteriormente en la descripción del data set. Sin embargo, vemos suficientemente relevante repetir y describir de nuevo los atributos para completar el apartado de Contenido. Las columnas nos ofrecen la siguiente información:

- País. - En esta columna contiene el nombre de país al cual se le aplica el censo.
- Total de casos. – Esta columna contiene el total de casos confirmados por país.
- Casos nuevos. - Esta columna contiene el total de casos nuevos, tomando como línea de tiempo todos los casos del mes anterior al actual.
- Total de muertes. - Esta columna contiene el total de muertes confirmadas por país
- Muertes nuevas. - Esta columna contiene el total de muertes nuevas, tomando como línea de tiempo todas las muertes del mes anterior al actual.
- Recuperados. - Esta columna contiene el total de personas recuperadas por país.
- Casos activos. - Esta columna contiene el total de casos activos por país.
- Casos críticos. - Esta columna contiene el total de casos críticos por país.

La información que proporciona el sitio Web es actualizada mes a mes con los datos de la OMS, donde se ofrece el censo a nivel mundial de esta pandemia y la evolución de la misma en los distintos países del mundo.

6. Agradecimientos.

Presentar al propietario del conjunto de datos. Es necesario incluir citas de análisis anteriores o, en caso de no haberlas, justificar esta búsqueda con análisis similares.

Los datos han sido recopilados desde el sitio web de cambio político el cual ofrece las estadísticas mundiales del COVID-19 proporcionada por la OMS, se ha hecho uso del lenguaje de programación Python y técnicas de WEB SCRAPING para extraer la información alojada en su sitio web

7. Inspiración.

Explique por qué es interesante este conjunto de datos y qué preguntas se pretenden responder.

El interés de analizar este conjunto de datos reside en las pocas veces que se pueden obtener estadísticas de un virus que afecta a todos los países de maneras diferentes. Debido a la pandemia que está azotando al mundo, es de relevancia importancia la información proporcionada, dado que así podemos ver y analizar como esta evolucionado el COVID 19 en los distintos países del mundo y encontrar así las diferencias entre casos y fallecidos según otros parámetros externos.

Algunas de las preguntas que podemos responder son: ¿Cuáles son los países con mayor número de casos? , ¿Dónde se están produciendo más contagios a día de hoy?, ¿ Cuáles son los países en los que se ha producido mayor número de muertes?, ¿Hay algún país que tenga más recuperados que contagios?

Todas las respuestas a estas preguntas las podemos obtener de los siguientes gráficos:

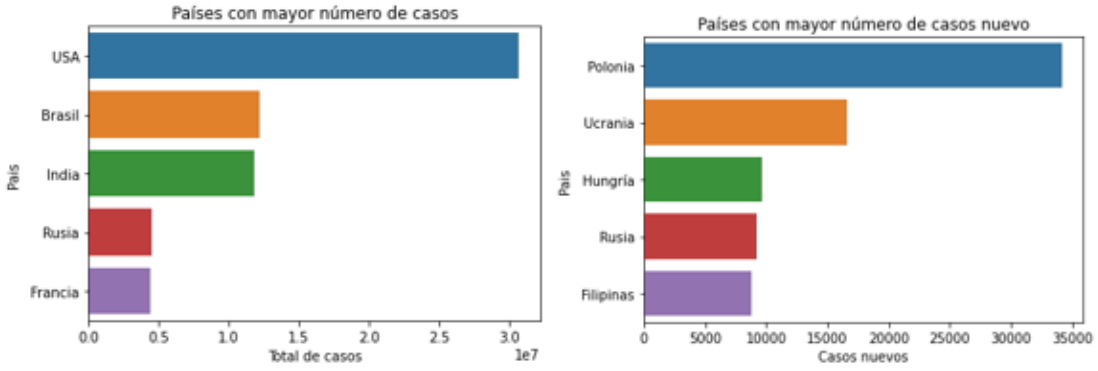

Imagen 3. Países con mayor número de casos. Totales y nuevos.

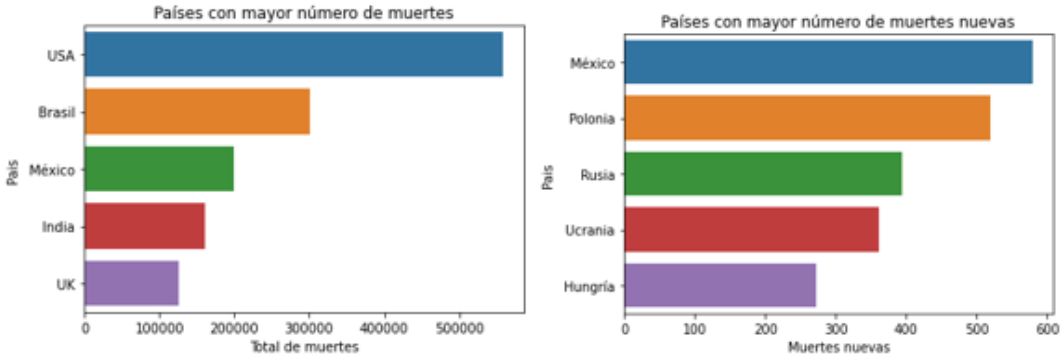

Imagen 4. Países con mayor número de muertes. Totales y nuevas.

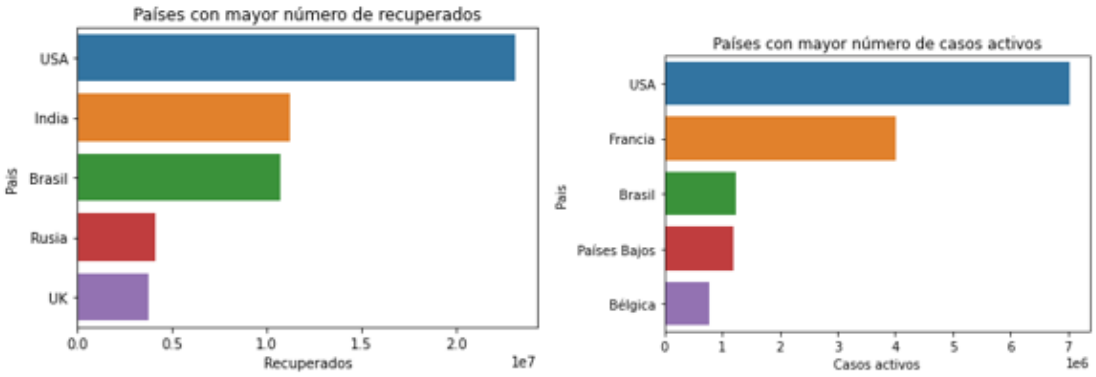

Imagen 5. Países con mayor número de recuperados y de casos activos.

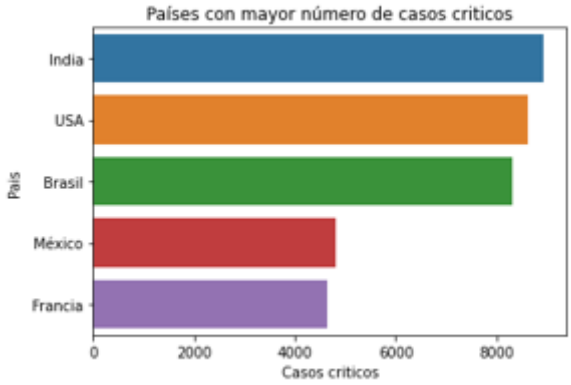

Imagen 6. Países con mayor número de casos críticos.

Como se puede observar de un simple vistazo a los diagramas de barras podemos ver la evolución de esta pandemia en los distintos países. Podemos ver que Estados Unidos junto y Brasil son los países con el mayor número de casos (Imagen 3), Polonia y Ucrania los países con mayor número de casos nuevos, Repite Estados Unidos junto y Brasil con el mayor número de muertes, México y Polonia son los países con mayor número de muertes nuevas (Imagen 4), los países con mayor número de casos recuperados son Estados Unidos e India etc.

En conclusión, podemos plantearnos, con esta información, cualquier estudio estadístico y saber cómo está afectando la pandemia a todas las naciones del mundo y que están haciendo los gobiernos para contenerla.

8. Licencia.

Seleccione una de estas licencias para su dataset y explique el motivo de su selección.

La licencia escogida para este set de datos es CC BY-SA 4.0 LICENSE dado que esta licencia permite a otros remezclar, modificar y basarse en su trabajo incluso con fines comerciales, siempre y cuando le acrediten y os autoricen sus nuevas creaciones bajo los mismos términos. Esta licencia se compara a menudo con licencias de software gratuitas y de código abierto "copyleft". Todas las obras nuevas basadas en la suya llevarán la misma licencia, por lo que cualquier derivado también permitirá el uso comercial. Esta es la licencia utilizada por Wikipedia, y se recomienda para materiales que se beneficiarían de la incorporación de contenido de Wikipedia y proyectos con licencia similar.

https://creativecommons.org/licenses/by-sa/4.0/

https://creativecommons.org/licenses/by-sa/4.0/legalcode

9. Código.

Adjuntar el código con el que se ha generado el data set, preferiblemente en Python o, alternativamente, en R.

El código se encuentra en la carpeta de GitHub, en el archivo .ipynb, así como adjunto al final de este documento. Es un código de Python con la librería SCRAPY implementada.

La carpeta donde se han subido los archivos a GitHub es:
https://github.com/raulec1982/Practica_1_Raul_Sanchez_Campos.git

10. Dataset.

Presentar el dataset en formato CSV

El data set en formato CSV se encuentra en la carpeta Git. También se ha publicado el data set formato CSV en ZENODO, incluyendo una breve descripción: https://zenodo.org/record/4637936#.YHHfAegzaUm

Contribuciones	Firma
Investigación previa	R. P.
Redacción de las respuestas	R. P.
Desarrollo código	R. P.

Código

```
!pip install scrapy

"""
OBJETIVO:
  - Extraer las estadísticas a nivel mundial del covid-19
  - Practica 1 Web scraping
CREADO POR: Raúl Sánchez Campos y Pablo Santos Ramos
ULTIMA VEZ EDITADO: 24 MARZO 2021
"""

from scrapy.spiders import Spider
from scrapy.selector import Selector
from scrapy.loader.processors import MapCompose
from scrapy.crawler import CrawlerProcess

# CLASE CORE - SPIDER
class EstadisticaCovidSpider(Spider):
  name = "EstadisticaCovidSpider"
  custom_settings = {
      'USER_AGENT': 'Mozilla/5.0 (X11; Linux x86_64)
AppleWebKit/537.36 (KHTML, like Gecko) Ubuntu Chromium/71.0.3578.80
Chrome/71.0.3578.80 Safari/537.36',
  }
  start_urls = ['https://cambiopolitico.com/estadisticas-covid-19']

  def parse(self, response):
        sel = Selector(response)
        table = sel.xpath('//*[@id="wpcv-table-3"]//table//tbody//tr')
        for row in table:
            yield {
```

```
                'Pais' : row.xpath('td[1]//text()').extract_first(),
                'Total de casos':
row.xpath('td[2]//text()').extract_first(),
                'Casos nuevos' :
row.xpath('td[3]//text()').extract_first(),
                'Total de muertes' :
row.xpath('td[4]//text()').extract_first(),
                'Muertes nuevas' :
row.xpath('td[5]//text()').extract_first(),
                'Recuperados' :
row.xpath('td[6]//text()').extract_first(),
                'Casos activos' :
row.xpath('td[7]//text()').extract_first(),
                'Casos criticos' :
row.xpath('td[8]//text()').extract_first(),
            }



process = CrawlerProcess({
    'FEED_FORMAT': 'csv',
    'FEED_URI': 'covid-19.csv'
})
process.crawl(EstadisticaCovidSpider)
process.start()



import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
df = pd.read_csv('covid-19.csv', encoding = 'utf8')
df.head(5)

df.loc[df['Casos nuevos'] == '-', 'Casos nuevos'] = "0"
df.loc[df['Muertes nuevas'] == '-', 'Muertes nuevas'] = "0"
df.loc[df['Total de muertes'] == '-', 'Total de muertes'] = "0"
df.loc[df['Muertes nuevas'] == '-', 'Muertes nuevas'] = "0"
df.loc[df['Recuperados'] == '-', 'Recuperados'] = "0"
df.loc[df['Casos activos'] == '-', 'Casos activos'] = "0"
df.loc[df['Casos criticos'] == '-', 'Casos criticos'] = "0"

df["Total de casos"] = df["Total de casos"].str.replace(",",
"").astype(float)
df["Casos nuevos"] = df["Casos nuevos"].str.replace(",",
"").astype(float)
df["Total de muertes"] = df["Total de muertes"].str.replace(",",
"").astype(float)
df["Muertes nuevas"] = df["Muertes nuevas"].str.replace(",",
"").astype(float)
df["Recuperados"] = df["Recuperados"].str.replace(",",
"").astype(float)
df["Casos activos"] = df["Casos activos"].str.replace(",",
"").astype(float)
df["Casos criticos"] = df["Casos criticos"].str.replace(",",
"").astype(float)

df.dtypes
missing_data = df.isnull()
```

```
missing_data.head()

for column in missing_data.columns.values.tolist():
    print(column)
    print (missing_data[column].value_counts())
    print("")

df.head(5)

fig, ax = plt.subplots()
ax.set_title("Países con mayor número de casos")
sns.barplot(x = 'Total de casos',
            y = 'Pais',
            orient = 'h',
            data = df.sort_values(by=['Total de casos'],
ascending=False).head(5))
plt.show()

fig, ax = plt.subplots()
ax.set_title("Países con mayor número de casos nuevo")
sns.barplot(x = 'Casos nuevos',
            y = 'Pais',
            orient = 'h',
            data = df.sort_values(by=['Casos nuevos'],
ascending=False).head(5))
plt.show()

fig, ax = plt.subplots()
ax.set_title("Países con mayor número de muertes")
sns.barplot(x = 'Total de muertes',
            y = 'Pais',
            orient = 'h',
            data = df.sort_values(by=['Total de muertes'],
ascending=False).head(5))
plt.show()

fig, ax = plt.subplots()
ax.set_title("Países con mayor número de muertes nuevas")
sns.barplot(x = 'Muertes nuevas',
            y = 'Pais',
            orient = 'h',
            data = df.sort_values(by=['Muertes nuevas'],
ascending=False).head(5))
plt.show()

fig, ax = plt.subplots()
ax.set_title("Países con mayor número de recuperados")
sns.barplot(x = 'Recuperados',
            y = 'Pais',
            orient = 'h',
            data = df.sort_values(by=['Recuperados'],
ascending=False).head(5))
plt.show()

fig, ax = plt.subplots()
ax.set_title("Países con mayor número de casos activos")
sns.barplot(x = 'Casos activos',
            y = 'Pais',
            orient = 'h',
            data = df.sort_values(by=['Casos activos'],
ascending=False).head(5))
```

```
plt.show()

fig, ax = plt.subplots()
ax.set_title("Países con mayor número de casos criticos")
sns.barplot(x = 'Casos criticos',
            y = 'Pais',
            orient = 'h',
            data = df.sort_values(by=['Casos criticos'],
ascending=False).head(5))
plt.show()
```