PRAC1 Fundamentos de la ciencia de los datos

Daniel Gerald Orbegoso Barrantes
12.04.2021

Descripción

Se ha recolectado un conjunto de datos relacionados con la pandemia del COVID-19, con información a nivel mundial del impacto que ha tenido sobre la población de los distintos países. Algunas de las variables que se recogen en el dataset son el número de casos totales, muertes totales, personas recuperadas o nuevos casos, entre otros.

Imagen identificativa

Contexto

La página https://www.worldometers.info/ ha sido elegida porque proporciona estimaciones y estadísticas en tiempo real de diversos temas. Uno de estos temas actuales hoy en día es el COVID-19, tema que se va a usar para llevar a cabo la creación del dataset.

Titulo del dataset

El título escogido para el dataset ha sido estadisticasCorona.csv

Contenido

- **Nr:** Hemos usado un numero como identificador para la información de los distintos países.

- **Country:** Nombre del país del cual se está mostrando la información.
- **Total Cases:** Numero total de casos del país del cual estamos mostrando la información.
- **New Cases:** Nuevos casos de infectados en el país desde ayer.
- **Total Deaths:** Muertes totales del país.
- **New Deaths:** Nuevas muertes en el país desde ayer.
- **Total Recovered:** Gente total recuperada.
- **New Recovered:** Nuevas personas recuperadas desde ayer
- **Active Cases:** Casos activos de coronavirus.
- **Serious/Critical:** Casos críticos en el país de coronavirus
- **Tot Cases per 1M Pop:** Casos de coronavirus totales por cada millón de personas.
- **Total Tests:** Número total de tests realizados.
- **Tests per 1M pop:** Número total de casos por cada millón de habitantes.
- **Population:** Población del pais.
- **1 Case every X ppl:** Un caso por X población.
- **1 Death every X ppl:** Una muerte por X población.
- **1 Test every X ppl:** Un test por X población.

Esta información ha sido recopilada de distintas fuentes, como las webs oficiales del ministerio de salud o las instituciones gubernamentales de distintos países, live streams de prensa, noticias, etc. las 24 horas del día, los 7 días de la semana. La información la podemos constatar en la página web https://www.worldometers.info/coronavirus/about/

En ella se nos informa que ciertos valores como *Total Recovered* puede ser algo impreciso debido a falta de reportes de parte de los distintos gobiernos, incompletos o incorrectos.

Agradecimientos

Los datos han sido recopilados desde la website *Worldometers*. Se ha realizado un código en Python para extraer dicha información y las técnicas de Web Scraping aprendidas en clase para extraer la información mediante el uso de librerías como *BeautifulSoup*.

Inspiración

Los datos han sido recopilados desde la website *Worldometers*. Se ha realizado un código en Python para extraer dicha información y las técnicas de Web Scraping aprendidas en clase para extraer la información mediante el uso de librerías como *BeautifulSoup*. Esta extracción de datos se ha realizado siguiendo análisis similares realizados para otros parámetros en esta página sobre la población, como la población mundial total actual que se puede encontrar bajo el siguiente enlace : https://www.worldometers.info/world-population/

Inspiración

En la situación actual que estamos viviendo, podríamos dar un gran uso a este dataset en muchos ámbitos y sectores y para distintos casos de uso. Tanto las universidades podrían implementarlo en distintos estudios debido al carácter actual que tiene el dataset, como por ejemplo empresas del ámbito turístico, las cuales podrían usarlo para comprender que países

tienen unas perspectivas más positivas en cara al verano para implantar una estrategia de marketing.

Si se deseara, se podría recolectar información en live de esta página web, y crear un dataset mayor para poder aplicar técnicas de minería de datos, creando distintos modelos para predecir variables de nuestro interés, como puede ser la evolución de contagios por habitantes. De esta manera podría la empresa turística aconsejar a los clientes que vuelen a distintos destinos de las precauciones que deberían tomar en cada zona durante su estancia.

Licencia

He considerado que la licencia adecuada para este dataset sería la licencia **CC BY-SA 4.0** (Reconocimiento – Compartir Igual). Esto se debe a los siguientes motivos:

- La capacidad de poder compartir esta información con otras personas, sin necesidad de pedir permisos para realizar cambios o crear aportaciones al dataset ya creado.
- Se debe reconocer siempre al autor de dicha obra e indicar las modificaciones realizadas, usando la misma licencia sobre la nueva obra producida.
- Se puede usar de manera comercial, en el caso de que alguna empresa desee el uso de los datos, podrán ser implementados en cualquier proyecto.
- Este tipo de licencia impulsa la cultura libre y la ideología del Open Data, algo que ha sido de gran ayuda para combatir la pandemia del COVID-19

Puesto que es un tema actual y la información que posee en ciertos casos puede ser necesario implementarlo directamente, pienso que este tipo de licencia es la adecuada.

Código fuente y dataset

El código realizado para extraer la información se presenta en el siguiente repositorio:

- https://github.com/danielGOB/TCVD.git

En él podemos encontrar un fichero *README.md* con una pequeña descripción del contenido del repositorio y la finalidad del proyecto.

A parte también tendremos una carpeta *PRAC1*, la cual contiene tres subcarpetas:

- **src**: código Python usado en la práctica.
- **pdf**: pdf con las respuestas a las preguntas planteadas.
- **csv**: dataset generado mediante el código usado en Python.

Podremos encontrar el dataset en Zenodo bajo el identificador ***10.5281/zenodo.4682077***

Contribuciones	Firma
Investigación previa	D.O
Redacción de las respuestas	D.O
Desarrollo código	D.O