**M2.851 -**

**BRAIS SANTOS NEGREIRA**

**JORGE ÁLVAREZ GRACIA**

**Aula: M2.851**

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

A la hora de elegir la información a recoger hemos seguido las siguientes premisas:

* Datos dinámicos.
* Datos dentro de la web bien estructurados.
* Datos cuyo método de obtención más eficiente sea el web scrapping.
* Datos cuyo análisis tenga una utilidad práctica.

Por otro lado y en referencia al último punto, tras varios años de pandemia del covid-19 que ha afectado a toda la población mundial, y donde los análisis de datos para su seguimiento y prevención han sido imprescindibles, pensamos que realizar una auditoria de los distintos lugares y sus determinadas características es fundamental para poder conocer que grupos de países con sus respectivas características son aquellos que mejor han combatido la enfermedad.

Para realizar este data set hemos elegidos varios sitios web como fuente de los datos:

Estas son:

* <https://www.worldometers.info/coronavirus/>

Esta es la página web donde se sustenta la información más relevante de nuestro proyecto, la hemos elegido debido a que:

* Lleva a cabo actualizaciones de datos de manera constante
* Los datos que necesitamos esta ordenados en una tabla web
* <https://datosmacro.expansion.com/otros/coronavirus-vacuna/>
* <https://datosmacro.expansion.com/pib>

Hemos elegido estas webs para comentar con información relevante el dataset. Al igual que para la anterior, las hemos elegido por su tipología y su constante actualización.

1. **Título.** Definir un título que sea descriptivo para el dataset.

El titulo elegido para el dataset es: covid\_19.*csv (Tal vez lo podríamos renombrar con un nombre más específico)*

1. **Descripción del dataset.** Desarrollar una descripción breve del conjunto de datos que se ha extraído. Es necesario que esta descripción tenga sentido con el título elegido.

En el dataset covid-19 encontramos tres tipos de variables diferenciados:

* El primer grupo hace referencia a los datos totales relacionados con los casos de covid-19 , muertes y personas recuperadas por países y con su respectiva población.
* En el segundo grupo, encontramos aquellas variables que tienen o podrían tener relación directa con la incidencia del virus y su respuesta ante este. En un primer momento, hemos seleccionado, variables relacionadas con la vacunación y el producto interior bruto de los países. Es este grupo el que hace nuestro proyecto escalable y donde podríamos ir añadiendo más variables para poder realizar modelos de minería de datos mas eficaces. *(Si se considera que el proyecto no es lo suficiente completo en relación con la tarea de web scrapping)*
* En el tercer grupo, encontramos aquellas variables que hemos obtenido a través de la combinación de las anteriores y que serán las más representativas del dataset y nos servirán para realizar los análisis necesarios.

1. **Representación gráfica.** Dibujar un esquema o diagrama que identifique el dataset visualmente y el proyecto elegido.

*Realizar el diagrama cuando avabemos el dataet*

1. **Contenido.** Explicar los campos que incluye el dataset, el periodo de tiempo de los datos y cómo se ha recogido.

*A continuación, vamos a describir cada variable del dataset:*

*Describir las variables cuando acabemos el dataset*

1. **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. Justificar qué pasos se han seguido para actuar de acuerdo a los principios éticos y legales en el contexto del proyecto.

Los propietarios de los datos obtenidos son los siguientes:

Worldometer, para las métricas relacionadas con el covid-19.

Para más detalle, https://www.worldometers.info/coronavirus/about/

En el caso de las vacunaciones por países y del producto interior bruto, los datos pertenecen a <https://datosmacro.expansion.com/>,

Adjuntamos una captura de los archivos robot.txt de las diversas páginas web utilizadas. *(No lo encuentro)*

1. **Inspiración.** Explicar por qué es interesante este conjunto de datos y qué preguntas se pretenden responder. Es necesario comparar con los análisis anteriores presentados en el apartado 6.

Identificar aquellos países con características similares que han conseguido contener de manera más o menos eficiente el covid-19, así como aquellos países donde la vacunación ha sido más rápida.

Realizar un ranking de las distintas variables por países enterminos absolutos y por habitante

*(Añadir más)*

1. **Licencia.** Seleccionar una de estas licencias para el dataset resultante y justificar el motivo de su selección:
   * Released Under CC0: Public Domain License.
   * Released Under CC BY-NC-SA 4.0 License.
   * Released Under CC BY-SA 4.0 License.
   * Database released under Open Database License, individual contents under Database Contents License.
   * Other (specified above).
   * Unknown License.
2. **Código.** Adjuntar en el repositorio Git el código con el que se ha generado el dataset, preferiblemente en Python o, alternativamente, en R.

Código:

<https://github.com/bsantosn/COVID-19_WEB_SCRAPPING.git>

Ejecución del programa:

$ git clone <https://github.com/bsantosn/COVID-19_WEB_SCRAPPING.git>

$ cd COVID-19\_WEB\_SCRAPPING

$ pip3 install -r requirements.txt

$ python3 covid\_19.py

1. **Dataset.** Publicar el dataset obtenido*(\*)* en formato CSV en Zenodo con una breve descripción. Obtener y adjuntar el enlace del DOI.

* <https://zenodo.org/record/5594958#.YXVFR55BxhE>

|  |  |
| --- | --- |
| **Contribuciones** | **Firma** |
| Investigación previa | Integrante 1, Integrante 2 |
| Redacción de las respuestas | Integrante 1, Integrante 2 |
| Desarrollo del código | Integrante 1, Integrante 2 |

Enlace Video