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 variables en el tiempo.
   - Datos dentro de la web bien estructurados.
   - Datos cuyo método de obtención más eficiente sea el web scraping.
   - 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 auditoría 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 están ordenados en una tabla web

   - https://datosmacro.expansion.com/otros/coronavirus-vacuna/
   - https://datosmacro.expansion.com/pib

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

De forma completaría y para profundizar en técnicas de web scraping hemos añadido la variable *temperatura media,* la cual podríamos a ver obtenido aplicando técnicas más simples.

Para ello. hemos creado un script que accede a la página de inicio de sesión de Wikipedia y se identifica como usuario, posteriormente utiliza el menú de búsqueda para acceder a la web de Wikipedia que necesitamos y de ahí, obtenemos los datos.

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

   El titulo elegido para el dataset es: Covid 19 Dataset, Cases by Country, Vaccination, GDP and Average Temperature

3. **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 o la respuesta ante este. Hemos seleccionado, variables relacionadas con la vacunación, el producto interior bruto y la temperatura media de los países. Es este grupo el que hace nuestro proyecto escalable y en él podríamos ir añadiendo más variables para poder realizar modelos de minería de datos más eficaces.

   - Para la siguiente PRAC, añadiríamos un tercer grupo donde encontraríamos aquellas variables que obtendremos a través de la combinación de las anteriores, que serán las más representativas del dataset y nos servirán para realizar los análisis necesarios.

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

   Ver Anexo 1 o archivo Esquema gráfico.png

5. **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:

Variable	Descripción	Periodo de los datos	Fuente
Country	Países	Diario	1*
Total Cases	Número Total de Casos del Covid-19	Diario	
New Cases	Nuevos Casos	Diario	
Total Deaths	Total muertes causadas por Covid-19	Diario	
New Deaths	Nuevas muertes	Diario	
Total Recovered	Recuperados Totales	Diario	
New Recovered	Nuevos Recuperados	Diario	
Active Cases	Casos Activos	Diario	
Serious Critical	Personas en situación crítica	Diario	
Deaths 1M pop	Muertes por 1M de habitantes	Diario	
Total Tests	Test Totales realizados	Diario	
Population	Población	Diario	
Continent	Continente	Diario	
% Deaths COVID/Population	% de Muertes por covid-19 respecto a la población total	Diario	
Date vaccinated update	Fecha de actualización de los datos de vacunaciones	Diario	2*
Fully vaccinated	Personas vacunadas	Diario	
% Fully vaccinated	% de personas vacunadas respecto al total de la población	Diario	

Year_GDP_update	Año de actualización del PIB	Anual	3*
GDP_Annual (M)	PIB Anual	Anual	
Temperature_1961_1990	Temperatura media	Sin actualización desde 1990	4*

1*: https://www.worldometers.info/coronavirus/
2* https://datosmacro.expansion.com/otros/coronavirus-vacuna/
3*:https://datosmacro.expansion.com/pib
4*:https://en.wikipedia.org/wiki/List_of_countries_by_average_yearly_temperature

En total el dataset se compone de 20 variables, de distintas fuentes y todas ellas obtenidas mediante técnicas de web scraping.

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. 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/

Para los datos relacionados con las vacunaciones por países y del producto interior bruto, los datos pertenecen a https://datosmacro.expansion.com/.

En el caso de los datos obtenidos de la temperatura pertenecen a Wikipedia

Debido a la importancia del Covid-19 durante estos últimos años existen numerosos estudios de este tema.

Entre ellos, podemos encontrar el estudio del SERGAS (Servizo galego de saúde) https://coronavirus.sergas.gal/datos/#/gl-ES/galicia. En su página web muestra información referente a casos nuevos, total de muertes y total de contagios (entre otros campos).

Adjuntamos una captura de los archivos robot.txt de las diversas páginas web utilizadas.

1. https://datosmacro.expansion.com/otros/coronavirus-vacuna/ y https://datosmacro.expansion.com/pib

Para obtener la información referente a las vacunas y PIB, hemos accedido a la siguiente url: https://datosmacro.expansion.com/robots.txt. Se ha comprobado que las rutas a las que accedemos están habilitadas y por lo tanto cumplimos con los principios éticos y legales.

```
Sitemap: https://datosmacro.expansion.com/sitemap.xml.gz
Sitemap: https://datosmacro.expansion.com/sitemap1.xml.gz
User-agent: *
# CSS, JS, Images
Allow: /core/*.css$
Allow: /core/*.css?
Allow: /core/*.js$
Allow: /core/*.js?
Allow: /core/*.gif
Allow: /core/*.jpg
Allow: /core/*.jpeg
Allow: /core/*.png
Allow: /core/*.svg
Allow: /profiles/*.css$
Allow: /profiles/*.css?
Allow: /profiles/*.js$
Allow: /profiles/*.js?
Allow: /profiles/*.gif
Allow: /profiles/*.jpg
Allow: /profiles/*.jpeg
Allow: /profiles/*.png
Allow: /profiles/*.svg
# Directories
Disallow: /core/
Disallow: /profiles/
# Files
Disallow: /README.txt
Disallow: /web.config
# Paths (clean URLs)
Disallow: /admin/
Disallow: /comment/reply/
Disallow: /filter/tips
Disallow: /node/add/
Disallow: /search/
Disallow: /user/register
Disallow: /user/password
Disallow: /user/login
Disallow: /user/logout
Disallow: /node/
# Paths (no clean URLs)
Disallow: /index.php/admin/
Disallow: /index.php/comment/reply/
Disallow: /index.php/filter/tips
Disallow: /index.php/node/add/
Disallow: /index.php/search/
Disallow: /index.php/user/password
Disallow: /index.php/user/register
Disallow: /index.php/user/login
Disallow: /index.php/user/logout
User-agent: MauiBot (crawler.feedback+wc@gmail.com)
Disallow: /
```

2. Para las páginas referentes a casos de coronavirus y temperatura de media de países no se ha encontrado nada referente a robots.txt.

7. **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.

   A continuación, enumeramos las principales respuestas que queremos responder con este estudio:

   - 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 en términos absolutos y por habitante que nos aporten información relevante sobre el covid-19.

   - Comprobar si la riqueza de un país influye y en qué medida en los casos y características del covid-19, así como la temperatura media

   A diferencia del ejemplo descrito en el apartado anterior, nuestro estudio esta enfocado a nivel global. Por otro lado, nuestro modelo esta enfocado para en un futuro poder ir añadiendo nuevas variables de forma que sea escalable y podamos llevar a cabo análisis de minería de datos más profundos.

8. **Licencia.** Seleccionar una de estas licencias para el dataset resultante y justificar el motivo de su selección:

   La licencia escogida para la publicación de este conjunto de datos ha sido **CC BY-SA 4.0 License** ya que tanto los derechos y las restricciones de su uso nos parecen adecuados.

   Entre los derechos de uso encontramos:

   - Libertad de uso y de desarrollo
   - Libertad de uso comercial
   - Libertad de distribuir obras parecidas o similares

   Las restricciones de esta licencia son:

   - No se debe restringir el acceso al trabajo utilizando medidas técnicas, o intentar imponer limitaciones a las libertades mencionadas anteriormente.
   - Se debe otorgar la debida atribución al autor y conservar el aviso de licencia.
   - Se debe publicar trabajos derivados bajo una licencia idéntica o similar.

9. **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

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

    - https://doi.org/10.5281/zenodo.5635766

Anexo 1


Esquema Gráfico del Proyecto
Web Scraping
https://www.worldometers.info/
Datos Covid-19
Datos Población
https://datosmacro.expansion.com/
Datos Población
Datos GDP
https://en.wikipedia.org/..
Datos Temperatura
DataSet
Country
Total Cases
New Cases
Total Deahts
Total Recovered
New Recovered
Active Cases
Serious Critical
Deaths 1M pop
Total Tests
Population
Continent
% Deaths COVID/Population
Date vaccinated update
Fully vaccinated
% Fully vaccinated
Year_GDP_update
GDP_Annual (%)
Temperature_1961_1990