Nuria García Fuentes y Alicia Perdices Guerra

Tipología de datos. Práctica\_1 Estudio preliminar

**TIPOLOGÍA DE DATOS. PRÁCTICA 1\_ESTUDIO PRELIMINAR**

# **Descripción de la Práctica.**

En el contexto de la práctica propuesta, y motivadas sobre todo por las ventajas y beneficios que el estudio de datos procedentes de enfermedades y de desastres naturales puede aportar a la comunidad científica, no solo para la prevención, sino para establecer protocolos, hemos decidido estudiar y hacer web scraping de los siguientes recursos:

[https://data.europa.eu/euodp/es/data/publisher/ecdc](https://www.google.com/url?q=https://data.europa.eu/euodp/es/data/publisher/ecdc&sa=D&source=hangouts&ust=1584813221920000&usg=AFQjCNFXaBDxR7QEoA9NpKMW7cIcEjJQbQ)

<https://reliefweb.int/disasters>

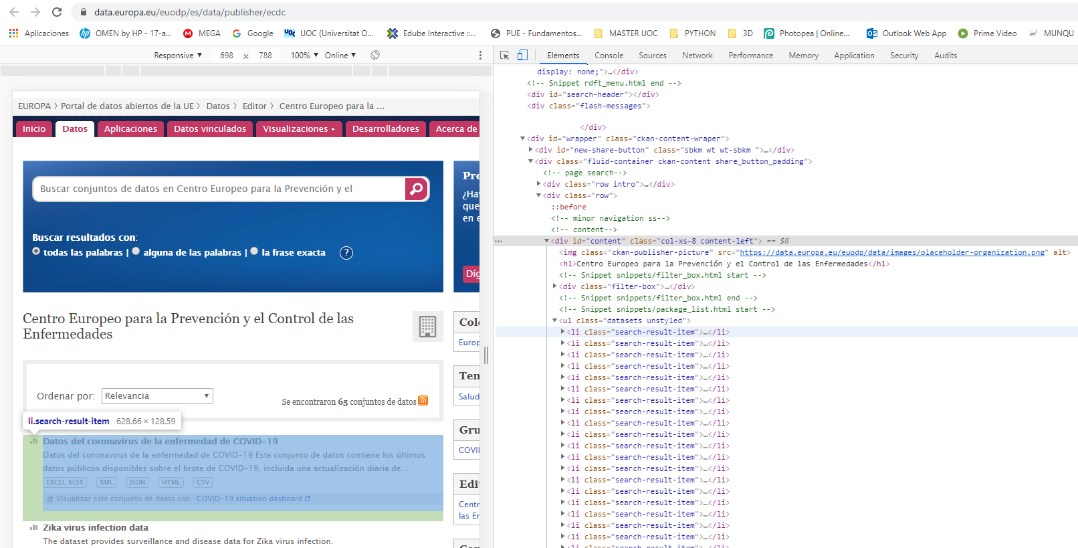
Posteriormente, a partir de los datos obtenidos, nos decantaremos por un estudio u otro. Creíamos interesante ver las dos formas de obtener datos y valorar qué datos obtenidos podrían darnos más información o que tema nos interesaría explotar más.

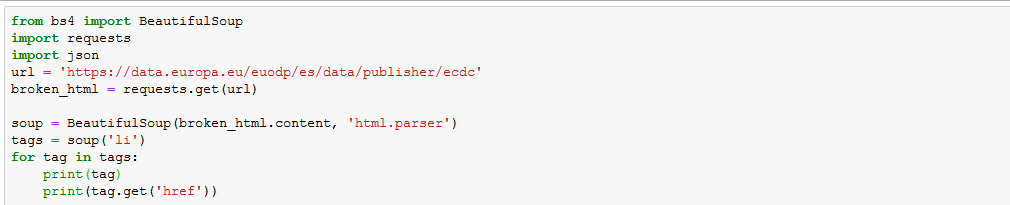
Expliquemos en que consiste cada caso:

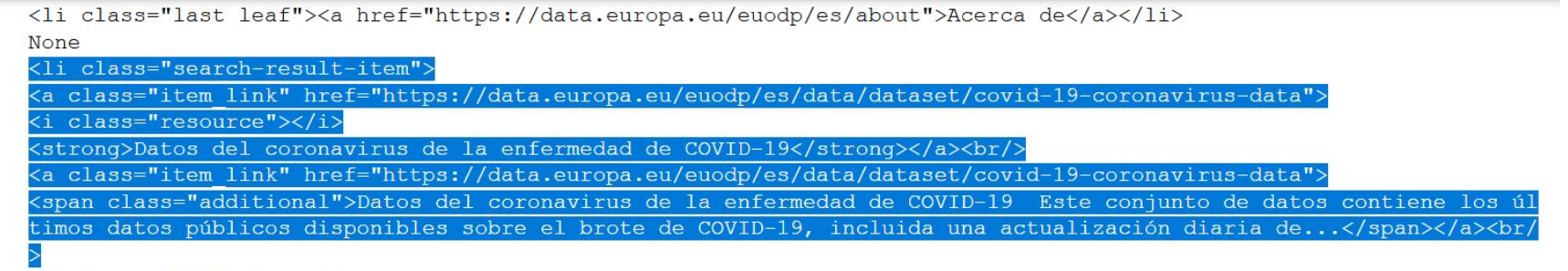
**Caso enfermedades: Descarga de un archivo xlsx**

De este modo, y con el fin, de descargar, analizar y organizar estos datos de interés, hemos completado el proceso de web scraping de la siguiente forma:

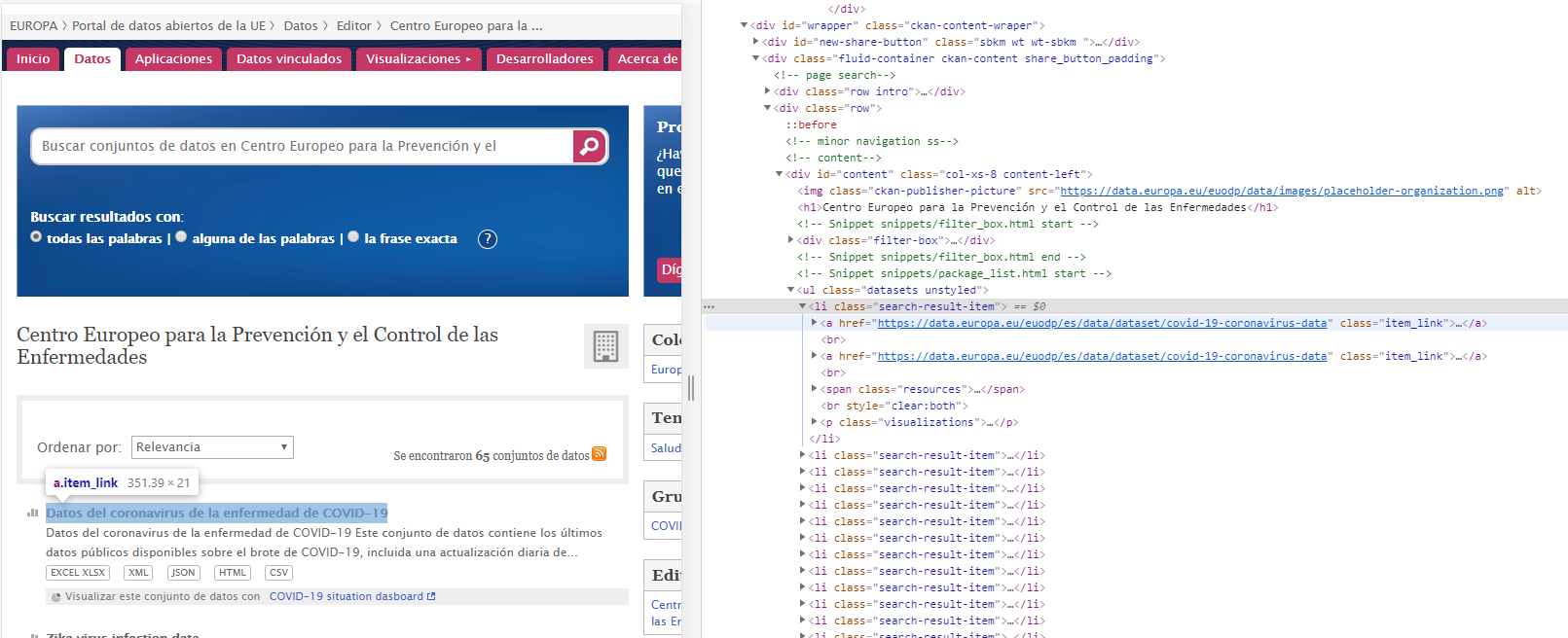
* Se ha evaluado el mapa de los sitios web para localizar el contenido del recurso, poniendo el foco en la estructura de la página, y así poder averiguar dónde y cómo descargar los datos que necesitamos para el estudio.
* Se ha convertido la página en HTML a estructura anidada.
* A continuación, se ha descargado la página web mediante las librerías **Requests** y **BeautifulSoup** en Python.
* Tras examinar las etiquetas, tags, clases de la página, hemos navegado por la estructura hasta el enlace que contiene los datos para posteriormente descargarlos y analizarlos.
  + Obtenemos las etiquetas **<li> con class: search-result-item,** que es donde tenemos las enfermedades y sus links.





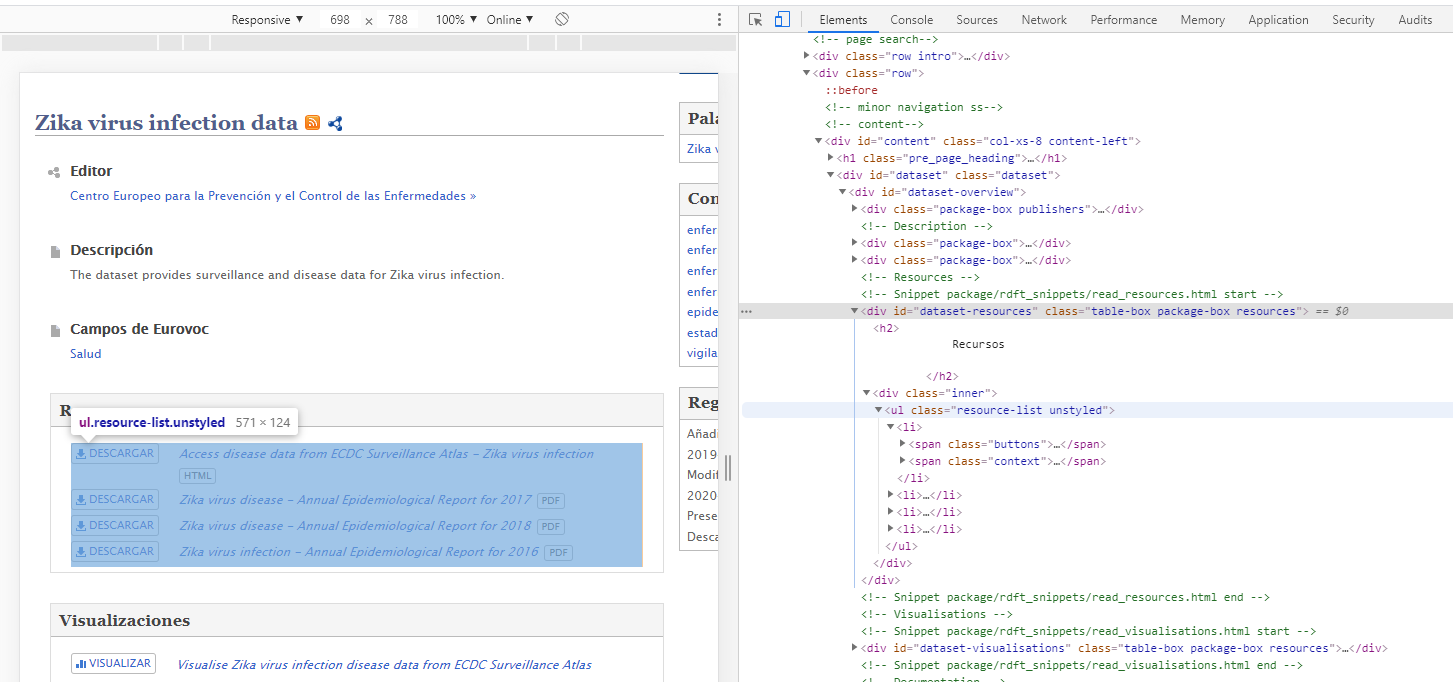


* + Repetimos el proceso para cada link de enfermedad elegida con **tag <a>** y **clase 'item\_link'**.

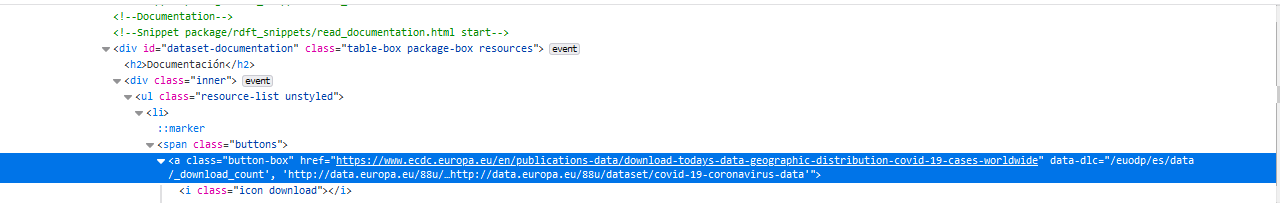


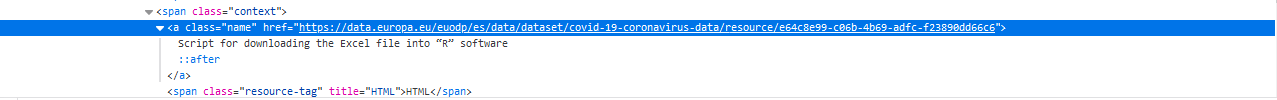


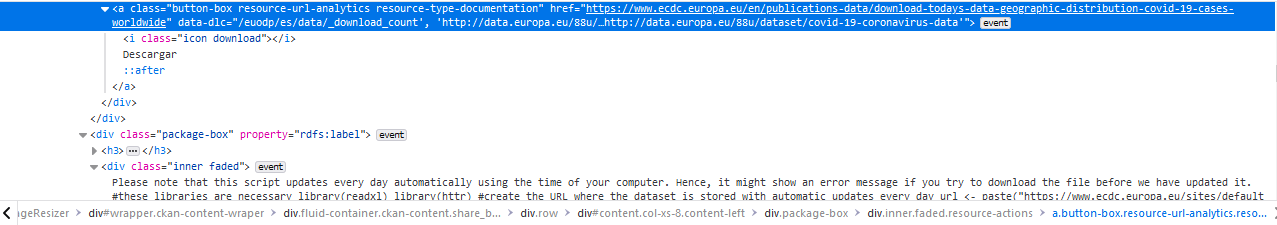
* + Recuperamos las entradas donde tenemos los links de descarga con tag **<ul>** y clase **'resource-list unstyled'**

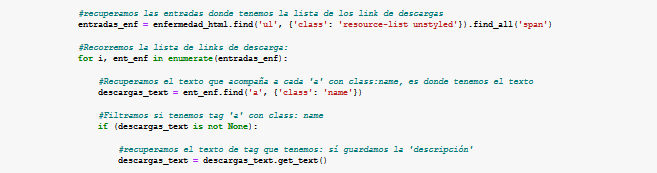


* + Recorremos la lista de links de descarga **con tag <a>** y **class =’name’**, accediendo primero a la descripción.

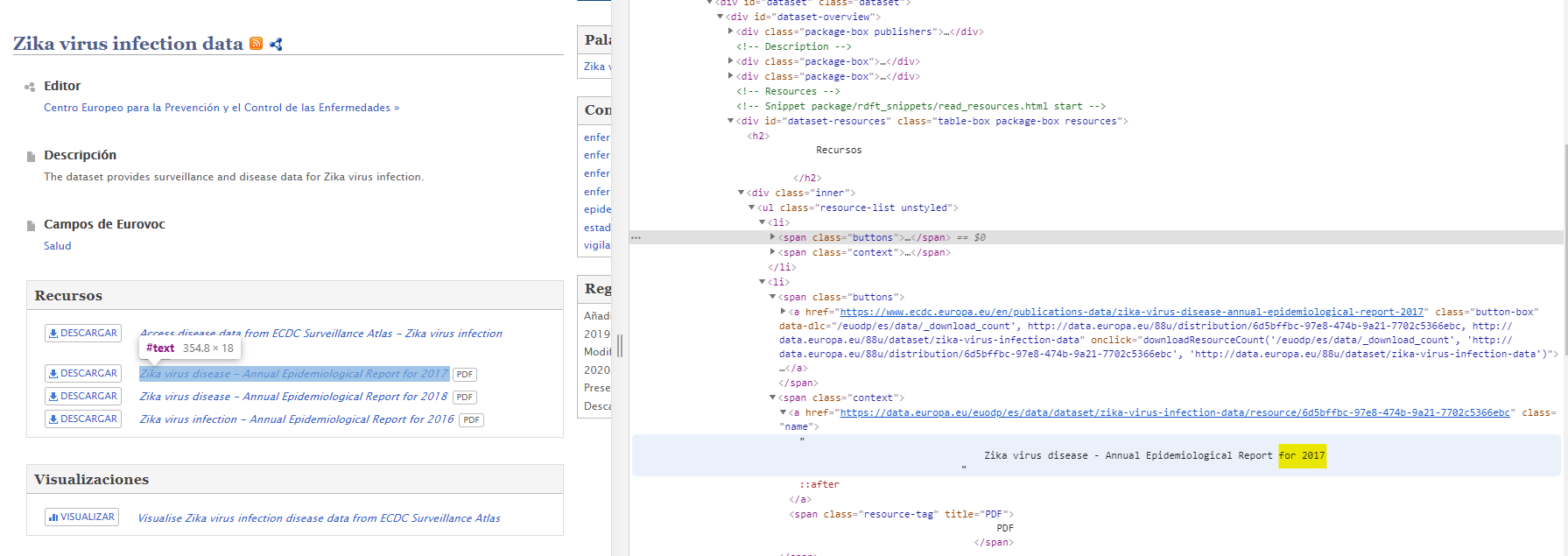


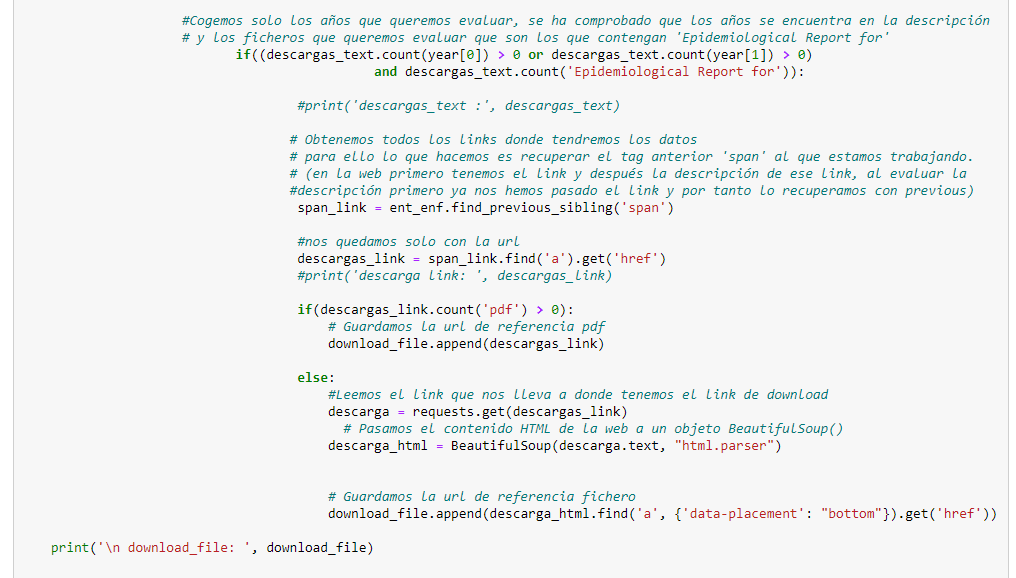






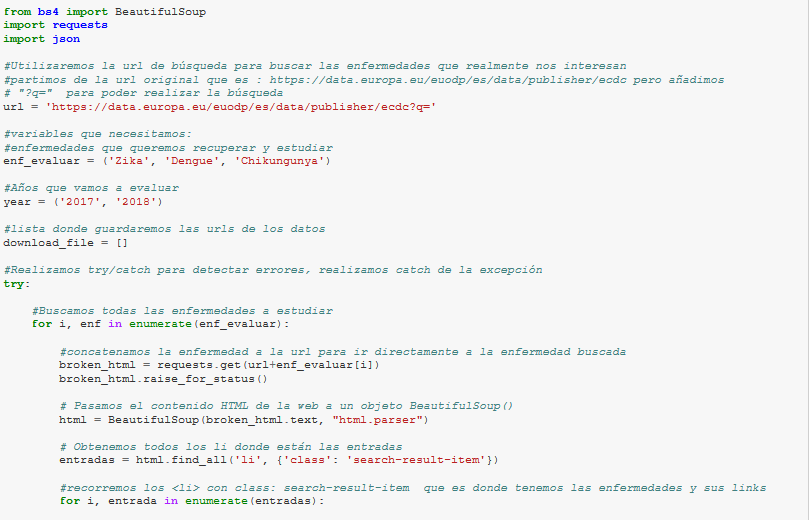
* + Obtenemos los años que queremos evaluar, dato que se encuentra en la descripción. A continuación, obtenemos con **previous\_sibling** los links donde se encuentras los datos (**tag ‘span’ anterior** **a esta descripción**).

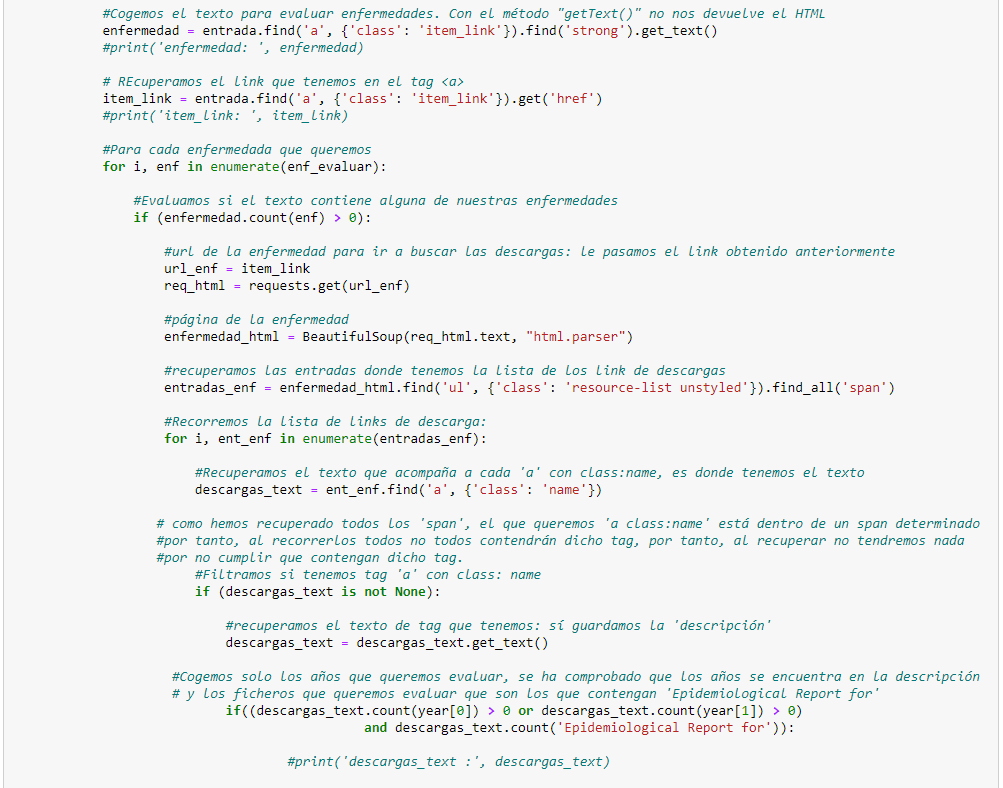




* Una vez obtenidos los enlaces desde los que nos descargaremos los datos, procedemos con su descarga y su análisis.

En definitiva, tenemos:



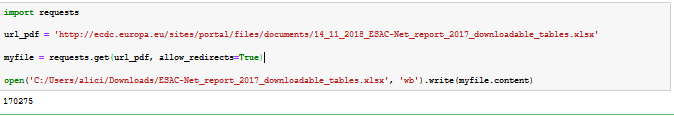




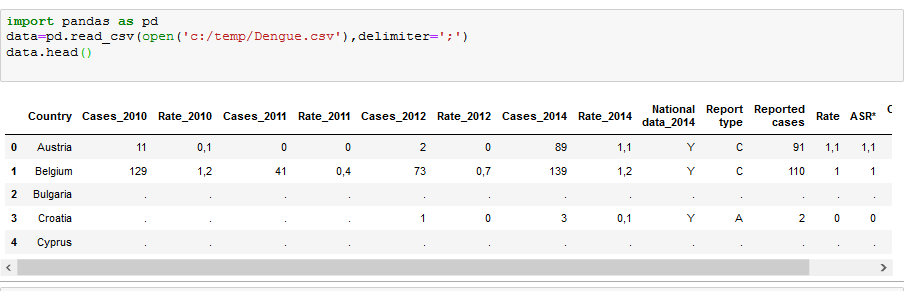
A continuación, vamos a hacer un pequeño análisis para los datos obtenidos del enlace de la enfermedad del **dengue (ESAC-Net\_report\_2017\_downloadable\_tables.xlsx, Dengue\_1.csv)**.

**ARCHIVO XLSX**

* Descargamos los datos en local.



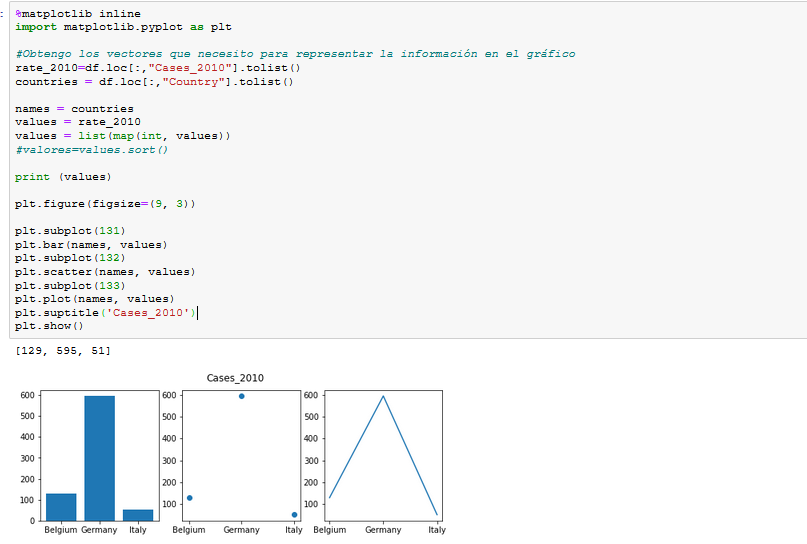
* Importamos el archivo descargado:



* Creamos un dataframe para hacer una comparativa de los casos de dengue en Italia, Alemania y Bélgica en 2010.







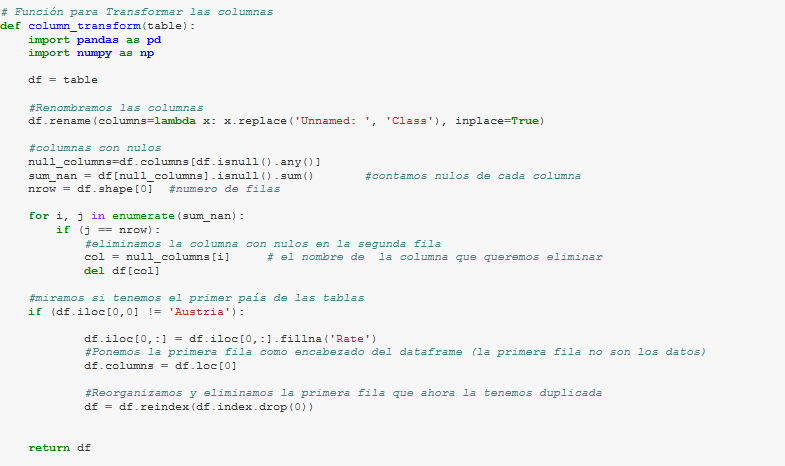


Podemos comprobar que en Alemania se han diagnosticado más casos de Dengue que en Italia y Bélgica. Observamos también que en 2011 hay un descenso de los casos en todos los países.

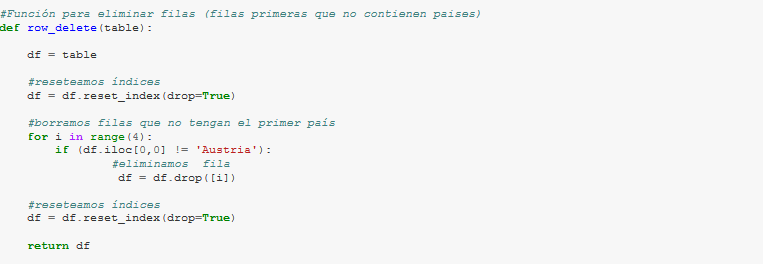
**ARCHIVO PDF**

Para poder hacer uso de los archivos descargados para su posterior análisis, vamos a crear una serie de funciones de transformación.

* Función para transformar las columnas.



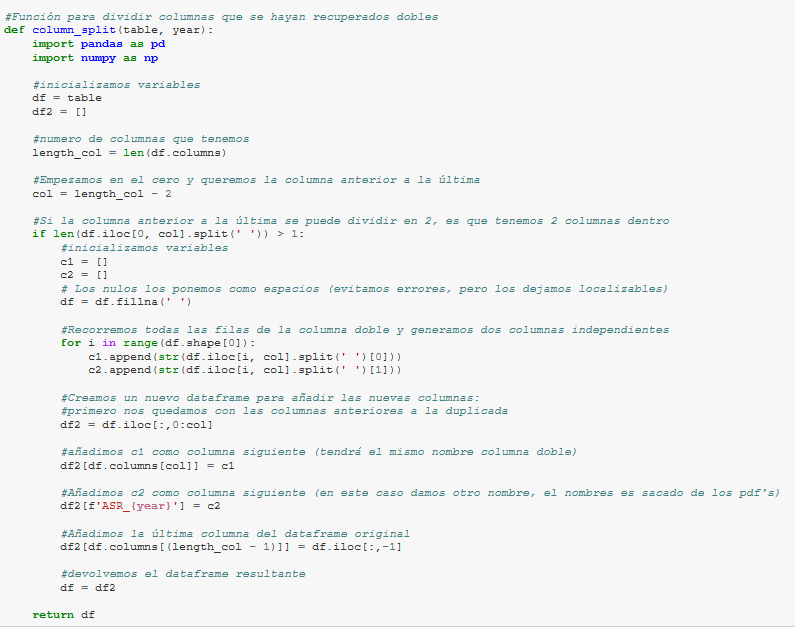
* Función para eliminar filas (filas primeras que no contienen países)



* Función para renombrar las columnas iguales con el año que le toca.



* Función para dividir columnas que se hayan recuperado dobles

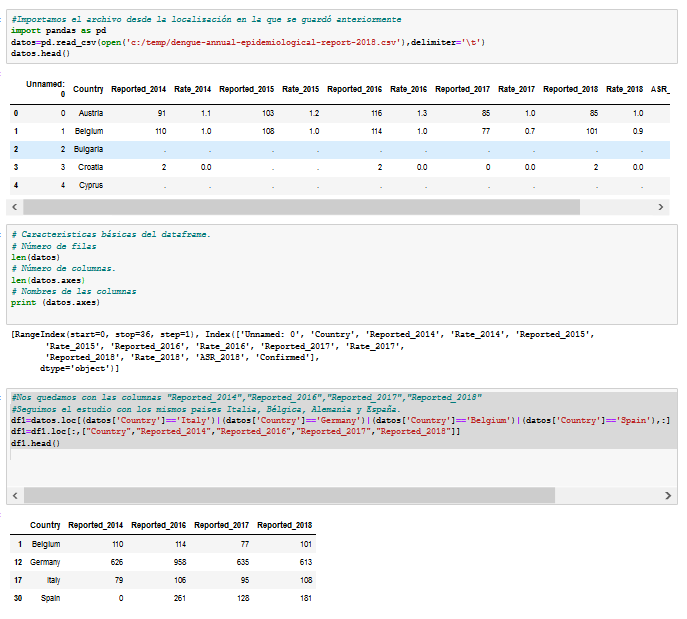


* Recuperamos la tabla.

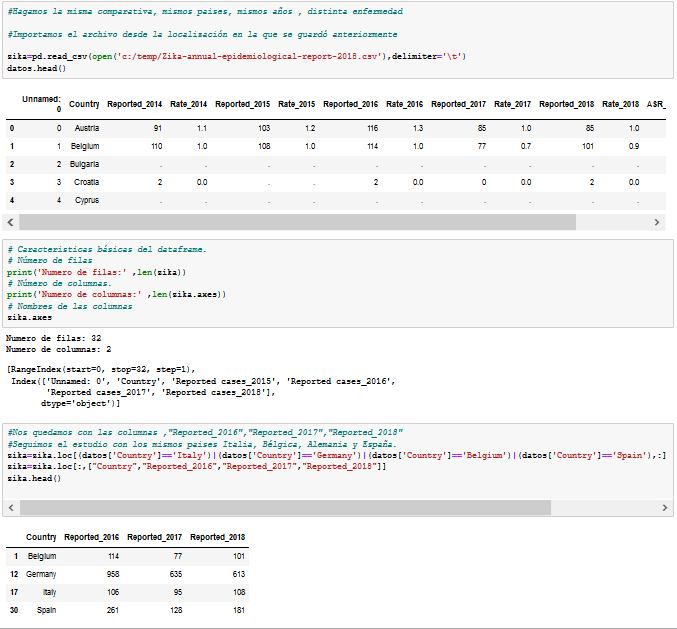


A continuación, vamos a hacer un pequeño análisis para los datos obtenidos del enlace de la enfermedad del **dengue y zika,** pero esta vez para los años 2016, 2017 y 2018 **(dengue-annual-epidemiological-report-2018.csv, Zika-annual-epidemiological-report-2018.csv)**.

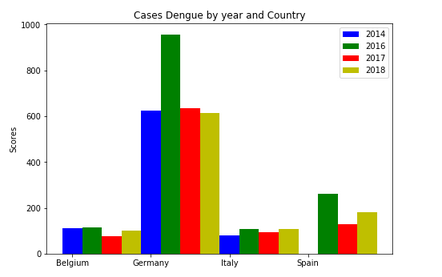
* Importamos los archivos con pandas.
* Exploramos el dataset, obteniendo número de filas, columnas y nombre de atributos.
* Seleccionamos años y países para poder comparar infecciones de zika y dengue.
* Representamos la información en varios gráficos, obteniendo:

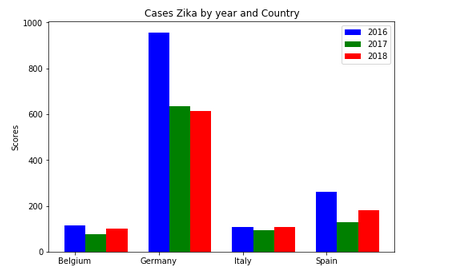












En Alemania y España los números pacientes diagnosticados de zika y dengue son muy similares, entre los años 2016 y 2018 contando Alemania con un mayor número de infectados en cómputo global. Observamos que en 2016 hubo un repunte de infecciones. Bélgica e Italia son los países que cuentan con menor número de infectados en el periodo de tiempo estudiado.

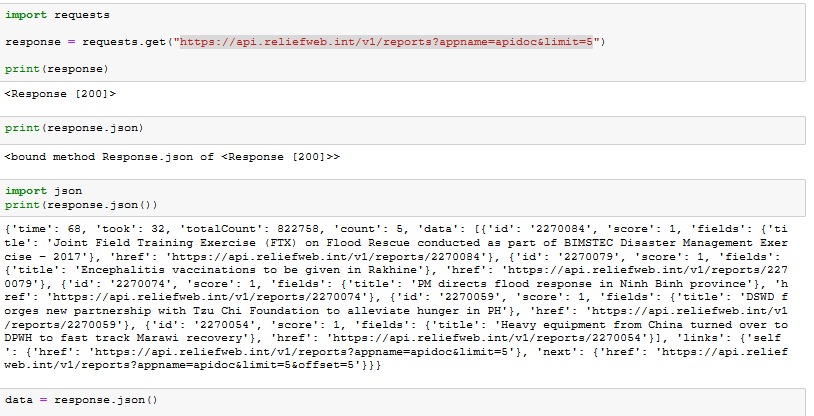
**Caso desastres naturales. Descarga pdf con mapa de la actividad del volcán Vanuatu en diciembre de 2018.**

En este caso utilizaremos las técnicas del web scraping, para obtener imágenes sobre elementos geográficos que puedan ser potenciales áreas peligrosas.

Por ejemplo, el volcán Monte Yasur que se encuentra en la isla de Tanna, y forma parte del estado de Vanuatu en el Pacífico Sur, es uno de los pocos volcanes activos que podemos encontrar. Nos podría interesar obtener mapas de actividad, para establecer protocolos de salvamento en el área.

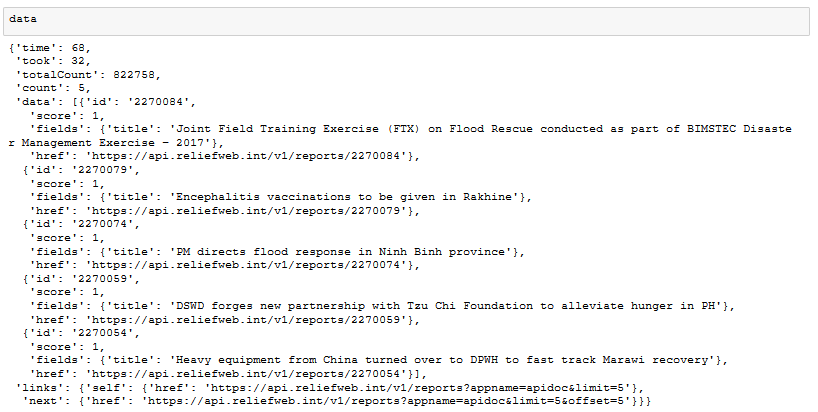
Para ello y desde el recurso <https://reliefweb.int/disasters>, comenzamos el proceso de web scraping del siguiente modo:

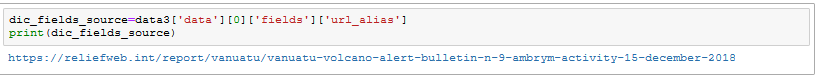
Comprobamos que todo el contenido de la página ReliefWeb es accesible mediante su API (<https://apidoc.rwlabs.org/>). Es de acceso público mediante solicitudes HTTP y devuelve datos JSON.



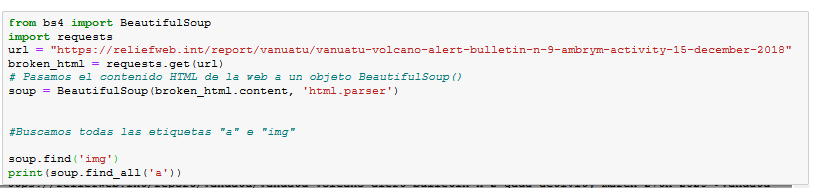
A continuación, navegamos por la estructura JSON, que consta de una colección de pares nombre /valor (objeto) y una lista ordenada de valores (array), hasta obtener el enlace con la información deseada.

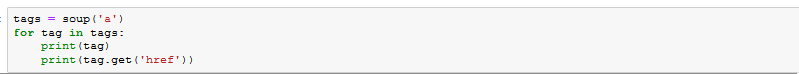
(Ver archivo **web\_scarping\_desastres\_Garcia\_Perdices\_Prac1.ipynb)**





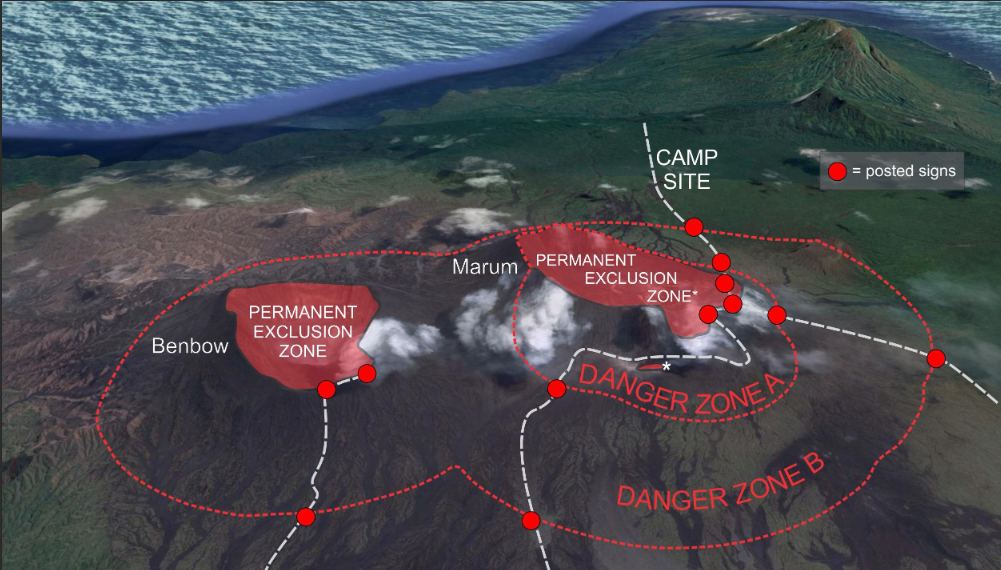
Después, paseamos el contenido HTML a un objeto BeautifulSoup(). Buscamos todos **los enlaces e imágenes** con etiqueta **‘a’**.





Y, por último, descargamos el enlace con el mapa que nos interesa.





Para información detallada del scraping, revisad archivo: **web\_scarping\_desastres\_Garcia\_Perdices\_Prac1.ipynb**

Descripción de los datos.

Datos de enfermedades obtenidos de la página del **Centro Europeo para la Prevención y el Control de las Enfermedades**:

<https://data.europa.eu/euodp/es/data/publisher/ecdc>

* **Zika-annual-epidemiological-report-2018**

|  |  |  |
| --- | --- | --- |
| Atributos | Descripción | Tipo de datos |
| Country | País analizado | String |
| Reported\_cases\_2015 | Casos de zika reportados en el año 2015 | String |
| Reported\_cases\_2016 | Casos de zika reportados en el año 2016 | String |
| Reported\_cases\_2017 | Casos de zika reportados en el año 2017 | String |
| Reported\_cases\_2018 | Casos de zika reportados en el año 2018 | String |

* **AER\_for\_2017-Zika-virus-disease**

|  |  |  |
| --- | --- | --- |
| Atributos | Descripción | Tipo de datos |
| Country | País analizado | String |
| 2015 Reported cases\_2015 | Casos de zika reportados en el año 2015 | String |
| 2015 Reported cases\_2016 | Casos de zika reportados en el año 2016 | String |
| 2015 Reported cases\_2017 | Casos de zika reportados en el año 2017 | String |

* **dengue-annual-epidemiological-report-2017**

|  |  |  |
| --- | --- | --- |
| Atributos | Descripción | Tipo de datos |
| Country | País analizado | String |
| Reported\_cases\_2013 | Casos de dengue reportados en 2013 | String |
| Rate\_2013 | Ratio dengue/ 100.000 hab en 2013 | String |
| Reported\_cases\_2014 | Casos de dengue reportados en 2014 | String |
| Rate\_2014 | Ratio dengue/ 100.000 hab en 2014 | String |
| Reported\_cases\_2015 | Casos de dengue reportados en 2015 | String |
| Rate\_2015 | Ratio dengue/ 100.000 hab en 2015 | String |
| Reported\_cases\_2016 | Casos de dengue reportados en 2016 | String |
| Rate\_2016 | Ratio dengue/ 100.000 hab en 2016 | String |
| Reported\_cases\_2017 | Casos de dengue reportados en 2017 | String |
| Rate\_2017 | Ratio dengue/ 100.000 hab en 2017 | String |
| Confirmed\_cases | Casos confirmados en 2017 | String |

* **dengue-annual-epidemiological-report-2018**

|  |  |  |
| --- | --- | --- |
| Atributos | Descripción | Tipo de datos |
| Country | País analizado | String |
| Reported \_2014 | Casos de dengue reportados en 2014 | String |
| Rate\_2014 | Ratio dengue/ 100.000 hab en 2014 | String |
| Reported \_2015 | Casos de dengue reportados en 2015 | String |
| Rate\_2015 | Ratio dengue/ 100.000 hab en 2015 | String |
| Reported \_2016 | Casos de dengue reportados en 2016 | String |
| Rate\_2016 | Ratio dengue/ 100.000 hab en 2016 | String |
| Reported \_2017 | Casos de dengue reportados en 2017 | String |
| Rate\_2017 | Ratio dengue/ 100.000 hab en 2017 | String |
| Reported \_2018 | Casos de dengue reportados en 2018 | String |
| Rate\_2018 | Ratio dengue/ 100.000 hab en 2018 | String |
| ASR\_2018 | Age standarized rate | String |
| Confirmed\_cases | Casos confirmados en 2018 | String |

* **chikungunya-virus-disease-annual-epidemiological-report-2017**

|  |  |  |
| --- | --- | --- |
| Atributos | Descripción | Tipo de datos |
| Country | País analizado | String |
| Reported \_2013 | Chikungunyas reportados en 2013 | String |
| Rate\_2013 | Chikungunyas / 100.000 hab en 2013 | String |
| Reported \_2014 | Chikungunyas reportados en 2014 | String |
| Rate\_2014 | Chikungunyas / 100.000 hab en 2014 | String |
| Reported \_2015 | Chikungunyas reportados en 2015 | String |
| Rate\_2015 | Chikungunyas / 100.000 hab en 2015 | String |
| Reported \_2016 | Chikungunyas reportados en 2016 | String |
| Rate\_2016 | Chikungunyas / 100.000 hab en 2016 | String |
| Reported \_2017 | Chikungunyas reportados en 2017 | String |
| Rate\_2017 | Chikungunyas / 100.000 hab en 2017 | String |
| Confirmed | Casos confirmados en 2017 | String |

* **chikungunya-annual-epidemiological-report-2018**

|  |  |  |
| --- | --- | --- |
| Atributos | Descripción | Tipo de datos |
| Country | País analizado | String |
| Reported \_2014 | Chikungunyas reportados en 2014 | String |
| Rate\_2014 | Chikungunyas / 100.000 hab en 2014 | String |
| Reported \_2015 | Chikungunyas reportados en 2015 | String |
| Rate\_2015 | Chikungunyas / 100.000 hab en 2015 | String |
| Reported \_2016 | Chikungunyas reportados en 2016 | String |
| Rate\_2016 | Chikungunyas / 100.000 hab en 2016 | String |
| Reported \_2017 | Chikungunyas reportados en 2017 | String |
| Rate\_2017 | Chikungunyas / 100.000 hab en 2017 | String |
| Reported \_2018 | Chikungunyas reportados en 2018 | String |
| Rate\_2018 | Chikungunyas / 100.000 hab en 2018 | String |
| ASR\_2018 | Age standarized rate | String |
| Confirmed | Casos confirmados en 2018 | String |

Datos de desastres obtenidos de la página del **Disasters. Alert and Ongoing Disasters.**

<https://reliefweb.int/disasters>

* **500\_registros\_desastres**

|  |  |  |
| --- | --- | --- |
| Atributos | Descripción | Tipo de datos |
| date\_changed | Fecha última modificación | String |
| date\_created | Fecha creación | String |
| date\_original | Fecha original | String |
| country\_href | URL pais de estudio | String |
| country\_name | Nombre pais de estudio | String |
| country\_location\_lon | Longitud | String |
| country\_location\_lat | Latitud | String |
| country\_id | Id del país | String |
| country\_shortname | Nombre abreviado del país | String |
| country\_iso3 | Nomenclatura iso3 del país | String |
| country\_primary | País principal | String |
| primary\_country\_href | URL pais principal de estudio | String |
| primary\_country\_name | Nombre del país principal de estudio | String |
| primary\_country\_location | Localización país principal | String |
| primary\_country\_id | Id del país principal | String |
| primary\_country\_shortname | Nombre abreviado país principal | String |
| primary\_country\_iso3 | Nomenclatura iso3 pais principal | String |
| format\_name | Formato de presentación | String |
| format\_id | Id formato de presentación | String |
| disaster\_type\_code | Código de tipo de desastre | String |
| disaster\_type\_name | Nombre tipo de desastre | String |
| disaster\_type\_id | Id tipo de desastre | String |
| language\_code | Código de idioma | String |
| language\_name | Nombre de idioma | String |
| language\_id | Id de idioma | String |
| source\_href | URL de la fuente | String |
| source\_longname | Nombre completo de la fuente | String |
| source\_name | Nombre de la fuente | String |
| source\_id | Id de la fuente | String |
| source\_type | Tipo de fuente | String |
| source\_shortname | Nombre de la fuente abreviado | String |
| source\_homepage | URL de la página de inicio | String |
| theme\_name | Nombre del tema | String |
| theme\_id | Id del tema | String |
| vulnerable\_groups\_name | Nombre de grupos vulnerables | String |
| vulnerable\_groups\_id | Id de grupos vulnerables | String |
| source\_spanish\_name | Nombre español de la fuente | String |
| file\_preview | URL previsualización | String |
| file\_filename | Nombre del archivo previsualización | String |
| file\_description | Descripción del archivo. | String |
| file\_mimetype | Tipo de archivo de previsualización | String |
| file\_id | Id del archivo de previsualización | String |
| file\_filesize | Tamaño (MB) archivo. | String |
| file\_url | URL previsualización | String |
| source\_disclaimer | Descargo de responsabilidad de la fuente | String |
| disaster\_name | Nombre del desastre | String |
| disaster\_id | Id del desastre | String |
| disaster\_glide | **Global identifier Number** del desastre | String |
| disaster\_type | Tipo de desastre | String |
| disaster\_status | Status del desastre | String |
| image\_copyright | Derechos de imagen | String |
| image\_filename | Nombre de la imagen | String |
| image\_url-thumb | URL de la imagen (thumb size) | String |
| image\_url-small | URL de la imagen (small size) | String |
| image\_width | Ancho de la imagen | String |
| image\_caption | Lugar de captura de la imagen | String |
| image\_mimetype | Tipo de imagen | String |
| image\_url-large | URL de la imagen (large size) | String |
| image\_id | Id de la imagen | String |
| image\_filesize | Tamaño del archivo de imagen | String |
| image\_url | URL de la imagen | String |
| image\_height | Altura de la imagen | String |
| ocha\_product\_name | Nombre del producto OCHA | String |
| ocha\_product\_id | Id del producto OCHA | String |
| headline\_summary | Resumen | String |
| headline\_image | Titular de la imagen | String |
| headline\_title | Título del encabezado | String |

Referencias

**Básicos**

* Subirats, L., Calvo, M. (2019). **Web Scraping**. Editorial UOC.

**Complementarios**

* Lawson, R. (2015). **Web Scraping with Python**. Packt Publishing Ltd. Chapter 2

Web

* **Centro Europeo para la prevención y control de las enfermedades**. Recuperado de:

<https://data.europa.eu/euodp/es/data/publisher/ecdc>

* **Disasters**. Recuperado de: <https://reliefweb.int/disasters>
* **Scraping en Python con ejemplos.** Recuperado de:

<https://jarroba.com/scraping-python-beautifulsoup-ejemplos/>

* **Monte Yasur en Vanuatu: El volcán en activo más accesible del mundo.** Recuperado de:

<https://tiempodexplorar.com/monte-yasur-en-vanuatu-el-volcan-en-activo-mas-accesible-del-mundo/>

* **All about the ReliefWeb API :** <https://apidoc.rwlabs.org/#doc-use>
* **Tabula documentation:**

<https://tabula-py.readthedocs.io/en/latest/tabula.html>

* Mar 08 2020**, “Tabula-py”.**

<https://readthedocs.org/projects/tabula-py/downloads/pdf/latest/>

* MedinePlus **“Virus del Zika”.**

<https://medlineplus.gov/spanish/zikavirus.html>

* Stackoverflow, 2017-11-28**. “How to extract table as text from the PDF using Python”**

<https://stackoverflow.com/questions/47533875/how-to-extract-table-as-text-from-the-pdf-using-python>

* Pythonic Excursions, Feb 02, 2019**. “Parse PDF Files While Retaining Structure with Tabula-py”**

<https://aegis4048.github.io/parse-pdf-files-while-retaining-structure-with-tabula-py>

* Datacamp**,** November 2nd, 2018**.** Derrick Mwitti **“JSON Data in Python”**

<https://www.datacamp.com/community/tutorials/json-data-python>

* **RIP tutorial:** <https://riptutorial.com/es/pandas/example/5621/eliminar-una-columna-en-un-dataframe>
* BBC.2 abril 2014,Lizzie Crouch y Paula McGrath. **“Las enfermedades que produce la picadura del mosquito”**

<https://www.bbc.com/mundo/noticias/2014/04/140402_salud_enfermedades_mosquitos_gtg>

* **Beautiful Soup Documentation:**

<https://www.crummy.com/software/BeautifulSoup/bs4/doc/#find-parents-and-find-parent>