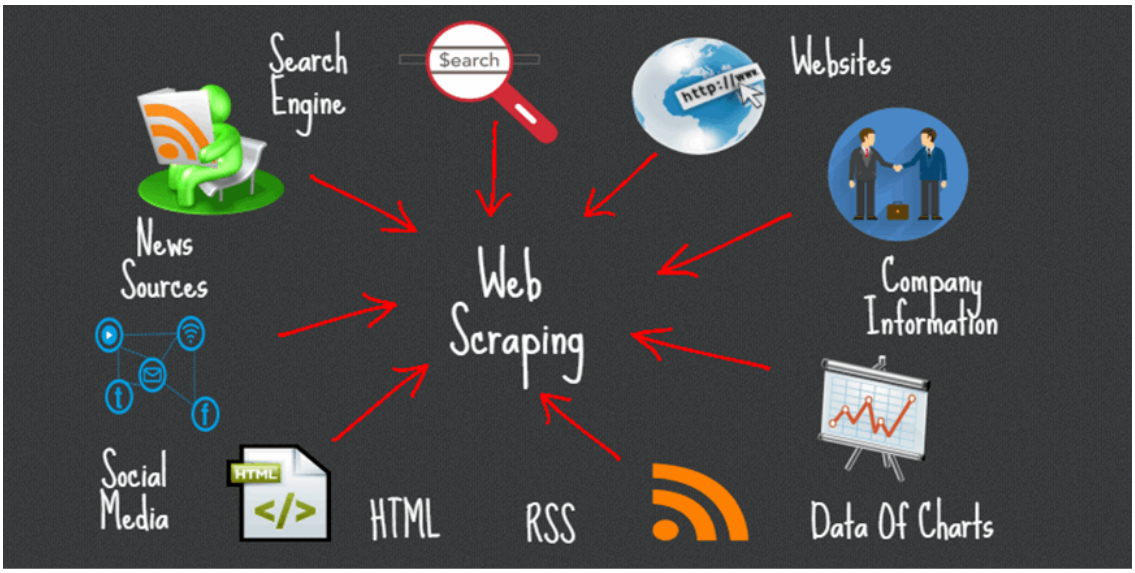
Web Scraping

Tipología y ciclo de vida de los datos



https://www.octoparse.com/blog/9-free-web-scrapers-that-you-cannot-miss

Alberto Quesada

Fernando Sevilla

Marzo 2020

1. Introducción

La expansión del Coronavirus (Covid-19) en el mundo está generando un impacto social sin precedentes. La rápida expansión del virus y la contribución de los países afectados mediante la documentación de nuevos casos diarios, ofrece la posibilidad de evaluar la magnitud del impacto de la epidemia de Covid-19.

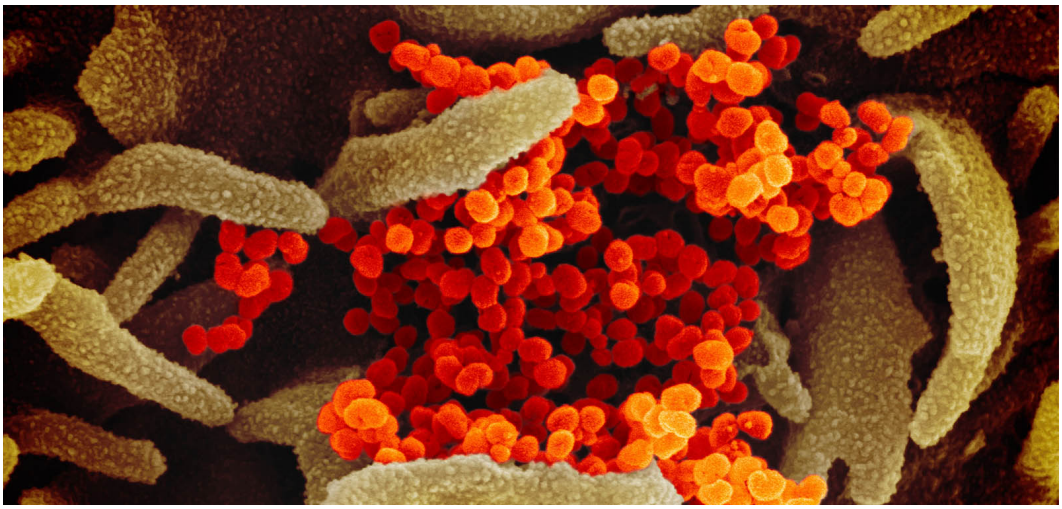


Foto: NIAID-RML - Imagen del coronavirus SARS-Cov-2 (amarillo) en medio de células humanas (rosa) obtenida por microscopía electrónica.

2. Investigación previa y Contexto

Dada la magnitud de la epidemia, en la actualidad, y dada la importancia a nivel global del tema, existe una gran cantidad de información referente a los casos afectados por el Covid-19.

Las webs muestran datos detallados de la expansión del Covid-19 con aportes casi al instante. Esta característica del flujo de información es una ventaja para la realización de esta práctica, dado que permite, una vez extraídos los datos, realizar una comprobación de la calidad de la extracción. De esta manera, se

puede poner en práctica, paralelamente, la representación de los datos mediante el uso de librerías específicas y comparar los resultados con la propia web.

Durante el proceso de investigación se ha constatado la gran diversidad de las páginas que informan del impacto del Covid-19. No todas las páginas consultadas eran idóneas para la realización de la práctica debido a que, o bien no eran páginas que mostraban datos oficiales o bien proporcionaban directamente el dataset con los datos en formato *.csv*.

Por tanto, la elección del repositorio de datos se ha basado en en las siguientes premisas:

1. Debe mostrar información tanto pormenorizada como global de los casos de Covid-19 en el mundo entero.
2. Debe actualizarse diariamente.
3. No debe existir una API para extraer los datos.
4. Los datos deben ser accesibles y públicos.
5. Las fuentes de datos que alimentan la web deben ser veraces, oficiales y actuales.
6. No debe proporcionar directamente un dataset.

La información que vamos a recolectar en esta práctica se corresponde con los datos de infecciones de Covid-19 en los distintos países del mundo. Esta información se presenta sesgada por países, así como por otras variables como los casos activos, casos críticos, etc.

Está granularidad en los datos nos permitirá realizar análisis descriptivos bastante detallados y generar reportes automáticos o dashboards con los que realizar una monitorización de la propagación del virus.

La página seleccionada (https://www.worldometers.info/coronavirus/) permite extraer esta información de una forma sencilla ya que cuenta con un sitemap bien estructurado.

3. Título del Dataset

El archivo .csv será guardado bajo el nombre *"Covid-19_global_spread.csv"*.

4. Descripción del Dataset

El dataset contiene información de la propagación del virus Covid-19 en el mundo. Cada registro representa los datos asociados a un país afectado por el virus. En el momento en el que se realiza este documento, hay un total de 162 países afectados y, por tanto, 162 registros en el dataset.

Como ya hemos comentado anteriormente, los datos han sido extraídos de la siguiente página web: https://www.worldometers.info/coronavirus/. El equipo que da soporte a esta web está formado por desarrolladores e investigadores de todo el mundo que trabajan para unir y publicar estadísticas sobre diferentes temas. Esta página web pertenece a Dadax, una empresa independiente. Podemos encontrar más información de la página web en el siguiente enlace: https://www.worldometers.info/about/.

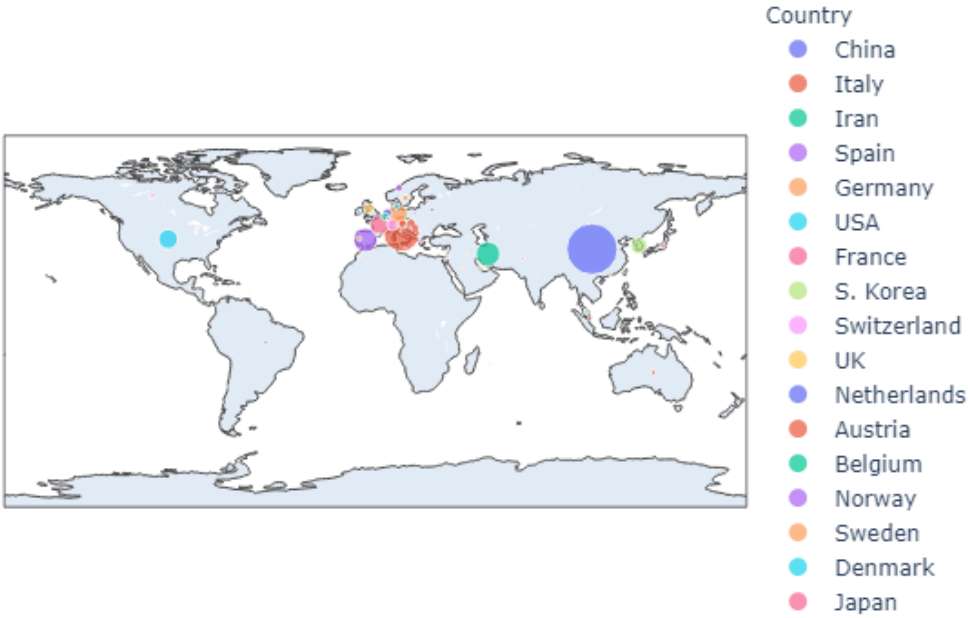
Con respecto a los datos mostrados de propagación del Covid-19, esta página lleva recogiendo la información desde el comienzo de la propagación, por lo que también están disponibles todos los históricos para trabajos futuros.

Las fuentes de esta página web actualizan la información diariamente. Entre las fuentes principales destacan organismos oficiales como:

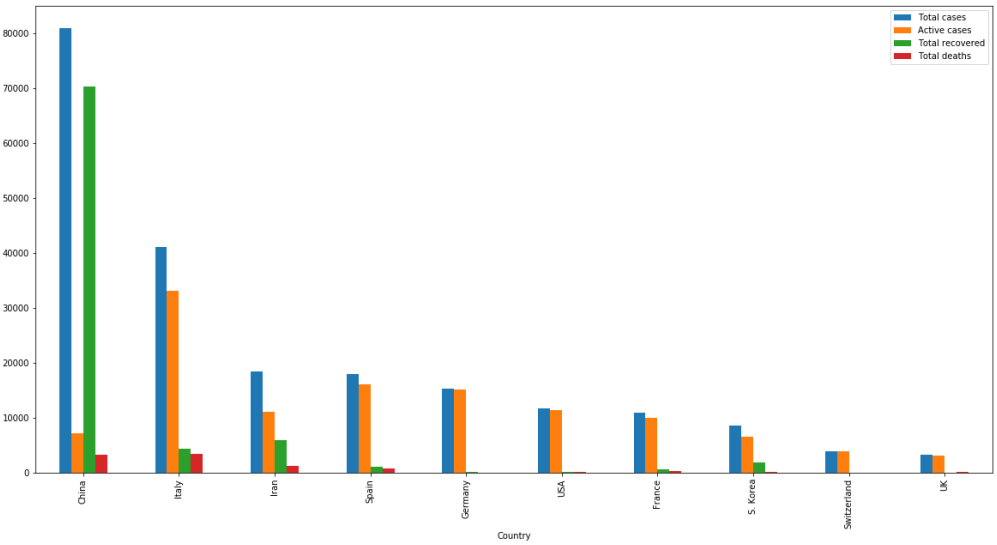
1. Novel Coronavirus (2019-nCoV) situation reports - World Health Organization (WHO)
2. 2019 Novel Coronavirus (2019-nCoV) in the U.S. -. U.S. Centers for Disease Control and Prevention (CDC)
3. Outbreak Notification - National Health Commission (NHC) of the People's Republic of China

Más información referente a las fuentes puede consultarse en la propia web, apartado "Sources".

5. Representación gráfica



Países afectados por Covid-19



Índices principales por países afectados por Covid-19

6. Contenido

A continuación, se muestra una descripción de las variables almacenadas.

Nombre	Descripción	Tipo
Country	País	Nominal
Total cases	Número total de casos	Float
New cases	Número de casos en el último día	Float
Total deaths	Número total de muertes	Float
New deaths	Número de muertes en el último día	Float
Total recovered	Número total de casos curados	Float
Active cases	Número de casos activos	Float
Critical cases	Número de casos en estado crítico	Float
Lat	Latitud del país	Float
Lon	Longitud del país	Float

Los datos se actualizan diariamente en la página web. La extracción se realiza bajo demanda. Los datos del dataset se actualizan cada vez que se ejecute el script.

La extracción de los datos se lleva a cabo utilizando la biblioteca *Beautiful Soup*. Es una biblioteca de Python para analizar documentos HTML (incluyendo los que tienen un marcado incorrecto). Esta biblioteca crea un árbol con todos los elementos del documento y puede ser utilizado para extraer información. Por lo tanto, esta biblioteca es útil para realizar web scraping y extraer información de sitios web.
Mediante el inspector de páginas web proporcionado por Firefox-Developer se localizó, en el código HTML de la página web, el nodo que incluía la tabla a extraer (<td>). Una vez delimitada la búsqueda, se extrae el texto contenido, el cual corresponde a los valores de cada cada columna de la tabla.

El proceso de extracción está documentado en el Notebook que se incluye en esta práctica.

```
# Send request
wurl = requests.get(url).text
soup = BeautifulSoup(wurl,'lxml')

# Get rows
rows = soup.find('table').find('tbody').find_all('tr')
```

Jupyter Notebook. PR1. Web Scraping

7. Agradecimientos

Worldometer está dirigido por un equipo internacional de desarrolladores, investigadores y voluntarios con el objetivo de hacer que las estadísticas mundiales estén disponibles en un formato que invite a la reflexión y que sea relevante para el tiempo a una amplia audiencia en todo el mundo. Worldometer es propiedad de Dadax, una empresa independiente y no tienen ninguna afiliación política, gubernamental o corporativa.

Worldometer fue votado como uno de los mejores sitios web de referencia gratuita por la Asociación Americana de Bibliotecas (ALA), la asociación de bibliotecas más antigua y grande del mundo. Se cita como fuente en más de 10.000 libros publicados, en más de 6.000 artículos de revistas profesionales, y en más de 1000 páginas de Wikipedia.

Los contadores en vivo muestran la estimación en tiempo real calculada por su algoritmo patentado, que procesa los últimos datos y proyecciones proporcionados por las organizaciones y oficinas de estadística más reputadas del mundo.

Se ha consultado el proceso de investigación de datos y de análisis estadístico y se puede resumir en los siguientes tres puntos principales:

1. Los datos son recogidos de las fuentes más fiables. Estas fuentes son continuamente monitoreadas para rastrear cualquier cambio y actualizar el algoritmo en consecuencia.

2. Se realiza un análisis estadístico para extrapolar la estimación actual en base a los datos y proyecciones disponibles en otras webs o servicios.

3. Se desarrolla un algoritmo avanzado para cada contador. Los contadores cargan los datos correspondientes del algoritmo alojado en su servidor central y muestran la estimación actual a medida que cambia en tiempo real. Su algoritmo incluye varias características avanzadas que se aplican según sea necesario, tales como:

    - Fórmulas exponenciales compuestas cada segundo (el valor asignado a las variables de la fórmula cambia cada segundo).
    - Contadores sensibles a la hora del día y/o al día de la semana/final de semana/feriado y cuentas regresivas.
    - Fórmulas complejas que proporcionan una sincronización exacta entre los contadores relacionados.
    - Contadores específicos de zona horaria, basados en el reloj de la PC, o contadores consistentes a nivel mundial.

Existen multitud de análisis realizados mediante los datos que componen este dataset. Un total de 25 referencias a los datos se pueden consultar en el apartado “Sources" del sitio web. La propia página emite diariamente un conjuntos de visualizaciones que muestran la evolución del Covid-19.

Una búsqueda a través de Google utilizando "***Worldometers*** Algorithm (RTS) coronavirus" devuelve más de 4.000 resultados que hecn referencia a los datos de este dataset.

8. Inspiración

El contexto de la situación actual en la que nos encontramos sirvió de inspiración para seleccionar este tema.

Monitorear el avance del virus y ofrecer públicamente esta información a través de GitHub es nuestra forma de contribuir y de ayudar.
Supimos desde el principio que esta información estaría disponible en multitud de organismos oficiales, los cuales ofrecen análisis muy similares entre ellos, a pesar de ello hemos querido plantear esta práctica como el inicio inicio de un análisis desde otro punto de vista utilizando las herramientas aprendidas hasta este momento en las asignaturas del Máster "Ciencia de Datos" sobre un conjunto de datos "vivo".

Extraer los datos es solo el principio, la visualización proporciona respuestas a preguntas directas, tales como:

- ¿Cuántos casos por países?
- ¿Cuántos casos nuevos cada día?
- ¿Qué país es el más afectado?
- etc…

Pero su análisis extrae realmente el valor de los datos:

- ¿Cuál es el ratio de infección?
- ¿A qué velocidad se expande?
- Comparando los datos históricos, ¿están sirviendo las medidas implantadas por los gobiernos?
- Mediante el ratio de muertes y los datos oficiales se puede deducir los casos reales, ¿existe una correspondencia entre ellos?, ¿son los gobiernos conscientes de las cifras reales?

9. Licencia

Los datos se publican bajo la licencia CC-BY (*Creative Commons Attribution 4.0 International*). Esta licencia es una de las licencias abiertas de Creative Commons y permite a los usuarios compartir y adaptar tu conjunto de datos siempre y cuando te den crédito. (Fuente: https://help.data.world/hc/en-us/articles/115006114287-Common-license-types-for-datasets)

Las motivo para la elección de esta licencia es el aumentado de las posibilidades de difusión de los datos, o que otros construyan sobre la base de los mismos, o por la perspectiva de contribuir al trabajo intelectual común.
Además de ser gratuitas, estas licencia tiene la ventaja de entregar información fácil de entender al autor y al usuario, respecto de lo que está o no autorizado a hacer con la obra intelectual.

El uso de esta licencia, por tanto, puede suponer una interesante alternativa para la distribución de datos a través de Internet, otorga libertad al usuario para que pueda realizar determinados usos del código, permitiendo experimentar con nuevas formas de promoción.

10. Código fuente y Dataset

El acceso al código de extracción (Notebook Jupyter Python 3.x) y al Dataset (archivo .csv) completo puede llevarse a cabo través del a GitHub que se incluye a continuación:
Link GitHub

11. Recursos

- Lawson, R. (2015). Web Scraping with Python. Packt Publishing Ltd. Chapter 2. Scraping the Data.
- Subirats, L., Calvo, M. (2019). Web Scraping. Editorial UOC.
- Masip, D. (2010). El lenguaje Python. Editorial UOC.