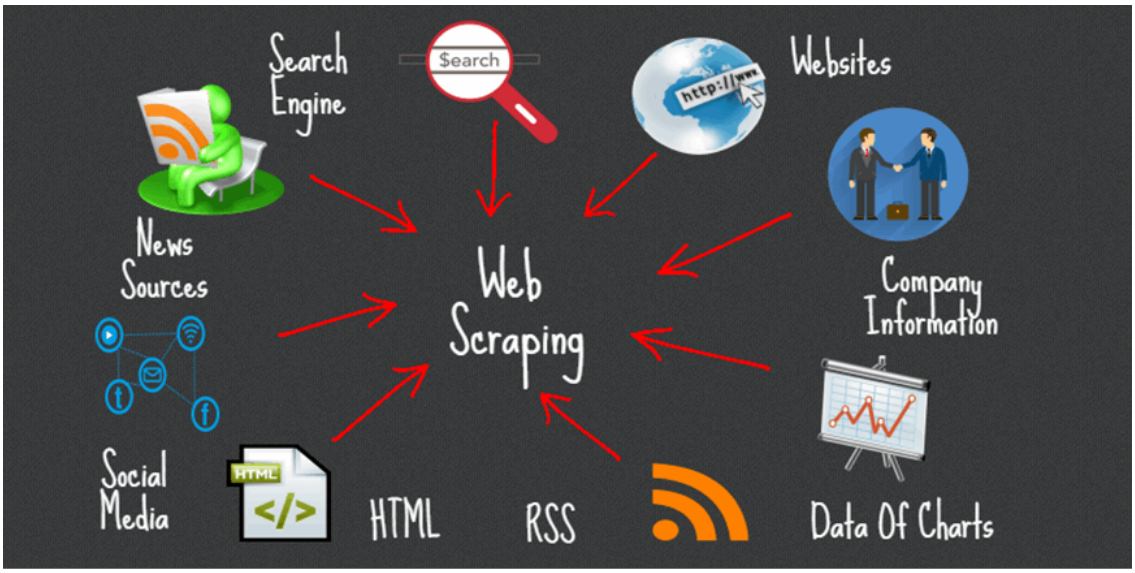
Web Scraping

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. Elección de los datos

La expansión del Coronavirus (COV-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 COV-19.

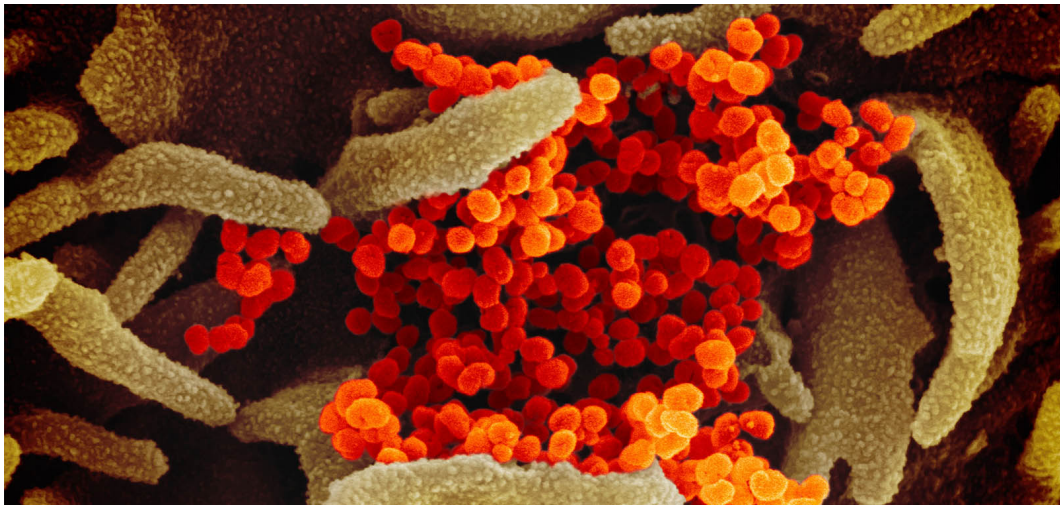


Foto: NIAID-RML - Imagen del coronavirus SARS-Cov-2 (amarillo) en medio de células humanas (rosa) obtenida por microscopía electrónica.

2. Investigación previa

Dada la magnitud de la epidemia, actualmente existe una gran cantidad de información referente a los casos afectados por el COV-19.

Las webs muestran datos detallados de la expansión del COV-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 COV-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 COV-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.*

3. Página Web seleccionada

La página seleccionada es:

https://www.worldometers.info/coronavirus/

Esta página muestra la siguiente información actual por países:

- Casos totales
- Casos nuevos
- Muertes totales
- Muertes nuevas
- Casos recuperados
- Casos activos
- Casos críticos
- Cota relativa por millón de habitantes

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*" .