

Práctica 1: Web Scraping - Dataset: Evolución COVID-19

Patricia García Suarez^{*}

Adrián Arnaiz-Rodríguez^{**}

10/4/2020

Índice

1. Contexto	2
1.1. Link a Worldometers - COVID-19	3
2. Título del dataset	3
3. Descripción del dataset	3
4. Representación gráfica	3
5. Contenido	3
5.1. Dataset	3
5.2. Cómo se ha recogido	4
6. Agradecimientos	4
7. Inspiración	4
8. Licencia	4
9. Código fuente y dataset	4
10. Bibliografía	5

^{*}Perfil Github:, <https://github.com/Kadashi>

^{**}Perfil Github:, <https://github.com/AdrianArnaiz/>

1. Contexto

En el contexto actual de la situación del virus **COVID-19** es imprescindible poder tener datos sobre la situación actual de los países para poder actuar en consecuencia. Si se disponen de buenos datos y un buen modelo, se puede incluso predecir datos futuros para poder prevenir o gestionar de manera más efectiva tanto recursos como medidas a tomar. Sin embargo, se produce una gran acumulación de datos de diversas fuentes. Existen muchas y muy diversas fuentes de los datos de personas contagiadas, casos activos, pacientes recuperados, muertes...

En este contexto de falta de homogeneidad y rigurosidad en los datos, nos adentramos a buscar fuentes de datos que pudieran ser candidatas a llevar un conteo diario (*serie temporal*) por país sobre: contagiados, casos activos, recuperados, muertes y tests realizados. El objetivo es realizar series temporales de estos datos por país con la función de realizar en un futuro análisis sobre los mismos. Nuestro contexto particular sería el de obtener estos datos tanto para **explicar comportamientos pasados, presentes como predecir futuros**. Incluso analizarlo minuciosamente comparando los datos de test realizados y casos totales o cualquier pareja de atributos, incluso analizando las series temporales de datos del COVID-19 con otros datos: PIB, exportaciones, valores bursátiles, densidades de población...

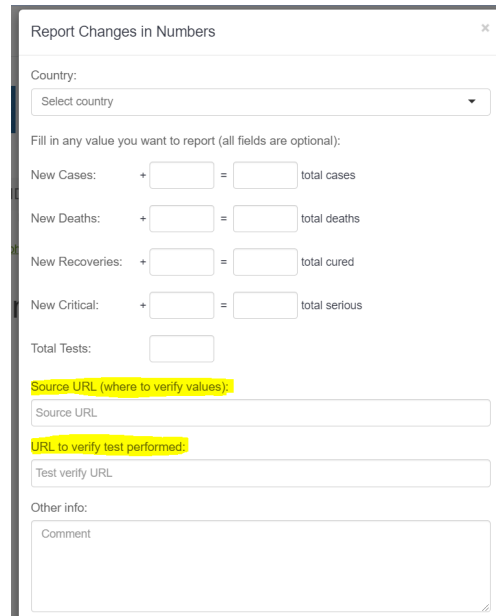
En **resumen**, digamos que nos queremos poner en la piel de instituciones como el CSIC o el INE y realizar estudio minucioso sobre el desarrollo de la enfermedad (tanto de la propia enfermedad, como su relación con datos externos). Por ello, para comenzar **necesitamos los datos centralizados y rigurosos de la evolución contagiados, casos activos, recuperados, muertes y tests realizados por país y por día**.

Hemos analizado muchas posibles fuentes de datos:

- *Oficiales (gobiernos)*: dificultad de recopilación de datos. Habría que buscar la vía por la que cada uno de los países realiza el comunicado y sacar los datos de ese comunicado (algunos comunicados son textos reales en pdf, se necesitaría una labor difícil de *NLP*.)
- *OMS*: En este caso, se debería pensar que es la fuente más fiable de los datos. Sin embargo, en el momento que empezamos a realizar esta práctica, no tenían una plataforma oficial donde se indicaran los datos por países. A esto se añade la poca fiabilidad que han tenido los datos de esta organización en las primeras fases de la epidemia.
- *European Centre for Disease Prevention and Control*: Permitía descargar un gran csv con datos. El motivo por el que no hemos elegido es porque sólo tiene datos de contagiados y muertes.
- *Universidad de Oxford - Our world in data*: Se trata de un estudio interactivo y actualizado en tiempo real que está realizando la Universidad de Oxford sobre la pandemia. Intentan analizar múltiples aspectos de la enfermedad, sin embargo obtienen los datos a través de varias organizaciones (los casos de la OMS, las muertes de la Johns Hopkins e incluso tienen acuerdos). Hemos visto que obtiene los datos a través de la universidad **Johns Hopkins**.
- **Johns Hopkins University**: Esta universidad ha tenido una visualización muy importante [Dong et al., 2020], la cual ha aparecido en todos los medios de comunicación. Sin embargo, los datos los tienen en un repositorio Github donde guardan los csv. También hemos visto que una de sus fuentes principales es **wolrdometers**.
- **Worldometers**: Nos hemos decantado por worldometers porque cumple con todas las funcionalidades que buscábamos: variedad en los datos, centralizados, rigurosos y actualizados dinámicamente. Deducimos que es **rigurosa** y de calidad por dos aspectos: el primero es que una universidad tan prestigiosa como la Johns Hopkins University la utiliza como fuente para su visualización. El segundo aspecto es por la completitud de datos (todos los atributos que queríamos, de muchos países y cada vez van añadiendo más). También, como vista de que es **dinámica** la propia página obtiene datos de diversas fuentes, pero si alguien da datos con justificantes de veracidad: los datos se actualizan (Ver Figura 1). Por ello, consideramos esta página como un repositorio central de los datos del COVID.

1.1. Link a Worldometers - COVID-19

- Link a la página que contiene la tabla: <https://www.worldometers.info/coronavirus>.



The screenshot shows a web form titled "Report Changes in Numbers" with a close button (X) in the top right corner. The form contains the following fields:

- Country:** A dropdown menu with the text "Select country".
- Fill in any value you want to report (all fields are optional):** A section with four rows, each containing a label, a plus sign, an input field, an equals sign, another input field, and a label:
 - New Cases: + [input] = [input] total cases
 - New Deaths: + [input] = [input] total deaths
 - New Recoveries: + [input] = [input] total cured
 - New Critical: + [input] = [input] total serious
- Total Tests:** A single input field.
- Source URL (where to verify values):** A label followed by an input field containing the text "Source URL".
- URL to verify test performed:** A label followed by an input field containing the text "Test verify URL".
- Other info:** A label followed by a large text area containing the text "Comment".

Figura 1: Como insertar nuevos datos con referencias de veracidad

2. Título del dataset

Evolución de contagio del COVID-19 por países.

3. Descripción del dataset

El dataset tendrá la evolución temporal de 5 datos relativos al covid por países. Guardaremos los datos relativos a **contagiados, casos activos, recuperados, muertes y tests realizados**. Es decir para cada uno de los países en los que haya casos registrados, guardaremos un dato al día (de manera automática) cada uno de los datos recién enumerados. Al final, reflejamos la serie temporal de cada uno de esos datos por países.

Como podemos ver es un dataset de 3 dimensiones (Tipo de Dato, País y fecha). En la sección 5.1, explicaremos cómo lo hemos resuelto.

4. Representación gráfica

5. Contenido

5.1. Dataset

Campos y periodo de tiempo

5.2. Cómo se ha recogido

Contar el retardo, porque no se han tenido más estrategias contra el bloqueo.

6. Agradecimientos

7. Inspiración

8. Licencia

La licencia escogida para la publicación de este conjunto de datos ha sido **CC BY-SA 4.0 License**. Los motivos que han llevado a la elección de esta licencia tienen que ver con la idoneidad de las cláusulas que esta presenta en relación con el trabajo realizado:

- *Se debe proveer el nombre del creador del conjunto de datos generado, indicando los cambios que se han realizado.* De esta manera, se reconoce el trabajo ajeno y en qué medida se han realizado aportaciones en relación con el trabajo original.
- *Se permite un uso comercial.* Esto haría que incrementen las probabilidades de que una empresa utilice los datos generados y realicen trabajos de calidad que reporten cierto reconocimiento al autor original.
- *Las contribuciones realizadas a posteriori sobre el trabajo publicado bajo esta licencia deberán distribuirse bajo la misma.* Esto hace que el trabajo del autor original continúe distribuyéndose bajo los términos que él mismo planteó.

9. Código fuente y dataset

El código fuente del scraping, actualización de datos y automatización mediante *Travis* se encuentra en este enlace. Además hay archivos readme en los directorios que dan más información del proyecto.

10. Bibliografía

**AÑADIR RECURSOS DE LIBROS USADOS PARA LAS BUENAS PRACTICAS DEL SCRAPPING (LAWSON, POPRIO DE UOC). AÑADIR REFERENCIAS A LOS TRABAJOS DE LA JHU, DE OUR WORLD IN DATA DE WORLDOMETERS...*

Referencias

Ensheng Dong, Hongru Du, and Lauren Gardner. An interactive web-based dashboard to track covid-19 in real time. *The Lancet infectious diseases*, 2020.