

Universidad Técnica Particular de Loja

Ciencias de la Computación

Nombre:

Aguilar Valarezo Milton Leonardo

Masache Ramón Royer José

Componente:

Fundamentos de Análisis de Datos

Ciclo académico:

abril / agosto 2021

ÍNDICE DE CONTENIDO:

[Extracción de Datos 3](#_Toc72421113)

[Importación de librerías 3](#_Toc72421114)

[Primera fuente 3](#_Toc72421115)

[Lectura y extracción de datos 4](#_Toc72421116)

[Segunda Fuente 4](#_Toc72421117)

[Lectura de datos 5](#_Toc72421118)

[Proceso de extracción de datos 5](#_Toc72421119)

[Conversión a DataFrame 7](#_Toc72421120)

[Limpieza de datos 8](#_Toc72421121)

[Tercera Fuente 8](#_Toc72421122)

[Lectura de datos 8](#_Toc72421123)

[Proceso de extracción de datos 9](#_Toc72421124)

[Conversión a DataFrame 9](#_Toc72421125)

[Limpieza de datos 9](#_Toc72421126)

ÍNDICE DE FIGURAS:

[Ilustración 1: Importación de librerías 4](#_Toc72421141)

[Ilustración 2: Lectura de archivo csv 5](#_Toc72421142)

[Ilustración 3: Lectura de datos 6](#_Toc72421143)

[Ilustración 4: Proceso de extracción de datos 8](#_Toc72421144)

[Ilustración 5: Conversión a DataFrame 8](#_Toc72421145)

[Ilustración 6: Lectura de datos 9](#_Toc72421146)

[Ilustración 7: Proceso de extracción de datos 10](#_Toc72421147)

[Ilustración 8: Conversión a DataFrame 10](#_Toc72421148)

# Extracción de Datos

Se presenta una explicación general acerca de la extracción, lectura y tratamientos utilizados para los datos obtenidos de las diversas fuentes que componen el presente proyecto.

## Importación de librerías

Se realiza la importación de las librerías que nos permitirán trabajar en la construcción de objetos Python, así como el proceso de *Scrapy* para fuentes web elegidas:

* **Requests:** Librería HTTP elegante y simple para Python que permite enviar solicitudes HTTP con mucha facilidad.
* **Beatiful Soup:** Librería de Python para extraer datos de archivos HTML y XML.
* **Pandas:** Biblioteca que proporciona estructuras de datos y herramientas de análisis de datos fáciles de usar.

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Ilustración 1: Importación de librerías

Al tener datos de distintas fuentes, el análisis exploratorio de datos se realizará de forma separada.

## Primera fuente

La fuente proviene del sitio Our World in Data y contiene un archivo .csv con registros acerca de la pandemia del COVID-19 en la mayoría de los países del mundo.

### Lectura y extracción de datos

Para la lectura y extracción de datos se hace uso de la librería *Pandas*, importada previamente y se lo realiza a través del método **read\_csv**, que permite leer el tipo de archivo y lo almacena de forma automática en un DataFrame.

Una captura de pantalla de una computadora

Descripción generada automáticamente

Ilustración 2: Lectura de archivo csv

La librería *Pandas* asigna automáticamente los tipos de datos presentes en el archivo leído, pero en este caso, se realizó un proceso particular para su asignación.

## Segunda Fuente

La fuente proviene del sitio web Worldmeter y presenta una página web que contiene una tabla con datos que corresponden a la población de cada país del mundo, así como datos relacionados y actualizados hasta el año 2020.

### Lectura de datos

Para la lectura y extracción de los datos presentes en esta página web, se hace uso de las librerías *Requests* y *Beautiful Soup*, previamente importadas.

Se realiza el siguiente procedimiento:

* Para enviar la solicitud HTTP y obtener la página, se usa el método **get** a través del envío de un parámetro (URL) y se almacena el resultado en una variable declarada.
* Para codificar la información extraída en un formado adecuado, el método **encoding** lo facilita, siendo “UTF-8” nuestro formado deseado.
* Se realiza el llamado de la variable y se mostrará un status, si se muestra 200 significa que la solicitud se ha generado correctamente.
* Se transforma la variable en un objeto HTML para permitir la navegación a través de la librería *Beautiful Soup.*

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Ilustración 3: Lectura de datos

### Proceso de extracción de datos

Para convertir nuestro objeto HTML en un DataFrame, se obtiene una lista con los encabezados y una lista que contendrá en cada una de sus posiciones, listas con los datos de la población de un país a través del siguiente procedimiento:

* Se hace uso del método **find** y **findAll**, los cuales permiten extraer el primer valor o una lista de valores a partir de la etiqueta con la clase deseada y a través de estructuras repetitivas se extraerán los valores del objeto HTML y se añadirán a las listas declaradas previamente.
* Se crean los arreglos que almacenarán las listas deseadas y mediante el método **find** extraemos el HTML de la tabla, ubicándola mediante la etiqueta **table** y el nombre de la clase establecido en el sitio web. Posterior a ello, se extrae de la tabla el encabezado, ubicado en la etiqueta **thead** y a este encabezado se aplica el método **findAll** extrayendo las etiquetas **th (**cuyos valores son los títulos del encabezado)en una lista, consecuentemente se implementa un ciclo repetitivo para recorrer las posiciones de la lista extraída y se irán guardando en nuestra lista creada para los encabezados, para ello se aplica el método **text** a cada uno de los valores extraídos, así se extraerá únicamente el valor sin la etiqueta HTML.

**Nota:** Se definió el nombre de “Global Rank” en la primera posición del encabezado, pues la que se extraía no era nada descriptiva.

* Se realiza el mismo proceso con las filas con la diferencia de que existe un cambio a las etiquetas correspondiente y la agregación de un ciclo repetitivo más, pues se extraerán todas las filas (a diferencia del **thead** que era una sola fila).

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Ilustración 4: Proceso de extracción de datos

### Conversión a DataFrame

Con las estructuras creadas, se hace uso de la librería *Pandas* y a través del método **pd.DataFrame**, se creará un DataFrame a partir de dichas estructuras recibidas como parámetros.

Aplicación, Calendario

Descripción generada automáticamente

Ilustración 5: Conversión a DataFrame

### Limpieza de datos

Procedimiento:

Al haber utilizado el método **text** para extraer los valores, los tipos de datos presentes en el DataFrame son String, por tanto, se requiere de un proceso de limpieza debido a que diversos datos poseen “,” para separar las milésimas y millonésimas. Primero las reemplazamos por un espacio vacío, el mismo procedimiento se realiza con los datos de tipo Strings “%”. Para finalizar, se cambian los tipos de datos de las columnas a los requeridos.

## Tercera Fuente

La fuente proviene del sitio web Worldmeter y presenta una página web que contiene una tabla con los datos correspondientes al Producto Interno Bruto (PIB) de cada país del mundo, actualizados hasta el 2017.

### Lectura de datos

Para la lectura y extracción de datos se realiza el mismo proceso descrito en la fuente 2, utilizando las mismas librerías, métodos y procedimientos para con la diferencia de que se utiliza la URL de la fuente respectiva

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Ilustración 6: Lectura de datos

### Proceso de extracción de datos

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Ilustración 7: Proceso de extracción de datos

### Conversión a DataFrame

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

Ilustración 8: Conversión a DataFrame

### Limpieza de datos

Al igual que en la fuente anterior, se realiza un proceso de limpieza para cambiar los tipos de datos a los adecuados a través de procedimientos similares. En esta fuente, se debe eliminar, además, el símbolo **$** presente en algunas columnas de la tabla para poder asignar el tipo de dato correspondiente.