UANL

UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN

FCFM

FACULTAD DE CIENCIAS FÍSICO MATEMÁTICAS

UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN

FACULTAD DE CIENCIAS FÍSICO MATEMÁTICAS

(FCFM)

Unidad de Aprendizaje: Programación Básica

PIA E2. Extracción, Limpieza y Estructuración de los Datos

Docente: Perla Marlene Viera González

Equipo 11:

Carolina Berlanga Dávila – 2162840

Alan Ronaldo Morales Ortíz - 2121366

San Nicolás de los Garza, Nuevo León, 25 de abril 2025

Documentación

API Seleccionada:

World Bank API:
https://datahelpdesk.worldbank.org/knowledgebase/topics/125589?ref=public_apis

1. Métodos de extracción de datos y herramientas empleadas

Para la extracción de datos, se utiliza el link de la API, se le agregan los parámetros y se solicita la información requerida. En este caso, la API proporciona diversos tipos de datos sobre los países, pero solo se requiere el PIB de 5 países de 2010 a 2020.

Primero se creó una lista con el ID de los países seleccionados. Después se definió una función, que tiene como parámetros los años que se van a estar analizando. Lo que hace esta función es cambiar ciertos parámetros para que acceda a cierta información. Con el uso de “for”, recorre la lista de los IDs y para cada uno de estos hace lo siguiente. Primero que tiene es un ”try” para asegurarse de que se pueda acceder a la información del URL. Luego, se creó una variable llamada respuesta que por medio de solicitudes web va a extraer la información de la API. Después añadimos una variable llamada datos que a convertir los datos a formato JSON.

Para la impresión de los nombres de los países, se añadió un diccionario llamado “países”, que guarda la correspondencia del ID con cada país. Al momento de la impresión se usa el mismo contador que recorre la lista de IDs.

Se crean diccionarios que van a almacenar el nombre del país, el año y el PIB correspondiente. Se usa un “for” para recorrer los datos de la API e irlos guardando en el diccionario. Se agregan los diccionarios a una lista vacía, creada al inicio del código, con el uso de “append”. Y por último se imprime la lista.

El código usa errores y excepciones que lanzaran mensajes al usuario en caso de que pase alguno de los siguientes:

- Error de conexión: se pedirá al usuario que revise el internet.
- Error “Timeout”: se notificará que la solicitud tardó demasiado en responder.
- Error HTTP: se va a especificar el tipo de error
- Error de valor: se le notificará al usuario que los datos de entrada no fueron válidos y le pedirá volver a ingresarlos.

2. Técnicas de limpieza aplicadas

En cuanto a la limpieza de los datos, no se necesito ningún procedimiento. El link utilizado para la extracción de información ya viene con los parámetros necesarios para que sólo proporcione los datos que se están buscando. (En indicator/NY. GDP.MKTP. CD?) Y al momento de la impresión, únicamente se imprimen la fecha y el año. Por lo que no hay datos que se deban eliminar.

3. Estructura de los datos optimizados y su diseño lógico

Para la impresión de datos se hace, primero imprimiendo el nombre del país y luego el año y el PIB correspondiente.

Para el almacenamiento de datos, se decidió que era mejor guardarlo en un archivo csv directamente, esto para la creación de tablas, y posibilidad de usar pandas o openpyxl para poder manipular la información. Además de que una tabla nos ayudará a representar los datos en gráficas con el uso de librerías como matpltolib.

Esto se hizo almacenando los datos en una lista de diccionarios. Cada diccionario contiene el apartado de “Nombre” (del país), “Año” y “PIB”. Esto es ventajoso porque facilita la clasificación o búsqueda de datos por alguno de estos campos. En este caso, se utiliza para crear una tabla que contiene en una columna el país, en otra el año y en otra el PIB.

El archivo csv se creó con el uso de “with open”, se le asignó un nombre, se dieron permisos para escritura y se eliminaron los espacios entre los saltos de línea que no fueran necesarios. Añadimos una lista que almacena los títulos de los apartados de la tabla y se añadieron los nombres de las listas.

4. Detallar las validaciones implementadas y las transformaciones realizadas.

A pesar de no llevar a cabo un proceso de limpieza, se llevó a cabo un procedimiento para que los datos devueltos por la API fueran de tipo entero flotante. Esto se hizo mediante el uso de expresiones regulares, verificando que los datos devueltos fueran únicamente dígitos entre 0 y 9 y que algunos podrían estar separados por un punto decimal. En caso de que no cumpliera con esto, se devuelve el mensaje “Dato invalido”. Al pedir los años, sólo se admiten enteros entre 1960 y 2023, por lo que no fue necesario verificar los datos de entrada.

En caso de que la API no devolviera “None”, se envía un mensaje al usuario diciendo que no se encontró un valor para ese año. Esto mediante el uso de la excepción Value Error.