



Test Backend Developer  
2022 - Season 1.2

El objetivo de la evaluación es el de implementar un proceso que permita scrapear la información de ciertos reportes estadísticas (dataset) desde la Plataforma Nacional de Datos Abiertos de Perú (<https://www.datosabiertos.gob.pe/> )

En esta plataforma se consolidan los diferentes reportes digitales que permite encontrar, explorar y reutilizar datos gubernamentales.

Estos scraper tendrán ciertos parámetros de configuración como la URL desde donde se obtendrá la data. <https://www.datosabiertos.gob.pe/>



## Actividades a realizar

### Parte 1:

Implementar uno o varios procesos que permitan scrapear la información desde la plataforma (<https://www.datosabiertos.gob.pe/>).

Para descargar el reporte en la opción de **Tipos de Contenido** seleccionar **Dataset**(Imagen 1), en **Categoría** seleccionar **Economía y Finanzas**(imagen 2) y por último en la opción **formato** seleccionar **csv**(imagen 3). El siguiente paso es buscar en el input **Search** el nombre del reporte en este caso “donaciones” y dar click en el botón **consultar**. Luego dar click en el reporte “Donaciones COVID-19 - [Ministerio de Economía y Finanzas - MEF]”(imagen 4). Finalmente debemos descargar el archivo dando click en el botón **descargar** de la opción **Data - Donaciones Covid-19**(imagen 5).

Imagen 1: Tipo de contenido

Tipos de contenido	
	Recurso (4296)
	Dataset (2268)
	Entidades (65)
	Harvest Source (19)
	Página (3)

Imagen 2: Categorías

Categorías	
	Economía y Finanzas (791)
	Gobernabilidad (456)
	Desarrollo Social (160)
	Transporte (130)
	Salud (60)

Imagen 3: Formato

Formato	
	xlsx (1183)
	csv (478)
	html (407)

Imagen 4: Resultado al seleccionar las opciones

The screenshot shows a web interface for data distribution. On the left, a sidebar contains filters: 'Tipos de contenido' (Dataset), 'Categorías' (Economía y Finanzas), 'Etiquetas', 'Formato' (csv, xls), and 'data (1)'. The main area displays the title '1 Distribución de Datos' and search results for 'donaciones'. The search bar contains 'donaciones', and the results show 'Donaciones COVID-19 - [Ministerio de Economía y Finanzas - MEF]'. Below the title, there are tags for 'COVID-19', 'Gobernabilidad', and 'Economía y Finanzas'. A description states: 'Datos registrados por las Unidades Ejecutoras del Sector Salud - MINSA, en el Modulo de Almacén del SIGA, sobre las bienes ingresados al Almacén con la denominación donación, procedente de personas naturales, jurídicas'. At the bottom, there are buttons for 'csv', 'xls', and 'data'.

Imagen 5: Descargar archivo

### Dato y Medio de Distribución

The screenshot shows the 'Dato y Medio de Distribución' section. It lists three data items: 'Muestra de Datos- Donaciones Covid-19' (Datos registrados por las Unidades Ejecutoras...), 'Diccionario de Datos - Donaciones Covid-19' (Diccionario de Datos de Donaciones), and 'Data - Donaciones Covid-19' (Data - Donaciones Covid-19). Each item has a 'Previsualizar' button and a 'Descargar' button. The 'Data - Donaciones Covid-19' item also has a 'Descargar' button.

Estos procesos de scrapear la plataforma debe tener los siguientes inputs que serán necesarios para descargar un archivo dinámicamente:

- **Tipo contenido:** Ejemplo: "dataset"
- **Categoría:** Ejemplo "Economía y Finanzas"
- **Formato:** Ejemplo "csv"
- **Nombre del reporte:** Ejemplo "donaciones"

## Parte 2:

Crear un proceso que tenga como entrada el nombre del archivo y leerlo con el paquete de pandas pandas. Luego se debe filtrar por la columna **REGIÓN**, y guardar un archivo con la información de cada región, el nombre del archivo csv debe ser el de la región en minúsculas. Por ejemplo "lima.csv" y la información que contiene debe pertenecer solo de la región Lima.

A	B	C	D	E	F	G	H	I	J
ANO	EJE	REGION	TIPO_GOB	GOBIERNO	SECTOR	SECTOR	PLIEGO	SEC	EJEC
1	2020	LIMA	E	GOBIERNO NACIONAL	11	11	11	1726	148
2	2020	LIMA	E	GOBIERNO NACIONAL	11	11	11	1726	148
3	2020	LIMA	E	GOBIERNO NACIONAL	11	11	11	1726	148
4	2020	LIMA	E	GOBIERNO NACIONAL	11	11	11	1726	148
5	2020	LIMA	E	GOBIERNO NACIONAL	11	11	11	1726	148
6	2020	LIMA	E	GOBIERNO NACIONAL	11	11	11	1726	148
7	2020	LIMA	E	GOBIERNO NACIONAL	11	11	11	1726	148
8	2020	LIMA	E	GOBIERNO NACIONAL	11	11	11	1726	148
9	2020	LIMA	E	GOBIERNO NACIONAL	11	11	11	1726	148
10	2020	LIMA	E	GOBIERNO NACIONAL	11	11	11	1726	148
11	2020	LIMA	E	GOBIERNO NACIONAL	11	11	11	1726	148
12	2020	LIMA	E	GOBIERNO NACIONAL	11	11	11	1726	148
13	2020	LIMA	E	GOBIERNO NACIONAL	11	11	11	1726	148
14	2020	LIMA	E	GOBIERNO NACIONAL	11	11	11	1726	148
15	2020	LIMA	E	GOBIERNO NACIONAL	11	11	11	1726	148
16	2020	LIMA	E	GOBIERNO NACIONAL	11	11	11	1726	148

## Proyecto

Para resolver la tarea es necesario utilizar obligatoriamente el lenguaje python. Se puede utilizar cualquier framework de Python como Flask, Django o Python puro.

## Restricciones

- No se puede utilizar la API disponible en la plataforma, el scrapeo debe ser vía HTML con alguna librería.
- Se puede utilizar cualquier librería disponible en pip.
- Los scrapers deben de ejecutarse como comandos o desde el shell.
- Incluir instrucciones de cómo ejecutar el script.

## Evaluación

- El proyecto se evaluará en función de los siguientes parámetros :
- Calidad del código (PEP8, facilidad de lectura, etc.)
- Facilidad de mantenimiento del código.
- Mecanismo elegido para scrapeado.
- Dockerizar el código es un plus.
- Se deben implementar pruebas unitarias como parte del código con las librería de su elección (unittest, nose, pytest, etc)
- Opcionalmente implementar otras técnicas de testing son un plus (Doctests, Hipótesis, Mutation Testing, etc)
- El proyecto se debe entregar a más tardar el Lunes 23 de Mayo del 2022 a las 2:00 pm Hora Chile - GMT -4. (1:00 pm Hora Perú)
- Por favor enviar el proyecto o repositorio git al correo: [marco@teamcore.net](mailto:marco@teamcore.net)