Laboratorio de Inteligencia Artificial y Métodos Computacionales en Ciencias Sociales (Q-LAB)

Taller de Web Scraping para la Investigación Social

SÍLABO 2025-I

I. INFORMACIÓN GENERAL

Nombre del curso	: Web Scraping para la Investigación Social
Número de sesiones	: 5 sesiones
Número de horas	: 10 horas
Encargado	: Alexander Noblejas alexander.noblejas@pucp.edu.pe
Horario clases	: 10:00 am - 12:00 pm
Inicio de clases	: Lunes 17 de agosto

II. SUMILLA

Este taller introductorio combina teoría y práctica para dotar a estudiantes e investigadores en ciencias sociales de herramientas clave en la extracción y procesamiento de datos. Abarca el uso de Python y RStudio para automatizar la descarga de información, permitiendo obtener datos de manera más eficiente y construir bases propias para el análisis cuantitativo.

Los participantes aprenderán a extraer información de páginas web (principalmente del Estado), recuperar y procesar datos y utilizar GitHub para el desarrollo colaborativo. Al finalizar, estarán capacitados para optimizar su recolección de datos y aplicar técnicas de Web Scraping en sus investigaciones.

III. CONTENIDO DEL CURSO

Sesión 1: Introducción a GitHub

a. Creación y uso de repositorios
b. Clonación y Commits
c. Ciclo de Branches-Pull Requests-Merge

Sesión 2: Web Scraíng con Selenium - Parte I

a. Creación de environments en Anaconda
b. Conceptos básicos de HTML y CSS

c. Inicialización de Web Driver

Sesión 3: Web Scraíng con Selenium - Parte II

a. Métodos de localización y envíos dinámicos
b. Esperas explícitas e implícitas
c. Configuraciones adicionales (Headless mode y adaptaciones a Google Colab)

Sesión 4: Web Scraíng con Rstudio - Parte I

a. Introducción a la extracción de datos con R
b. Uso de *rvest* para la recolección de información
c. Identificación y selección de elementos en páginas web.
d. Manejo de datos extraídos y estructuración en data frames

Sesión 5: Repaso

a. Comparación de enfoques en Python (Selenium) y RStudio (rvest) para Web Scraping
b. Limpieza y organización de los datos extraídos para su análisis.
c. Manejo de errores y estrategias para evitar bloqueos en páginas web.
d. Espacio para dudas, consultas y orientación en proyectos personales.

IV. REQUISITOS

Para el desarrollo del taller los estudiantes necesitan:

- Tener instalada la plataforma Anaconda en su PC. Más información aquí.
- Tener una cuenta personal activa de GitHub y tener instalado GitHub Desktop en su PC. Más información aquí.

Por último, la asistencia es obligatoria y solo están permitidas un 20% de faltas.

V. METODOLOGÍA

Las clases se desarrollarán de manera sincrónica mediante la plataforma Zoom. Asimismo, desarrollarán actividades que complementarán su aprendizaje a lo largo del taller. Al terminar el curso, los participantes serán capaces de manejar todos los contenidos del taller.

VI. EVALUACIÓN

Actividades	**Porcentaje**
Actividades calificadas (I y II)	30 % (15 % c/u)
Trabajo final	50 %
Asistencia	20 %

VII. CRONOGRAMA

El cronograma de las clases es el siguiente:

Sesiones	Fecha	Temas	Actividades
Sesión 1	17 de marzo (10:00 - 12:00m)	Introducción a GitHub y conocimiento de estructuras de las páginas web	- Clase
Sesión 2	18 de marzo (10:00 - 12:00m)	Primera parte de Web Scraping con Python	- Clase
Sesión 3	19 de marzo (10:00 - 12:00m)	Segunda parte de Web Scraping con Python	- Clase - Actividad I
Sesión 4	20 de marzo (10:00 - 12:00m)	Web Scraping con Rstudio	- Clase - Actividad II
Sesión 5	21 de marzo (10:00 - 12:00m)	Repaso	- Clase - Trabajo final