# Práctica 10: Web data extraction

Contenido

[Práctica 10: Web data extraction 1](#_Toc468111189)

[Contenido 1](#_Toc468111190)

[Objetivos 1](#_Toc468111191)

[Propósito del curso al que apoya la práctica 1](#_Toc468111192)

[Actividades 1](#_Toc468111193)

[Práctica: Extracción de precios del sitio Finanzas Yahoo 1](#_Toc468111194)

[1 Descarga de la práctica e instalación del SW Screen-Scraper 1](#_Toc468111195)

[2 Diagnóstico el sitio Finanzas Yahoo para la acción VITROA.MX 2](#_Toc468111196)

[3 Extracción de precios de la acción VITROA.MX 2](#_Toc468111197)

[4 Variaciones para extraer ahora la acción BIMBOA.MX 2](#_Toc468111198)

[5 Ejercicio de tarea 3](#_Toc468111199)

## Objetivos

1. Entender el concepto de extracción de datos web en el contexto de la integración de datos.

## Propósito del curso al que apoya la práctica

Proporcionar herramientas que permitan ampliar las capacidades de obtención de datos a partir de repositorios web públicos y privados.

## Actividades

* Conocer una herramienta especializada en extracción de datos web.
* Extraer información a partir de un sitio web y transformarla en un archivo csv delimitado por comas.

## Práctica: Extracción de precios del sitio Finanzas Yahoo

## Descarga de la práctica e instalación del SW Screen-Scraper

1.- Descarga los datos requeridos para la práctica haciendo clic en la opción “Clone or Download” y descomprime el archivo en algún directorio:

<https://github.com/mariofonseca/webextractiondemo.git>

2.- Descarga la versión **ENTERPRISE** de la herramienta Screen-Scraper de acuerdo con tu sistema operativo:

<https://www.screen-scraper.com/developers>

3.- Instala la herramienta Screen-Scraper Enterprise. Tendrá una vigencia de 30 días. Para utilizarlo después de la fecha de expiración sólo hay que desinstalarlo y volverlo a instalar.

## Diagnóstico el sitio Finanzas Yahoo para la acción VITROA.MX

1.- Búsqueda de la acción VITROA.MX en el “Buscador de Cotizaciones” del sitio de Yahoo finanzas, usando un navegador web.

2.- Ir a la sección de “Precios Históricos” y seleccionar información de precios diarios sólo ocho años.

3.- Analizar y discutir el url, los parámetros y sus valores.

4.- Iniciar el programa Screen-Scraper. Ejecutar la opción **File->New Scraping Session **. Asignarle el nombre “YahooFinanzas”. Hacer clic izquierdo en el “Scraping Session” y seleccionar la opción “Add Scrapable file from URL”. Copiar el url descrito en (3). Asignarle el nombre “VITROA”. En la pestaña “Advanced” seleccionar “Jericho” en la opción de *Tidy Html*. Hacer clic en el ícono “Save”.

5.- Seleccionar el  Scraping sesión “YahooFinanzas” y ejectuarlo . Seleccionar el scrapeable file “VITROA” y seleccionar la pestaña “Last Response”. Luego hacer clic en la opción “Find”  y escribrir “precio apertura”. Seleccionar el patrón que tiene el primer renglón de datos, hacer clic derecho y seleccionar “Generate extractor pattern from selected text”. Asignar en la etiqueta “Identifier” el valor “preciosAccion”.

6.- Reemplazar cada uno de los valores del renglón por una variable para almacenar su valor. Para ello hacer clic derecho “Generate extractor pattern token from selected text”. Asignarles nombres significativos. Revisar las opciones  y . Ejecutar de nuevo la Sraping sesión “YahooFinanzas” .

## Extracción de precios de la acción VITROA.MX

1.- Importar los archivos script “IniciarVariables”, ”CerrarVariables”, “GuardarPreciosAccion” de la carpeta ../Scripts/ss/scripts/, usando la opción “File->Import”

2.- Configurar la variable “workingDir” en el archivo “IniciarVariables” para que apunte al directorio local donde instalaron la práctica.

3.- Configurar la Scraping Session “YahooFinanzas” para incorporar los scripts “IniciarVariables” y “CerrarVariables”. Configurar el scrapable file “ACCION” para incorporar el script “GuardarPreciosAccion”. Ejecutar la Scraping Session “YahooFinanzas” . Revisar el archivo Csv con los resultados.

## Variaciones para extraer ahora la acción BIMBOA.MX

1.- Cambiar la variable “stockName” en el script “IniciarVariables”.

2.- En el scrapeable file “ACCION”, en la pestaña “Parameters”, cambiar el valor “VITROA.MX” por el valor “~#stockName#~”.

3.- En la pestaña “Properties” del scrapable file “ACCION” habilitar el check box “This scrapeable file will be invoked manually…”.

4.- En el script “IniciarVariables” descomentar la línea 14. Ejecutar el Scraping Session “YahooFinanzas”. Revisar el archio csv con los resultados.

## Ejercicio de tarea

* Modificar el código para recuperar todas las páginas con los precios de las acciones “VITROA.MX” y “BIMBOA.MX”.
* El entregable será un documento con un párrafo donde se expliquen los cambios realizados al código actual para lograr dicho objetivo.