Centro de Investigación de Computación

Profesor: Alan Badillo Salas

Alumno: Esteban Dario Marroquin Tetlalmatzi

Correo: estebandmt@gmail.com

07/2023

Programación Python en el Ámbito Científico

Introducción

En la actualidad el comparar productos con la competencia se ha convertido en una tarea difícil de realizar, debido a la enorme cantidad de información que se encuentra en las diversas páginas web. Tiempo que aumenta estrepitosamente si esa información se desea guardar. Debido a esto se han generado una gran variedad de herramientas que facilitan este tipo de labores como herramientas que realizan el Escrapeado de información

Justificación

El realizar un Escrapeado de información permite la recolección masiva de información de diferentes lugares, permitiendo que diversas áreas de estudio sean beneficiadas. Como el área de comercio. La cual al obtener datos de algunos productos desde una pagina web ajena puede facilitar la comparación de información entre diferentes paginas permite tener un mercado mas competitivo. Permitiendo que los clientes tengan mejores ofertas al momento de elegir algún producto.

Pasos realizados

Para obtener los productos se realizaron los siguientes pasos:

1. Analisis de la pagina web, donde se obtendrá la información deseada
2. Analisis de los nodos a nivel HTML para conocer los xpaths de donde se obtendrá la información
3. Se crea la conexión con la pagina web que se quiere “Scrapear”
4. Se toman los diferentes nodos y son guardados
5. La información obtenida es acumulada y posteriormente se guarda en un archivo CSV

Configuración y conexión a la pagina deseada

```
#Configuracion del Selenium
navegador = webdriver.Chrome()

#Realiza la busqueda de la pagina web
navegador.get("https://www.sanborns.com.mx/")
```

Se obtienen los nodos principales que tienen la información de los productos

```
#Se obtienen los nodos deseados
nodos = navegador.find_elements(By.XPATH,"//div[starts-with(@class,'CardProduct_contDataCard')]")
```

Obtiene la información de cada nodo principal

```
#Se realiza el raspado en cada nodo y s eobtienen los datos deseados
for articulo in nodos:
    nombre = articulo.find_element(By.XPATH,"./h3").text
    imagen = articulo.find_element(By.XPATH,"../picture/img").get_attribute("src")
    precio = articulo.find_element(By.XPATH,"./p[starts-with(@class, 'CardProduct_precio1')]").text

    if len(nombre) == 0 or len(precio) == 0:
        continue

    lista_nombres.append(nombre)
    lista_precios.append(precio)
    lista_imagenes.append(imagen)
```

Se crea el DataFrame y se guarda la información en un archivo CSV

```
#Se crea el Data Frame con los datos obtenidos
articulosDT = pd.DataFrame({
    "Nombre": lista_nombres,
    "Precio": lista_precios,
    "Imagen": lista_imagenes
    })

#Se guarda la informacion en un csv
articulosDT.to_csv("Precios.csv")
```

Se cierra la conexión

```
#Se cierra la conexion
navegador.close()
```

Muestra del archivo CSV

		Nombre	Precio	Imagen
1	0	Pantalla LG 77 pulga...	$59,995MXN	https://resources.san...
2	1	Huawei Matepad 10....	$4,369MXN	https://resources.san...
3	2	Cámara Nikon Z50 1...	$20,789MXN	https://resources.san...
4	3	Audífonos Sony WF-1...	$4,299MXN	https://resources.san...
5	4	The lord of the rings ...	$489MXN	https://resources.san...
6	5	Motorola G60 128GB ...	$5,999MXN	https://resources.san...
7	6	Base de Maquillaje M...	$144MXN	https://resources.san...
8	7	Ipad Pro 12.9 Wi-Fi 2...	$25,359MXN	https://resources.san...
9	8	Preventa - Beyond th...	$499MXN	https://resources.san...
10	9	Xbox One FIFA 22	$299MXN	https://resources.san...
11	10	Papel KP-108 Canon	$759MXN	https://resources.san...
12	11	Pantalla Samsung 5...	$7,595MXN	https://resources.san...
13	12	Smartwatch Redmi ...	$1,299MXN	https://resources.san...
14	13	Samsung Galaxy Wat...	$4,499MXN	https://resources.san...
15	14	Smartwatch Huawei ...	$5,699MXN	https://resources.san...
16	15	Apple Watch Ultra 49 ...	$19,999MXN	https://resources.san...
17	16	Cámara Bosma XCG...	$1,499MXN	https://resources.san...
18	17	Monitor XTREME 2 USB	$499MXN	https://resources.san...
19	18	Pack focos inteligent...	$299MXN	https://resources.san...
20	19	Dron DJI Tello Boost ...	$4,099MXN	https://resources.san...

Conclusiones

Se pudo obtener la mayoría de los datos de los productos mediante la herramienta d selenium. Se pudo realizar el escrapeado de información facilitando la obtención de datos. Permitiendonos utilizar dicha información para comparacion de precios y mejorar la experiencia del cliente. Para futuras mejoras se detecto que no se obtuvieron todos los precios, mostrando que los no todos los xpaths son iguales por lo que se deben añadir los precios faltantes. El scrapeado nos permite el obtener una gran cantidad de información para comparar precios o incluso el revisar si algún producto aumenta o reduce su precio y hasta se puede saber las fechas y lugares donde conviene comprar.