Exploración de Instituciones Educativas en El Salvador mediante Web Scraping y Selenium.

Exploration of Educational Institutions in El Salvador through Web Scraping and Selenium.

Andrés Javier Guzmán Lovo
Estudiante de Ingeniería en Desarrollo de Software;
andr35guzman2001@gmail.com
Immner Nehemias Guevara Ramos
Estudiante de Ingeniería en Desarrollo de Software;
neheguevara14@gmail.com

Resumen

En este estudio, se empleó técnicas de web scraping y la biblioteca Selenium en Python para recopilar información sobre instituciones educativas en El Salvador. Se utilizó un script que extrae datos de una página web que lista diferentes sedes educativas, y se presentan los resultados en un formato estructurado. Se ofrece una explicación del código utilizado, destacando las bibliotecas clave y el proceso de extracción de datos. El artículo aborda la metodología utilizada, los resultados obtenidos y algunas consideraciones relevantes.

Abstract

In this study, web scraping techniques and the Selenium library in Python were used to collect information about educational institutions in El Salvador. A script was used that extracts data from a web page that lists different educational locations, and the results are presented in a structured format. An explanation of the code used is provided, highlighting key libraries and the data extraction process. The article addresses the methodology used, the results obtained and some relevant considerations.

Palabras Clave

Web Scraping, Selenium, Datos, Python, Extracción de Datos.

Keywords

Web Scraping, Selenium, Data, Python, Data Extraction.

Introducción

En este estudio, se adoptó una estrategia utilizando web scraping y Selenium en Python para recopilar información específica sobre sedes educativas en El Salvador. La elección de estas herramientas se fundamenta en su capacidad para

automatizar la navegación web y extraer datos de manera eficiente.

Metodología

El código implementado en Python utiliza dos bibliotecas clave: BeautifulSoup y Selenium. BeautifulSoup se utiliza para analizar y extraer información de la estructura HTML de las páginas web, mientras que Selenium se encarga de la automatización del navegador. El script accede a una página que lista instituciones educativas, navega por la información y extrae datos como nombres, códigos, direcciones y números de teléfono. La información se almacena en un formato tabular para su análisis posterior.

```
# Extracto del código Python
from bs4 import BeautifulSoup as bs
from selenium import webdriver
import pandas as pd
import random
import time

# Configuración del navegador
browser = webdriver.Firefox()
urlSedes = "https://portal.siges.sv/pp/sedes"
browser.get(urlSedes)

# Extracción de enlaces de instituciones
html = browser.page_source
soup = bs(html, 'lxml')
li_s = soup.find('ul', {'id': 'form:basicDT_list'}).find_all('li')
id_instituciones = [li.find('a').get('href') for li in li_s]

# Recopilación de datos de instituciones
institucion_list = []
for id_institucion in id_instituciones:
    # ... Código de extracción detallada ...

# Cierre del navegador
browser.quit()
```

Este código realiza las siguientes acciones:

1. **Importación de bibliotecas:**

Se importan las bibliotecas necesarias: BeautifulSoup para analizar el HTML de la página web, Selenium para la automatización del navegador, y pandas para manejar y estructurar los datos.

2. **Configuración del navegador:**

Se configura el navegador, en este caso, se utiliza Firefox. Es importante asegurarse de tener el controlador (geckodriver) instalado y en el PATH del sistema.

3. **Acceso a la página web:**

Se especifica la URL del portal de instituciones educativas y se utiliza Selenium para abrir el navegador y acceder a esa URL.

4. **Extracción de enlaces de instituciones:**

Se extraen los enlaces de las instituciones de la página utilizando BeautifulSoup. Se busca un elemento <ul> con el ID form:basicDT_list y se obtienen los enlaces de cada <li> dentro de esa lista.

5. **Almacenamiento de IDs en un DataFrame y CSV:**

En este paso del proceso, se realiza la creación de un DataFrame utilizando la biblioteca pandas en Python para almacenar las IDs de las instituciones educativas recopiladas mediante web scraping. Posteriormente, este DataFrame se guarda en un archivo CSV llamado **instituciones.csv**. El objetivo es organizar y estructurar de manera tabular las IDs para su posterior análisis y referencia

Este archivo CSV puede servir como referencia para futuros análisis y también proporciona una forma estructurada de almacenar las IDs de las instituciones educativas recopiladas

6. **Iteración sobre las instituciones:**

Se crea una lista (institucion_list) para almacenar información detallada de cada institución.

Se define una función parsear_institucion que toma una ID de institución, accede a la página específica de esa institución y extrae información como nombre, código, dirección y teléfono.

Se ejecuta esta función para la primera institución y se crea un DataFrame con esta información.

Luego, se ejecuta la función para el resto de las instituciones en un bucle, y la información se agrega al DataFrame.

Se añade un tiempo de espera aleatorio entre 1 y 3 segundos entre cada solicitud para evitar la detección de actividades automatizadas por parte del sitio web.

7. **Mostrar información detallada y guardar en un archivo CSV:**

Se imprime en la consola la lista de información detallada de cada institución (institucion_list).

Se guarda el DataFrame con la información de todas las instituciones en un archivo CSV llamado instituciones.csv.

8. **Cerrar el navegador:**

Finalmente, se cierra el navegador utilizando el método quit () de Selenium.

Resultados

Se obtuvo un conjunto de datos estructurado que contiene información detallada sobre varias instituciones educativas en El Salvador. Los resultados incluyen nombres de instituciones, códigos, direcciones y números de teléfono. Estos datos proporcionan una visión general de la distribución geográfica y características de las sedes educativas en el país.

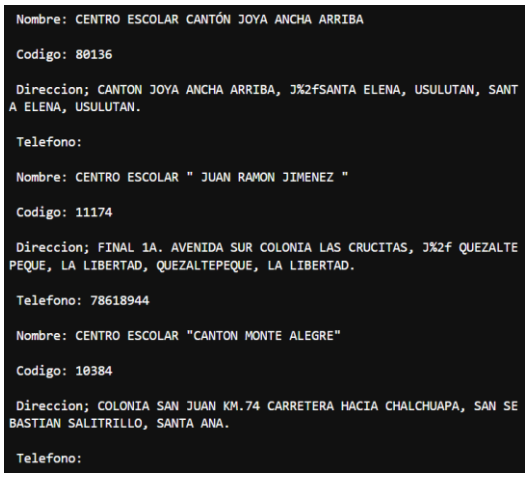

Ilustración 1Muestra de datos en tiempo de ejecución.

	Nombre	Codigo	Direccion	Telefono
0	CENTRO ESCOLAR MARCO RENE REVELO	90002	CENTRO PENAL APANTEOS, SANTA ANA, SANTA ANA.	2484-2800
0	CENTRO ESCOLAR MARCO RENE REVELO	90002	CENTRO PENAL APANTEOS, SANTA ANA, SANTA ANA.	2484-2800
0	CENTRO ESCOLAR CANTÓN SANTA RITA ALMENDRO	86376	CANTON SANTA RITA ALMENDRO, J%2f SANTIAGO NONU...	
0	CENTRO ESCOLAR CASERÍO CHICUMA, CANTÓN EL GAVILÁN	66091	CASERIO CHICUMA, C%2f EL GAVILAN, J%2f NUEVA C...	
0	CENTRO ESCOLAR "DE LA COLONIA SANTA MARTA"	10738	4ª CALLE ORIENTE NO. 2-6, COLONIA SANTA MARTA,...	2451-1661
...	...	...	...	...
0	COMPLEJO EDUCATIVO "CANTÓN LA FLOR"	11516	KM. 19 1%2f2 CARRETERA PANAMERICANA C%2flA FLO...	
0	CENTRO ESCOLAR CASERÍO SAN NICOLÁS, CANTÓN LAS...	86403	CASERIO SAN NICOLAS, CANTON LAS MARIAS, J%2f l...	
0	ESCUELA DE EDUCACIÓN PARVULARIA "DE SAN SEBAST...	12391	BARRIO SAN ANTONIO, CALLE PRINCIPAL, SAN SEBAS...	23967644

Ilustración 2Muestra de datos en un DataFrame de pandas

Conclusiones

El uso de técnicas de web scraping y Selenium facilitó la recopilación de datos sobre instituciones educativas. Los resultados obtenidos pueden ser útiles para analizar la distribución geográfica de las sedes educativas y para identificar patrones o tendencias en la información recopilada.

La combinación de web scraping y Selenium en Python ha demostrado ser una estrategia efectiva para recopilar datos detallados sobre instituciones educativas. Este enfoque no solo facilita la automatización del proceso, sino que también permite una adaptabilidad a diferentes estructuras de páginas web.

La elección de BeautifulSoup y Selenium se justifica por su flexibilidad y eficacia en la manipulación de contenido web dinámico. BeautifulSoup facilita la extracción de información a partir del HTML, mientras que Selenium automatiza la interacción con el navegador, asegurando la estabilidad en entornos web dinámicos.

Recomendaciones

Se sugiere realizar análisis más detallados utilizando los datos recopilados, como la identificación de áreas con mayor concentración de instituciones educativas o la comparación de características entre diferentes tipos de instituciones.

Promover la creación de espacios de colaboración entre instituciones educativas, donde puedan compartir recursos, metodologías exitosas y experiencias. Fomentar una cultura

Referencias

Foundation, P. S. (2023). Obtenido de https://www.python.org/

Selenium. (2023). *Selenium*. Obtenido de https://www.selenium.dev/documentation/en/

Soup, B. (s.f.). *Beautiful Soup*. Obtenido de https://www.crummy.com/software/BeautifulSoup/bs4/doc/