Universitat Oberta de Catalunya

Master in Data Science

M2.851 - Tipología y ciclo de vida de los datos

Práctica 1 (Web Scraping)

Abril 2018

Contenido

Presentación

En esta práctica se elabora un caso práctico orientado a aprender a identificar los datos relevantes por un proyecto analítico y usar las herramientas de extracción de datos.

Objetivos

El objetivo de esta actividad será la creación de un dataset a partir de los datos contenidos al web. Teneis que indicar las siguientes características del dataset general:

1. Titulo del dataset.

Aplicación de Web Scraping para la obtención de datos del mercado inmobiliario en la Cd. De México

2. Subtitulo del dataset.

Caso práctico para mostrar el uso de Web Scraping y Python para la obtención y elaboración de un conjunto de datos sobre el mercado inmobiliario de la Cd. De México a fin de conocer los precios de los inmuebles por localidad y zona dentro de la Cd. De México.

3. Imagen.

4. Contexto.

El caso práctico se enfoca en la información que proporciona un sitio dedicado a promover la adquisición y/o renta de bienes inmuebles a fin de conocer los precios de los inmuebles de casa habitación por localidad y zona dentro de la Cd. De México.

El caso puede extenderse hacia otras ciudades dentro del país y puede ser útil para conocer el costo de las viviendas entre una ciudad u otra.

El sitio utilizado no es un sitio del gobierno, sino que es un sitio comercial el cual el ejemplo utilizado puede ser de utilidad para que cierto desarrollador inmobiliario pueda conocer información sobre

sus competidores o inclusive conocer donde la competencia esta desarrollando proyectos y conocer los precios basado en la oferta y demanda de ciertas ciudades.

5. Contenido

Como caso práctico realizado por primera vez en el conocimiento del web scraping se tomaron los siguientes datos de 685 anuncios para el objetivo mencionado al principio del presente trabajo.

- Categoría de la Vivienda (Casa o Departamento)
- Municipio o Localidad (ejemplo. Coyoacán)
- Colonia o Vecindario dentro de la localidad (Ejem. Los Alpes)
- Precio.

Dentro del análisis del sitio, se pudo encontrar que no esta muy bien estructurado en cuando a su codificación HTML a fin de que se pueda facilitar la obtención de otros datos que pueden ser interesantes analizar junto a los mencionados como pueden ser:

- Superficie de la Vivienda
- Número de habitaciones
- Cajón para estacionamiento.

El sitio no menciona fecha de la publicación de los datos, por lo que se puede asumir que por la naturaleza de un sitio que cuenta con mucha publicidad y con menús de búsqueda de información se puede decir que hay buena probabilidad de que el sitio de encuentre al día en cuanto a la información que muestra.

6. Agradecimientos

Se agradece al sitio : https://www.avisosdeocasion.com/reforma para la realización de este trabajo con fines académicos, así como el tema y conocimientos proporcionados por la Universitat Oberta de Cataluña dentro del master de Ciencia de Datos.

7. Inspiración

Dentro de la creciente demanda de inmuebles en la Ciudad además de que los costos de las hipotecas no son baratas en el país puede ser interesante mostrar los costos de las viviendas por localidad para una persona que desea cambiar de localidad o vecindario dentro de la ciudad.

Así mismo, una de las aplicaciones interesantes del uso de web scraping es obtener información de sitios o páginas para poder analizar la información de algún área en específico.

8. Licencia

9. Código.

El código utilizado para la práctica fue en Python así como la experiencia de ser el primer código desarrollado con el uso de este lenguaje.

Para su ejecución se requirieron además de Python, la instalación de los siguientes paquetes que se incluyen con Python pero se requiere ejecutar por aparte su instalación.

BeuatifulSoap4

Requests

```
import requests
import csv
from bs4 import BeautifulSoup

#salida del archivo
f = csv.writer(open('inmuebles_mx.csv', 'w'))
f.writerow(['Categoria', 'Localidad', 'Colonia', 'Precio'])

paginas=[]

for i in range(1,10):
    url = "https://www.avisosdeocasion.com/reforma/venta/casas/casas.aspx?ntext=venta-casas-casas-Distrito-Federal&PlazaBusqueda=1&Plaza=1&pagina=" +str(i)
    paginas.append(url)

    for pagina in paginas:

        page = requests.get(pagina)
        soup = BeautifulSoup(page.content, 'html.parser')

        lista_anuncios = soup.find_all('td', class_="tituloresultchico")

        for anuncios_items in lista_anuncios:
            titulos = anuncios_items.contents[0]
            categoria = anuncios_items.find('h3').get_text()
```

```
localidad=anuncios_items.find('h4').get_text()

colonia = anuncios_items.find('h5').get_text()

detalle = anuncios_items.find('tr')

precio = detalle.find('td').get_text()

nprecio = precio.replace ("|", "")


# Agrega cada información del anuncio de inmuebles a un registro del archivo de salida


f.writerow([categoria, localidad, colonia, nprecio])


#print(titulos)

#print(localidad)

#print(colonia)

#print (nprecio)
```

10. Dataset en formato CSV

inmuebles_mx.csv

Referencias

El lenguaje Python. Universidad Oberta de Cataluña. David Masip Rodo.

Lawson, R. (2015). Web Scraping with Python. Packt Publishing Ltd. Chapter 2. Scraping the Data. Simon Munzert, Christian Rubba, Peter Meisner, Dominic Nyhuis. (2015).