Instituto Tecnológico de Costa Rica

Sede San Carlos.

Escuela de Computación

Curso: Introducción al Desarrollo de Aplicaciones para Web.

Profesor(a): Efrén Jiménez Delgado.

Estudiantes: Alvarado Gómez Cristian.

Grupo: 50

Periodo: 1 semestre 2017

# Resumen ejecutivo

Para poder desarrollar un buen web scraping hay que hacer una investigación para saber que plataformas utilizar, Python un lenguaje el cual tenga una amplia comunidad, por su estilo de codificar es uno de los más adecuados para implementar un webScraper.

Para realizar este trabajo si hizo uso de Python desde el IDE Pycharm, además se instalaron librerías como BeautifulSoup, re, lxml. Y por otro lado esta MySQL en cual se usará para almacenar estos datos.

# Tabla de contenidos

[Resumen ejecutivo 2](#_Toc479019744)

[Tabla de contenidos 3](#_Toc479019745)

[Introducción 4](#_Toc479019746)

[Desarrollo 5](#_Toc479019747)

[Conclusiones 7](#_Toc479019748)

[Recomendaciones 7](#_Toc479019749)

[Bibliografía. 8](#_Toc479019750)

# Introducción

Web scraping es una técnica que está tomando mucho auge en el ambiente computacional ya que brinda una gran facilidad para trabajar con la información ya sea de un sitio web o muchos sitios web en conjunto. Un punto importante con esta técnica es la privacidad de la información ya de cierta forma se está violando los derechos de copyright.

Para la implementación de un web scraping hay una gran cantidad de lenguajes, y métodos que se pueden utilizar, esto depende del problema que se quiera resolver. En este caso se implementara utilizando el lenguaje Python ya que tiene un amplia comunidad de desarrollo y además MSQL para almacenar todos los datos extraídos.

# Desarrollo

Para la implementación del web Scraping se hizo sobre el sitio web <https://www.encuentra24.com/> , de la cual se extraerán toda la información requerida:

* Localización.
* Marca.
* Modelo.
* Enviado.
* Precio.
* Año.
* Kilómetros.
* Motor.
* Imágenes.
* Descripción.
* etc.

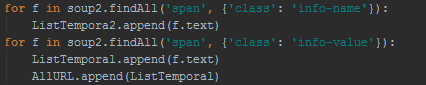
Para la implementación de esto se hará uso de la librería BeautifulSoup las cual nos permite extraer cualquier información del sitio web y guardarlo en una variable, seguidamente mediante una consulta a la base de datos en MySQL se inserta la información obtenida.

La información que se encuentra en esta página web es referente a la venta de diferentes tipos de vehículos, para este WebScraping se aplicó un filtro para solo obtener información sobre motos.

Inicialmente se hace un escaneo del sitio web en el cual se buscan todos los “href” los cuales re direccionan a la página web de cada Moto de la cual se obtendrá la información, para esto usamos:



Seguidamente sobre cada link obtenido aplicamos otro filtro para obtener los datos que necesitamos sobre la Moto:



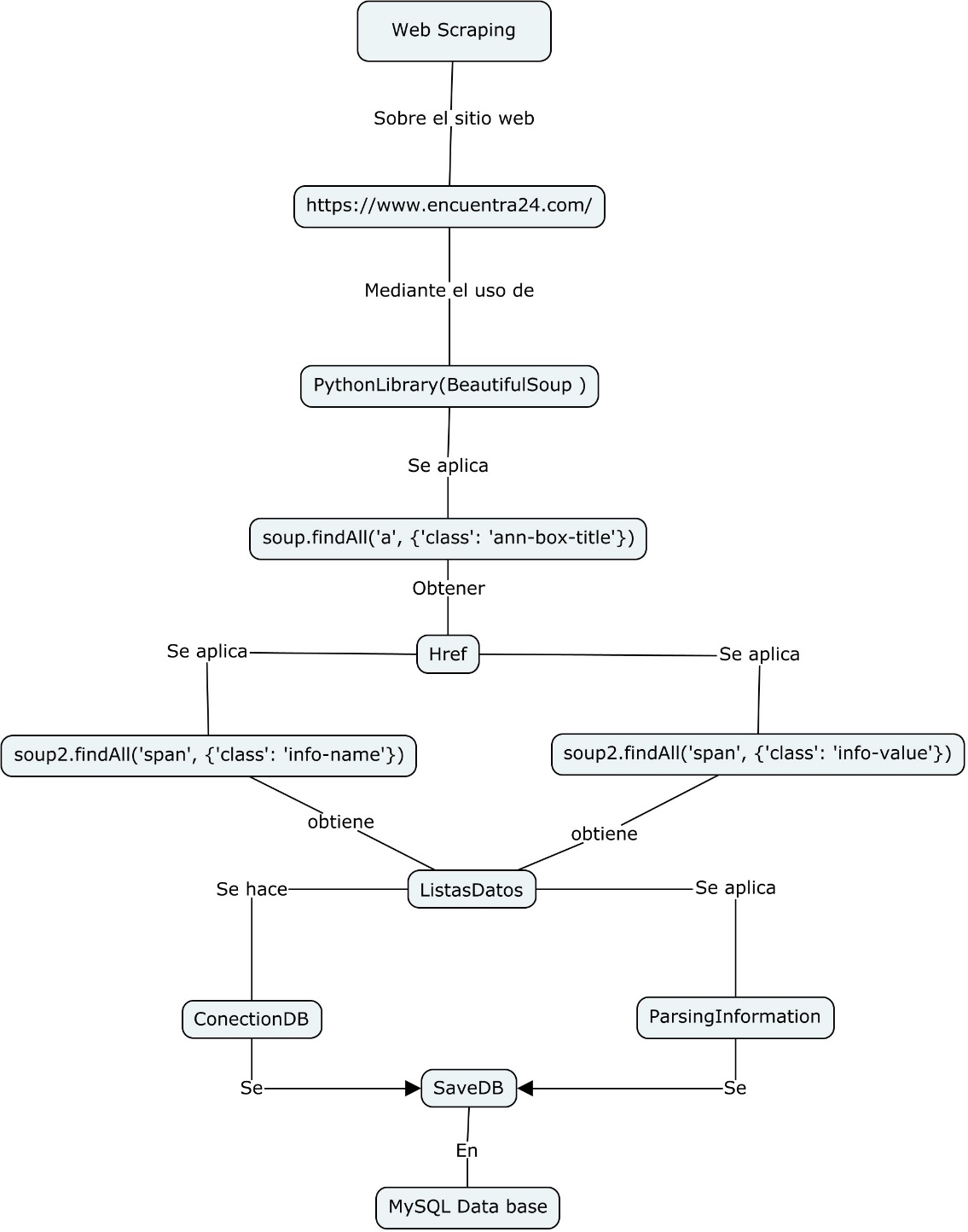
Para guardar en la base datos, se realiza mediante la libraría “**import** MySQLdb **as** mdb”, la cual permite hacer la conexión a la base datos “con = mdb.connect(DB\_HOST, DB\_USER, DB\_PASSWORD, DB\_NAME)”, con esto ya se puede crear la base datos y las tablas que se requieran.

Consulta para crear la tabla donde estarán los datos almacenados:

*sql = "CREATE TABLE encuentra24 (id INT PRIMARY KEY AUTO\_INCREMENT, localizacion VARCHAR(25), marca VARCHAR(25), modelo VARCHAR(25), Enviado VARCHAR(25), precio VARCHAR(25), ano VARCHAR(25), kilometros VARCHAR(25),motor VARCHAR(25),imagen VARCHAR(100))"*

consulta para insertar los datos obtenidos de la pagina web:

*cur.execute("INSERT INTO encuentra24 (localizacion,marca,modelo,Enviado,precio,ano,kilometros,motor,imagen) VALUES ('%s','%s','%s','%s','%s','%s','%s','%s','%s')" % (i[0],i[1],i[2],i[3],i[4],i[5],i[6],i[7],i[8]))*



# Conclusiones

A la hora de implementar este WebScraping logre entender el funcionamiento de la librería BeautifulSoup y algunas de sus propiedades. Un punto a considerar es antes de escoger una página para realizar este proceso, es realizar pruebas a la página que ver si permite obtener los datos requeridos, ya que hay sitios web que tiene políticas que no permiten este proceso.

# Recomendaciones

Antes de realizar este proceso investigar adecuadamente los lenguajes disponibles y las librerías, también realizar pruebas sobre la página para comprobar si es posible realizarle el proceso.

# Bibliografía.

© encuentra24.com/. (2017). *Motos Costa Rica - motos usados*. [online] Available at: https://www.encuentra24.com/costa-rica-es/autos-motos.11#search=f\_currency.CRC|number.50&page=1 [Accessed 29 Mar. 2017].

Python?, H. (2017). *How can I get href links from HTML using Python?*. [online] Stackoverflow.com. Available at: http://stackoverflow.com/questions/3075550/how-can-i-get-href-links-from-html-using-python [Accessed 27 Mar. 2017].

YouTube. (2017). *Introduction To Web Scraping (with Python and Beautiful Soup)*. [online] Available at: https://www.youtube.com/watch?v=XQgXKtPSzUI [Accessed 28 Mar. 2017].

YouTube. (2017). *Web scraping and parsing with Beautiful Soup & Python Introduction p.1*. [online] Available at: https://www.youtube.com/watch?v=aIPqt-OdmS0 [Accessed 25 Mar. 2017].