*Instituto Tecnológico de Costa Rica*

*Escuela de Ingeniería en Computación*

*Tarea #4 – Investigación “web scraping”*

*Estudiante: Luis Javier Ramírez Torres*

*Profesor: Efrén Jiménez*

*Sede San Carlos*

*15 Mayo 2017*

# Resumen Ejecutivo

El web scraping es un tema muy llamativo actualmente en el área de la informática y la comercial, la capacidad de extraer automáticamente mucha información de un sitio web sin la necesidad de estar despierto es un concepto que atrae cada vez más a las personas. En esta investigación se abordará la temática del web scraping, aportando datos específicos sobre algunas de las técnicas que se pueden abordar a la hora de realizar este proceso web scraping. Además, se desarrollará el ámbito legal que rodea esta temática, como también, algunas herramientas que llegan a facilitar el desarrollo del web scraping. Se mencionarán algunos lenguajes de programación que permiten el desarrollo de un buen web scraping, como también, una metodología para la extracción de información de una página web donde se evidencia el planteamiento inicial de un proyecto de web scraping, Algunas conclusiones y recomendaciones brindarán una guía a futuro para todas aquellas investigaciones o proyectos que involucren la temática del web scraping.

Contenido

[Resumen Ejecutivo 2](#_Toc482615851)

[Introducción 4](#_Toc482615852)

[Desarrollo 4](#_Toc482615853)

[Técnicas 4](#_Toc482615854)

[Cuestiones legales 5](#_Toc482615855)

[Herramientas 5](#_Toc482615856)

[Lenguajes de programación 6](#_Toc482615857)

[Metodología de extracción para una página web 7](#_Toc482615858)

[Elaboración de la extracción de información de una página web 9](#_Toc482615859)

[Almacenamiento de los datos y alojamiento de la base de datos en la nube 12](#_Toc482615860)

[Conclusiones 14](#_Toc482615861)

[Recomendaciones 14](#_Toc482615862)

[Bibliografía 15](#_Toc482615863)

# Introducción

Cada página web que existe en internet brinda mucha información, unas más que otras por supuesto. Dicha información se nos presenta en forma de documentos HTML y en todo momento. Esta información puede que no signifique mucho si es de solo un momento dado, pero cuando se recopila bloques de información en todo momento, estos datos comienzan a tener bastante valor si se les da el uso adecuado. El “web scraping” es una técnica que toma como base todos esos datos en forma de documento html para transfórmalos en una pila de estructurada de información que se puede almacenar en una base datos para su futuro análisis o para su uso inmediato en otra aplicación o página web.

# Desarrollo

El “web scraping” es una técnica que permite extraer información de los sitios web mediante la utilización de software, dicho software simulará la navegación humana dentro del sitio en cualquier momento y de forma automática.

## Técnicas

Algunas de las técnicas existentes para realizar “web scraping” son las siguientes:

* Expresiones regulares: Esta técnica ayuda que el contenido que queremos extraer sea el más preciso posible.
* Protocolo HTTP: Algunas páginas web estáticas y dinámicas pueden ser obtenidas utilizando peticiones HTTP al servidor remoto.
* Parsers de HTML: Algunos lenguajes permiten parsear, recuperar y transformar el contenido de documentos HTML.

## Cuestiones legales

En el ámbito legal hay que tener en cuenta 2 factores, la legalidad vigente en el país de origen y las políticas de uso o condiciones legales del sitio web al que se le está haciendo “web scraping”. Desde este punto es importante conocer si mediante el “web scraping” se estaría violentando sus derechos de autor, su propiedad intelectual y si al utilizar dichos datos se podría considerar como competencia desleal, imitación, denigración de reputación, entre otros.

Ahora bien, si dicha extracción de datos es para consumo propio, es decir, los datos no se van a hacer públicos, ni se va a lucrar con ellos, se puede afirmar que dicha extracción de datos es completamente legal.

## Herramientas

Algunas herramientas existentes que facilitan la realización de “web scraping” son las siguientes:

* Webscraper.io: Esta herramienta es un plugin para el navegador Google Chrome, se utiliza un poco de expresiones regulares y maquetación web, por lo tanto es importante tener conocimientos en programación y maquetación web.
* Import.io: Esta herramienta es un poco más básica y fácil de utilizar, desde el panel de control web se pueden realizar “scrapeos” básicos, no obstante, si se desean realizar operaciones más complejas es necesario descargar el programa. En resumen, es una herramienta que se podría utilizar para todos aquellos que apenas se vienen adentrando al mundo del “web scraping” debido a que no se necesita de conocimiento, solo estar un poco relacionado con el mundo web y la visualización de datos en herramientas como Excel.
* Scrapy.org: Esta herramienta está abierta a todo el mundo y es mucho más compleja de utilizar que las 2 anteriores, en este caso, es necesario tener conocimientos avanzados del lenguaje Python ya que solo funciona con dicho lenguaje, además, esta herramienta está pensada para proyectos que no requieran mucho de visualizar los datos extraídos ya sea en hojas de cálculo de Excel u otra herramienta.

## Lenguajes de programación

Algunos lenguajes de programación que se pueden utilizar para hacer “web scraping” son los siguientes:

* Python
* PHP
* Java
* Javascript
* XQuery
* HTQL
* NodeJS
* R
* GO
* C++

## Metodología de extracción para una página web

Equipo a utilizar:

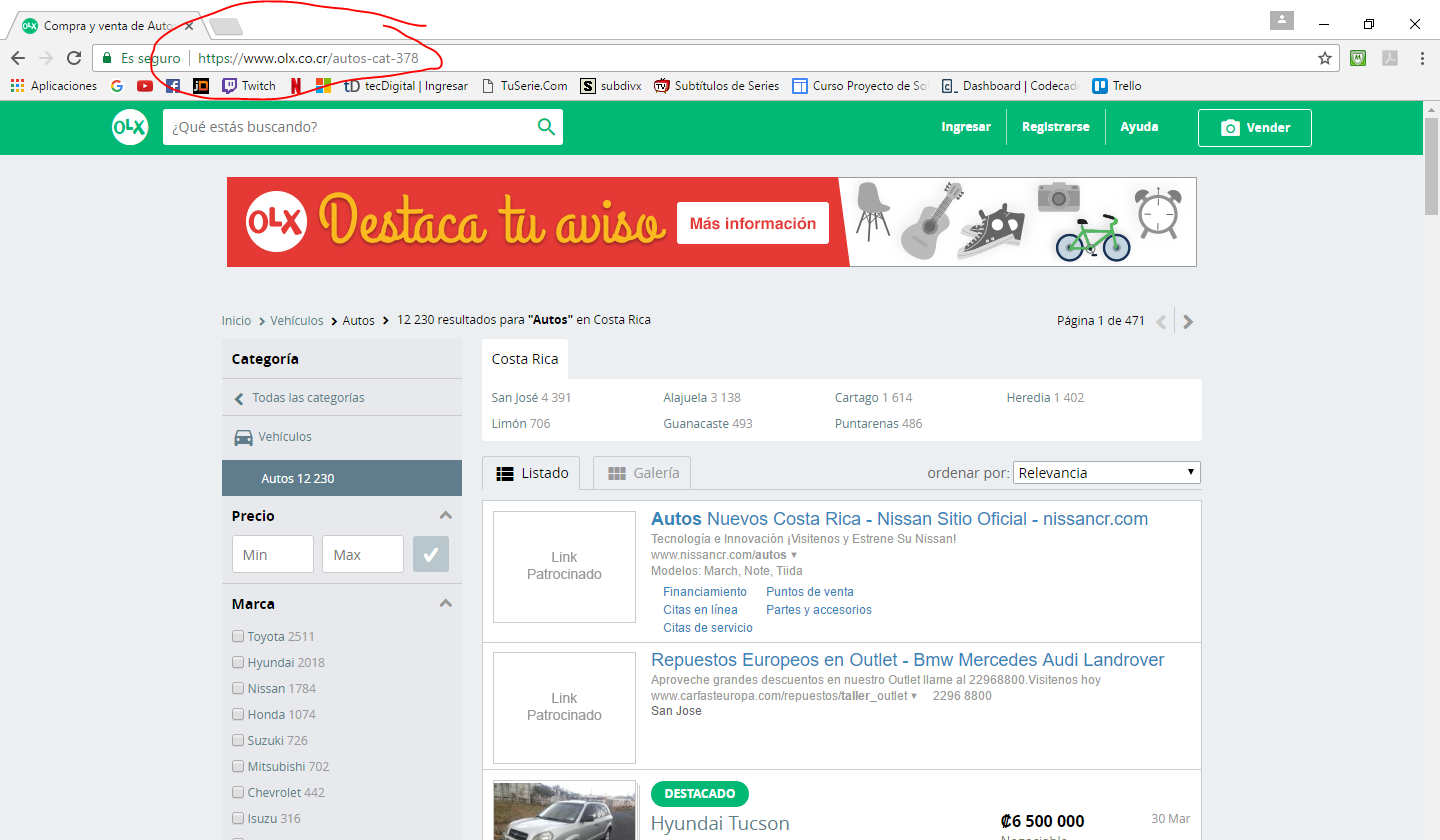
Página web: olx.co.cr

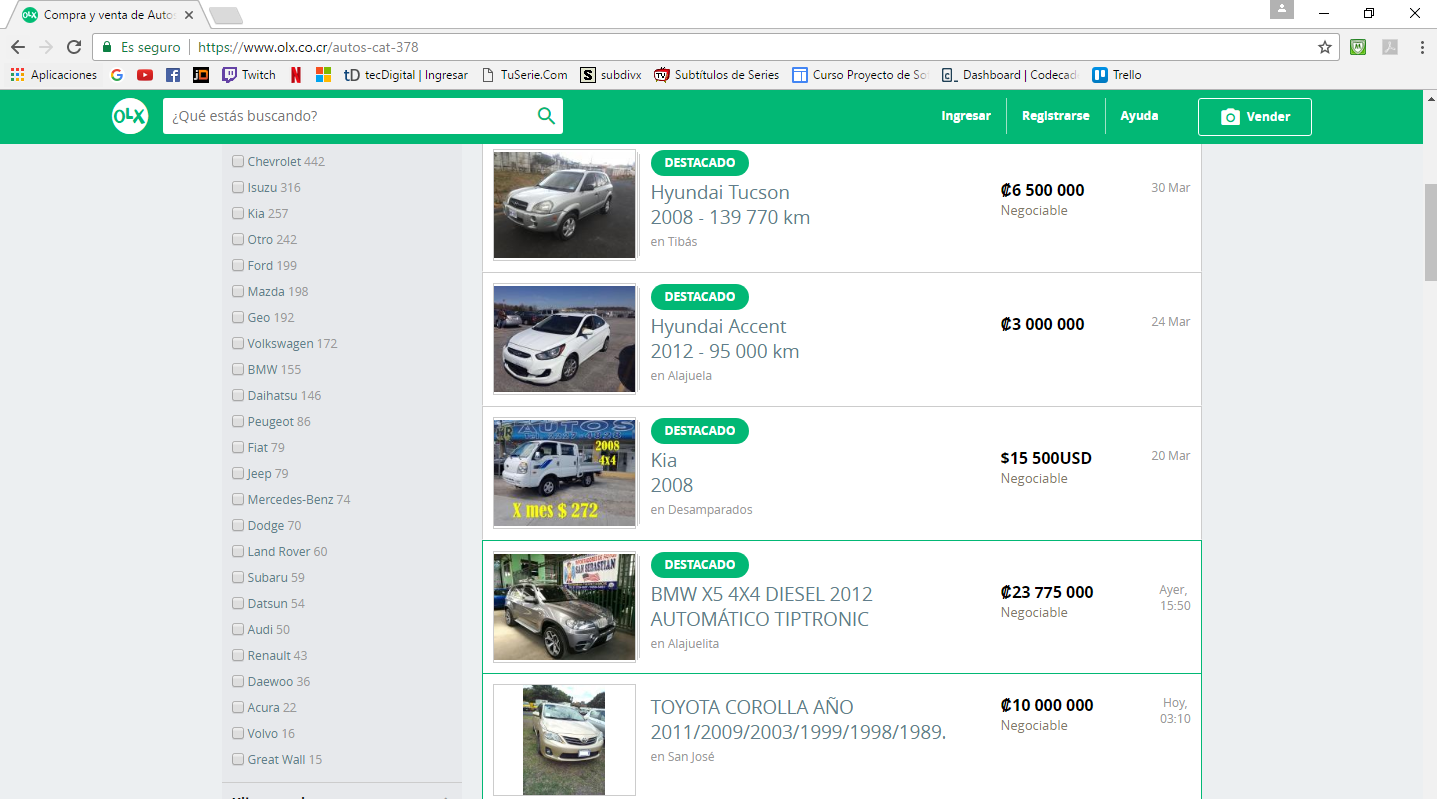
Lenguaje a utilizar: Python

Base de datos a utilizar: PostgreSQL

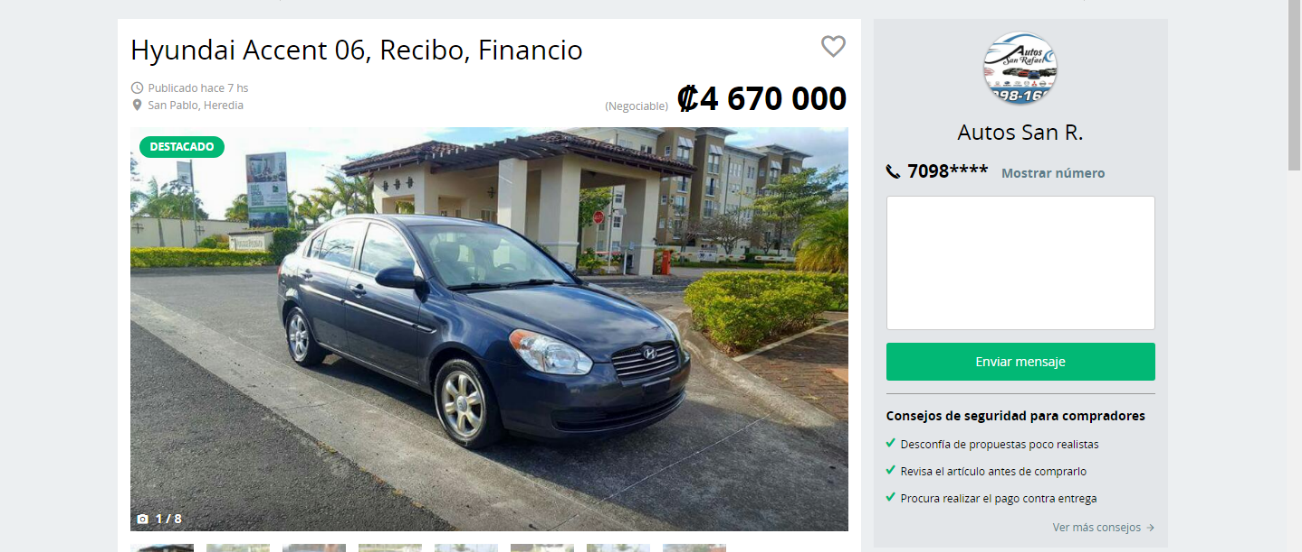
Librerías: Selenium 3.3.1, psycopg2 2.7.1  
  
Pasos a realizar:

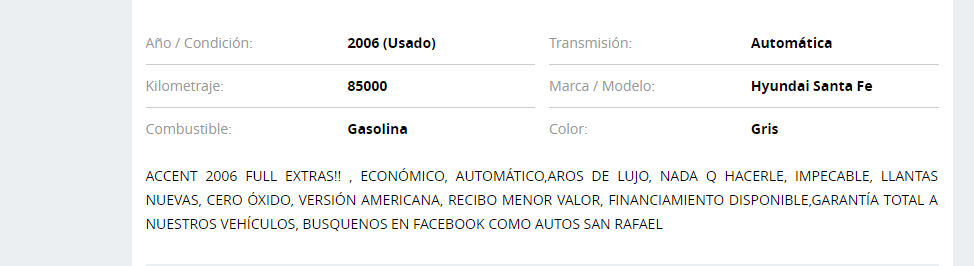
Primeramente se ingresará a la página web olx.co.cr, específicamente a la sección de autos.



Seguidamente aparecerá una lista con todos los autos que están en venta.

Se seleccionará el primer auto que aparece en la lista para que nos re-direccione a un apartado donde aparecerán los datos del auto.





Una vez extraídos los datos de la información general del auto, se comenzará el proceso de parsear dichos datos en estructuras y atributos definidos, para que así se puedan almacenar en la base de datos.

Luego, nos devolveremos a la lista anterior con los autos usados, y seguiremos con los siguientes autos, extrayendo su información general, parseando sus datos y almacenándolos en la base de datos.

Estos datos se utilizarán para sacar estadísticas que muestren las características de los autos usados que se ponen a la venta en mayor cantidad en la página de olx.co.cr, además de filtrar por medio de datos específicos como kilometraje y lugar de venta para así analizar de donde son puestos más carros a la venta entre otras cosas más.

## Elaboración de la extracción de información de una página web

En la página de olx.co.cr se pueden encontrar a la venta distintos tipos de artículos, desde dispositivos electrónicos hasta propiedades como casas o lotes. Para la elaboración de esta investigación se delimitó solo a extraer la información de los autos puestos a la venta en esta página.

Información de los atributos a extraer y su XPATH de selección:

* **Título**: Atributo que se utiliza para darle un nombre a la publicación del auto.  
  XPATH: "//\*[@id='item\_index\_view']/article/div/section[1]/header/div[1]/h1"
* **Hora Publicación**: Atributo que indica hace cuanto se realizó la publicación del auto.  
  XPATH: "//\*[@id='item\_index\_view']/article/div/section[1]/header/div[2]/p[1]/time"
* **Ubicación**: Atributo que indica la ubicación donde se puso a la venta el auto.  
  XPATH: "//\*[@id='item\_index\_view']/article/div/section[1]/header/div[2]/p[1]/span"
* **Precio**: Atributo que indica el precio al que se está vendiendo el automóvil.  
  XPATH: "//\*[@id='item\_index\_view']/article/div/section[1]/header/div[2]/p[2]/strong"
* **Año**: Atributo que indica el año del auto.  
  XPATH: "//\*[@id='item\_index\_view']/article/div/section[1]/ul/li[1]/span"
* **Uso**: Atributo que indica si el auto está usado o no.  
  XPATH: "//\*[@id='item\_index\_view']/article/div/section[1]/ul/li[1]/span"
* **Kilometraje**: Atributo que indica el kilometraje que posee el auto que está a la venta.  
  XPATH: "//\*[@id='item\_index\_view']/article/div/section[1]/ul/li[3]/span"
* **Combustible**: Atributo que indica el tipo de combustible que utiliza el vehículo.  
  XPATH: "//\*[@id='item\_index\_view']/article/div/section[1]/ul/li[5]/span"
* **Transmisión**: Atributo que indica el tipo de transmisión que utiliza el vehículo.  
  XPATH: "//\*[@id='item\_index\_view']/article/div/section[1]/ul/li[2]/span"
* **Modelo**: Atributo que indica el modelo que tiene el vehículo.  
  XPATH: "//\*[@id='item\_index\_view']/article/div/section[1]/ul/li[4]/span"
* **Color**: Atributo que indica el color que posee el vehículo.  
  XPATH: "//\*[@id='item\_index\_view']/article/div/section[1]/ul/li[6]/span"
* **Descripción**: Atributo que indica alguna descripción que el vendedor quiera aportar en la publicación de la venta del auto.  
  XPATH: "//\*[@id='item\_index\_view']/article/div/section[1]/p"
* **Vendedor**: Atributo que indica el nombre de usuario del que realizó la publicación de la venta del automóvil.  
  XPATH: "//\*[@id='item\_index\_view']/article/aside/div[1]/div[2]/p[1]"
* **Teléfono**: Atributo que indica el teléfono del usuario que realizó la publicación de la venta del automóvil.  
  XPATH: "//\*[@id='item\_index\_view']/article/aside/div[1]/div[3]/a"
* **Imagen**: Atributo que indica la imagen del auto que está a la venta.  
  XPATH: "//\*[@id='item\_index\_view']/article/div/section[1]/section/figure/a[1]/img"

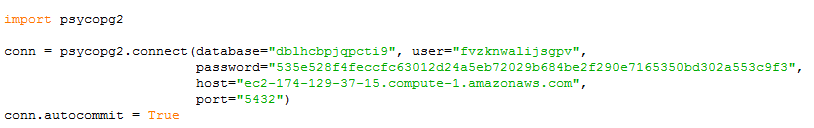
## Almacenamiento de los datos y alojamiento de la base de datos en la nube

Para la base de datos se utilizó Postgresql y se alojó en la nube en la plataforma de heroku.

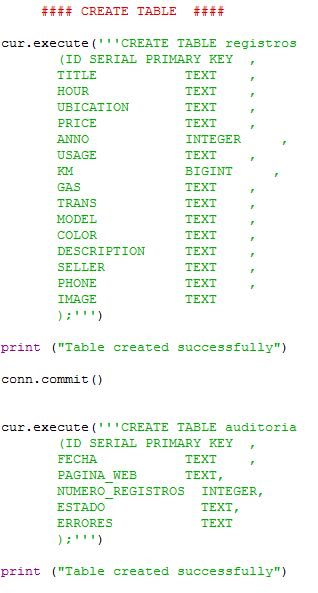
Primeramente se crea una base de datos en heroku de manera gratuita utilizando el add-on que nos brinda llamado “Heroku Postgres”, una vez hecho esto la plataforma nos proveerá con los credenciales de la base de datos para que la podamos acceder desde cualquier aplicación.

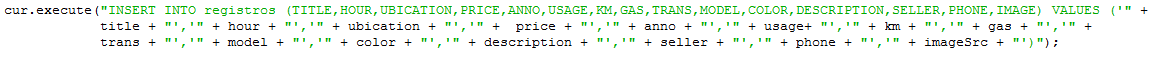


Ahora con estas credenciales la aplicación web scraping en Python se puede conectar a la base de datos en la nube para poder extraer e insertar datos a la misma.



Seguidamente se crean las tablas de la base de datos desde el código y se insertan los datos extraídos con el web scraping.





# Conclusiones

* No cabe duda alguna que este mundo de la web es gigantesco, la cantidad de datos que se puede encontrar parece ser ilimitado. Ahora bien, sin técnicas como el web scraping, la recopilación de estos datos es muy difícil, tediosa y puede costar mucho tiempo.
* El web scraping se puede hacer con varias técnicas, ahora bien, dependiendo del proyecto que se esté realizando se puede combinar dichas técnicas para provocar un resultado más favorable.
* Respecto a las cuestiones legales, si bien es cierto que muchos pueden considerar el web scraping como una estafa o un “hack”, hay que tener en cuenta que lo que va a definir si el web scraping realizado está dentro o fuera de los ámbitos legales son las políticas vigentes en el país de origen y las condiciones legales del sitio web al que se le realizó el scraping.
* Actualmente existe un conjunto de herramientas que hacen del web scraping una técnica más sencilla de realizar, pero al mismo tiempo, existe otro conjunto de herramientas que permite trabajar el web scraping de una forma más compleja y elaborada, permitiendo así que se obtenga el mejor resultado posible al tratar de recopilar la información que se requiera.
* El web scraping es muy amplio en el panorama de los lenguajes de programación que se pueden utilizar para realizarlo.

# Recomendaciones

* Investigar bien de las técnicas webscraping que más te convengan a la hora de realizar tu proyecto.
* Informarse de las políticas vigentes en tu país de origen respecto a este ámbito y de las condiciones legales del sitio al que tiene tu proyecto pensado realizarle web scraping.
* Buscar la herramienta que te ayude a realizar el web scraping de tu proyecto de la mejor manera posible, ya sea que andes buscando facilidad de uso o rendimiento en operaciones complejas.

# Bibliografía

Martí, M. (08 de 04 de 2016). *Site Labs*. Obtenido de Site Labs: https://sitelabs.es/web-scraping-introduccion-y-herramientas/

Neuman, A. (22 de 11 de 2013). *Vida Digital*. Obtenido de Vida Digital: https://vidadigital.com.pa/que-es-el-web-scraping-o-screen-scraping-y-por-que-nos-debe-importar/

Ramírez, W. (s.f.). *Tribuna Hacker*. Obtenido de Tribuna Hacker: http://www.tribunahacker.com.ar/2015/03/hacking-para-arrancar-tecnicas-para-web-scrapping/

**Referencias de repositorios indexados**

Myers, D., & McGuffee, J. (2015). Choosing Scrapy. *Journal of Computing Sciences in Colleges*, 83-89.

Kanaoka, K., Fujii, Y., & Toyama, M. (2014). Ducky: a data extraction system for various structured web documents. *Proceedings of the 18th International Database Engineering & Applications Symposium*, 342-347.