Instituto Tecnológico de Costa Rica

Escuela de Ingeniería en Computación.

Nombre del curso:

Introducción al desarrollo de aplicaciones WEB

Informe de Investigación Web Scraping

Profesor: Efrén Antonio Jiménez Delgado

Estudiante: Kevin Alvarado Varela.

Fecha de entrega: Semana 12

Grupo: 50

Periodo: 1 semestre 2017

# Resumen Ejecutivo

El presente documento brinda información referente al web scraping, ¿qué es?, ¿Qué técnicas existen? ¿Qué herramientas existen?, beneficios, ejemplos de aplicación y condiciones legales para su adecuada implementación.

Además, se explicará e implementará una metodología específica para un sitio web y el almacenaje de información seleccionada en una base de Datos. También se almacena online para su libre acceso. Los lenguajes y herramientas a utilizar son Python, PyCharm, beautifulSoup,json y firebase.

El proyecto se encuentra alojado en el siguiente link de un repositorio de GitHab:

<https://github.com/kevinAndres126/webscrapping>

Tabla Contenidos

[Resumen Ejecutivo 2](#_Toc477205868)

[Introducción 2](#_Toc477205869)

[Desarrollo 2](#_Toc477205870)

[Ejemplos de aplicación 3](#_Toc477205871)

[Técnicas 3](#_Toc477205872)

[Manualmente: 3](#_Toc477205873)

[Patrones de Texto: 3](#_Toc477205874)

[Programación HTTP: 3](#_Toc477205875)

[HTML parsing: 4](#_Toc477205876)

[DOM parsing: 4](#_Toc477205877)

[Reconocimiento de anotaciones semánticas: 4](#_Toc477205878)

[Implicaciones legales 4](#_Toc477205879)

[Herramientas existentes 4](#_Toc477205880)

[Metodología 5](#_Toc477205881)

[Conclusiones 5](#_Toc477205882)

[Recomendaciones 5](#_Toc477205883)

[Bibliografía 5](#_Toc477205884)

# Introducción

EL análisis de datos es uno de los principales campos tecnológicos en los que se ha enfocado el siglo XXl. En el mundo de la tecnología se empieza a generar un pensamiento universal; “Todo en este mundo es información”, si se es capaz de analizar toda esta información seremos capaces de realizar cualquier cosa.

Gracias a la estadística y las probabilidades es posible sacar el mayor provecho al análisis de toda esta información. Este informe busca dar a conocer una pequeña parte de lo que el análisis de datos permite, y aun bajo el análisis de datos no relevantes o de gran importancia como lo son los de una página de ventas.

# Desarrollo

Web Scraping, la traducción al español seria “escarbar en la web”. Es decir, es una técnica que permite extraer información de la Web. Esta técnica es aplicada de forma automatizaba, lo cual permite extraer cantidades masivas de información, mediante la aplicación de algoritmos de búsqueda se extraer solo la información que se necesita y se analiza.

Por otra parte, podemos extraer esos datos de manera estructurada y almacenarla en una base de datos. Un ejemplo de funcionamiento seria extraer el título de 500 páginas que se encuentran en un mismo sitio, es decir, una página con elementos repetitivos como páginas de ventas, alquiler, información, etc.

# Ejemplos de aplicación

* Rastrear la posición en google de las entradas de un nuestro.
* Crear un bot para rastrear los twitteres relacionados a un hashtag especifico.
* Para escapear una determinada página y generar nuestra propia base de datos.

El web scraping no es algo sencillo de realizar, desde el punto de vista de beneficio, es necesario identificar y reconocer cual es exactamente la información que se quiere analizar, para que nos brinden los datos que representen un mayor beneficio; y por el otro lado, en la parte técnica es necesario conocimiento básico de las estructuras html y de regex para realizar el scraping de la manera más óptima posible. También es de vital importancia el manejar de herramientas de software que nos permitan visuales los datos recolectados.

# 

# Técnicas

Manualmente: realizando un copiado y pegado de la información que se desea (no sería automatizado, pero sigue siendo una técnica).

Patrones de Texto: Los patrones permitirán identificar la información que necesitamos y nos permitirán extraerla.

Programación HTTP: Mediante la utilización de un socket programado de peticiones podremos recuperar las páginas.

HTML parsing: Mediante el uso de un wrapper se accede a la información desestructurada y se traduce a información relacionada o estructurada.

DOM parsing: Se puede scrapear la página atreves de árbol generado por el DOM en exploradores como Internet Explorer y Mozilla.

Reconocimiento de anotaciones semánticas: Se scrapea la página mediante el reconocimiento de metadatos o anotaciones.

# 

# Implicaciones legales

Para la utilización legal del web scraping se debe cumplir las siguientes condiciones:

Ausencia de aceptación expresa de los términos y condiciones que prohíban el scraping.

Ausencia de una base de datos en el sentido de la ley de propiedad intelectual.

Evitar la extracción, definida como a transferencia permanente o temporal de toda o de una parte sustancial de la base de datos a otro soporte distinto de la base de datos original.

# Herramientas existentes

En la actualidad existen muchas herramientas entre ellas:

Data Toolbar

Heritrix

HTTrack

Kantu

Octoparse

Lenguajes como Python permiten muy fácilmente la implementación de web scraping, No obstante, también se puede realizar en Node.js, java,R y otros.

# Metodología

Este informe presenta un caso de prueba de web scraping con python y la librería beautifulsoup, esta permite fácilmente la realización de este.

Requerimientos:

* Instalar Python: Descargas instalador oficial de la página de Python la versión 2.X.
* Instalar pip: Descargas el archivo de instalación para Windows y lo ejecutas desde consola.
* Instalar un IDE para Python: Puede ser cualquier IDE, en este caso se utilizó PyCharm, este lo puedes descargar desde la página oficial de JetBrains
* Instalar Beautifulsoup: Una vez instalado pip descargas beautifulsoup, mediante el comando ‘pip install beatifulsoup’
* instalar requests: Este se instala de la misma manera que BeatifulSoup pero cambiado beautifulsoup por ‘requests’ en el comando.
* FireBase: Este también se instala de la misma manera.

Pasos a seguir

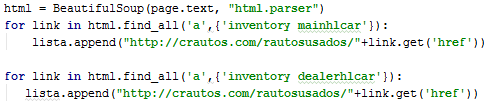
* Crear nuevo proyecto desde tu IDE
* Importar paquete requests y json de la siguiente manera



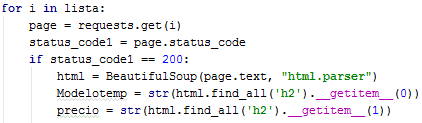
* Crear método para obtener la página, mediante el método get de la librería requests.



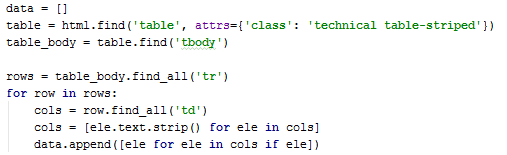
* Recorremos la página en busca de los links necesarios para obtener los datos que deseamos de cada artículo. Esto se implementó de la siguiente manera. Html es nuestra página parseada por beautiful soup y mediante métodos como find\_all que nos brinda beautifulsoup podemos encontrar lo que necesitamos en este caso las etiquetas ‘a’ de la clase ‘inventory mainhlcar’ y ‘inventory dealerhlcar’.



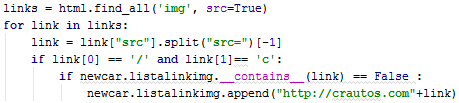
* Teniendo almacenados nuestro link podremos realizar el mismo procedimiento anterior, pero en cada link, extrayendo la información que deseamos. En este caso los ‘h2’.

.

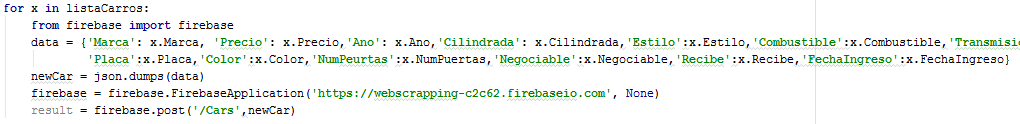
* En esta página es necesario recorrer una página para obtener los datos, para esto se utilizó el siguiente código, el cual nos permite almacenar los datos en forma de matriz, para luego accederlos.



* También se desea obtener imágenes, para ello se implementa el siguiente código:



Este código solo almacena la dirección web de la imagen, no la imagen como objeto

* Para almacenar los datos de manera online se utilizó firebase, de la siguiente manera:

Los datos fueron guardados en memoria mediante una clase; listaCarros contiene todos los datos. Se crea un diccionario para guardar los datos; se les aplica la función dumps de json para transformarlos, esto debido a que firebase solo maneja archivos json,por ultimo realizamos un post() del json a la base de datos en firebase.

* Para obtener los datos de firebase realizamos una petición mediante el método get() de firebase, especificando lo que deseamos obtener.

# Conclusiones

* La implementación de web scraping no es fácil, se necesita conocimiento básico en programación html y manejo masivo de datos.
* El uso inadecuado puede acarrear consecuencias legales a los implicados, es importante informarse al respecto antes de realizar cualquier tipo de scraping en la web.
* Es necesario realizar un análisis previo a la página, para reconocer la mejor técnica de web scraping a utilizar.
* La investigación es necesaria, debido que supone el uso de muchas herramientas de programación para su implementación.

# Recomendaciones

* Tener conocimiento básico en programación web.
* Dominar regex(regular expression) para mejorar la precisión de las búsquedas y el filtrado de las mismas.
* Dominar el tema en el cual se está realizando el scraping.
* Tomar en cuenta que la mayoría de las veces los datos no se obtienen de la manera que se desea, por lo que es necesario aplicar métodos y técnicas que nos permitan transformar esos datos a la forma deseada.
* La investigación es fundamental antes de realizar un proyecto de este tipo, se debe investigar todo lo posible al respecto.

# Bibliografía

Elnuevoderecho.blogspot.com. (2017). *Implicaciones legales del uso del web scraping (o screen scraping)*. [online] Available at: http://elnuevoderecho.blogspot.com/2015/05/implicaciones-legales-del-uso-del-web.html [Accessed 13 Mar. 2017].

Google Books. (2017). *Patent US7072890 - Method and apparatus for improved web scraping*. [online] Available at: https://www.google.com/patents/US7072890 [Accessed 11 Mar. 2017].

Search.proquest.com. (2017). *Web Content Mining Techniques: A Survey - ProQuest*. [online] Available at: http://search.proquest.com/openview/10c85aa791b7834617b53dfd771ed486/1?pq-origsite=gscholar&cbl=136216 [Accessed 12 Mar. 2017].

Sitelabs. (2017). *Qué es el Web scraping? Introducción y herramientas*. [online] Available at: https://sitelabs.es/web-scraping-introduccion-y-herramientas/ [Accessed 13 Mar. 2017].

Vargiu, E. and Urru, M. (2017). *Exploiting web scraping in a collaborative filtering- based approach to web advertising*. [online] Sciedu.ca. Available at: http://www.sciedu.ca/journal/index.php/air/article/view/1390/1115 [Accessed 12 Mar. 2017].

Docs.python.org. (2017). *18.2. json — JSON encoder and decoder — Python 2.7.13 documentation*. [online] Available at: https://docs.python.org/2/library/json.html [Accessed 1 May 2017].

Firebase. (2017). *Firebase*. [online] Available at: https://firebase.google.com [Accessed 10 May 2017].

Pypi.python.org. (2017). *python-firebase 1.2 : Python Package Index*. [online] Available at: https://pypi.python.org/pypi/python-firebase/1.2 [Accessed 10 May 2017].