Instituto Tecnológico de Costa Rica Unidad de computación

Primera investigación.

Estudiante:

Kimberly Morales Arias 2014096395

Sede San Carlos.

Fecha de entrega: 15 de mayo de 17

[Introducción 3](#_Toc482648798)

[Desarrollo: 3](#_Toc482648799)

[Herramientas: 3](#_Toc482648800)

[BeautifulSoup: 3](#_Toc482648801)

[Planteamiento: 3](#_Toc482648802)

[Conclusiones: 3](#_Toc482648803)

[Bibliografía 4](#_Toc482648804)

# Introducción

Actualmente en internet la transferencia de datos se ve reflejada en niveles impresionantes, en un minuto millones de correos electrónicos son enviados, en 24 horas la cifra es impresionantemente grande. Existen muchas páginas en internet, algunas estáticas que únicamente brindan información al usuario y otras dinámicas que permiten la interacción de los usuarios, por ejemplo: páginas de compras en línea. Toda esta información es pública y puede consultarse con diversos fines desde cualquier parte del mundo, las grandes empresas o negocios pagan por esta información, por datos importantes que ayudan a construir estadísticas, a raíz de la necesidad de obtener esta información lo más rápido y exacta posible nace el “web scraping” que consiste en una técnica que logra extraer información de las páginas de forma automatizada.

# Desarrollo:

## BeautifulSoup:

Librería para Python que permite la extracción sencilla de datos específicos de un sitio web en HTML sin mucha programación, una de las ventajas es que toda la

## Planteamiento:

Se eligió una página web costarricense dedicada a la venta de autos llamada crautos, estos autos poseen muchísimas características que son de gran relevancia. Mediante el lenguaje de programación Python y el uso de la herramienta BeautifulSoup se logró extraer alguna información, aunque no toda.

# Conclusiones:

Existen varias formas de realizar la extracción de información de sitios y páginas web, sin embargo se debe analizar muy bien el manejo que se harán con los datos, tomando en cuenta que aunque la información sea pública esta tiene un autor quien es el propietario legítimo, al hacer web scraping estamos extrayendo una información ajena y empleándola en fines propios, aunque a veces pareciera inofensivo a veces puede ser muy valiosa esta captura de información por lo cual se debe ser muy responsable.

# Bibliografía

Debina Laishram,Merin Sebastian. (2015). Extraction of Web News from Web Pages Using a Ternary Tree Approach. *Advances in Computing and Communication Engineering (ICACCE), 2015 Second International Conference.* IEEE.

Deepak Kumar Mahto,Lisha Singh. (2016). A Dive into Web Scraper World . *International Conference on Computing for Sustainable Global Development (INDIACom).*

Eloisa Vargiu1, 2, Mirko Urru1. (2012). Exploiting web scraping in a collaborative filteringbased. *Artificial Intelligence Research, 2013, Vol. 2, No. 1.*

Malik, Sanjay Kumar ; , SAM Rizvi;. (s.f.). Information Extraction Using Web Usage Mining, Web Scrapping and Semantic Annotation. *Computational Intelligence and Communication Networks (CICN), 2011 International Conference.* IEEE.

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

Rizqi Putri Nourma Budiarti , Nanang Widyatmoko , Mochamad Hariadi and Mauridhi Hery Purnom. (2016). Web Scraping for Automated Water Quality. *International Seminar on Intelligent Technology and Its Application.*