

**Instituto Tecnológico de Costa Rica**

**Introducción al desarrollo de aplicaciones web**

**Tarea de Investigación 1**

**Profesor:**

**Efrén Jiménez**

**Integrantes:**

**Andrés Pérez Badilla**

**Sede San Carlos**

**Fecha de entrega: lunes 13 de marzo de 2017**

**Implementación Web Scrapping**

Retomando lo avanzado y dicho anteriormente sobre la propuesta de desarrollo para el web scrapping el cual era una técnica utilizada para extraer datos de sititos sitios web, la cual implementa navegación en los sitios web simulando usuarios, lo que le permite extraer información del sitio web.

Según la idea implementación plantee el uso de web scrapping en el sitio web Amazon, pero al intentar concretar dicha misión propuestas surgieron problemas de camino, el cual genero el fracaso de intentar el uso de web scrapping sobre esta página ya que Amazon no permite la implementación de web scrapping en su sitio web.

En cuanto a las tecnologías planteadas para aplicar dicha metodología al estar haciendo ciertas pruebas me incline por el uso de Python para poder aplicar estas lo dicho, y al no poder hacer uso de Amazon como página fuente de datos decide probar con el sitio web de noticia la nación de Costa Rica de la cual se puede extraer variedad de contenido como lo son audios, videos, imágenes, texto, fechas entre otros datos.

Una vez en Python para poder implementar dicha tecnología se utilizará una librería llamada beautifulsoup la cual nos permite hacer la tarea de cargar la página web ya propuesta que es [www.nacion.com](http://www.nacion.com) una vez ahí mediante uso de funciones predefinidas por la librería como lo son urlopen() la cual nos permite cargar el contenido del HTML de un link en especifico también tenemos otras como en find() and findAll() que nos permiten hacer búsquedas dentro del HTML por nombres de contenedores, id y seleccionarlos hasta por clases para poder desfragmentar al máximo lo que es HTML y así obtener los datos necesarios para nuestra base.

Ya llegando a la parte final del desarrollo de la práctica de este reto mediante el cual al aplicar web scrapping podíamos acceder a información de un sitio x en mi caso con la nación tuve varias y muchas dificultades ya que me sucedió que utilizan muchos contenedores de tipo div para todo ejemplo un div para un título y si ese título le correspondía un texto ese texto se encontraba dentro de otro contenedor diferente. Y eso me ocasiono problemas para el acceso de información de manera más directa.

Como conclusión final se recomienda leer bastante documentación al respecto del tema y de las distintas tecnologías disponibles para realizarlo así seleccionando la que vea que le va a permitir solucionar su problema de una manera más rápida eficiente y eficaz.

**Link de repositorio Git**

<https://github.com/andrez96/webScraping.git>

**Bibliografia**

Mitchell, R. (n.d.). *Web Scrapping with python.* Retrieved from http://dl.finebook.ir/book/6f/13125.pdf

Moya, R. (n.d.). *Scraping en Python (BeautifulSoup), con ejemplos*. Retrieved from https://jarroba.com/scraping-python-beautifulsoup-ejemplos/

Python, M. D. (n.d.). *Web Scraping con Python y BeautifulSoup*. Retrieved from http://www.pythondiario.com/2016/08/web-scraping-con-python-y-beautifulsoup.html

S.A., G. N. (n.d.). *La nacion*. Retrieved from www.nacion.com