TEC | Tecnológico de Costa Rica

Sede San Carlos.

Escuela de Computación

Plan de estudio: Semestral

Curso: Introducción al desarrollo de aplicaciones para web

Profesor: Efrén Antonio Jiménez Delgado

Informe de investigación Web Scraping

Estudiante: Pérez Barquero Gabriel.

Grupo: 50

Fecha de entrega: 03/04/2017.

Periodo: 1 semestre 2017

Índice

Resumen ejecutivo

En este resumen de investigación se van a mencionar ciertas herramientas, metodologías y lenguajes utilizados en el web scraping, así como sus implicaciones legales con respecto a las políticas de uso que tienen las páginas web.

Además se llevará a cabo un ejemplo demostrativo de donde consumiremos datos de una página web y los almacenaremos en una base de datos, para el ejemplo se utilizará el lenguaje R que es útil en estos temas, como parte de esta demostración se crearemos un diagrama conceptual de la metodología que se utilizará.

Introducción

La realidad en cuanto a la manipulación de datos hoy en día, es muy tedioso tener que manejar manualmente esta información, ya que nos quita tiempo valioso; el navegar en la red o revisar la información de un determinado hardware, para realizar investigaciones, comparaciones entre otras, ahora existen muchas formas alternativas para poder automatizar dichas tareas, a continuación se demostrará una de estas alternativas.

Desarrollo

- Metodología de web scraping a modo demostrativo

  A continuación haremos un pequeño ejemplo, por medio del lenguaje de programación R, de extracción de datos de una página web, esto utilizando las etiquetas de CSS de la estructura HTML de dicha página, esto gracias a una librería que nos ofrece el lenguaje R, esta librería es **rvest**. Además de esta librería se utiliza: **jsonlite** en este caso para crear jsons a partir de un data.frame de R; **mongolite** que se utiliza para conectarse con una base de datos mongodb y almacenar en esta collections, los cuales son los jsons generados desde los data.frames.

  Para este ejemplo se utilizó la página web "http://www.usados.cr/?gclid=CLrg7eiohtMCFQgDhgod3wQD6Q&f_ucr_s_marca=Chevrolet&page=1". En esta la información que se encuentra son las caracteristicas de autos que están a la venta, en este caso de autos Chevrolet que se encuentran distribuidos en páginas, por lo cual se debe de recorrer la paginación, además por cada auto que se encuentra en esta página, se debe acceder a otra, donde se muestran todas las caracteristicas del vehiculo, de los cuales se extraen los siguientes objetos:

  Imagen: Atributo donde se guarda el url de la imagen del automotor.
  - id = "#dtlist_imgVehiculo_0", atributo html = 'src'

  Características específicas:

  Marca: Atributo donde se guarda la marca.
  Modelo: Atributo donde se guarda el modelo.
  Transmisión: Atributo donde se guarda la transmisión.
  Combustible: Atributo donde se guarda el tipo de combustible del motor.
  Año: Atributo donde se guarda el año del vehículo.
  Estilo: Atributo donde se guarda el estilo.

Precio: Atributo donde se guarda el valor monetario.
Kilometraje: Atributo donde se guarda los KMs recorridos.
Color Exterior: Atributo donde se guarda el color del vehículo.
Color Interior: Atributo donde se guarda el color interior.

- Class = ".technical-data ul li"

Datos de contacto:

AGENCIA: Atributo donde se guarda la agencia que vende el vehículo.
Teléfono: Atributo donde se guarda el teléfono de contacto.
Email: Atributo donde se guarda el email.

- CSS atributos = "h1 span"

Conclusiones

El tema del web scraping no es tan nuevo, pero se han creado grandes herramientas que nos pueden facilitar el consumo de datos de manera sencilla, estos datos nos pueden ser de gran utilidad y sobre todo sin ningún riesgo si sabemos utilizarlos.

Python ha sido uno de los lenguajes que se está utilizando actualmente para la realización de muchos cosas, y por supuesto en cuanto a lo del web scraping no se queda atrás, R es otro lenguaje que tampoco se queda atrás, con su poder en la minería de datos, resulta ser muy útil en el web scraping.

Recomendaciones

Es necesario saber sobre las estructuras de HTML, para poder hacer un uso más efectivo de estas técnicas.

Para evitar el web scraping se deberían de tomar en cuenta los captcha en las páginas web, ya que requieren la interacción de un humano para poder continuar navegando en los sitios web. Además tomar en cuenta las políticas de uso de la páginas que se desean consultar por medio de estos métodos de obtención de datos, para así no tener ningún inconveniente con nuestros trabajos.

Referencias

concatenated?, H. (2016). *How can 2 strings be concatenated?*. [online] Stackoverflow.com. Available at: http://stackoverflow.com/questions/7201341/how-can-2-strings-be-concatenated [Accessed 1 Apr. 2017].

Docs.mongodb.com. (n.d.). *db.collection.drop() — MongoDB Manual 3.4*. [online] Available at: https://docs.mongodb.com/manual/reference/method/db.collection.drop/ [Accessed 1 Apr. 2017].

dummies. (n.d.). *How to Split Strings in R - dummies*. [online] Available at: http://www.dummies.com/programming/r/how-to-split-strings-in-r/ [Accessed 1 Apr. 2017].

json, c. (2016). *convert data frame to json*. [online] Stackoverflow.com. Available at: http://stackoverflow.com/questions/25550711/convert-data-frame-to-json [Accessed 1 Apr. 2017].

package, I. (2016). *Inserting json object into MongoDB using mongolite R package*. [online] Stackoverflow.com. Available at: http://stackoverflow.com/questions/35979720/inserting-json-object-into-mongodb-using-mongolite-r-package [Accessed 1 Apr. 2017].

R, r. (2016). *remove all line breaks (enter symbols) from the string using R*. [online] Stackoverflow.com. Available at: http://stackoverflow.com/questions/21781014/remove-all-line-breaks-enter-symbols-from-the-string-using-r [Accessed 1 Apr. 2017].

Stackoverflow.com. (2013). *Append value to empty vector in R?*. [online] Available at: http://stackoverflow.com/questions/22235809/append-value-to-empty-vector-in-r [Accessed 1 Apr. 2017].

Stackoverflow.com. (2016). *How can I remove an element from a list?*. [online] Available at: http://stackoverflow.com/questions/652136/how-can-i-remove-an-element-from-a-list [Accessed 1 Apr. 2017].

Stackoverflow.com. (2016). *Using 'rvest' to extract links*. [online] Available at: http://stackoverflow.com/questions/35247033/using-rvest-to-extract-links [Accessed 1 Apr. 2017].

Stat.ethz.ch. (n.d.). *R: Data Frames*. [online] Available at: https://stat.ethz.ch/R-manual/R-devel/library/base/html/data.frame.html [Accessed 1 Apr. 2017].

YouTube. (2017). *R Tutorial - Using the Data Frame in R*. [online] Available at: https://www.youtube.com/watch?v=9f2g7RN5N0I [Accessed 1 Apr. 2017].