Instituto Tecnológico de Costa Rica

Ingeniería en Computación

Tarea de Investigación: Web Scraping Segunda Parte

Keylor Jesús Guevara Jiménez

Sede San Carlos

03 – 04 – 17

Resumen Ejecutivo

Dentro de este segundo documento se estarán analizando los diferentes procesos que son necesarios realizar para poder alcanzar con éxito le ejecución de la metodología descrita en el entregable número uno de esta misma investigación sobre como poder realizar un proceso de web scraping adecuado de una página de compra y muestreo de libros.

Es necesario aplicar conocimientos relacionados con el uso del lenguaje de programación Python, el cual fue seleccionado entre las diferentes opciones para poder realizar la toma de datos de la página. Dentro de Python contamos con una de las mejores librerías para poder realizar la extracción de datos entre todas las que el mercado de internet nos da a disposición, la cual es Beautiful Soup. Sin duda de las mejores herramientas capaces de convertir la información de html en instancias de su clase y así poder iniciar con todas las propiedades tan poderosas que esta librería tiene para nosotros.

Tabla de Contenidos

Introducción

Dentro del presente documento de investigación se estará realizando la puesta en práctica de la primer parte de la metodología descrita en el enunciado número uno de esta serie de documentos entregables.

Para esta ocasión primeramente se define la página X para la extracción de la información, en específico, la página x corresponde a: http://www.e-libro.com/ una página dedicada a ofrecer textos completos, desde textos de cátedra, libros, hasta artículos, pasando por investigaciones científicas de todas las disciplinas académicas.

Para lo que corresponde a los objetos que se están analizando se puede mencionar que todos los elementos presenten en ellos son utilizados para el proceso de web scraping. Dentro de los elementos que son importantes de mencionar está el lenguaje de programación y sus librerías a utilizar:

Lenguaje de programación	Python
Librerías utilizadas	
Urllib	pymysql
urllib2	BeautifulSoup4
re	lmxl
json	sys

Desarrollo

En el internet existe un sinfín de páginas con información, con información de todo tipo, de aquel que deseamos ver y de aquel que también no deseamos ver, se nos ofrece de una forma sutil y estratégica. La información que se encuentra en el internet no tiene un fin en específico, es decir no está hay solo par a “informar” sino que es un mercado completo de oportunidades para las grandes y/o porque no, las pequeñas, compañías de dar sus productos a conocer, de aumentar la popularidad de un nuevo producto, o para promocionar un nuevo centro de atención turística en alguna zona poco conocida.

La información que hoy en día está en internet tiene un poder incalculable, pues esta información está en constante aprendizaje, y nosotros sin desearlo somos sus maestros, les damos la información a las compañías para que sepan cuáles son nuestros gustos y preferencias para luego ellos poder vender esa información a terceros y devolvérnosla en sugestivos anuncios de “ofertas y promociones”.

Ese ejemplo que se acaba de mencionar es solo uno de los miles de ejemplos que se pueden mencionar, otro podría ser el poder de la minería de datos que actualmente está tomando un fuerte impulso en la sociedad, donde con el potencial de la información se pueden conocer datos tan interesantes como lo son los sentimientos o reacciones a situaciones determinadas por parte de las masas, como por ejemplo, por medio de las reacciones de los usuarios de una red social, saber cuál iba a ser el candidato de una bancada política de un país.

Para este proyecto la página que suministrará la información corresponde a una página que se encarga de brindar información y vender libros, corresponde a http://www.e-libro.com/ la cual es una página dedicada a ofrecer textos completos, desde textos de cátedra, libros, hasta artículos, pasando por investigaciones científicas de todas las disciplinas académicas.

Dentro de la página usted podrá encontrar todo tipo de material educativo posible ya que es una plataforma con una serie de bibliografías básicas como por ejemplo computación, derecho y medicina, por solo mencionar algo, cuenta con una

gran variedad de editoriales y algunos beneficios especiales como el contar con un acceso simultaneo, pode realizar la descarga de la información y realizar investigaciones avanzadas.

Análisis de atributos:

Nombre	Descripción
Cover_book	Es la imagen de la portada del libro
Title	El título del libro
Autor	Incluye el nombre del autor del libro
Editorial	Incluye el nombre de la editorial a la que pertenece el libro.
Temas	Incluye los temas a los que se encuentra asociado el libro
Identificación del documento	Incluye información acerca de datos correspondientes al libro
pISBN	Número de rastreo de los libros digitales
eISBN	Número de identificación del libro.
Precio de acceso exclusivo	Ofrece un precio único que hay para el libro.
Three user	Precio disponible para tres usuarios
Precio para múltiples usuarios	Precios que tienen múltiples usuarios al querer adquirir el libro.

Link de acceso repositorio:

El link de acceso al repositorio es el siguiente:

`https://github.com/KeylorGuevara/SP-WebScrapingPython.git`

Conclusiones y Recomendaciones

Conclusiones

1. Realizar búsqueda exhaustivas de diferentes herramientas antes de tomar la decisión de qué lenguaje implementar dentro del proceso de Web Scraping.
2. Indagar páginas y comunidades confiables, como la misma comunidad de Python la cual cuenta con gran cantidad de datos importantes sobre librerías y demás herramientas sumamente eficientes según nuestras necesidades.

Recomendaciones

1. Controlar de una forma adecuada todas las diferentes versiones de código que estemos implementando en nuestro trabajo.
2. Diseñar diferentes estrategias para poder identificar cuáles son los segmentos de HTML que realmente estamos tratando de identificar, por ejemplo, existen extensiones dentro de ciertos buscadores diseñadas específicamente para identificar las partes del HTML.

Bibliografía

Crummy.com. (2004). Beautiful Soup Documentation — Beautiful Soup 4.4.0 documentation. [online] Disponible desde: https://www.crummy.com/software/BeautifulSoup/bs4/doc/ [Accesado 1 Abr. 2017].

Lawson, R. (2015). Web Scraping with Python. Packt Publishing Ltd.

Mitchell, R. (2015). Web Scraping with Python: Collecting Data from the Modern Web. California: O'Reilly Media, Inc.

Munzert, S., Rubba, C., Meißner, P., & Nyhuis, D. (2014). Automated Data Collection with R – A Practical Guide to Web Scraping and Text Mining. En S. Munzert, C. Rubba, P. Meißner, & D. Nyhuis, *Journal of Statistical Software* (pág. 480). Chichester: John Wiley & Sons.