6-6-2017

Web scraping Python

Sistemas Distribuidos

Jhon James Cano Sánchez – Juan Carlos Serna Gómez – Jheison Andrés Velásquez Sánchez
CORPORACIÓN DE ESTUDIOS TECNOLÓGICOS DEL NORTE DEL VALLE

WEB SCRAPING PYTHON

INTRODUCCIÓN

Web Scraping es una técnica utilizada mediante programas de software para extraer información de sitios web, está muy relacionado con la indexación de la web, la cual indexa la información de la web utilizando un robot y es una técnica universal adoptada por la mayoría de los motores de búsqueda. Sin embargo, el web scraping se enfoca más en la transformación de datos sin estructura en la web (como el formato HTML) en datos estructurados que pueden ser almacenados y analizados en una base de datos central, en una hoja de cálculo o en alguna otra fuente de almacenamiento.

FINALIDAD

La finalidad de este proyecto es mostrar cómo hacer las conexiones HTTP y como tratar y extraer el contenido importante o clave de los HTML que se obtienen a través de la librería BeautifulSoup.

Se pueden encontrar los elementos necesarios para el análisis de un sitio y sus vulnerabilidades, además de poder extraer el árbol de etiquetas del sitio a través de simples comandos que se conectan de forma transparente para el usuario.

MANUAL DE USO

El programa consta de validación de usuarios para ingresar al aplicativo, esto significa que una vez ingresado los datos, el sistema validará los datos y dependiendo de los privilegios asignados al usuario mostrará el menú de administrador o cliente, en caso que los datos sean erróneos mostrará error de validación.

RECOMENDACIONES INICIALES

- El usuario debe tener conocimientos básicos en el manejo de lenguaje Python, Pyro y conexión a base de datos usando MysqlDB.
- Antes de iniciar con la ejecución es necesario abrir la terminal de Python y ejecutar el siguiente comando:

  python -m Pyro4.naming
- Este comando permitirá la ejecución y reconocimiento de los sitios creados a través de Pyro y las funciones registradas para el sitio.
- Es necesario que el servidor de base de datos creado en la red se encuentre en ejecución y tenga los privilegios necesarios de acceso para realizar la conexión.

MENÚ ADMINISTRADOR

El menú de administrador cuenta con las siguientes funciones:

```
******** Administrador ********
-----------------------------------------
1. Registro de Empresas (Páginas Web)
2. Agregar Palabras Clave
3. Consultar Empresas Registradas (Páginas)
4. Cerrar Sesión
-----------------------------------------
Ingrese una Opción:
```

1. Registro de Empresas (Páginas Web): permite el registro de páginas web en la base de datos para su posterior análisis.

   ```
   Ingrese una Opción: 1
   Digite Nombre de la Página sin http:// : www.eltiempo.com/deportes
   ```

2. Agregar Palabras Clave: Permite al administrador registrar las palabras clave coincidentes con el sitio web creado.

   ```
   -----------------------------------------
   Ingrese una Opción: 2
   _____________________
   Páginas Registradas
   ID = 3, Pagina = http://www.espn.com.co, Palabras Clave = Deportes, ciclismo, fútbol, champions League
   ID = 5, Pagina = http://www.cotecnova.edu.co, Palabras Clave = educación, cotecnova, cartago, universidad, ingenieria de sistemas, producción agropecuaria
   ID = 6, Pagina = http://www.eltiempo.com/deportes, Palabras Clave =

   Digite id de la Página: 6
   Ingrese Palabras Claves: deportes, noticias, colombia
   ```

3. Consultar: permite listar el total de páginas creadas en el sistema con sus palabras claves.

   ```
   -----------------------------------------
   Ingrese una Opción: 3
   ID = 3, Pagina = http://www.espn.com.co, Palabras Clave = Deportes, ciclismo, fútbol, champions League
   ID = 5, Pagina = http://www.cotecnova.edu.co, Palabras Clave = educación, cotecnova, cartago, universidad
   ID = 6, Pagina = http://www.eltiempo.com/deportes, Palabras Clave = deportes, noticias, colombia
   ```

4. Cerrar sesión: Permite desconectar la sesión de forma segura.

MENU CLIENTE

El menú del cliente cuenta con las siguientes funciones:

```
******** Cliente ********
-----------------------------------
1. Listado de Páginas
2. Eliminar Páginas Web
3. Agregar Nueva Página
4. Información SEO de un sitio Web
5. Ranking de Todos los Sitios Agregados
6. Listado de Páginas Penalizadas
7. Cerrar Sesión
-----------------------------------
Ingrese una Opción:
```

1. Listado de Páginas: permite listar el total de páginas creadas en el sistema con sus palabras claves.

```
-----------------------------------------
Ingrese una Opción: 3
ID = 3, Pagina = http://www.espn.com.co, Palabras Clave = Deportes, ciclismo, fútbol, champions League
ID = 5, Pagina = http://www.cotecnova.edu.co, Palabras Clave = educación, cotecnova, cartago, universidad
ID = 6, Pagina = http://www.eltiempo.com/deportes, Palabras Clave = deportes, noticias, colombia
```

2. Eliminar Páginas Web: permite al cliente eliminar páginas que considere inapropiadas o que hayan sido descartadas.
3. Agregar Nueva Página: permite el registro de páginas web en la base de datos para su posterior análisis.

```
Ingrese una Opción: 1
Digite Nombre de la Página sin http:// : www.eltiempo.com/deportes
```

4. Información SEO de un sitio Web: esta opción contiene un submenú en el cual se realiza el análisis a las páginas agregadas.

```
-----------------------------------
a. Contar Palabras (MapReduce)
b. Diccionario de Palabras Clave
c. Contar Imágenes
d. Contar Enlaces Internos y Externos
e. Analizar URL
f. Analizar Palabras Clave (Keywords)
h. Estructura del Sitio Web
i. Penalizar Contenido No Apto
j. Penalizar Contenido de Dudosa Reputación
k. Penalizar Malas Prácticas de Desarrollo Web
l. Información de Librerias Usadas en el Sitio
m. Comprobar Enlaces Externos
n. Si se enlaza a una página web almacenada en el servidor debe dar más puntuación
o. Regresar a Menú Principal
-----------------------------------
Ingrese una Opción:
```

- Contar Palabras (MapReduce): permite contar el top 20 de palabras en un sitio web buscado, evitando el listado de palabras comunes descartadas.

```
Digite id de la Página: 3
http://www.espn.com.co
PoolWorker-1 reading archivo.txt

TOP 20 principales palabras

madrid     :    11
real       :    11
seleccion  :    10
espana     :     9
colombia   :     8
espn       :     7
colombiano :     7
tigre      :     7
jugador    :     6
crack      :     6
cucuteno   :     6
gol        :     6
fue        :     5
final      :     5
tras       :     5
le         :     5
pero       :     5
champions  :     5
nexo       :     4
maximo     :     4
```

- Diccionario de Palabras Clave: Muestra el total de palabras clave registradas en la base de datos para los sitios agregados.

```
-----------------------------------
Ingrese una Opción: b
Palabras Clave = Deportes, ciclismo, fútbol, champions League
Palabras Clave = educación, cotecnova, cartago, universidad, ingeniería de sistemas, producción agropecuaria
Palabras Clave = deportes, noticias, colombia
```

- Contar Imágenes: Cuenta el número de imágenes de un sitio.

```
-----------------------------------
Ingrese una Opción: c

_____________________________________
Páginas Registradas
ID = 3, Pagina = http://www.espn.com.co
ID = 5, Pagina = http://www.cotecnova.edu.co
ID = 6, Pagina = http://www.eltiempo.com/deportes

_____________________________________
Digite id de la Página: 3
http://www.espn.com.co
Total Imágenes = 50
```

- Contar Enlaces Internos y Externos: muestra el total de enlaces de una página web con los enlaces externos si los tiene.

```
_____________________________________
Digite id de la Página: 3
http://www.espn.com.co
Enlaces de Página
Total Enlaces = 319
Conteo de Enlaces Externos = 52
```

- Analizar URL: muestra el nombre de la página, título, estado, palabras clave.
- Analizar Palabras Clave: muestra el listado de palabras clave de un sitio registrado, estos datos se sacan directamente de las Keywords de la página.

```
------------------------------------
Ingrese una Opción: f

_____________________________________
Páginas Registradas
ID = 3, Pagina = http://www.espn.com.co
ID = 5, Pagina = http://www.cotecnova.edu.co
ID = 6, Pagina = http://www.eltiempo.com/deportes

_____________________________________
Digite id de la Página: 3
http://www.espn.com.co
espn, ESPN, deportes, básquetbol, fútbol, béisbol, rugby, pelota, baloncesto, boxeo
```

- Estructura del sitio web: muestra todo el contenido HTML del sitio.

- Penalizar contenido no apto: se crea un listado de palabras prohibidas y busca su coincidencia en el texto de la página, si se encuentra dicho contenido se penaliza la página con puntuación negativa.

```
Ingrese una Opción: i

_____________________________________
Páginas Registradas
ID = 3, Pagina = http://www.espn.com.co, Palabras Clave = Deportes, ciclismo, fútbol, champions League
ID = 5, Pagina = http://www.cotecnova.edu.co, Palabras Clave = educación, cotecnova, cartago, universidad
ID = 6, Pagina = http://www.eltiempo.com/deportes, Palabras Clave = deportes, noticias, colombia

_____________________________________
Digite id de la Página: 3
http://www.espn.com.co
Página NO tiene contenido no Apto
```

- Penalizar contenido de dudosa reputación: el sistema revisa las URL ingresadas en la página y se encarga de verificar que no contenga información dudosa, de ser así se penaliza la página.

- Penalizar malas prácticas de Desarrollo Web: revisa el contenido en busca de código sospechoso o prácticas que puedan afectar al visitante del sitio, de ser encontrado se hace la respectiva advertencia.
- Información de librerías usadas en el sitio: se obtienen todas las librerías usadas en el sitio para su posterior análisis.
- Comprobar enlaces externos: revisa los enlaces que redireccionan fuera del sitio web y los lista de forma tal que puedan ser accedidos.
- Si se enlaza a una página web almacenada en el servidor debe dar más puntuación: si la página contiene enlaces a un sitio web almacenado en la base de datos se entrega más puntuación.
- Regresar a Menú Principal: retorna al menú de cliente.

5. Ranking de Todos los Sitios Agregados: permite ver cuáles sitios tienen mejor reputación y cumplen con los criterios de análisis.

```
Ingrese una Opción: 5
ID = 3, Pagina = http://www.espn.com.co, Puntos = 7
ID = 6, Pagina = http://www.eltiempo.com/deportes, Puntos = 0
ID = 5, Pagina = http://www.cotecnova.edu.co, Puntos = -2
```

6. Listado de Páginas Penalizadas: muestra el listado de las páginas que han sido penalizadas por contener enlaces maliciosos o contenido no apto.
7. Cerrar sesión: Permite desconectar la sesión de forma segura.