CENTRO DE TECNOLOGÍA Y EDUCACIÓN PERMANENTE

DIPLOMADO EN ARQUITECTURA Y DISEÑO DE SOLUCIONES WEB Y MÓVIL

Asignación:

Reporte Práctica 3

Entregado a:

Ing. Carlos Camacho

Entregado por:

Dariana Tavera

Santiago, Sábado 9 de abril 2016

Introducción

El Web scraping es una técnicas utilizadas para obtener datos de forma automática de un sitio web. En este se puede transformar datos sin estructuras en la web en datos estructurados que pueden ser almacenados y analizados, además se puede utilizar para automatizar tareas dentro de la internet.

Jsoup es una librería de Java que resulta ser una alternativa para trabajar web scraping, este nos permite extraer y manipular la data utilizando los métodos DOM y selectores similares a los jquery.

En la siguiente práctica se utilizara este parser para realizar algunas consultas.

Desarrollo de la práctica:

Crea un programa (consola) que pida por la entrada estándar una URL válida, una vez consultada realice las siguientes operaciones:

a) Indicar la cantidad de líneas del recurso retornado.
b) Indicar la cantidad de párrafos (p) que contiene el documento HTML .
c) Indicar la cantidad de imágenes (img) que contiene el archivo HTML.
d) indicar la cantidad de formularios (form) que contiene el HTML.
e) Para cada formulario mostrar los campos del tipo input que contiene el HTML.

El primer paso (después de agregar la librería a netbeans) fue crear un objeto tipo Scanner que me permitiera leer la url que ingresa el usuario, esto se realiza dentro de un bloque try y catch. En este bloque de código se hace una petición http, pasando la url introducida por el usuario al método connect de Jsoup y obteniendo como resultado un objeto de tipo Document con el contenido html del sitio.

```
try {
        System.out.println("\nIntroduzca una Url valida,\nFormato aceptado-http://ejemplo.com : ");
        url = leer.nextLine();
        doc = Jsoup.connect(url).get();
    }
    catch (Exception e) {

       }
```

Luego con el objeto Document se podrá obtener el contenido entre las etiquetas HTML. Para obtener los datos específicos se crean objetos tipo Elements en los cuales se guardará el contenido de etiquetas HTML requeridas. En este caso para obtener los datos se escogió usar la sintaxis de selector para obtener los elementos.Ejemplo:

```
Elements parrafos = doc.select("p");
```

Estos selectores funcionan de manera similar a los selectores CSS o los de Jquery.

Por último se utilizó la propiedad size para saber la cantidad de elementos existentes de cada tipo de selector. Ejemplo:

```
System.out.println("Número de lineas: " + lineas.size());
```

Conclusión

En esta práctica se realizaron consultas basicas de una pagina web haciendo uso de web scraping por medio de Jsoup.

Entre los componentes utilizados están:

- Los métodos utilizados para realizar las peticiones HTTP
- Los elementos que nos permite manipular los elementos utilizando selectores similares a los de CSS y Jquery.