UTN.BA
UNIVERSIDAD TECNOLÓGICA NACIONAL
FACULTAD REGIONAL BUENOS AIRES

Ingeniería en Sistemas de Información

Sintaxis y Semántica de los Lenguajes

Curso K2054

TRABAJO PRÁCTICO N° 2

Web Scraping

Alumno: Cesar Mejia
Legajo: 1418713
Correo: cmejia@frba.utn.edu.ar
Usuario Github: comejia

Repositorio: https://github.com/comejia/utn-ssl-examples

Docente: Ing. Pablo D. Mendez

Fecha de entrega:		Nota:		Fecha Aprobación:	

Comentarios:

Tabla de contenido

Consignas

Realizar el web scraping sobre archivos alojados en un servidor web que simula la página de acciones líderes de bolsar.info (https://bolsar.info/lideres.php). Se deben emitir los siguientes reportes:

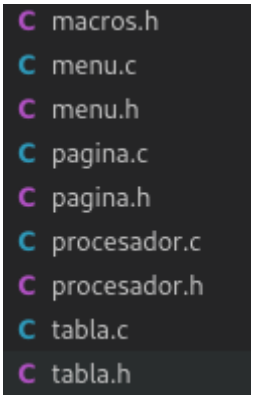
A) Listar en pantalla las especies cuyo % de variación es negativo.
B) Listado de las cotizaciones de compra y de venta en un archivo .CSV (Comma Separated Values) importable desde excel. La hoja debe incluir encabezados y las columnas deben ser:
   Especie; Precio de compra; Precio de venta; Apertura; Precio Máximo; Precio Mínimo.
   (Un archivo CSV no es más que un archivo de texto con valores separados por caracteres ‘,’ o ‘;’. También se suelen utilizar tabulaciones como separadores. Debe tenerse en cuenta que debe establecer uno sólo de ellos como separador).
C) Mismo reporte que el listado A pero en una tabla html, indicando en color verde las filas de las especies cuyo precio de compra y precio de venta es menor al precio de apertura.
D) Imprimir código fuente de página (extra)

El programa debe contener un menú que permita elegir al usuario el reporte deseado. Al presionar la opción deben buscarse los datos por wget en tiempo real.
Cabe destacar, que el sitio web tiene protección para evitar el scrapping, por lo que se cuenta con un sitio alternativo con un archivo fijo de donde puedan leerlo.

Recomendaciones:
En etapas de prueba, pueden bajarse el archivo HTML en cuestión y abrirlo directamente desde el FILE SYSTEM. Esto ahorra tiempo y además permite analizar la estructura del HTML de manera más fácil.

Desarrollo

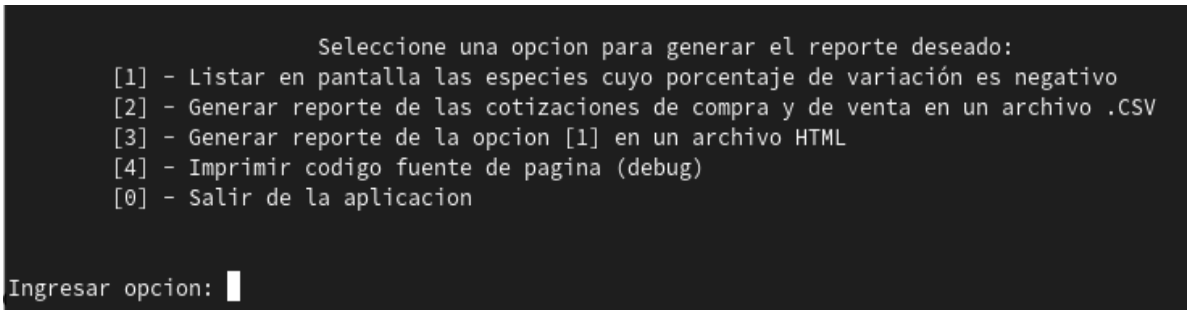
El desarrollo del presente trabajo práctico se realizó en lenguaje C, y se trabajó desde un inicio de forma modular. Es decir separar en diferentes archivos ".c" el desarrollo del código. Se optó esta forma porque el código iba a ser un poco extenso para un solo archivo y además difícil de mantener y entender para otra persona.
Algunos de los archivos mencionados anteriormente son:



donde cada uno cumple alguna funcionalidad específica. Por ejemplo "menu.c" muestra el menú con el cual va a interactuar el usuario, "tabla.c" tiene estructuras y funciones para el manejo de los datos de las acciones.

Otra consideración importante que se hizo, fue trabajar de manera ***offline*** siguiendo la recomendación de utilizar un archivo HTML de prueba con el código fuente de la página del bolsar. Esto no solo agilizo el desarrollo sino que también las pruebas fueron compatibles con la página en tiempo real (como era de esperar).

Aclarado estos puntos, se muestra a continuación las salidas que se obtuvieron en los distintos casos al seleccionar una opción del menú.



- **Punto D / Opcion 4**

Se arranca por este punto, y aunque no era un requisito del TP, me pareció importante tener un caso que pruebe rápidamente si la lectura de la página, tanto online como offline, se está haciendo de forma correcta.

```
Ingresar opcion: 4
<!doctype html>

<html lang="es" class="no-js">

<head>

<title>Resumen de Mercado - BCBA</title>

<meta http-equiv="Content-Type" content="text/html; charset=utf-8" />

<meta charset="utf-8">

<meta http-equiv="X-UA-Compatible" content="IE=edge">

<meta name="keywords" content="bolsar">

<meta name="viewport" content=" initial-scale=1, maximum-scale=1">

<meta http-equiv="Content-Security-Policy" content="upgrade-insecure-requests">
```

En la imagen se puede ver que la consola muestra el contenido HTML que trae de la página del bolsar.

- **Punto A / Opción 1**

Al presionar esta opción se observa la siguiente salida con las especies en negativo:

```
      Especies en negativo
Especie-Vencimiento-Cant. Nominal-Compra-Venta-Cant. Nominal-Ultimo-Variacion-Apertura-Min-Max-Cierre Anterior-Volumen-Monto-Operacion-Hora
HARG - 48hs - 8 - 181,00 - 184,50 - 500 - 181,00 - -0,14%  - 182,00 - 180,00 - 182,00 - 181,25 - 73,00 - 13.192,00 - 9 - 11:03:55
TECO2 - 48hs - 50 - 190,50 - 193,15 - 2.765 - 192,80 - -0,03%  - 189,80 - 188,00 - 194,00 - 192,85 - 7.440,00 - 1.419.343,00 - 23 - 11:05:21
```

la cual se podría pensar en un principio que hay un error por la poca cantidad de datos. Pero hay que considerar el momento en que se está haciendo la lectura de las acciones.

Para reforzar lo antedicho se muestra la salida de la página de prueba guardada con anterioridad:

```
        Especies en negativo
Especie-Vencimiento-Cant. Nominal-Compra-Venta-Cant. Nominal-Ultimo-Variacion-Apertura-Min-Max-Cierre Anterior-Volumen-Monto-Operacion-Hora
BBAR - 48hs - 4 - 188,10 - 198,50 - 8 - 192,50 - -0,39%  - 190,60 - 188,20 - 192,90 - 193,25 - 83.995,00 - 16.045.371,00 - 288 - 17:00:01
BYMA - 48hs - 44 - 750,00 - 780,00 - 4 - 751,00 - -1,89%  - 765,00 - 746,00 - 767,00 - 765,50 - 9.907,00 - 7.446.979,00 - 216 - 17:00:01
CEPU - 48hs - 2 - 40,10 - 45,30 - 1 - 43,55 - -1,14%  - 44,00 - 41,75 - 45,15 - 44,05 - 97.082,00 - 4.159.996,00 - 325 - 17:00:01
COME - 48hs - 1.150 - 4,31 - 4,50 - 6 - 4,42 - -0,67%  - 4,48 - 4,37 - 4,48 - 4,45 - 2.299.003,00 - 10.115.471,00 - 562 - 17:00:01
CRES - 48hs - 48 - 92,00 - 102,00 - 61 - 93,75 - -4,09%  - 97,90 - 92,80 - 97,95 - 97,75 - 314.877,00 - 29.757.897,00 - 1.399 - 17:00:01
HARG - 48hs - 10 - 175,00 - 204,25 - 42 - 183,50 - -3,42%  - 190,50 - 178,75 - 190,50 - 190,00 - 15.045,00 - 2.760.893,00 - 119 - 17:00:01
LOMA - 48hs - 1 - 245,00 - 265,00 - 420 - 263,15 - -2,95%  - 269,90 - 256,00 - 269,90 - 271,15 - 222.961,00 - 57.983.454,00 - 889 - 17:00:01
MIRG - 48hs - 11 - 1.895,00 - 2.368,00 - 5 - 2.225,50 - -1,48%  - 2.267,00 - 2.218,00 - 2.267,00 - 2.259,00 - 4.334,00 - 9.667.469,00 - 221 - 17:00:01
SUPV - 48hs - 7.000 - 68,10 - 74,00 - 143 - 69,65 - -1,49%  - 69,10 - 69,00 - 70,70 - 70,70 - 153.170,00 - 10.653.448,00 - 292 - 17:00:01
TECO2 - 48hs - 2.187 - 174,10 - 183,30 - 50 - 177,75 - -1,52%  - 181,10 - 174,20 - 181,10 - 180,50 - 46.011,00 - 8.124.263,00 - 292 - 17:00:01
TGNO4 - 48hs - 254 - 39,00 - 62,00 - 25 - 58,30 - -1,19%  - 59,00 - 56,20 - 59,00 - 59,00 - 270.850,00 - 15.553.069,00 - 471 - 17:00:01
TRAN - 48hs - 2 - 31,20 - 33,45 - 98 - 31,85 - -3,04%  - 32,70 - 31,40 - 32,70 - 32,85 - 421.072,00 - 13.328.000,00 - 442 - 17:00:01
VALO - 48hs - 2.500 - 22,25 - 25,00 - 1 - 22,65 - -0,88%  - 22,60 - 22,45 - 23,25 - 22,85 - 250.915,00 - 5.684.047,00 - 604 - 17:00:01
YPFD - 48hs - 6 - 768,00 - 800,00 - 15 - 784,95 - -3,09%  - 793,95 - 769,00 - 794,20 - 809,95 - 201.953,00 - 158.049.185,00 - 1.564 - 17:00:01
```

- **Punto B / Opción 2**

En este caso, en vez de imprimir en consola, se guarda el listado de las cotizaciones en un archivo .csv, que abriendolo con algun editor de texto que formatee la salida, vemos algo como:

Especie	Precio de compra	Precio de venta	Apertura	Precio Minimo	Precio Maximo
ALUA	67,90	68,00	67,90	67,50	69,00
BBAR	258,65	258,75	254,05	254,05	262,00
BMA	331,45	331,95	324,00	324,00	332,00
BYMA	821,00	822,00	820,00	819,00	825,00
CEPU	49,75	49,95	49,30	49,30	50,00
COME	5,37	5,38	5,32	5,28	5,39
CRES	103,15	103,25	101,50	101,50	103,50
CVH	374,50	375,00	365,50	365,50	380,00
EDN	65,35	65,75	65,00	65,00	66,25

Ahora si comparamos la salida con la página de bolsar, los valores de las columnas correspondientes coinciden (salvo aquellas que de un instante a otro vario).

ESPECIES	48hs	Cant. Nominal	Compra	Venta	Cant. Nominal	Último	Variación	Apertura	Mín	Max	Cierre Anterior	Volumen	Monto	Oper	Hora
ALUA	48hs	200	67,80	68,00	10.411	68,00	1,04% →	67,90	67,50	69,00	67,30	171.946,00	11.711.388,00	225	11:32:07
BBAR	48hs	1	258,65	258,80	100	258,75	5,61% →	254,05	254,05	262,00	245,00	62.225,00	15.998.557,00	149	11:31:12
BMA	48hs	69	331,40	331,95	10	331,45	3,18% →	324,00	324,00	332,00	321,25	32.429,00	10.697.963,00	126	11:31:27
BYMA	48hs	12	820,00	821,00	22	821,00	0,18% ↓	820,00	819,00	825,00	819,50	202,00	166.273,00	31	11:31:46
CEPU	48hs	1.274	49,75	49,90	567	49,75	3,22% ↓	49,30	49,30	50,00	48,20	44.712,00	2.220.645,00	97	11:32:03
COME	48hs	458	5,37	5,39	116.805	5,39	1,51% →	5,32	5,28	5,39	5,31	318.319,00	1.711.239,00	157	11:31:59
CRES	48hs	49	103,15	103,20	160	103,25	2,13% →	101,50	101,50	103,50	101,10	11.697,00	1.194.207,00	53	11:30:45
CVH	48hs	188	375,00	376,00	427	376,00	3,16% ↑	365,50	365,50	380,00	364,50	8.371,00	3.140.090,00	131	11:32:02
EDN	48hs	373	65,35	65,75	1.135	65,75	1,39% ↑	65,00	65,00	66,25	64,85	21.081,00	1.387.443,00	90	11:31:47

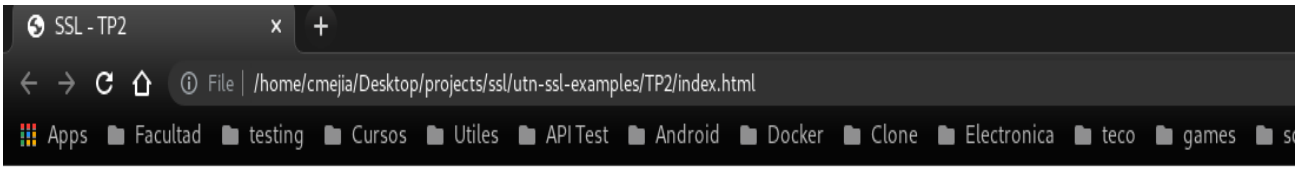
- **Punto C / Opción 3**

Este punto es igual al punto A, solo que la salida se muestra en una tabla html.

Para resolver este punto se trabajó con una plantilla “template.html” el cual contiene la estructura de la tabla a mostrar:

```
<!DOCTYPE html>
<html>
    <head>
        <title>SSL - TP2</title>
    </head>
    <body>
        <h1>Listado de especies en negativo</h1>
        <table>
            <tr style="font-size:15px; text-align:center;">
                <th>Especie</th>
                <th>Vencimiento</th>
                <th>Cant. Nominal</th>
                <th>Compra</th>
                <th>Venta</th>
                <th>Cant. Nominal</th>
                <th>Ultimo</th>
                <th>Variación</th>
                <th>Apertura</th>
                <th>Minimo</th>
                <th>Máximo</th>
                <th>Cierre Anterior</th>
                <th>Volumen</th>
                <th>Monto</th>
                <th>Operacion</th>
                <th>Hora</th>
            </tr>
            %s
        </table>
    </body>
</html>
```

La línea 27 que tiene ‘**%s**’ es la parte que se va a completar, desde el código en C, con los datos procesados de la página de bolsar. Generando así un nuevo archivo “index.html” que al abrirlo con algún navegador se obtiene lo pedido (y resaltando en verde cuando corresponda).



Listado de especies en negativo

Especie	Vencimiento	Cant. Nominal	Compra	Venta	Cant. Nominal	Ultimo	Variación	Apertura	Minimo	Máximo	Cierre Anterior	Volumen	Monto	Operacion	Hora
BYMA	48hs	48	819,00	822,00	32	819,00	-0,06%	820,00	819,00	825,00	819,50	3.207,00	2.630.444,00	89	12:10:59
LOMA	48hs	2.400	275,50	275,60	2.407	275,30	-0,34%	277,00	275,00	278,80	276,25	33.356,00	9.217.281,00	404	12:11:13
TECO2	48hs	845	190,50	191,20	1.469	190,25	-1,35%	189,80	188,00	194,00	192,85	61.016,00	11.679.624,00	181	12:10:50