Juan Ignacio Falcone

Sintaxis y Semántica de los Lenguajes
171.555-0 - jfalcone@frba.utn.edu.ar
Github: FalcFalcFalc - Repositorio del TP2

TP 2 - Web Scraping

3 de octubre de 2021

Objetivos

I. Crear un repositorio de git.
II. Utilizar las herramientas que nos brinda el lenguaje C, sobre la página bolsar.info para:

   A. Listar en pantalla las especies cuyo porcentaje de variación es negativo
   B. Listar las cotizaciones de compra y de venta en un archivo **.CSV**
   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

Web Scraping

El web scraping es la acción de analizar un archivo web, mayoritariamente html. Para realizarlo, debemos primero bajar la página, utilizando el comando **WGET**:

```
FILE *a = popen("wget -q - https://bolsar.info/lideres.php -O html.html", "r");
pclose(a);
```

Esto lo que hace es bajar la pagina bolsar.info/lideres.php en el archivo html.html. Lo que hacemos después es abrir el archivo resultante:

```
FILE *f = fopen("html.html","rt");
```

Ahora el archivo esta abierto y está a nuestra disposición para analizar.

Especie	Vto	Cant. Nominal Compra	Compra	Venta	Cant. Nominal	Último	Variación	Apertura	Mín	Max	Cierre Anterior	Volumen	Monto	Oper	Hora
ALUA	Cdo.	241	65,20	72,00	50	65,80	-0,30%	67,80	63,00	67,80	66,00	73.356,00	4.747.237,00	114	16:00:00

```
while (fgets(buffer,sizeof(buffer),f))
{
    if(tablaEncontrada){
        if(strstr(buffer, "<table ")){
            tablaEncontrada = false;
        }
    }
```

Lo que hacemos es no arrancar el proceso hasta encontrar una ocurrencia del texto '**<table** ', ya que en la página únicamente hay una sola tabla. Luego de esto, podemos empezar:

```
if ((pchar = strstr(buffer, "onclick=")) && tablaEncontrada)
```

Lo primero que hacemos es, si encontramos la tabla, saltar hasta la siguiente ocurrencia de 'onclick=' ya que cada fila de la tabla tiene una de estas al principio.

```
i = 0;

while (*pchar != '>')
    pchar++;
pchar++;

while(*pchar != '<'){
    tabla[fila].especie[i] = *pchar;
    i++;
    pchar++;
}

tabla[fila].especie[i] = '\0';
```

Luego adelantamos el puntero hasta encontrar un símbolo **'>'**, lo que indica el cierre de la etiqueta TD de apertura, y lo adelantamos una vez más para pararnos sobre el nombre de la especie. Leemos y almacenamos todos los caracteres hasta encontrar un símbolo **'<'**, que indica la apertura de la etiqueta TD de cierre. Luego escribimos en el último caracter de la especie el código **'\0'**, que indica el final de la cadena de chars.

A la hora de copiar números, la solución que encontré fue almacenar el texto de la página en una cadena de chars, luego corregir la puntuación:

```
i = 0;

while (*pchar != '>')
    pchar++;
pchar++;

while(*pchar != '<'){
    recoleccion[i] = *pchar;
    i++;
    pchar++;
}
recoleccion[i] = '\0';
corregirPuntuacion(recoleccion);

tabla[fila].compra = atof(recoleccion);
```

```
void corregirPuntuacion(char * string){
    int i;
    char retorno[20];
    for(i = 0;i<sizeof(string)/sizeof(string[0]);i++){
        if(string[i] == '.'){
            char * pDest = string;
            while(*string){
                if(*string != '.'){
                    *pDest++ = *string;
                }
                string++;
            }
        }
        if(string[i] == ','){
            string[i] = '.';
        }
    }
}
```

Recorremos el string en búsqueda de puntos o comas. En caso de encontrar un punto, utilizamos un algoritmo de corrección de direcciones para hacer que ninguno de los chars contenidos en el string dirija a un punto, así eliminándolo. En caso de encontrar una coma, simplemente la reemplazamos por un punto. Esto nos da el formato que la función **atof()** reconoce, que convierte un char[] a un float. Esto se repite para cada una de las variables numéricas que nos interesa tomar, que son precio de compra y de venta, precio de apertura, máximo y mínimo.

Punto A

El algoritmo requerido en A es simple, recorremos la tabla en búsqueda de especies con variación negativa. En caso de ser negativa, se la escribe. Utilizo la expresión **'%.2f'** para indicarle al printf que solo imprima dos dígitos decimales del float de variación. Un pequeño detalle que quise agregar es el último caracter que coloco al final de la impresión, con código **25**, que es una flecha para abajo.

```
for(i = 0; i<fila;i++){

    if(tabla[i].variacion < 0){
        if(i < 9) printf(" %d",i+1);
        else printf("%d",i+1);

        printf("| %s      %.2f%c %c\n", tabla[i].especie, tabla[i].variacion, '%', 25);
    }
}
```

```
MOSTRANDO ESPECIES CUYAS VARIACIONES SON NEGATIVAS
N.| Nombre      | Variacion
=================================================
 1| ALUA          -0.16% ↓
 3| ALUA          -0.63% ↓
 4| BBAR          -0.79% ↓
 6| BBAR          -0.24% ↓
10| BYMA          -1.24% ↓
12| BYMA          -0.31% ↓
15| CEPU          -0.16% ↓
18| COME          -0.20% ↓
22| CVH           -0.81% ↓
24| CVH           -1.10% ↓
34| LOMA          -0.05% ↓
36| LOMA          -0.92% ↓
39| MIRG          -0.27% ↓
40| PAMP          -0.70% ↓
41| PAMP          -0.44% ↓
42| PAMP          -0.70% ↓
43| RICH          -0.10% ↓
45| RICH          -0.42% ↓
51| TECO2         -3.03% ↓
64| VALO          -0.21% ↓
67| YPFD          -1.52% ↓
69| YPFD          -0.10% ↓
=================================================
```

Punto B

La imagen es pequeña, pero lo que se hace es escribir por cada línea guarda de las tablas del web scraping una fila en formato csv, lo único que exige es escribir los valores separándolos con ‘,’ o ‘;’.

```
csv = fopen("listadoDeCotizaciones.csv","wt");

fprintf(csv,"Especie; Precio de compra; Precio de venta; Apertura; Precio Máximo; Precio Mínimo\n");
for(i=0;i<fila;i++){
    fprintf(csv,"%s; %.2f; %.2f; %.2f; %.2f; %.2f;\n", tabla[i].especie, tabla[i].compra, tabla[i].venta, tabla[i].apertura, tabla[i].max, tabla[i].min);
}

fclose(csv);
```

Especie	Precio de compra	Precio de venta	Apertura	Precio Máximo	Precio Mínimo
ALUA	62.40	70.00	64.00	64.00	62.50
ALUA	62.40	64.90	63.00	63.50	63.00
ALUA	62.10	65.00	63.80	64.00	62.30
BBAR	248.50	1400.00	252.00	252.00	246.05
BBAR	244.00	0.00	0.00	0.00	0.00
BBAR	235.00	256.00	247.95	251.50	245.00
BMA	334.55	337.00	328.05	333.00	328.00
BMA	325.20	0.00	0.00	0.00	0.00
BMA	326.30	342.00	330.00	336.00	327.00
BYMA	796.00	1500.00	819.00	823.50	795.00
BYMA	796.00	809.50	818.00	818.00	818.00
BYMA	781.00	805.00	800.00	818.00	796.00
CEPU	60.00	66.00	58.80	61.10	58.80
CEPU	60.25	0.00	0.00	0.00	0.00
CEPU	58.00	63.20	62.00	62.00	59.60
COME	4.33	5.20	4.92	5.04	4.92

Punto C

En el punto C, generamos la misma tabla que en el punto A, pero con dos variaciones: tiene que ser en formato HTML, y hay que resaltar aquellos cuyos precios de compra y venta sean menores que su precio de apertura.

```
html = fopen("reporteVariaciones.html","wt");

fprintf(html,"<html><head><title>Listado de Variaciones</title></head><body><h1>Reporte de Variaciones</h1><table border=\"1px\">\n");

for(i = 0; i<fila;i++){
    if(tabla[i].variacion < 0){
        if(tabla[i].compra < tabla[i].apertura && tabla[i].venta < tabla[i].apertura){
            fprintf(html,"<tr style=\"background-color:#336633; color:#00ff00;\">");
        }
        else{
            fprintf(html,"<tr>");
        }
        fprintf(html,"<td>%i</td><td>%s</td><td>%.2f%\%</td></tr>\n", i+1, tabla[i].especie, tabla[i].variacion);
    }
}

fprintf(html,"</table></body></html>");

fclose(html);
```

1	ALUA	-0.16%
3	ALUA	-0.63%
4	BBAR	-0.79%
6	BBAR	-0.24%
10	BYMA	-1.24%
12	BYMA	-0.31%
15	CEPU	-0.16%
18	COME	-0.20%
22	CVH	-0.81%
24	CVH	-1.10%
34	LOMA	-0.05%
36	LOMA	-0.92%
39	MIRG	-0.27%
40	PAMP	-0.70%