Universidad de Costa Rica
Facultad de Ingeniería
Escuela de Ciencias de la Computación e Informática

CI-0117 Programación Paralela y Concurrente
Grupo 01
I Semestre

II Tarea programada: Contador de Etiquetas HTML

Profesor:
Francisco Arroyo

Estudiantes:
Rodrigo Vílchez Ulloa | B78292

12 de junio del 2020

Índice

1. Introducción

El tema de esta tarea programada consiste en implementar un programa que manipule y analice archivos externos mediante procesos ejecutados de forma paralela, ya sea para leerlos o interpretar su contenido, de manera que se controlen las condiciones de carrera, se asignen porciones del archivo a procesos específicos y que se haga de forma eficiente.

2. Objetivos

Contador de Etiquetas HTML:

- Construir un programa en C++ para contar todas las etiquetas que aparecen en un conjunto de archivos HTML que se pasará como parámetros. El programa debe contar todas las etiquetas contenidas en los archivos indicados y mostrar un listado de esas etiquetas y su cantidad de apariciones.
- Debe construir una clase C++ (FileReader) que sea capaz de procesar un archivo de texto HTML, con **H** líneas, siguiendo varias estrategias de acuerdo con la cantidad de trabajadores (**t**) que el usuario pretenda utilizar.

3. Descripción

Contador de Etiquetas HTML:

Se debe crear un programa capaz de recibir varios archivos HTML como parámetros, donde a cada uno se le debe indicar una estrategia para que contar sus etiquetas HTML, así como la cantidad de trabajadores/procesos que van a realizar esta tarea. Para lograr esto, debe existir un programa o hilo principal que cree tantos hilos como archivos indique el usuario, y estos hilos deberan crear igualmente tantos hilos como trabajadores haya indicado el usuario para ese archivo. Los trabajadores para cada archivo deberán seguir una estrategia para leer el archivo, es decir, las líneas del archivo HTML se le van a asignar a los trabajadores de una forma específica, a partir de la estrategia seleccionada. Las estrategias son las siguientes:

- **Mapeo por bloques**: Dividir el total de líneas del archivo HTML entre los t trabajadores y entregar una porción a cada trabajador (H/t).
- **Mapeo cíclico**: Entregar al primer trabajador (0) todas las líneas cuyo resto de dividir el número de línea entre t sea 0; entregar al segundo trabajador todas las líneas cuyo resto de dividir el número de línea entre t sea 1 y así sucesivamente.
- **Mapeo dinámico**: Entregar una línea del archivo a cada trabajador por demanda, cada vez que un trabajador requiere una línea le es entregada la siguiente línea disponible del archivo, el lector debe saltar a la siguiente línea. Debe sincronizar los trabajadores para que dos de ellos no reciban la misma línea y el conteo de etiquetas no sea correcto.
- **Mapeo personalizado**: Esta estrategia fue diseñada para esta tarea, consiste en un mapeo similar al por bloques, la diferencia es que la asignación de líneas es inversa, es decir, al primer trabajador se le asigna la última línea, al segundo trabajador la penúltima línea y así sucesivamente.

4. Diseño

Contador de Etiquetas HTML:

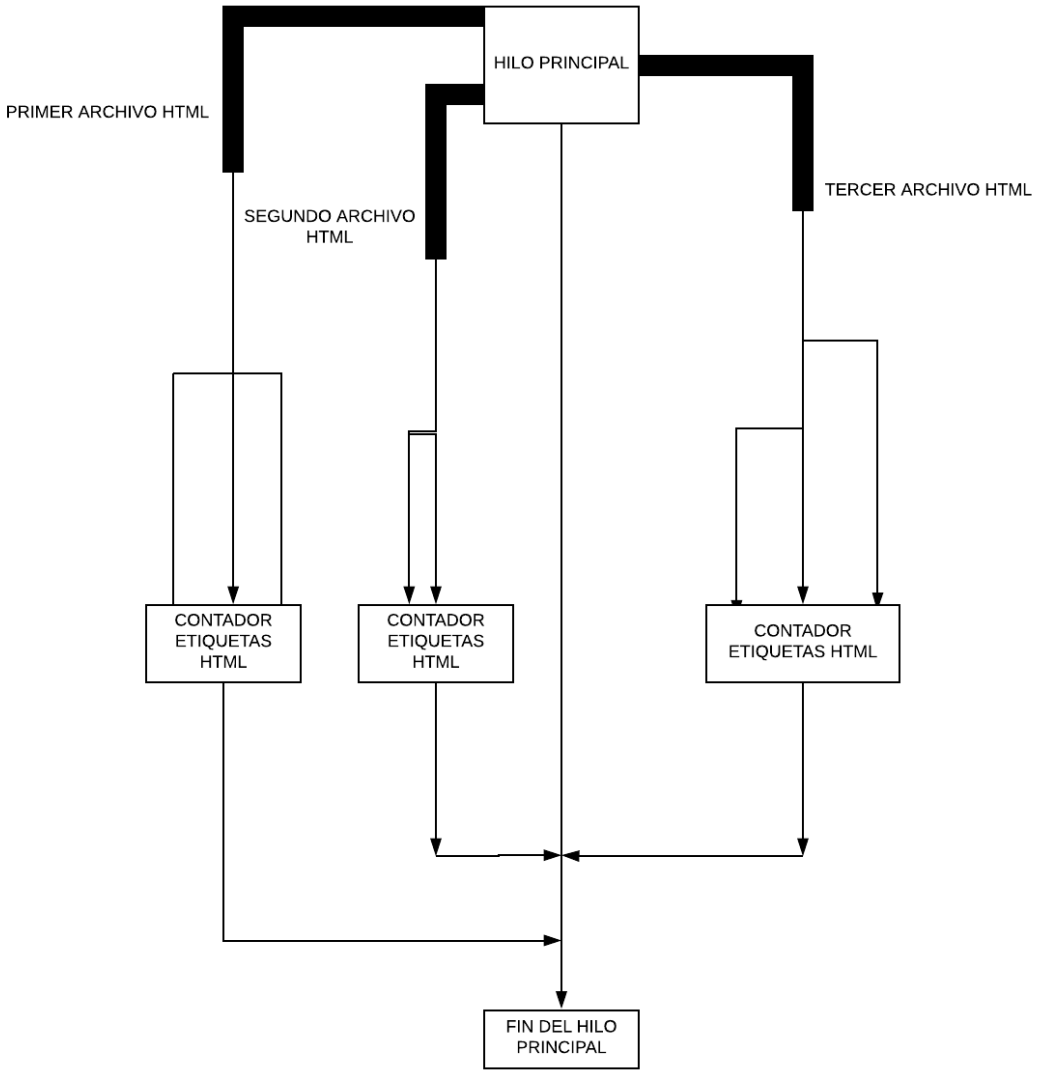

Figura 1: Diseño del Contador de Etiquetas HTML.

5. Desarrollo

Contador de Etiquetas HTML

Para implementar el programa, se utilizaron los hilos que ofrece la librería **thread** mediante *std::thread*. El primer paso es que el usuario defina, en tiempo de ejecución, la cantidad de archivos que desea procesar, y para cada uno de estos, debe indicar también la cantidad de trabajadores que van a procesar ese archivo, así como la estrategia a seguir por parte de estos trabajadores. De esta forma, el hilo principal va a crear **n** hilos, que corresponde a la cantidad de archivos que el usuario indicó. Cada uno de estos hilos va a ejecutar una función en paralelo, que se encarga de crear una instancia de la clase *Parser*, el cual ya fue implementado para el Laboratorio 7, que se encarga de recibir una línea de caracteres, y este separa las etiquetas HTML que hayan en él utilizando expresiones regulares. Una modificación en relación con el Laboratorio 7 es que cada instancia de esta clase tiene un *std::map* donde se van a almacenar las etiquetas HTML que se encuentren, así como la cantidad de apariciones, de manera que los trabajadores que se creen a partir del hilo respectivo a un archivo, podrán ir sumando las etiquetas HTML que procesen, sin necesidad de pasar estructuras o referencias entre los trabajadores, el cual es una restricción para esta tarea. Otra instancia creada es de la clase *FileReader*, el cual fue implementado para el Laboratorio 8, con los cambios necesarios para que sus métodos funcionen de manera paralela. También, se va a crear un array de semáforos, uno para cada trabajador, el cual permite que, según la estrategia utilizada, cada trabajador pueda leer una línea del archivo cuando el corresponda, garantizando que no hayan condiciones de carrera ni líneas asignada dos veces a distintos trabajadores. Por motivos de diseño, el semáforo del primer trabajador será inicializado en 1 y los demás en 0, para que haya un proceso que comience el procesamiento y que, según la estrategia, vaya avisando a los otros trabajadores que pueden proceder a leer una línea. Esto se hizo de esta manera ya que, como restricción del proyecto, el archivo que se está procesando solo puede estar abierto una vez por cada instancia de FileReader, además de que la lectura de un archivo en paralelo es técnicamente imposible pues existe un puntero que indica que byte del archivo se está leyendo, de manera que se pueden esperar resultados incorrectos cuando dos procesos solicitan leer una línea al mismo tiempo. Por este motivo es que los semáforos se utilizan como mecanismos de sincronización para los procesos, pues *FileReader* simplemente recibe la instrucción de leer una línea y este la devolverá, de forma que, siguiendo el orden de asignación de líneas para cada trabajador por parte de la estrategia, los trabajadores se dormirán cuando intenten leer su línea y se despertarán cuando el trabajador anterior haya leído su línea correspondiente. Una vez creado el arreglo de semáforos, cada uno de estos primeros procesos (uno por archivo HTML) creará tantos hilos como el usuario le haya indicado, donde cada uno va a tener un ID único, y ejecutará otra función en paralelo que se encarga de leer líneas del archivo a través de la instancia de *FileReader* anteriormente creada. Una vez que estos trabajadores terminen, el proceso que los creó mostrará el resultado del conteo de etiquetas y el hilo finalizará, regresando así al hilo principal. Para cada estrategia de mapeo se le asigna un número de línea inicial a cada trabajador, y mediante semáforos, el proceso n le avisará al proceso $n+1$ cuando haya terminado de leer su línea para que este pueda proseguir, cuando el último trabajador lee su línea, le avisará al primero que puede proseguir y así sucesivamente. Esto solamente no sucede en el **mapeo dinámico**, donde las líneas son asignadas por demanda, de manera que simplemente se utiliza un *std::mutex* cada vez que un trabajador vaya a realizar un *std::getline*. En el caso de **mapeo por bloques** y **mapeo personalizado**, por su naturaleza de implementación, hará que el trabajador n lea una cantidad de líneas de manera consecutiva hasta que haya leído todas las líneas que le corresponde, y le avisará al trabajador $n+1$ que puede seguir leyendo el bloque de líneas que le corresponde. Cada línea es almacenada una sola vez por trabajador, de manera que un proceso lee-cuenta etiquetas-lee-cuenta etiquetas-... Para el **mapeo cíclico**, en el trabajador n lee UNA sola línea y le avisará al trabajador $n+1$ que la línea siguiente y así sucesivamente. Toda comunicación se realiza entre *std::mutex* y semáforos. Una vez que cada hilo haya mostrado la cantidad de etiquetas encontradas por sus trabajadores, se regresará al hilo principal para finalizar el programa. El espacio en memoria utilizado por las instancias es liberado al terminar la ejecución del hilo.

6. Manual de usuario

Requerimientos de Software

- **Sistema Operativo:** Linux
- **Arquitectura:** 32/64 bits
- **Ambiente:** Terminal

Compilación

- **Contador de Etiquetas HTML**
  Para compilar el programa, se utiliza el comando `make`:

```
$ make $
```

Especificación de las funciones del programa

Para comenzar a correr el programa, solamente se ejecuta el siguiente comando:

```
$ ./etiquetaHTML $
```

Es necesario recalcar que, el programa va a preguntar el usuario la cantidad de archivos que desea procesar, y para cada uno de ellos, le va a consultar por el número de trabajadores así como la estrategia a utilizar. La mayoría de posibilidades para corromper el programa a partir de las respuestas del usuario están cubiertas. También se ofrecen tres archivos HTML para probar el programa, estos son: **ecci.html**, **mediacion.html** y **ori.html**, los cuales tienen un código fuente dentro de lo esperable por parte de HTML. De igual forma, es posible que parezca que el output entre una prueba y otra sea distinta para un archivo en específico, esto sucede porque en algunas ocasiones, se imprime que una etiqueta aparece 0 veces, pero también se imprime indicando la cantidad de apariciones, esto no afecta el resultado de la cantidad de ocurrencias real para una etiqueta en específico.

7. Casos de Prueba

Contador de Etiquetas HTML sin tomar en cuenta el tiempo de procesamiento

Prueba 1:

ecci.html: 4 trabajadores y estrategia 1 (mapeo por bloques).
mediacion.html: 8 trabajadores y estrategia 3 (mapeo dinámico).

```
rigovil@rodrigo:~/Escritorio/UCR/I Semestre 2020/CI-0117/Tareas programadas/II TP$ ./etiquetasHTML
Ingrese el numero de archivos: 2

Ingrese el nombre del archivo: ecci.html
Ingrese la cantidad de trabajadores: 4
Ingrese la estrategia: (1 = bloques, 2 = ciclico, 3 = dinamico, 4 = personalizado): 1

Ingrese el nombre del archivo: mediacion.html
Ingrese la cantidad de trabajadores: 8
Ingrese la estrategia: (1 = bloques, 2 = ciclico, 3 = dinamico, 4 = personalizado): 3

ETIQUETAS DEL ARCHIVO "mediacion.html"
/a --> 142              /body --> 1             /button --> 10          /div --> 23
/footer --> 1           /h2 --> 10              /h3 --> 3               /head --> 1
/header --> 1           /html --> 1             /li --> 135             /p --> 5
/script --> 1           /section --> 16         /span --> 10            /style --> 1
/title --> 1            /ul --> 10              a --> 142               body --> 1
br --> 1                button --> 10           div --> 24              footer --> 1
h2 --> 10               h3 --> 3                head --> 1              header --> 1
hr --> 1                html --> 1              img --> 4               li --> 135
link --> 3              meta --> 4              p --> 5         script --> 1
section --> 15          span --> 10             style --> 1             title --> 1
ul --> 10

ETIQUETAS DEL ARCHIVO "ecci.html"
/a --> 158              /article --> 1          /body --> 1             /button --> 3
/div --> 126            /footer --> 1           /form --> 1             /h2 --> 4
/h4 --> 10              /h5 --> 3               /head --> 1             /header --> 2
/html --> 1             /i --> 21               /li --> 117             /nav --> 4
/p --> 14               /script --> 12          /section --> 14         /span --> 52
/strong --> 1           /style --> 1            /title --> 1            /ul --> 31
a --> 158               article --> 1           body --> 1              br --> 4
button --> 3            div --> 126             footer --> 1            form --> 1
h2 --> 4                h4 --> 10               h5 --> 3                head --> 1
header --> 2            html --> 1              i --> 21                img --> 20
input --> 3             li --> 117              link --> 13             meta --> 4
nav --> 4               p --> 14                script --> 13           section --> 14
span --> 52             strong --> 1            style --> 1             title --> 1
ul --> 31
rigovil@rodrigo:~/Escritorio/UCR/I Semestre 2020/CI-0117/Tareas programadas/II TP$
```

Prueba 2:

mediacion.html: 10 trabajadores y estrategia 2 (mapeo cíclico).
ori.html: 3 trabajadores y estrategia 1 (mapeo por bloques).

```
rigovil@rodrigo:~/Escritorio/UCR/I Semestre 2020/CI-0117/Tareas programadas/II TP$ ./etiquetasHTML
Ingrese el numero de archivos: 2

Ingrese el nombre del archivo: mediacion.html
Ingrese la cantidad de trabajadores: 10
Ingrese la estrategia: (1 = bloques, 2 = ciclico, 3 = dinamico, 4 = personalizado): 2

Ingrese el nombre del archivo: ori.html
Ingrese la cantidad de trabajadores: 3
Ingrese la estrategia: (1 = bloques, 2 = ciclico, 3 = dinamico, 4 = personalizado): 1

ETIQUETAS DEL ARCHIVO "ori.html"
/a --> 52               /article --> 1          /b --> 1                /body --> 1
/button --> 6           /div --> 99             /figcaption --> 4               /figure --> 4
/footer --> 1           /form --> 1             /h2 --> 13              /h3 --> 4
/head --> 1             /header --> 1           /html --> 1             /i --> 12
/label --> 1            /li --> 26              /main --> 1             /nav --> 7
/p --> 17               /script --> 37          /section --> 9          /span --> 39
/svg --> 12             /time --> 8             /title --> 1            /ul --> 7
a --> 52                article --> 1           b --> 1         body --> 1
button --> 6            div --> 99              figcaption --> 4                figure --> 4
footer --> 1            form --> 1              h2 --> 13               h3 --> 4
head --> 1              header --> 1            html --> 1              i --> 12
img --> 17              input --> 1             label --> 1             li --> 26
link --> 12             main --> 1              meta --> 6              nav --> 7
p --> 17                path --> 4              rect --> 8              script --> 37
section --> 9           span --> 38             svg --> 12              time --> 8
title --> 1             ul --> 7

ETIQUETAS DEL ARCHIVO "mediacion.html"
/a --> 142              /body --> 1             /button --> 10          /div --> 23
/footer --> 1           /h2 --> 10              /h3 --> 3               /head --> 1
/header --> 1           /html --> 1             /li --> 135             /p --> 5
/script --> 1           /section --> 16         /span --> 10            /style --> 1
/title --> 1            /ul --> 10              a --> 142               body --> 1
br --> 1                button --> 10           div --> 24              footer --> 1
h2 --> 10               h3 --> 3                head --> 1              header --> 1
hr --> 1                html --> 1              img --> 4               li --> 135
link --> 3              meta --> 4              p --> 5         script --> 1
section --> 15          span --> 10             style --> 1             title --> 1
ul --> 10
rigovil@rodrigo:~/Escritorio/UCR/I Semestre 2020/CI-0117/Tareas programadas/II TP$
```

Prueba 3:

ori.html: 5 trabajadores y estrategia 4 (mapeo personalizado).
ecci.html: 1 trabajador (demuestra el funcionamiento serial) y estrategia 2 (mapeo cíclico).

```
rigovil@rodrigo:~/Escritorio/UCR/I Semestre 2020/CI-0117/Tareas programadas/II TP$ ./etiquetasHTML
Ingrese el numero de archivos: 2

Ingrese el nombre del archivo: ori.html
Ingrese la cantidad de trabajadores: 5
Ingrese la estrategia: (1 = bloques, 2 = ciclico, 3 = dinamico, 4 = personalizado): 4

Ingrese el nombre del archivo: ecci.html
Ingrese la cantidad de trabajadores: 1
Ingrese la estrategia: (1 = bloques, 2 = ciclico, 3 = dinamico, 4 = personalizado): 2

ETIQUETAS DEL ARCHIVO "ori.html"
/a --> 52               /article --> 1          /b --> 1                /body --> 1
/button --> 6           /div --> 99             /figcaption --> 4               /figure --> 4
/footer --> 1           /form --> 1             /h2 --> 13              /h3 --> 4
/head --> 1             /header --> 1           /html --> 1             /i --> 12
/label --> 1            /li --> 26              /main --> 1             /nav --> 7
/p --> 17               /script --> 37          /section --> 9          /span --> 39
/svg --> 12             /time --> 8             /title --> 1            /ul --> 7
a --> 52                article --> 1           b --> 1         body --> 1
button --> 6            div --> 99              figcaption --> 4                figure --> 4
footer --> 1            form --> 1              h2 --> 13               h3 --> 4
head --> 1              header --> 1            html --> 1              i --> 12
img --> 17              input --> 1             label --> 1             li --> 26
link --> 12             main --> 1              meta --> 6              nav --> 7
p --> 17                path --> 4              rect --> 8              script --> 37
section --> 9           span --> 38             svg --> 12              time --> 8
title --> 1             ul --> 7

ETIQUETAS DEL ARCHIVO "ecci.html"
/a --> 158              /article --> 1          /body --> 1             /button --> 3
/div --> 126            /footer --> 1           /form --> 1             /h2 --> 4
/h4 --> 10              /h5 --> 3               /head --> 1             /header --> 2
/html --> 1             /i --> 21               /li --> 117             /nav --> 4
/p --> 14               /script --> 12          /section --> 14         /span --> 52
/strong --> 1           /style --> 1            /title --> 1            /ul --> 31
a --> 158               article --> 1           body --> 1              br --> 4
button --> 3            div --> 126             footer --> 1            form --> 1
h2 --> 4                h4 --> 10               h5 --> 3                head --> 1
header --> 2            html --> 1              i --> 21                img --> 20
input --> 3             li --> 117              link --> 13             meta --> 4
nav --> 4               p --> 14                script --> 13           section --> 14
span --> 52             strong --> 1            style --> 1             title --> 1
ul --> 31
rigovil@rodrigo:~/Escritorio/UCR/I Semestre 2020/CI-0117/Tareas programadas/II TP$
```

Contador de Etiquetas HTML tomando en cuenta el tiempo de procesamiento

Prueba 4:

ecci.html: 8 trabajadores y estrategia 1 (mapeo por bloques).

```
rigovil@rodrigo:~/Escritorio/UCR/I Semestre 2020/CI-0117/Tareas programadas/II TP$ ./etiquetasHTML
Ingrese el numero de archivos: 1

Ingrese el nombre del archivo: ecci.html
Ingrese la cantidad de trabajadores: 8
Ingrese la estrategia: (1 = bloques, 2 = ciclico, 3 = dinamico, 4 = personalizado): 1

ETIQUETAS DEL ARCHIVO "ecci.html" Y TIEMPO DE DURACIÓN

Tiempo de procesamiento: 0.40218

/a --> 158              /article --> 1          /body --> 1             /button --> 3
/div --> 126            /footer --> 1           /form --> 1             /h2 --> 4
/h4 --> 10              /h5 --> 3               /head --> 1             /header --> 2
/html --> 1             /i --> 21               /li --> 117             /nav --> 4
/p --> 14               /script --> 12          /section --> 14         /span --> 52
/strong --> 1           /style --> 1            /title --> 1            /ul --> 31
a --> 158               article --> 1           body --> 1              br --> 4
button --> 3            div --> 126             footer --> 1            form --> 1
h2 --> 4                h4 --> 10               h5 --> 3                head --> 1
header --> 2            html --> 1              i --> 21                img --> 20
input --> 3             li --> 117              link --> 13             meta --> 4
nav --> 4               p --> 14                script --> 13           section --> 14
span --> 52             strong --> 1            style --> 1             title --> 1
ul --> 31
rigovil@rodrigo:~/Escritorio/UCR/I Semestre 2020/CI-0117/Tareas programadas/II TP$
```

Prueba 5:

ecci.html: 8 trabajadores y estrategia 2 (mapeo cíclico).

```
rigovil@rodrigo:~/Escritorio/UCR/I Semestre 2020/CI-0117/Tareas programadas/II TP$ ./etiquetasHTML
Ingrese el numero de archivos: 1

Ingrese el nombre del archivo: ecci.html
Ingrese la cantidad de trabajadores: 8
Ingrese la estrategia: (1 = bloques, 2 = ciclico, 3 = dinamico, 4 = personalizado): 2

ETIQUETAS DEL ARCHIVO "ecci.html" Y TIEMPO DE DURACIÓN

Tiempo de procesamiento: 0.644629

/a --> 158              /article --> 1          /body --> 1             /button --> 3
/div --> 126            /footer --> 1           /form --> 1             /h2 --> 4
/h4 --> 10              /h5 --> 3               /head --> 1             /header --> 2
/html --> 1             /i --> 21               /li --> 117             /nav --> 4
/p --> 14               /script --> 12          /section --> 14         /span --> 52
/strong --> 1           /style --> 1            /title --> 1            /ul --> 31
a --> 158               article --> 1           body --> 1              br --> 4
button --> 3            div --> 126             footer --> 1            form --> 1
h2 --> 4                h4 --> 10               h5 --> 3                head --> 1
header --> 2            html --> 1              i --> 21                img --> 20
input --> 3             li --> 117              link --> 13             meta --> 4
nav --> 4               p --> 14                script --> 13           section --> 14
span --> 52             strong --> 1            style --> 1             title --> 1
ul --> 31
rigovil@rodrigo:~/Escritorio/UCR/I Semestre 2020/CI-0117/Tareas programadas/II TP$
```

Prueba 6:

ecci.html: 8 trabajadores y estrategia 3 (mapeo dinámico).

```
rigovil@rodrigo:~/Escritorio/UCR/I Semestre 2020/CI-0117/Tareas programadas/II TP$ ./etiquetasHTML
Ingrese el numero de archivos: 1

Ingrese el nombre del archivo: ecci.html
Ingrese la cantidad de trabajadores: 8
Ingrese la estrategia: (1 = bloques, 2 = ciclico, 3 = dinamico, 4 = personalizado): 3

ETIQUETAS DEL ARCHIVO "ecci.html" Y TIEMPO DE DURACIÓN

Tiempo de procesamiento: 0.642434

/a --> 158              /article --> 1          /body --> 1             /button --> 3
/div --> 126            /footer --> 1           /form --> 1             /h2 --> 4
/h4 --> 10              /h5 --> 3               /head --> 1             /header --> 2
/html --> 1             /i --> 21               /li --> 117             /nav --> 4
/p --> 14               /script --> 12          /section --> 14         /span --> 52
/strong --> 1           /style --> 1            /title --> 1            /ul --> 31
a --> 158               article --> 1           body --> 1              br --> 4
button --> 3            div --> 126             footer --> 1            form --> 1
h2 --> 4                h4 --> 10               h5 --> 3                head --> 1
header --> 2            html --> 1              i --> 21                img --> 20
input --> 3             li --> 117              link --> 13             meta --> 4
nav --> 4               p --> 14                script --> 13           section --> 14
span --> 52             strong --> 1            style --> 1             title --> 1
ul --> 31
rigovil@rodrigo:~/Escritorio/UCR/I Semestre 2020/CI-0117/Tareas programadas/II TP$
```

Prueba 7:

ecci.html: 8 trabajadores y estrategia 4 (mapeo personalizado).

```
rigovil@rodrigo:~/Escritorio/UCR/I Semestre 2020/CI-0117/Tareas programadas/II TP$ ./etiquetasHTML
Ingrese el numero de archivos: 1

Ingrese el nombre del archivo: ecci.html
Ingrese la cantidad de trabajadores: 8
Ingrese la estrategia: (1 = bloques, 2 = ciclico, 3 = dinamico, 4 = personalizado): 4

ETIQUETAS DEL ARCHIVO "ecci.html" Y TIEMPO DE DURACIÓN

Tiempo de procesamiento: 0.431904

/a --> 158              /article --> 1          /body --> 1             /button --> 3
/div --> 126            /footer --> 1           /form --> 1             /h2 --> 4
/h4 --> 10              /h5 --> 3               /head --> 1             /header --> 2
/html --> 1             /i --> 21               /li --> 117             /nav --> 4
/p --> 14               /script --> 12          /section --> 14         /span --> 52
/strong --> 1           /style --> 1            /title --> 1            /ul --> 31
a --> 158               article --> 1           body --> 1              br --> 4
button --> 3            div --> 126             footer --> 1            form --> 1
h2 --> 4                h4 --> 10               h5 --> 3                head --> 1
header --> 2            html --> 1              i --> 21                img --> 20
input --> 3             li --> 117              link --> 13             meta --> 4
nav --> 4               p --> 14                script --> 13           section --> 14
span --> 52             strong --> 1            style --> 1             title --> 1
ul --> 31
rigovil@rodrigo:~/Escritorio/UCR/I Semestre 2020/CI-0117/Tareas programadas/II TP$
```

FIN