MODELOS DE COMPUTACION
Práctica de LEX
Curso 2018-19

Autor[1]: Víctor García Carrera

El ejercicio llevado a cabo pretende ser una aproximación al funcionamiento de un robot de indexación de páginas webs, como lo puede ser GoogleBot. Utilizando LEX como analizador léxico, partimos del análisis de ficheros HTML como el fichero provisto *ANALISIS*. El programa implementado busca los enlaces existentes en la página web y los va almacenando en un fichero *webs.txt*. Al mismo tiempo, busca determinadas palabras clave que nosotros establecemos, como lo han sido en este caso los términos *atomo, universo* y *luz*. Al encontrar cualquier coincidencia, la muestra por pantalla, junto con la línea analizada en la que se encontró y el número de coincidencias que lleva hasta el momento.

Esta aplicación puede ser realmente útil, pues se pueden implementar diversas funciones muy interesantes. Podríamos utilizar las urls almacenadas en el fichero *webs.txt* para descargar su fichero HTML y repetir este proceso de forma recursiva (debido al gran número de enlaces que suelen contener a día de hoy las páginas webs), partiendo de un fichero HTML inicial y buscando las palabras clave no solo en esa web, sino en todas aquellas que aparecen referenciadas en esa web de partida. Deberíamos establecer un número máximo de webs analizadas o coincidencias, a elección del objetivo que busquemos. Podemos mejorar la búsqueda y análisis de las urls utilizando un segundo fichero *analyzedwebs.txt* donde fueramos metiendo los enlaces de las webs analizadas a fin de no repetir búsquedas innecesarias. Finalmente, otro posible uso es el mencionado al principio del documento de crear un robot de indexación, donde mantenga registro del número de veces que aparece un dominio o url y al final el robot pueda determinar que webs son las más relevantes al aparecer un mayor número de veces o a través de cualquier otro posible método de cálculo de la relevancia.

Añadir que el análisis recursivo de nuevas páginas webs en las urls encontradas está casi totalmente implementada, solo que algunos problemas relativos al tamaño máximo YYLEN de la cadena yytext (cadena que contiene el texto que analiza LEX) y algunos campos HTML de longitud excesiva han impedido completarlo. Sin embargo, destacar el esfuerzo que ha supuesto y el trabajo realizado en esta parte, pues con algunas pequeñas modificaciones estaría completamente implementado ese buscador de palabras claves en páginas webs y analizador de todos los enlaces de la misma.

---

1 Como autor declaro que los contenidos del presente documento son originales y elaborados por mi. De no cumplir con este compromiso, soy consciente de que, de acuerdo con la "Normativa de evaluación y de calificaciones de los estudiantes de la Universidad de Granada" esto "conllevará la calificación numérica de cero … independientemente del resto de calificaciones que el estudiante hubiera obtenido ..."