Universidad de Oviedo

Prácticas de laboratorio

Sesión 2

Sistemas de Información para la Web

Creación de un crawler básico

Omar Teixeira González, *UO281847*

12/10/2023

Tabla de contenidos

Documentación sobre el *crawler*

Descripción

El crawler realizado para la práctica 2 de la asignatura representa una implementación básica de lo que sería un crawler real para la extracción de información y documentos en la web.

Forma de uso

Para utilizarlo es necesario el paso de 4 parámetros, los cuales son:

- *filename*: nombre del archivo donde se encuentran los enlaces para su posterior crawling correspondiente.
- *max_downloads*: número máximo de descargas permitidas.
  - valor por defecto: 10.
- *seconds*: número de segundos a esperar entre peticiones.
  - valor por defecto: 2.
- *search_type*: tipo de búsqueda a realizar, siendo este *breadth-first* (en anchura) o *depth-first* (en profundidad).
  - valor por defecto: false (*depth-first*).

Ejemplos y resultados obtenidos.

seed.txt contiene el siguiente enlace: https://www.game.es/. seeds.txt contiene los siguientes enlaces:

- https://en.wikipedia.org/wiki/Dromaeosauroides.
- https://en.wikipedia.org/wiki/Theropoda.
- https://en.wikipedia.org/wiki/Dinosaur.
- https://en.wikipedia.org/wiki/Early_Cretaceous.

Cabe mencionar que se intentó probar con otros enlaces, tales como: http://ingenieriainformatica.uniovi.es/, pero se encintraron errores relacionados con la configuración SSL, uno de ellos (el más predominante) era el siguiente:

<urlopen error [SSL: UNSAFE_LEGACY_RENEGOTIATION_DISABLED] unsafe legacy renegotiation disabled (_ssl.c:1006)>

Pasando a los resultados obtenidos, distingo entre el número de enlaces empleado.

Un único enlace

Recorrido en profundidad

Crawyler.py files/seed.txt

Resultados obtenidos

SISTEMAS DE INFORMACIÓN PARA LA WEB
- CRAWLER BÁSICO

Links a explorar:
-> https://www.game.es
Número de descargas permitidas: 10
Tiempo de espera entre peticiones: 2 sec
Recorrido en profundidad
--> 10 ----------
Link actual: https://www.game.es
--> 9 ----------
Link actual: https://www.game.es/tiendas
--> 8 ----------
Link actual: https://www.game.es/app
--> 7 ----------
Link actual: https://www.game.es/atencion-al-cliente
--> 6 ----------
Link actual: https://www.game.es/atencion-al-cliente#header-login
--> 5 ----------
Link actual: https://www.game.es/comprar
--> 4 ----------
Link actual: https://www.game.es/comprar#header-login
--> 3 ----------
Link actual: https://www.game.es/VIDEOJUEGOS
--> 2 ----------
Link actual: https://www.game.es/VIDEOJUEGOS#header-login
--> 1 ----------
Link actual: https://www.game.es/xbox-all-access

- FIN DEL CRAWLING

Recorrido en anchura.

Crawyler.py files/seed.txt -downloads 10 -seconds 2 -search

Resultados obtenidos.

SISTEMAS DE INFORMACIÓN PARA LA WEB
- CRAWLER BÁSICO

Links a explorar:
-> https://www.game.es
Número de descargas permitidas: 10
Tiempo de espera entre peticiones: 2 sec
Recorrido en anchura
--> 10 ----------
Link actual: https://www.game.es
--> 9 ----------
Link actual: https://www.game.es/tiendas
--> 8 ----------
Link actual: https://www.game.es/app
--> 7 ----------
Link actual: https://www.game.es/atencion-al-cliente
--> 6 ----------
Link actual: https://www.game.es#header-login
--> 5 ----------
Link actual: https://www.game.es/comprar
--> 4 ----------
Link actual: https://www.game.es/VIDEOJUEGOS
--> 3 ----------
Link actual: https://www.game.es/xbox-all-access
--> 2 ----------
Link actual: https://www.game.es/videojuegos/consolas
--> 1 ----------
Link actual: https://www.game.es/VIDEOJUEGOS/PS5

- FIN DEL CRAWLING

Múltiples enlaces

Recorrido en profundidad

Crawyler.py files/seeds.txt

Resultados obtenidos

SISTEMAS DE INFORMACIÓN PARA LA WEB
- CRAWLER BÁSICO

Links a explorar:
-> https://en.wikipedia.org/wiki/Dromaeosauroides
-> https://en.wikipedia.org/wiki/Theropoda
-> https://en.wikipedia.org/wiki/Dinosaur
-> https://en.wikipedia.org/wiki/Early_Cretaceous
Número de descargas permitidas: 10
Tiempo de espera entre peticiones: 2 sec
Recorrido en profundidad
--> 10 ----------
Link actual: https://en.wikipedia.org/wiki/Dromaeosauroides
--> 9 ----------
Link actual: https://en.wikipedia.org/wiki/Dromaeosauroides#bodyContent
--> 8 ----------
Link actual: https://en.wikipedia.org/wiki/Main_Page
--> 7 ----------
Link actual: https://en.wikipedia.org/wiki/Main_Page#bodyContent
--> 6 ----------
Link actual: https://en.wikipedia.org/wiki/Wikipedia:Contents
--> 5 ----------
Link actual: https://en.wikipedia.org/wiki/Wikipedia:Contents#bodyContent
--> 4 ----------
Link actual: https://en.wikipedia.org/wiki/Portal:Current_events
--> 3 ----------
Link actual: https://en.wikipedia.org/wiki/Portal:Current_events#bodyContent
--> 2 ----------
Link actual: https://en.wikipedia.org/wiki/Special:Random
--> 1 ----------
Link actual: https://en.wikipedia.org/wiki/Wikipedia:About

- FIN DEL CRAWLING

Recorrido en anchura

Crawyler.py files/seeds.txt -downloads 10 -seconds 2 -search

Resultados obtenidos

SISTEMAS DE INFORMACIÓN PARA LA WEB
- CRAWLER BÁSICO

Links a explorar:
-> https://en.wikipedia.org/wiki/Dromaeosauroides
-> https://en.wikipedia.org/wiki/Theropoda
-> https://en.wikipedia.org/wiki/Dinosaur
-> https://en.wikipedia.org/wiki/Early_Cretaceous
Número de descargas permitidas: 10
Tiempo de espera entre peticiones: 2 sec
Recorrido en anchura
--> 10 ----------
Link actual: https://en.wikipedia.org/wiki/Dromaeosauroides
--> 9 ----------
Link actual: https://en.wikipedia.org/wiki/Theropoda
--> 8 ----------
Link actual: https://en.wikipedia.org/wiki/Dinosaur
--> 7 ----------
Link actual: https://en.wikipedia.org/wiki/Early_Cretaceous
--> 6 ----------
Link actual: https://en.wikipedia.org/wiki/Dromaeosauroides#bodyContent
--> 5 ----------
Link actual: https://en.wikipedia.org/wiki/Main_Page
--> 4 ----------
Link actual: https://en.wikipedia.org/wiki/Wikipedia:Contents
--> 3 ----------
Link actual: https://en.wikipedia.org/wiki/Portal:Current_events
--> 2 ----------
Link actual: https://en.wikipedia.org/wiki/Special:Random
--> 1 ----------
Link actual: https://en.wikipedia.org/wiki/Wikipedia:About

- FIN DEL CRAWLING