
# Crawler Wikipedia Español

*REC-INF*

CARLOS BELTRÁN ROMERO

Crawler basado en descargas de la Wikipedia a raíz de una semilla

Curso 22/23

**Resumen**

Crawler escrito en Java que está diseñado para rastrear páginas de Wikipedia y extraer información específica de ellas.

Utiliza la biblioteca Apache Tika para analizar las páginas HTML y extraer la información, y expresiones regulares para filtrar los enlaces y extraer los datos relevantes.

El rastreador comienza con una URL semilla, que es una página de Wikipedia, y sigue recursivamente los enlaces en esa página a otras páginas de Wikipedia, evitando los enlaces que no son relevantes.

Las URLs visitadas se almacenan en un HashSet para evitar visitar la misma página más de una vez.

# Índice

# 1. Estructura principal del Crawler

El crawler, comienza pidiendo por teclado el idioma en el que se desea descargar las páginas web de la Wikipedia. Al ser un crawler específicamente para páginas en español, la introducción de otro lenguaje a éste dara fin al crawler.

Una vez estudiado como se forman los enlaces de la wikipedia, procedo a formar la esturctura que debe tener la semilla del crawler.

Con la semilla ya estructurada y completa, se procede a preguntar por teclado la URL sobre la que descargaremos su contenido y se verifica que cumple con la esturctura de la semilla. El crawler está preparado para, si la página web no existe o la URL introducida es érronea ( no se corresponde con la estructura de la semilla) muestra por pantalla el error y el programa finaliza. Este rastreador está preparado para descargar como máximo 200 URLs internas. Se procede a llamar al método crawling para tratar las demás URL internas.

## 1.1. Crawling

El método crawling es el proceso principal del crawler y es el encargado de llevar a cabo el recorrido de las páginas. Este método es el encargado de recibir la URL semilla y comenzar el proceso de recorrido. El método de crawling se repite de forma recursiva para cada URL en la lista de URLs a visitar. Los pasos que realiza son los siguientes:

1. Se verifica que la URL sea válida (es un link y existe en la wikipedia) .Si la resupuesta da como resultado 404 es que el link no existe en la wikipedia
2. Si el link es correcto, se procede a descagrar el HTML de la página correspondiente Uvez descargado el HTML se aplican los filtros (patrones regulares) para que puedan aparecer correctamente signos de puntuación y demás.
3. Los enlaces internos encontrados en la página se guardan en el HashMap url_lista que actúa como cola de enlaces que quedan por visitar.
4. Se agrega la URL actual al conjunto de URLs visitadas para evitar volver a visitarla en el futuro.
5. Se toma la siguiente URL de la lista y se repite el proceso desde el paso 1.

El proceso se repite hasta que se alcanza el número máximo de páginas a visitar o se hayan visitado 200 páginas.

El contenido de cada página aparece escrito en forma de doc en una carpeta denominada Archivos.

# 2. Patrones Regulares usados

El crawler utiliza varios patrones de expresiones regulares para filtrar información en las páginas HTML que recorre.

- "semilla":Este patrón es usado para comprobar que la URL semilla cumple con la estructura de una URL perteneciente a la wikipedia en español

- "filtro2": Este patrón se utiliza para buscar enlaces que no son relevantes en el HTML de la página y reemplazarlos con espacios vacíos.

- "filtro3": Este patrón se utiliza para buscar enlaces relevantes en el HTML de la página y extraerlos. Estos enlaces son agregados a una lista para ser visitados en el futuro.

Los filtros se utilizan en el método AplicaFiltros para limpiar el HTML y extraer los enlaces relevantes.

# 3. Métodos auxiliares usados

Para limpiar el código he realizado ciertas operaciones en métodos que son los siguientes:

- Método "GeneroHTML": Este método utiliza la biblioteca Apache Tika para analizar el HTML de una página dada y devolverlo como una cadena de texto.

- Método "HttpResponse": Este método recibe una URL y devuelve el código de respuesta HTTP de la página correspondiente.

- Método AplicaFiltros : Este método recibe el HTML de una página y lo procesa a través de varios patrones de expresiones regulares para limpiarlo y extraer los enlaces relevantes.

- Método "fill_URL_array": Su función es la de llenar el arreglo de url_lista con urls que el programa debe visitar.

- Método "WriteFile": Es un método que recibe como parámetro un objeto "Filez una cadena de texto, y se encarga de escribir la cadena de texto en el archivo especificado. Este método utiliza la clase "BufferedWriter"para escribir el texto en el archivo. En este caso específico es utilizado para escribir el contenido de cada pagina que se visita en un archivo,

# 4. Tika

Se ha utilizado la biblioteca Apache Tika en este código porque permite analizar fácilmente el contenido de archivos y páginas web en diferentes formatos y extraer información específica de ellos. Tika utiliza diferentes detectores y parseadores para procesar los distintos formatos de archivo, como HTML, PDF, Word, entre otros.

En este caso específico, Tika se utiliza para analizar las páginas HTML de Wikipedia y extraer su contenido como una cadena de texto. El objeto HtmlParser de Tika es utilizado para analizar el HTML y devolverlo como una cadena de texto, lo cual es útil para el rastreador ya que permite procesar el contenido de la página para extraer enlaces relevantes.