Crawler para Wikimedia Commons

Una de las técnicas que pueden utilizarse para rastrear e indexar contenido de la Web es conocida como "*crawling*". Consiste en descargar una pagina web y analizarla en busca de contenido especifico, o bien localizar errores, plagios. Esta técnica tiene dos partes. La primera de ellas es localizar los enlaces y descargar las paginas de interés. La segunda, que es la que nos va a ocupar en este ejercicio, consiste en analizar el contenido fuente para detectar aquello que es objeto de interés.

En concreto en este ejercicio se propone analizar el contenido de paginas de *Wikimedia Commons* en busca de imágenes y ficheros de audio, video y otros contenidos asociados a un determinado concepto.

Las paginas de esta web tienen una estructura fija generada a partir de plantillas, por lo que resulta fácil realizar el análisis mediante un programa en JFLEX. En concreto este ejercicio se descompone en varios apartados, cada uno de ellos independiente del otro y de complejidad creciente.

Para completar la descripción mediante un ejemplo, supongamos que estamos analizando la pagina correspondiente al piano, que se ha descargado de la dirección web https://commons.wikimedia.org/wiki/Piano, con el nombre de archivo piano.html

1. Obtener el número total de contenido multimedia de tipo imagen que hay en una pagina, (solo el numero de enlaces a contenido de tipo imagen).
2. Obtener el número total de contenido multimedia que hay en una pagina, para cada una de las tres clases , imágenes, audio y video. Nota, los ficheros de audio no siempre figuran directamente en la pagina, pero se pueden contabilizar.
3. Obtener los enlaces al contenido multimedia de tipo imagen. Atención, no se quieren las versiones reducidas (los "*thumb*") de las imágenes, sino los enlaces que permiten descargar la imagen en alta definición. Las extensiones que definen imágenes son `.jpg .jpeg .png .svg .gif` tanto en mayúscula como minúscula.
4. Obtener los enlaces al contenido multimedia de tipo video. Los ficheros de video tienen la extensión `.ogv`, en mayúsculas o minúsculas.
5. Obtener los enlaces de imágenes señaladas con algún tipo de marca como una estrellita  o una moneda  . Hay varios otros tipos de marcas. Cualquiera de ellas se considerara como "Destacada".

Se debe enviar un fichero `WikiCrawler.lex` que implementa el analizador léxico el cual se ejecuta mediante . Estos ficheros se compilaran y ejecutaran mediante las siguientes instrucciones:

- `jflex WikiCrawler.lex`
- `javac WikiCrawler.java`
- `java WikiCrawler <opción> <Fichero de entrada>`

Se proporciona el fichero `WikiCarwler.java` que no debe modificarse.
Para completar la descripción mediante ejemplos, supongamos que estamos analizando la pagina correspondiente al piano, que se ha descargado de la dirección web https://commons.wikimedia.org/wiki/Piano, con el nombre de archivo "piano.html".

- `java WikiCrawler -ni piano.html`
  `29  piano.html`

- `java WikiCrawler -nm piano.html`
  `29  5  1  piano.html`

- `java WikiCrawler -ei piano.html`
  `https://commons.wikimedia.org/wiki/File:Steinway_%26_Sons_upright..`
  `https://commons.wikimedia.org/wiki/File:Klavier_nah_offen.jpg`
  `https://commons.wikimedia.org/wiki/File:Piano_player.jpg`
  `https://commons.wikimedia.org/wiki/File:Schiedmayer_1851.jpg`
  `...`

- `java WikiCrawler -ev piano.html`
  `https://upload.wikimedia.org/wikipedia/commons/8/87/Steinway_`

- `java WikiCrawler -ed piano.html`
  `https://commons.wikimedia.org/wiki/File:Steinway_%26_Sons_upright..`
  `https://commons.wikimedia.org/wiki/File:Steinway_%26_Sons_concert_`
  `https://commons.wikimedia.org/wiki/File:Fortepian_-_schemat..`
  `https://commons.wikimedia.org/wiki/File:Fortepian_-_mechanizm_`
  `https://commons.wikimedia.org/wiki/File:DuplexScaling.jpg`
  `https://commons.wikimedia.org/wiki/File:Steinway_grand_piano_-_ped`

Hay ligeras diferencias de formato en otras paginas, como "guitarra.html" que habrá que tener en cuenta.

- `java WikiCrawler -ni guitarra.html`
  `73  piano.html`

- `java WikiCrawler -nm guitarra.html`
  `73  6  3  piano.html`

- `java WikiCrawler -ei guitarra.html`
  `https://commons.wikimedia.org/wiki/File:Guitar-1333353440ujZ.jpg`
  `https://commons.wikimedia.org/wiki/File:Classical_Guitar_two_views.`
  `https://commons.wikimedia.org/wiki/File:Classical_Guitar_labelled_g`
  `https://commons.wikimedia.org/wiki/File:Classical_Guitar_not_labell`
  `https://commons.wikimedia.org/wiki/File:Guitar_1.jpg`
  `https://commons.wikimedia.org/wiki/File:E-Gitarre.jpg`
  `...`

- `java WikiCrawler -nv guitarra.html`
  `https://upload.wikimedia.org/wikipedia/commons/a/a0/Bernd_Voss_-_Co`
  `https://upload.wikimedia.org/wikipedia/commons/f/f6/20091104_Sharon`
  `https://upload.wikimedia.org/wikipedia/commons/0/09/Alisa_Gladyseva`

- `java WikiCrawler -ed guitarra.html`

NOTA: Además de estos ficheros se probaran otros, sacados de la misma web, a fin de garantizar que la practica se ha hecho correctamente.