WikiCrawler.pdf

_david

Procesadores de Lenguajes

3º Grado en Ingeniería Informática

Escuela Técnica Superior de Ingeniería Informática
Universidad de Málaga

```
%%
%int
%xstate MEDIA, LINKS

/*
HAY DOS IMAGENES jpeg/JPEG en guitarra.html QUE NO SE CUENTAN
*/

FT_IMAGE                                = (png|PNG|jpg|JPG|GIF|gif|svg|SVG)
FT_AUDIO                                = (mp3|MP3|ogg|OGG)
FT_VIDEO                                = (ogv|OGV)
IMAGE                                   = (class\=\"image\")
AUDIO1                                  = (class\=\"unicode\ audiolink\")
AUDIO2                                  = (class\=\"mwPlayerContainer\
k-player\")
VIDEO                                   = (class\=\"PopUpMediaTransform\")
FAVS                                    =
(Featured\spictures|Commons\:Valued\simages)
LINK                                    = (href\=\"([^\"\s]*)\.)
LINK_I                                  = ({LINK}{FT_IMAGE}\")
LINK_A                                  = ({LINK}{FT_AUDIO}\")
LINK_V                                  = ({LINK}{FT_VIDEO}\")

%{
    public String tmp                   = new String("");
%}

%%
<YYINITIAL>{
    class\=\"toc\"                      { yybegin(MEDIA); }
    [^]                                 {}
}

<MEDIA>{
    {LINK_I}                            { WikiCrawler.nImg++; tmp =
yytext().substring(6, yytext().length()-1);
WikiCrawler.enlacesImagenes.add(tmp); }
    {IMAGE}                             { yybegin(LINKS); }
    {AUDIO1}                            { yybegin(LINKS); }
    {AUDIO2}                            { WikiCrawler.nAudio++; }
    {VIDEO}                             { yybegin(LINKS); }
    \<table[^]*\<\/table\>              {}
    [^]                                 {}
}

<LINKS>{
    {LINK_I}                            { WikiCrawler.nImg++; tmp =
yytext().substring(6, yytext().length()-1);
WikiCrawler.enlacesImagenes.add(tmp); yybegin(MEDIA); }
```

```
    {LINK_V}                            { WikiCrawler.nVideo++;
WikiCrawler.enlacesVideo.add(yytext().substring(6, yytext().length()-1));
yybegin(MEDIA); }
    {LINK_A}                            { WikiCrawler.nAudio++;
yybegin(MEDIA); }
    title\=\"{FAVS}\"                   {
if(!WikiCrawler.enlacesDestacados.contains(tmp))
WikiCrawler.enlacesDestacados.add(tmp); }
    [^]                                 {}
}
```