Sample Programs

The example programs included with the HTML Parser distribution are listed below, with some details.

Note: On unix systems if you used the Java jar command or some older unzip utility to extract the distribution zip file, the executable flag will not have been preserved on the files in the bin directory. You can fix this by issuing the following command:

```
chmod u+x bin/*
```

Parser

Parse a web page and print the tags in a simple loop.

org.htmlparser.Parser.main(String[] args)

```
bin/parser http://website_url [tag_name]
where tag_name is an optional tag name to be used as a filter, i.e.
    A - Show only the link tags extracted from the document
    IMG - Show only the image tags extracted from the document
    TITLE - Extract the title from the document
NOTE: this is also the default program for the htmlparser.jar, so the above could be:
java -jar lib/htmlparser.jar http://website_url [tag_name]
```

Lexer

Print the low level nodes of a web page.

org.htmlparser.lexer.Lexer

```
bin/lexer http://website_url
```

Filter Builder

Interactively generate source code to extract web site contents.

org.htmlparser.parserapplications.filterbuilder.FilterBuilder

```
bin/filterbuilder
```

Execute the FilterBuilder application online using Java Web Start:

org.htmlparser.parserapplications.filterbuilder.FilterBuilder

Link Extractor

Extract links/mail addresses from a web page.

org.htmlparser.parserapplications.LinkExtractor

```
bin/linkextractor http://website_url [-maillinks]
the optional -maillinks argument causes mailto: links to be printed
```

String Extractor

Extract text from a web page.

org.htmlparser.parserapplications.StringExtractor

```
bin/stringextractor http://website_url [-links]
the optional -links argument causes hyperlinks to be shown within the text
```

Site Capturer

Save a web site locally.

org.htmlparser.parserapplications.SiteCapturer

```
bin/sitecapturer http://source_website /target_directory/ [true|false]
the optional boolean argument determines whether resources such as images,
audio and video are to be captured
```

Execute the SiteCapturer application online using Java Web Start:

org.htmlparser.parserapplications.SiteCapturer

Thumbelina *View images behind thumbnails.*

org.htmlparser.lexerapplications.thumbelina.Thumbelina

```
bin/thumbelina [http://starting_website]
```

Execute the Thumbelina application online using Java Web Start:

org.htmlparser.lexerapplications.thumbelina.Thumbelina

BeanyBaby *Parser Java Bean demo.*

org.htmlparser.beans.BeanyBaby

```
bin/beanybaby [http://starting_website]
```

Translate *Numeric character reference and character entity reference to unicode codec.*

org.htmlparser.util.Translate

```
bin/translate [-encode] <input_file >output_file
```