txt2html/HTML::TextToHTML Sample Conversion

This sample is based hugely on the original sample.txt produced by Seth Golub for txt2html.

I used the following options to convert this document:

-titlefirst -mailmode -make_tables
--custom_heading_regexp '^ *--[\w\s]+-- *$'
--system_link_dict txt2html.dict
--append_body sample.foot --infile sample.txt --outfile sample.html

This has either been done at the command line with:

perl -MHTML::TextToHTML -e run_txt2html -- *options*

or using the script

txt2html *options*

or from a (test) perl script with:

```
use HTML::TextToHTML;
my $conv = new HTML::TextToHTML();
$conv->txt2html([*options*]);
```

==========

From bozo@clown.wustl.edu

Return-Path: <bozo@clown.wustl.edu>

Message-Id: <9405102200.AA04736@clown.wustl.edu>

Content-Length: 1070

From: bozo@clown.wustl.edu (Bozo the Clown)

To: kitty@example.com (Kathryn Andersen)

Subject: Re: HTML::TextToHTML

Date: Sun, 12 May 2002 10:01:10 -0500

Bozo wrote:

BtC> Can you post an example text file with its html'ed output?

BtC> That would provide a much better first glance at what it does

BtC> without having to look through and see what the perl code does.

Good idea. I'll write something up.

-=-

The header lines were kept separate because they looked like mail headers and I have mailmode on. The same thing applies to Bozo's quoted text. Mailmode doesn't screw things up very often, but since most people are usually converting non-mail, it's off by default.

Paragraphs are handled ok. In fact, this one is here just to

demonstrate that.

THIS LINE IS VERY IMPORTANT!

(Ok, it wasn't *that* important)

EXAMPLE HEADER

Since this is the first header noticed (all caps, underlined with an "="), it will be a level 1 header. It gets an anchor named "section_1".

Another example

This is the second type of header (not all caps, underlined with "="). It gets an anchor named "section_1_1".

Yet another example

This header was in the same style, so it was assigned the same header tag. Note the anchor names in the HTML. (You probably can't see them in your current document view.) Its anchor is named "section_1_2".

Get the picture?

-- This is a custom header --

You can define your own custom header patterns if you know what your documents look like.

Features of HTML::TextToHTML

==========

* Handles different kinds of lists

1. Bulleted

2. Numbered

- You can nest them as far as you want.
- It's pretty decent about figuring out which level of list it is supposed to be on.
- You don't need to change bullet markers to start a new list.

3. Lettered

A. Finally handles lettered lists

B. Upper and lower case both work

a) Here's an example

b) I've been meaning to add this for some time.

C. HTML without CSS can't specify how ordered lists should be indicated, so it will be a numbered list in most browsers.

4. Definition lists (see below)

* Doesn't screw up mail-ish things

* Spots preformated text

It just needs to have enough whitespace in the line. Surrounding blank lines aren't necessary. If it sees enough whitespace in a line, it preformats it. How much is enough? Set it yourself at command line if you want.

* You can append a file automatically to all converted files. This is handy for adding signatures to your documents.

* Deals with paragraphs decently.

Looks for short lines in the middle of paragraphs and keeps them short with the use of breaks (
). How short the lines need to be is configurable.

Unhyphenates split words that are in the middle of para-graphs. Let me know if trailing punctuation isn't handled "prop-erly". It should be.

One can also have multi-paragraph list items, like this one.

* Puts anchors at all headers and, if you're using the mail header features, at the beginning of each mail message. The anchor names

for headings are based on guessed section numbers.
- You can turn off this option too, if you don't like it.

* Groks Mosaic-style "formatted text" headers (like the one below)

* Can hyperlink things according to a dictionary file.
The sample dictionary handles URLs like http://www.aigeek.com/ and
<http://www.katspace.com/> and also shows how to do simpler
things such as linking the word txt2html the first time it appeared.
* One can also use the link-dictionary to define custom tags, for
example using the star character to indicate *italics*.

* Recognises and parses tables of different types:
o DELIM: A table determined by delimiters.
o ALIGN: No need for fancy delimiters, this figures out
a table by looking at the layout, the spacing of the cells.
o BORDER: has a nice border around the table
o PGSQL: the same format as Postgresql query results.

* Also with XHTML! Turn on the --xhtml option and it will ensure that
all paragraphs and list items have end-tags, all tags are in
lower-case, and the doctype is for XHTML.

Example of short lines

We're the knights of the round table
We dance whene'er we're able
We do routines and chorus scenes
With footwork impeccable.
We dine well here in Camelot
We eat ham and jam and spam a lot.

Example of varied formatting

If I want to *emphasize* something, then I'd use stars to wrap around the words, *even if there were more than one*, *that's* what I'd do. But I could also _underline_ words, so long as the darn thing was not a_variable_name, in which case I wouldn't want to lose the underscores in something which thought it was underlining. Though we might want to _underline more than one word_ in a sentence. Especially if it is _The Title Of A Book_.
For another kind of emphasis, let's go and #put something in bold#.
But it doesn't even need to be that simple. Something which is *really exciting* is coping with italics and similar things *spread across multiple lines*.

Example of Long Preformatting

(extract from Let It Rain by Kristen Hall)

I have given, I have given and got none
Still I'm driven by something I can't explain
It's not a cross, it is a choice
I cannot help but hear his voice
I only wish that I could listen without shame

Let it rain, let it rain, on me
Let it rain, oh let it rain,
Let it rain, on me

I have been a witness to the perfect crime
Wipe the grin off of my face to hide the pain
It isn't worth the tears you cry
To have a perfect alibi
Now I'm beaten at the hands of my own game

Let it rain, let it rain, on me
Let it rain, oh let it rain,
Let it rain, on me

Definition Lists

A definition list comprises the following:

Term:

The term part of a DL item is a word on a line by itself, ending with a colon.

Definition:

The definition part of a DL item is at least one paragraph following the term.

If one has more than one paragraph in the definition, the first line of the next paragraph needs to be indented two spaces from where the term starts, otherwise we don't know that it belongs to the definition.

Examples of Tables

ALIGN

~~~~~

Here is a simple ALIGN table:

-e File exists.

-z File has zero size.

-s File has nonzero size (returns size).

Here are some of the conditions of ALIGN tables:

#Context:# A table needs to be surrounded by blank lines.
~~~~~

#Length:# A table must contain at least two rows.

#Width:# A table must contain at least two columns.

#Spacing:# There needs to be at least two spaces between the columns, otherwise there might be some random paragraph which could have inter-word spacing that lined up by accident.

#Cell Size:# If you have more than one line (as just above) then you will simply get empty cells where the other column is empty.

#Alignment:# Alignment of cells is attempted to be preserved.

BORDER

This is a table with a border.

```
+---------+-----+
| Food    | Qty |
+---------+-----+
| Bread   |   1 |
| Milk    |   1 |
| Oranges |   3 |
| Apples  |   6 |
+---------+-----+
```

PGSQL

This is the same table like Postgresql would make it.

```
Food    | Qty
--------+-----
Bread   |   1
Milk    |   1
Oranges |   3
Apples  |   6
(4 rows)
```

DELIM

~~~~~

A delimited table needs to have its delimiters at the start and end, just to be sure that this is a table.

```
:Fred:Nurk:58:
:George:Washington:62:
:Mary:Quant:35:
```

And one can have almost any delimiter one wishes.

```
| Darcy, Fitzwilliam | hero    |
| Bennet, Elizabeth  | heroine |
| Wickham, George    | villain |
```
~~~~~

THINGS TO DO

There are some things which this module doesn't handle yet which I would like to implement.

A. I would like to be able to preserve lettered lists, that is:
a) recognise that they are letters and not numbers (which it already does)
b) display the correct OL properties with CSS so as to preserve that information.

The footer is everything from the end of this sentence to the </BODY> tag.