Downloading Text from the Internet

Here is a simple example of how to download and reformat text from the Internet.

Let's start by using `curl` to get some text data. We need the `-L` because this is a URL with a redirect built in.

In [21]:
```
!curl -L 'http://goo.gl/g3aE4' > tomsawyer.html
```

```
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100   336    0   336    0     0   5905      0 --:--:-- --:--:-- --:--:-- 22400
100  467k  100  467k    0     0   182k      0  0:00:02  0:00:02 --:--:--  255k
```

For single pages, `curl` is generally the best tool to use. For whole directory trees and mirroring, `wget` is what people usually use.

If we look at it, we got the page in HTML format.

In [22]:
```
!head tomsawyer.html
```

```
<HTML>
<HEAD>
<TITLE>The Adventures of Tom Sawyer </TITLE>
</HEAD>
<BODY BGCOLOR="#FFFFF2">
<CENTER><B>Twain, Mark, 1835-1910. The Adventures of Tom Sawyer </B>
<BR>
Electronic Text Center, University of Virginia Library</CENTER>
```

Now let's convert it to text format (we could also have used `lynx -dump URL` directly). There are several other tools for converting HTML to text; they may or may not work better.

In [23]:
```
!lynx -dump tomsawyer.html > tomsawyer.txt
```

The output is pretty good, but has some extra spaces at the beginning.

In [24]:
```
!head tomsawyer.txt
```

```
                Twain, Mark, 1835-1910. The Adventures of Tom Sawyer
               Electronic Text Center, University of Virginia Library

                       | [1]Table of Contents for this work |

               | [2]All on-line databases | [3]Etext Center Homepage |
   ____________________________________________________________________________

      About the electronic version
      The Adventures of Tom Sawyer
```

In [25]:
```
!sed '1000,1020!d' tomsawyer.txt
```

```
curtain of a second-story window. Was the sacred presence there? He
climbed the fence, threaded his stealthy way through the plants, till
he stood under that window; he looked up at it long, and with emotion;
then he laid him down on the ground under it, disposing himself upon
his back, with his hands clasped upon his breast and holding his poor
wilted flower. And thus he would

                    ---------------------------------------------------

                                      -43-

die -- out in the cold world, with no shelter over his homeless head,
no friendly hand to wipe the death-damps from his brow, no loving face
to bend pityingly over him when the great agony came. And thus she
would see him when she looked out upon the glad morning, and oh! would
she drop one little tear upon his poor, lifeless form, would she heave
one little sigh to see a bright young life so rudely blighted, so
untimely cut down?

   The window went up, a maid-servant's discordant voice profaned the
holy calm, and a deluge of water drenched the prone martyr's remains!
```

Let's fix these with sed.

In [26]:
```
!sed 's/^ *//' -i tomsawyer.txt
```

Also, there's a header at the beginning of the file (we can see that from the web page).

In [27]:
```
!sed '/About the print/,+10!d' tomsawyer.txt
```

```
About the print version
The Adventures of Tom Sawyer
Mark Twain
Harper and Brothers
New York and London
1903

Author's National Edition: The Writings of Mark Twain, Vol. XII

Spell-check not performed due to presence of dialect.
Published: 1876
```

In [28]:
```
!sed '1,/About the print/d' -i tomsawyer.txt
```

In [29]:
```
!head tomsawyer.txt
```

```
The Adventures of Tom Sawyer
Mark Twain
Harper and Brothers
New York and London
1903

Author's National Edition: The Writings of Mark Twain, Vol. XII

Spell-check not performed due to presence of dialect.
Published: 1876
```

Other Converters

If you don't like that kind of editing, there are a bunch of other converters you can use. To find them, use `apt-cache search` or `synaptic` on Ubuntu or Debian. Other Linux distributions have their own search tools. You can also check Freecode (AKA Freshmeat)

In [33]:
```
!apt-cache search html text converter
# install with "sudo apt-get install ..." if needed
```

```
html2text - advanced HTML to text converter
poppler-utils - PDF utilities (based on Poppler)
xmlto - XML-to-any converter
wap-wml-tools - Wireless Markup Language development and test tools
highlight - Universal source code to formatted text converter
highlight-common - source code to formatted text converter (architecture
independent files)
html2ps - HTML to PostScript converter
khmerconverter - converts between legacy Khmer encodings and Unicode
libghc-pandoc-dev - general markup converter
libghc-pandoc-doc - general markup converter
libghc-pandoc-prof - general markup converter
libhighlight-perl - perl bindings for highlight source code to formatted text
converter
mira-assembler - Whole Genome Shotgun and EST Sequence Assembler
pandoc - general markup converter
php-text-wiki - transforms Wiki and BBCode markup into XHTML, LaTeX or plain
text markup
pod2pdf - Plain Old Documentation to Portable Document Format converter
python-pdfminer - PDF parser and analyser
python-zope.app.renderer - Zope 3 Text Renderer Framework
src2tex - A converter from source program files to TeX format files
stx2any - Converter from structured plain text to other formats
t2html - text to HTML converter implemented in Perl
txt2html - Text to HTML converter
uni2ascii - UTF-8 to 7-bit ASCII and vice versa converter
unrtf - RTF to other formats converter
vilistextum - a HTML to text converter
wp2x - WordPerfect 5.x documents to whatever converter
yodl - Your Own Document Language (Yodl) is a pre-document language
yodl-doc - Documenation for Your Own Document Language (Yodl)
```

In [31]:
```
!html2text tomsawyer.html | sed '1000,1020!d'
```

```
his speckled straw hat. He now looked exceedingly improved and uncomfortable.
He was fully as uncomfortable as he looked; for there was a restraint about
whole clothes and cleanliness that galled him. He hoped that Mary would forget
his shoes, but the hope was blighted; she coated them thoroughly with tallow,
as was the

================================================================================

                                      -48-


custom, and brought them out. He lost his temper and said he was always being
made to do everything he didn't want to do. But Mary said, persuasively:
   "Please, Tom -- that's a good boy."
   So he got into the shoes snarling. Mary was soon ready, and the three
children set out for Sunday-school -- a place that Tom hated with his whole
heart; but Sid and Mary were fond of it.
   Sabbath-school hours were from nine to half-past ten; and then church
service. Two of the children always remained for the sermon voluntarily, and
the other always remained too -- for stronger reasons. The church's high-
backed, uncushioned pews would seat about three hundred persons; the edifice
```

In [32]:
```
!vilistextum tomsawyer.html - | sed '1000,1020!d'
```

```
pityingly over him when the great agony came. And thus she would see him
when she looked out upon the glad morning, and oh! would she drop one little
tear upon his poor, lifeless form, would she heave one little sigh to see a
bright young life so rudely blighted, so untimely cut down?

   The window went up, a maid-servant's discordant voice profaned the holy
calm, and a deluge of water drenched the prone martyr's remains!

   The strangling hero sprang up with a relieving snort. There was a whiz as
of a missile in the air, mingled with the murmur of a curse, a sound as of
shivering glass followed, and a small, vague form went over the fence and
shot away in the gloom.

   Not long after, as Tom, all undressed for bed, was surveying his drenched
garments by the light of a tallow dip, Sid woke up; but if he had any dim
idea of making any "references to allusions," he thought better of it and
held his peace, for there was danger in Tom's eye.

   Tom turned in without the added vexation of prayers, and Sid made mental
note of the omission.
```