#60 Tracking BBC News with lynx

As I mentioned earlier, one of the unsung heroes of the command-line Internet is unquestionably the `lynx` web browser (or its newer sibling `links`). Although you can use it to surf the Web if you dislike graphics, its real power is accessed on the command line itself, within a shell script.

The `-dump` flag, for example, produces the text but not the HTML source, as shown in the following when checking the BBC World Service website, tracking technology news:

```
$ url=http://news.bbc.co.uk/2/low/technology/default.stm
$ lynx -dump $url | head

   [1]Skip to main content
   BBC NEWS / TECHNOLOGY
   [2]Graphics version | [3]Change to UK Edition | [4]BBC Sport Home
     ____________________________________________________________

   [5]News Front Page | [6]Africa | [7]Americas | [8]Asia-Pacific |
   [9]Europe | [10]Middle East | [11]South Asia | [12]UK | [13]Business |
   [14]Health | [15]Science/Nature | [16]Technology | [17]Entertainment |
   [18]Have Your Say
```

This output is not very interesting, but it's easily fed to `grep` or any other command-line utility, because it's just a text stream at this juncture. Now we can easily check a website to see if there are any stories about a favorite news topic, computer company, or group of people. Let's see if there's any news about games, with a one-line context shown around each match, by using `grep`:

```
$ lynx -dump $url | grep -C1 -i games

   [21]Screenshot from Vice City [22]Britons' love affair with games
   Britain is turning into a nation of keen gamers, research by the UK
   games industry trade body suggests.

--
   line-up
   Many of the Nintendo games for the Christmas run-up return to familiar
   characters and brand names.
```

```
--
   Virtual pets fed by photos and pronunciation puzzles are just some of
   the mobile phone games popular in Japan.

   [28]Next gen consoles spark concern
   The next generation of consoles could shake up the games industry,
   with smaller firms going bust, say experts.
--

   [37]Text msgs play games with TV
   Your TV and mobile are coming closer together, with game shows played
--
   [38]Mobile gaming 'set to explode'
   Consumers will be spending millions of pounds to play games on their
   mobiles by next year, say experts.
```

The numbers in brackets are URL references listed later in the output, so to identify the [37] link, the page needs to be requested again, this time having `grep` find the associated link URL:

```
$ lynx -dump $url | grep '37\.'
  37. http://news.bbc.co.uk/2/low/technology/3182641.stm
```

Switch to `-source` rather than `-dump`, and the output of `lynx` becomes considerably more interesting.

```
$ lynx -source $url | grep -i 'PublicationDate'
<meta name="OriginalPublicationDate" content="2003/08/29 15:01:14" />
```

The `-source` flag produces the HTML source of the page specified. Pipe that source into a `grep` or two, and you can extract just about any information from a page, even information within a tag or comment. The `bbcnews` script that follows lets you easily scrape the top technology stories from the Beeb at any time.

The Code

```
#!/bin/sh

# bbcnews - Reports the top stories on the BBC World Service.

url="http://news.bbc.co.uk/2/low/technology/default.stm"

lynx -source $url | \
  sed -n '/Last Updated:/,/newssearch.bbc.co.uk/p' | \
  sed 's/</\
</g;s/>/>\
/g' | \
  grep -v -E '(<|>)' | \
```

```
  fmt | \
  uniq
```

How It Works

Although this is a short script, it is rather densely packed. These scraper scripts are best built iteratively, looking for patterns to filter in the structure of the web page information and then tuned line by line to produce just the output desired.

On the BBC website, this process is surprisingly easy because we're already looking at the low-bandwidth version of the site. The first task is to discard any HTML associated with the navigational menus, bottom material, and so forth, so that we just have the core content of the page, the stories themselves. That's what the first `sed` does — it reduces the data stream by preserving only the headline and body of the new stories between the "Last Updated" string at the top of the page and the `newssearch.bbc.co.uk/p` search box at the bottom of the page.

The next invocation of `sed` is uglier than the first, simply because it's doing something peculiar:

```
  sed 's/</\
</g;s/>/>\
/g'
```

Every time it finds an open angle bracket (`<`), it's replacing it with a carriage return followed by an open angle bracket. Close angle brackets (`>`) are replaced by a close angle bracket followed by a carriage return. If `sed` supported an `\n` notation to specify carriage returns, the second `sed` invocation would not need to be written across three lines and would read much more easily, as follows:

```
sed 's/</\n</g;s/>/>\n/g'
```

Once the added carriage returns put all the HTML tags on their own lines, the second invocation of `grep` strips out all the tags (`-v` inverts the logic of the `grep`, showing all lines that do *not* match the pattern, and the `-E` flag specifies that the argument is a complex regular expression), and the result is fed to `fmt` to wrap the resultant text lines better. Finally, the `uniq` command is used to ensure that there aren't multiple blank lines in the output: It removes all nonunique lines from the data stream.

Running the Script

This script has no arguments, and as long as the BBC hasn't changed its basic low-source page layout, it'll produce a text-only version of the top technology headlines. The first version of the `bbcnews` script was written around a layout that changed during the summer of 2003: The BBC originally had all its articles wrapped in `<div>` tags but has since changed it. Fortunately, the update to the script involved only about ten minutes of work.

The Results

Here's the top technology news at the end of August 2003:

```
$ bbcnews | head -20
Last Updated:  Friday, 29 August, 2003, 15:01 GMT 16:01 UK

                      Youth suspected of net attack

                      An American youth is suspected by the FBI
                      of being one of the authors of the crippling
                      MSBlast internet worm, say reports.

                      Britons' love affair with games

                      Britain is turning into a nation of keen
                      gamers, research by the UK games industry
                      trade body suggests.

                      Familiar faces in Nintendo's line-up

                      Many of the Nintendo games for the Christmas
                      run-up return to familiar characters and
                      brand names.
```

Hacking the Script

With a little more tuning, you could easily have the top technology story from the BBC News pop up each time you log in to your account. You could also email the results to your mailbox via a `cron` job every so often, if you wanted:

```
bbcnews | mail -s "BBC Technology News" peter
```

Don't send it to a list, though; there are some copyright and intellectual property issues to consider if you begin republishing Internet content owned by other people. There's a fine line between fair use and violation of copyright, so be thoughtful about what you do with content from another website.