National University of Singapore
School of Computing

CS2281 **Programming Assignment 6** Semester 3 04/05

Deadline

Fri Jun 4 17:00:00 GMT-8 2004

Learning Keywords

shell programming, awk, sed, regular expression

Your Task

In this assignment, you are asked to analyze log files from SoC's web server and display (i) the list of web pages that are accessed via Google, and (ii) the search terms in Google that lead to these web pages.

Background

Web Access Log Files

Today's log file from SoC web servers can be found at:

```
/var/log/httpd-logs/www.comp.nus.edu.sg/access_log
```

The log files from the last 4 days can be found in the same directory, under the name of `access_log.[0-3].gz`.

The log files contain information about HTTP requests to SoC's web server, including where the requests come from, the size of each requested objects etc. In this assignment, we are interested in two particular information, the URLs requested, and the referer's URLs.

The following line shows an example of an entry from the web access log.

```
209.84.81.97 - - [26/May/2004:03:26:56] "GET /~ahkow/ HTTP/1.1" 200 0 "http://www.foo.com/page.htm" "Mozilla/4.0" -
```

This example shows that 209.84.81.97 reached the web page of `/~ahkow/` by clicking on a link from `http://www.foo.com/page.htm`.

Google's URL

If you look through the current access log, you will most likely see some entries where the referer's URL comes from Google. For example, the following entry shows that somebody accessed `/~cs2281/` through Google.

```
... "GET /~cs2281/ HTTP/1.1" ... "http://www.google.com/search?hl=en&lr=&ie=UTF-8&q=2281+UNIX&btnG=Search" ...
```

From the Google URL, we can find out the search terms that has lead to the requested page. The search term is assumed to appear between "&q=" and the next "&". In the example above, the search term is `2281 UNIX`. (Note that a plain '+' in the search term represent a space character.).

URL Encoding and Decoding

Since URLs may contain special characters (such as spaces, quotes etc.), the URL standard specifies that such special characters should be encoded using their ASCII values in hexadecimal. For instance, the space character, whose decimal ASCII value is 32, is encoded as %20. Another example is '~', which is encoded as %7E. Hence, the URL `http://www/~ahkow/Kow'sStory.html` becomes `http://www/%7Eahkow/Kow%27sStory.html` after we apply URL encoding. Such encoding is used to encode international characters (such as chinese characters) as well.

What You Should Do

You should write a shell program (with the help of awk and sed) called `gg.sh`, that accepts the name of the web access log file and an *optional* home directory as arguments. `gg.sh` should analyze the access log, and prints all request URLs under the given home directory that are accessed via Google, and all Google search terms that lead to the URL.

If no home directory is given, print out all URLs that are accessed via Google and their respective search terms, in sorted order of the requested URL.

Duplicated search terms should be printed only once.

Any special ASCII characters in encoded URLs should be decoded in your output. You may ignore encoded characters whose values are larger than 127. A C program that decodes URLs have been written for you and can be found on the CS2281 web site.

Examples

```
$ gg.sh access_log cs2281
/~cs2281/archive/N234/w5.ps
    errno 254 string
    solaris error reading from pipe errno=4
/~cs2281/man/grep.html
    grep and options+examples
    /usr/xpg4/bin/grep
/~cs2281/man/xargs.html
    xargs replace-str
$
$ gg.sh access_log
/~abhik/
    abhik roychoudhury
    a. roychoudhury
```

```
/~abhik/pdf/ppdp99.pdf
    S. Etalle, M. Gabrielli and M.C. Meo. Unfold/Fold transformation of CCP programs
/~abhik/SVV03/
    software verification techniques
    SVV03
      :
      :
      :
/~zhuyi/resume.html
    resume
/~zhuyx/usoc04.pdf
    systemc uml
```

Additional Tips

- Google's URL is not limited to www.google.com. You will find referers URL from www.google.co.uk, www.google.com.sg, www.google.ca etc.
- You may ignore URL requests that comes from Google's image search.
- A copy of access_log that generates the above examples is available on CS2281's website for testing. You should also run your program on recent web access logs to check for robustness and correctness.
- Since web server access logs are normally huge in size, you should pay attention to the efficiency of your program. You can time your program using the `time` command.

Documentation

There are many possible ways to solve this assignment. You are required to document your approach clearly as comments in your script (it will be graded). If you use any regular expression, you should explain what each regular expression is suppose to match.

Submission Requirement

You are required to submit the encrypted version of your shell script `gg.sh`. Make sure you have read the submission instruction document posted on CS2281 website. For this assignment, create a subdirectory under `$HOME/CS2281_LABs/` called `a6` and put your encrypted files under the subdirectory. You must include your name as a comment in the *second* line of your files (Since first line is sha-bang). I will access your submission through the pathname `$HOME/CS2281_LABs/a6/gg.sh.pgp`. It is your responsibility to make sure that the filenames are correct and permissions are set properly according to the instructions given.