

comp.lang.perl.modules: Is there a module to spider Google usenet archives yet?

Is there a module to spider Google usenet archives yet?

Source: <http://coding.derkeiler.com/Archive/Perl/comp.lang.perl.modules/2004-02/0159.html>

From: ~greg (g_m_at_(remove_to_reply)comcast.net)

Date: 02/12/04

Date: Wed, 11 Feb 2004 23:47:47 -0500

Hello.

Is there a module to spider Google usenet archives yet?

I believe that subject line question has been asked before, and answered in the negative, which is why I say "yet".

I just want to get to all the posts to a particular usenet group, in their original usenet formats. For personal use only. And respecting all robot rules –if possible ;) In any case, there's no hurry. Just a few hindered per day with plenty of delays between requests is fine with me, if that's how they (–the Google team) prefers it be done.

I routinely write hacks to harvest sets of pages. However, my knowledge of true spidering is very limited.

The only way I can think of doing it manually is like this:

Starting from Google Advanced Groups Search, set the number of messages to 100 and sort by date.

Name the newsgroup.

Then a reasonable date limits.
(The date limits will be incremented in overlapping steps, and the redundant links returned will later be eliminated by script.).

The first returned pages seems to list

Is there a module to spider Google usenet archives yet?

comp.lang.perl.modules: Is there a module to spider Google usenet archives yet?

just thread-starting posts.
(And even to get all of them you need to follow the "Next # threads" links.)

Clicking on the thread-starting posts links has different consequences, depending on whether there is one or more than one post in the thread.

If there is more than one post in the thread, then the link leads to a frames set, with a tree-structure on the left listing the posts in the thread, and one particular post, in html, on the right.

>From the html of the post you next have to click "View This Article Only".

And then, finally, you click on the "Original Format" link to get to the usenet format, – which is the objective.

The question is, –how can all that be automated? Or is there a better way? Or is there a module (yet) that does it?

(Incidentally, I was one of the –probably many –who wrote to Google requesting that they provide access to the original usenet format, at a time when they didn't. They eventually wrote back to me: "Thanks for the suggestion! We really appreciate thoughtful feedback from our users, and we'll keep it in mind as we grow and evolve. Regards, The Google Team". And the access was provided very shortly after that!)

~Greg.