

Re: Read strings from one file and search for them in a directory containing htm files

Source: <http://unix.derkeiler.com/Newsgroups/comp.unix.shell/2005-11/1423.html>

From: Meghavvarnam (*meghsatish_at_yahoo.com*)

Date: 11/28/05

Date: 27 Nov 2005 23:37:09 -0800

Ed Morton wrote:

> *Meghavvarnam wrote:*

>

> <snip>

> > *Sample data does help a great deal. Here it is:*

> >

> > *allStrings.txt contains lines likes these –*

> > ===== *Begin allStrings.txt* =====

> > *WPA1*

> > *WPA2*

> > *Automatic (WPA2 or WPA1)*

> > *XyZ technology helps make home networking simple.*

> > *XyZ architecture offers network connectivity between personal*

> > *computers, printers, intelligent appliances and wireless devices.*

> > *XyZ architecture leverages ABC/DE and the Web to enable seamless*

> > *proximity networking in addition to control and data transfer among*

> > *networked devices in the home and office.*

> > *If you enable XYZ, then XYZ-enabled devices can print to this device.*

> > *Privacy*

> > *SampleText:
 Simpler, smarter online supplies ordering*

> > *Learn more about
XYZ SampleText*

> > *Transfer printer information to XYZ SampleText?*

> > ===== *End allStrings.txt* =====

> >

> > *Which means, the script will search for these lines in .htm files. Each*

> > *of these lines need to appear as is (case sensitive) to say that there*

> > *is a match. Now consider we read the 3rd line in the file above –*

> > *Automatic (WPA2 or WPA1).*

> >

> > > *From the .htm snippet pasted below, the third option tag contains the*

> > *search string –*

> >

> > *Automatic (WPA2 or WPA1)*

> >

> > *So when a match like this occurs, I simply need to write Automatic*

> > *(WPA2 or WPA1) in the files – usedStrings.*

comp.unix.shell: Re: Read strings from one file and search for them in a directory containing htm files

```
> if (index($0,">"string"<") {
> usedStrings[string]++
> delete strings[string] # for efficiency
> }
> }
> END { for (string in usedStrings)
> print string
> }' allStrings.txt directory/*.htm > usedStrings.txt
>
> Note that, since you said something in a previous posting about only
> wanting to look for text when it's part of an HTML tag (or something
> like that...) the search for ">"string"<" surrounds the line from
> "allStrings.txt" with ">" and "<" so it only matches when the text
> appears between those 2 characters. If you don't want that restriction,
> just get rid of the ">" and "<". Similairly for the grep solution.
>
```

This is the script that I tried –

```
# listused
# lists strings that are used in all .htm files

gawk 'NR==FNR{strings[$0]++;next} {
    for (string in strings) #}
print string
    if (index($0,">"string"<") || index($0,"\"string\""))
|| index($0,">"string"\n")) {
        usedStrings[string]++
        delete strings[string] # for efficiency
    }
}
END {
    for (string in usedStrings)
        print string
}' allStrings.txt htm/*.htm > usedStringsfile
```

Please let me know, if there is any mistake in this. I gave execute permission for the file that contained this script and ran it from the shell.

usedStringsfile was empty at the end of it.

Any pointers will be of great help.

```
> If you'd like the awk script to tell you which strings are/aren't used,
> that's trivial, e.g.:
>
> gawk 'NR==FNR{strings[$0]++;next}
> { for (string in strings)
> if (index($0,">"string"<") {
> usedStrings[string]++
> delete strings[string] # for efficiency
```

Re: Read strings from one file and search for them in a directory containing htm files

comp.unix.shell: Re: Read strings from one file and search for them in a directory containing htm files

```
> }
> }
> END {
> print "Used Strings:"
> for (string in usedStrings)
> printf "\t%s\n",string
> print "Unused Strings:"
> for (string in strings)
> printf "\t%s\n",string
> }' allStrings.txt directory/*.htm
>
>
```

I modified the script above to remove all parse errors. Here is the script that I used to try out –

```
gawk ' NR==FNR { strings[$0]++;next }
{ for (string1 in strings)
  string = sprintf("<%s>", string1)
  if (index($0,">"string"<")) {
    usedStrings[string]++
    delete strings[string] # for efficiency
  }
}
END {
  print "Used Strings:"
  for (string in usedStrings)
    printf "\t%s\n", string
  print "Unused Strings:"
  for (string in strings)
    printf "\t%s\n", string
}' allStrings.txt htm/*.htm
```

I see the same behaviour with this as with the earlier script. Would we need a different approach for this thing at all ??

What does the line – NR==FNR{strings[\$0]++;next} do.

Thank you in advance so much for your help.

Megh

```
> If there can be newlines in the strings you're trying to match in the
> HTML files, then we need to figure out what "match" means since there
> aren't newlines in the strings in "allStrings.txt" and we need to figure
> out a different record separator than a newline char.
>
> Ed.
```