Subject: Removing (or replacing) substrings in a string array
Posted by on Wed, 22 Jan 2014 14:13:08 GMT
View Forum Message <> Reply to Message

Say I have a string array where each string may or may not begin with a dot (".") and may or may not end with a dot. What is an efficient (non-loop?) way of removing those dots in IDL 7?

The reason I specify IDL 7 is that I realize this could be done with strsplit in IDL 8 but in IDL 7 straplit does not support string arrays.

Alternatively, if I could make it so those dots are not there in the first place, that's even better. What I'm trying to do it to parse an array of strings, where the fields are separated by dots but I cannot identify the substrings by their position. What I can do is to identify them with regular expressions because they all have different forms. And stregex supports string arrays also in IDL 7.

So I can do something like strmid(stregex(strlist,'\.[0-9]{5}\.',/extr),1,5) if I know I have a five-digit number surrounded by dots. I remove the dots with the strmid command.

But any field can also be first or last, so to find it I need to do stregex(strlist,'(\.|^)[0-9]{5}(\.|$)',/extr). But then one of the dots will be missing in those cases when the field really is first or last, so the simple strmid operation does not work.

And some fields can have variable lengths but be identified by an initial character, so I could find it with something like stregex(strlist,'(\.|^)f[+-][0-9]+(\.|$)',/extr). In this case strmid does not work even when there is a dot at the end, because I don't know its position.

Would be great if I could tell stregex to return everything but the dots (if present) but I don't know if that is possible. Can't you do something like that with regular expressions in the shell? Or was it in elisp?

I wanted to educate myself and went looking for “Making Regular Expressions Your Friends” by Mike Galloy but the links seem to be dead in http://michaelgalloy.com/2006/06/11/regular-expressions.html and the link in http://www.exelisvis.com/docs/Learning_About_Regular_E.html leads to a page telling me that "The mgunit project site has moved to Github" and when I follow the provided link to github I can't find the document.

Hmmm... I guess if I substitute '.'+strlist+'.' for strlist in all stregex calls I don't have to worry about the case where there are no dots. But that seems not so elegant and it still does not solve the problem with unknown lengths.

Subject: Re: Removing (or replacing) substrings in a string array

Posted by Matthew Argall on Wed, 22 Jan 2014 15:19:14 GMT
View Forum Message <> Reply to Message

I used the following site to learn about regular expressions. It is a bit wordy, but gets the job done.
http://www.regular-expressions.info/tutorial.html

```
;Strings
myStr1 = '.aldfa09741_{}+=!@#$%^&*(.'
myStr2 = 'aldfa09741_{}+=!@#$%^&*(.'
myStr3 = '.aldfa09741_{}+=!@#$%^&*('

;Stregex
regex1 = stregex(myStr1, '^\.?([^.]*)\.?$', /SUBEXP, /EXTRACT)
regex2 = stregex(myStr2, '^\.?([^.]*)\.?$', /SUBEXP, /EXTRACT)
regex3 = stregex(myStr3, '^\.?([^.]*)\.?$', /SUBEXP, /EXTRACT)

;Print
print, regex1[1]
print, regex2[1]
print, regex3[1]
```

'^\.?'   -- look for an optional (?) dot (\.) at the beginning of the string (^)
'([^.]*)' -- look for any character except the dot ([^.]) any number of times (*) and extract it ()
'\.?$'   -- look for an optional (?) dot (\.) at the end of the string ($)

Subject: Re: Removing (or replacing) substrings in a string array
Posted by Matthew Argall on Wed, 22 Jan 2014 15:33:12 GMT
View Forum Message <> Reply to Message

also, to test your regular expressions, I recommend the "Regex Tester App" for Google Chrome
https://chrome.google.com/webstore/detail/regexp-tester-app/cmmblmkfaijaadfjapjddbeaoffeccib?utm_source=chrome-ntp-icon

It comes in handy.

Subject: Re: Removing (or replacing) substrings in a string array
Posted by on Wed, 22 Jan 2014 15:34:26 GMT
View Forum Message <> Reply to Message

Den onsdagen den 22:e januari 2014 kl. 16:19:14 UTC+1 skrev Matthew Argall:
> I used the following site to learn about regular expressions. It is a bit wordy, but gets the job done.
> http://www.regular-expressions.info/tutorial.html

Thanks, I'll have a look at it.

```
>  ;Strings
>  myStr1 = '.aldfa09741_{}+=!@#$%^&*(.'
>  myStr2 = 'aldfa09741_{}+=!@#$%^&*(.'
>  myStr3 = '.aldfa09741_{}+=!@#$%^&*('
>
>  ;Stregex
>
>  regex1 = stregex(myStr1, '^\.?([^.]*)\.?$', /SUBEXP, /EXTRACT)
>  regex2 = stregex(myStr2, '^\.?([^.]*)\.?$', /SUBEXP, /EXTRACT)
>  regex3 = stregex(myStr3, '^\.?([^.]*)\.?$', /SUBEXP, /EXTRACT)
>
>  ;Print
>
>  print, regex1[1]
>  print, regex2[1]
>  print, regex3[1]
>
>  '^\.?'    -- look for an optional (?) dot (\.) at the beginning of the string (^)
>  '([^.]*)' -- look for any character except the dot ([^.]) any number of times (*) and extract it ()
>  '\.?$'    -- look for an optional (?) dot (\.) at the end of the string ($)
```

Maybe that is part of the solution. I hadn't realized you can use the subexp that way. But it fails when there are more fields. The dots are (in general) the separators between multiple fields.

```
IDL> mystr4= 'gag.aldfa09741_{}+=!@#$%^&*(.sdf'
IDL> regex4 = stregex(myStr4, '^\.?([^.]*)\.?$', /SUBEXP, /EXTRACT)
```

Then regex4 is an array of two empty strings.

Subject: Re: Removing (or replacing) substrings in a string array
Posted by Fabzi on Wed, 22 Jan 2014 15:57:50 GMT
View Forum Message <> Reply to Message

... if I might sneak into this post: is there a general solution for the problem:
"Removing (or replacing) all occurences of a substring of any length in a string array" ?

This sounds like such a trivial and usefull operation that I wonder why it is not native in the IDL language...

I rely on this solution which I found on the newsgroup:

https://groups.google.com/d/msg/comp.lang.idl-pvwave/4eq9yOP NTJ8/mDltqt72pogJ

There is also the Astro solution:

http://www.astro.washington.edu/docs/idl/cgi-bin/getpro/libr ary27.html?STRREPLACE

But it would be nice to have a unified and fast solution, no?

Is there one already?

Thanks!

Subject: Re: Removing (or replacing) substrings in a string array
Posted by on Wed, 22 Jan 2014 15:58:18 GMT
View Forum Message <> Reply to Message

Den onsdagen den 22:e januari 2014 kl. 16:34:26 UTC+1 skrev Mats Löfdahl:
>
> Maybe that is part of the solution.

Indeed it is!

In my own example I had these two calls:

```
stregex(strlist,'(\.|^)[0-9]{5}(\.|$)',/extr)
stregex(strlist,'(\.|^)f[+-][0-9]+(\.|$)',/extr)
```

If I rewrite them like this

```
stregex(strlist,'(\.|^)([0-9]{5})(\.|$)',/extr,/subexp)
stregex(strlist,'(\.|^)(f[+-][0-9]+)(\.|$)',/extr,/subexp)
```

I get 2D string arrays. And I can get a list of the field I'm interested in by doing

```
(stregex(strlist,'(\.|^)([0-9]{5})(\.|$)',/extr,/subexp))[2, *]
(stregex(strlist,'(\.|^)(f[+-][0-9]+)(\.|$)',/extr,/subexp)) [2,*]
```

Thank you Matthew!

Subject: Re: Removing (or replacing) substrings in a string array
Posted by wlandsman on Wed, 22 Jan 2014 18:14:38 GMT
View Forum Message <> Reply to Message

A very simple and fast (but not general) method to do this is to convert the string to a byte array and remove all appearance of the byte character for a period. This is what the routine remchar does
http://idlastro.gsfc.nasa.gov/ftp/pro/misc/remchar.pro

On Wednesday, January 22, 2014 9:13:08 AM UTC-5, Mats Löfdahl wrote:
> Say I have a string array where each string may or may not begin with a dot (".") and may or may not end with a dot. What is an efficient (non-loop?) way of removing those dots in IDL 7?
>
>
>
> The reason I specify IDL 7 is that I realize this could be done with strsplit in IDL 8 but in IDL 7 straplit does not support string arrays.
>
>
>
>
>
> Alternatively, if I could make it so those dots are not there in the first place, that's even better. What I'm trying to do it to parse an array of strings, where the fields are separated by dots but I cannot identify the substrings by their position. What I can do is to identify them with regular expressions because they all have different forms. And stregex supports string arrays also in IDL 7.
>
>
>
> So I can do something like strmid(stregex(strlist,'\.[0-9]{5}\.',/extr),1,5) if I know I have a five-digit number surrounded by dots. I remove the dots with the strmid command.
>
>
>
> But any field can also be first or last, so to find it I need to do stregex(strlist,'(\.|^)[0-9]{5}(\.|$)',/extr). But then one of the dots will be missing in those cases when the field really is first or last, so the simple strmid operation does not work.
>
>
>
> And some fields can have variable lengths but be identified by an initial character, so I could find it with something like stregex(strlist,'(\.|^)f[+-][0-9]+(\.|$)',/extr). In this case strmid does not work even when there is a dot at the end, because I don't know its position.
>
>
>
> Would be great if I could tell stregex to return everything but the dots (if present) but I don't know if that is possible. Can't you do something like that with regular expressions in the shell? Or was it in elisp?
>
>
>
>
>
> I wanted to educate myself and went looking for “Making Regular Expressions Your

Friends” by Mike Galloy but the links seem to be dead in http://michaelgalloy.com/2006/06/11/regular-expressions.html and the link in http://www.exelisvis.com/docs/Learning_About_Regular_E.html leads to a page telling me that "The mgunit project site has moved to Github" and when I follow the provided link to github I can't find the document.
>
>
>
>
>
> Hmmm... I guess if I substitute '.'+strlist+'.' for strlist in all stregex calls I don't have to worry about the case where there are no dots. But that seems not so elegant and it still does not solve the problem with unknown lengths.

Subject: Re: Removing (or replacing) substrings in a string array
Posted by jimuba on Wed, 22 Jan 2014 21:12:54 GMT
View Forum Message <> Reply to Message

FWIW, here is an approach that combines the strings in the array, adding in a separator character that doesn't appear in the original string (in this example a '\'), and then splitting up the joined string at one or more periods or at the added separator character:

```
s = STRJOIN(aStringArr, '\')
result = STRSPLIT(s, '\.+|\\', /REGEX, /EXTRACT)
```

I hope this helps,
Jim (Exelis VIS)

Subject: Re: Removing (or replacing) substrings in a string array
Posted by on Wed, 22 Jan 2014 21:43:37 GMT
View Forum Message <> Reply to Message

Den onsdagen den 22:e januari 2014 kl. 16:58:18 UTC+1 skrev Mats Löfdahl:
>
> Thank you Matthew!

And thank you for completely answering my first question. I almost forgot that I had asked it when I had written about what I was really trying to do. :o)

Subject: Re: Removing (or replacing) substrings in a string array
Posted by Matthew Argall on Wed, 22 Jan 2014 22:36:22 GMT
View Forum Message <> Reply to Message

>> Thank you Matthew!

>
> And thank you for completely answering my first question. I almost forgot that I had asked it when I had written about what I was really trying to do. :o)

My pleasure! Glad you were able to solve it.

Subject: Re: Removing (or replacing) substrings in a string array
Posted by Michael Galloy on Wed, 22 Jan 2014 22:57:46 GMT
View Forum Message <> Reply to Message

On 1/22/14, 7:13 AM, Mats Löfdahl wrote:
> Say I have a string array where each string may or may not begin with
> a dot (".") and may or may not end with a dot. What is an efficient
> (non-loop?) way of removing those dots in IDL 7?
>
> The reason I specify IDL 7 is that I realize this could be done with
> strsplit in IDL 8 but in IDL 7 straplit does not support string
> arrays.
>
>
> Alternatively, if I could make it so those dots are not there in the
> first place, that's even better. What I'm trying to do it to parse an
> array of strings, where the fields are separated by dots but I cannot
> identify the substrings by their position. What I can do is to
> identify them with regular expressions because they all have
> different forms. And stregex supports string arrays also in IDL 7.
>
> So I can do something like
> strmid(stregex(strlist,'\.[0-9]{5}\.',/extr),1,5) if I know I have a
> five-digit number surrounded by dots. I remove the dots with the
> strmid command.
>
> But any field can also be first or last, so to find it I need to do
> stregex(strlist,'(\.|^)[0-9]{5}(\.|$)',/extr). But then one of the
> dots will be missing in those cases when the field really is first or
> last, so the simple strmid operation does not work.
>
> And some fields can have variable lengths but be identified by an
> initial character, so I could find it with something like
> stregex(strlist,'(\.|^)f[+-][0-9]+(\.|$)',/extr). In this case strmid
> does not work even when there is a dot at the end, because I don't
> know its position.
>
> Would be great if I could tell stregex to return everything but the
> dots (if present) but I don't know if that is possible. Can't you do
> something like that with regular expressions in the shell? Or was it
> in elisp?

>
>
> I wanted to educate myself and went looking for “Making Regular
> Expressions Your Friends” by Mike Galloy but the links seem to be
> dead in http://michaelgalloy.com/2006/06/11/regular-expressions.html
> and the link in
> http://www.exelisvis.com/docs/Learning_About_Regular_E.html leads to
> a page telling me that "The mgunit project site has moved to Github"
> and when I follow the provided link to github I can't find the
> document.
>
>
> Hmmm... I guess if I substitute '.'+strlist+'.' for strlist in all
> stregex calls I don't have to worry about the case where there are no
> dots. But that seems not so elegant and it still does not solve the
> problem with unknown lengths.
>

Check out MG_STREPLACE:

https://github.com/mgalloy/mglib/blob/master/src/strings/mg_ streplace.pro

I have that "Making Regular Expressions Your Friends" article here somewhere too. I will update the link on my website and post here when I find it. Have to run now...

Mike
--
Michael Galloy
www.michaelgalloy.com
Modern IDL: A Guide to IDL Programming (http://modernidl.idldev.com)
Research Mathematician
Tech-X Corporation

Subject: Re: Removing (or replacing) substrings in a string array
Posted by Michael Galloy on Thu, 23 Jan 2014 02:50:05 GMT
View Forum Message <> Reply to Message

On 1/22/14, 3:57 pm, Michael Galloy wrote:
> On 1/22/14, 7:13 AM, Mats Löfdahl wrote:
>> Say I have a string array where each string may or may not begin with
>> a dot (".") and may or may not end with a dot. What is an efficient
>> (non-loop?) way of removing those dots in IDL 7?
>>
>> The reason I specify IDL 7 is that I realize this could be done with
>> strsplit in IDL 8 but in IDL 7 straplit does not support string
>> arrays.

>>
>>
>> Alternatively, if I could make it so those dots are not there in the
>> first place, that's even better. What I'm trying to do it to parse an
>> array of strings, where the fields are separated by dots but I cannot
>> identify the substrings by their position. What I can do is to
>> identify them with regular expressions because they all have
>> different forms. And stregex supports string arrays also in IDL 7.
>>
>> So I can do something like
>> strmid(stregex(strlist,'\.[0-9]{5}\.',/extr),1,5) if I know I have a
>> five-digit number surrounded by dots. I remove the dots with the
>> strmid command.
>>
>> But any field can also be first or last, so to find it I need to do
>> stregex(strlist,'(\.|^)[0-9]{5}(\.|$)',/extr). But then one of the
>> dots will be missing in those cases when the field really is first or
>> last, so the simple strmid operation does not work.
>>
>> And some fields can have variable lengths but be identified by an
>> initial character, so I could find it with something like
>> stregex(strlist,'(\.|^)f[+-][0-9]+(\.|$)',/extr). In this case strmid
>> does not work even when there is a dot at the end, because I don't
>> know its position.
>>
>> Would be great if I could tell stregex to return everything but the
>> dots (if present) but I don't know if that is possible. Can't you do
>> something like that with regular expressions in the shell? Or was it
>> in elisp?
>>
>>
>> I wanted to educate myself and went looking for “Making Regular
>> Expressions Your Friends” by Mike Galloy but the links seem to be
>> dead in http://michaelgalloy.com/2006/06/11/regular-expressions.html
>> and the link in
>> http://www.exelisvis.com/docs/Learning_About_Regular_E.html leads to
>> a page telling me that "The mgunit project site has moved to Github"
>> and when I follow the provided link to github I can't find the
>> document.
>>
>>
>> Hmmm... I guess if I substitute '.'+strlist+'.' for strlist in all
>> stregex calls I don't have to worry about the case where there are no
>> dots. But that seems not so elegant and it still does not solve the
>> problem with unknown lengths.
>>
>
> Check out MG_STREPLACE:

>
>
> https://github.com/mgalloy/mglib/blob/master/src/strings/mg_ streplace.pro
>
> I have that "Making Regular Expressions Your Friends" article here
> somewhere too. I will update the link on my website and post here when I
> find it. Have to run now...
>
> Mike

I updated the article at:

   http://michaelgalloy.com/2006/06/11/regular-expressions.html

with working links to the article and the example code. I would use
MG_STREPLACE from my Github repo over the STR_REPLACE linked to in the
article.

Mike
--
Michael Galloy
www.michaelgalloy.com
Modern IDL: A Guide to IDL Programming (http://modernidl.idldev.com)
Research Mathematician
Tech-X Corporation
