

Re: best method to perform operations on word lists

Source: <http://coding.derkeiler.com/Archive/Perl/comp.lang.perl.misc/2006-06/msg00811.html>

- *From:* "Francois Massion" <massion@xxxxxx>
 - *Date:* 12 Jun 2006 07:21:39 -0700
-

I have come up with something which seems to work (partially). For my current purpose it'll do the trick but any suggestion for optimization is welcome:

```
(...)  
chomp(my $prev = <TERMS>);  
  
my @reducedlist = $prev;  
  
while ( <TERMS> ) {  
  chomp;  
  push @reducedlist, $_ unless ( /^$prev/ && length($_) - length($prev) < 3 )  
  ; # I can set the maximum length of a suffix here  
  $prev = $_;  
}  
print "$_\n" for @reducedlist;
```

[Amazing to see how much time people can invest in a few lines of code when they are no professionals ;-)!]

Francois

Bart Van der Donck schrieb:

Francois Massion wrote:

[...]

```
#!/perl  
use strict; use warnings;  
my $list =  
"überzeugt  
überzeugt,  
überzogen  
überzogen,  
überzogen.
```

Re: best method to perform operations on word lists

```
üblich
übliche
üblichen
üblicherweise";
my @terms = split /\n/, $list;
my $prev = 'nonesuch584685542256RANOM58544';
```

This didn't modify the list.

I didn't mean to modify \$list; the new content is in @terms. If you want \$list to contain the new words, you can use something like this at the end of the program.

```
$list = join "\n", @terms;
```

Maybe the reason is the \$prev definition.

\$prev has no direct importance here, it's only required that it should not be present in @terms, because it is used to delete double entries from @terms.

```
s/(\.|,|e|en|e,|en,|e\.|en\.)$// for @terms;
```

I also tried Dr. Ruud's regex but it would have to be rewritten for each language.

That is correct, hence my thoughts about language files. My code is a very brute algorithm – it only strips out the following from the end of each line:

```
. , e en e en, e. en.
```

If you are planning to use this for different languages, you would obviously need to modify those patterns each time.

—
Bart