Regular Expression in SAS

Dajun Tian

Presented at GAUSS 2018 Fall

Realistic Problem

- Anonymize the person's identifiers in the data.

Obs	message
1	My phone is 314-xxx-xxxx.
2	cell is (314) 111-3456
3	mobile number: (322) 899-1234

Obs	message
1	My phone is 314-xxx-xxxx.
2	cell is [redacted]
3	mobile number: [redacted]

- In reality, the identifiers could be other forms
  - Date: 0000-00-00
  - SSN: 000-00-0000
  - Email: abc@def.gh
  - Etc.

Simplistic Solution To Anonymize Data

- The dataset was named as
  `find_phone_number`
- And the variable was named as
  `message`

Obs	message
1	My phone is 314-xxx-xxxx.
2	cell is (314) 111-3456
3	mobile number: (322) 899-1234

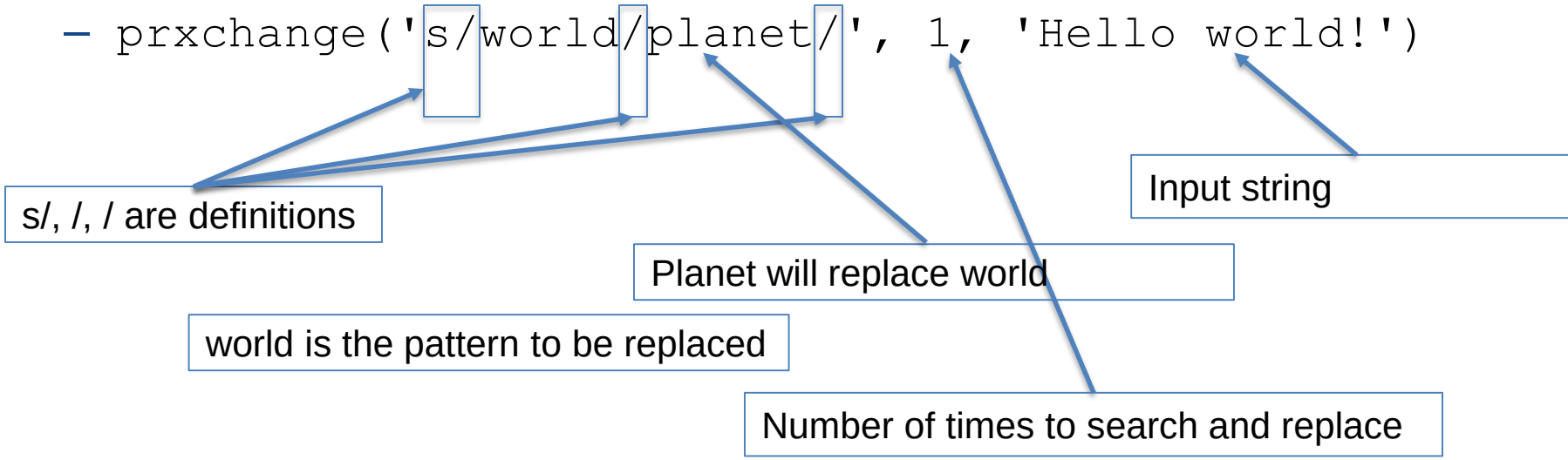
- Our task is to replace phone number with xxx

```
data check_number;
    set find_phone_number;
    message2 = prxchange("s/\(\d{3}\) ?\d{3}-\d{4}/xxx/", 1, message);
run;
```

Obs	message	message2
1	My phone is 314-xxx-xxxx.	My phone is 314-xxx-xxxx.
2	cell is (314) 111-3456	cell is xxx
3	mobile number: (322) 899-1234	mobile number: xxx

Find and replace in SAS

- Basic Syntax for Searching and Replacing Text
  - general form is `s/regular-expression/replacement-string/`
  - `prxchange('s/world/planet/', 1, 'Hello world!')`



  - The function returns "`Hello planet!`".

Build regular expression

- To extend the power of regular expression, we refer to Regular Expression (PRX) Metacharacters
- What is metachracter? A metacharacter is a character that has a special meaning to a regular expression. For example, * means the preceding character could occur 0 or more times.
- SAS observes these metacharacters:
  - { } [ ] ( ) ^ $. | * + ? \
  - To match metacharacter, use \ to override

Build Regular Expression: greedy/lazy repetition

Define Pattern	Description	Example
*	matches preceding zero or more times	zo* matches z, zo, zoo and so on
+	matches preceding one or more times	zo+ matches zo and zoo zo+ does not match “z”
?	matches preceding zero or one time	do(es)? matches do and dose only.
{n,}	matches a pattern at least n times	zo{1,} is equivalent to zo+ zo{0,} is equivalent to zo*
{n,m}	matches n-m times.	zo{0,1} is equivalent to zo?

Lazy Repetition is implemented by adding ?. For example,
zo*? will match a pattern by as few characters as possible.

Build pattern using metachracters

Define Pattern	Description
\d	matches a digit character
\D	matches a non-digit character
\w	matches any word character or alphanumeric character, including the underscore.
\s	matches any whitespace character, including space, tab, form feed, and so on, and is equivalent to [\f\n\r\t\v].
\S	matches any character that is not a whitespace character and is equivalent to [^\f\n\r\t\v].

Build Regular Expression: Groupings

Metacharacter	Description and example
[…]	matches any one of the enclosed characters: • [abc] matches the “a” in “apple”
[^…]	matches any character that is not encloded • [^abc] matches the “p” in “apple”
[[:alpha:]]	matches an alphabetic character.
[[:alnum:]]	matches an alphanumeric character.
[[:^alpha:]]	matches a nonalphabetic character.
[[:^alnum:]]	matches a non-alphanumeric character.
[[:space:]]	matches a space.
[[:^space:]]	matches a non-space character

Summary

- Use `prxchange` to search and replace
  - `s/pattern-to-match/pattern-to-replace/`
- How to use SAS documentation for RE
  - https://regexr.com
  - http://support.sas.com/documentation/cdl/en/lefunctionsref/63354/HTML/default/viewer.htm#p1vz3ljudbd756n19502acxazevk.htm
  - http://documentation.sas.com/?docsetId=lefunctionsref&docsetTarget=p0s9ilagexmjl8n1u7e1t1jfnzlk.htm&docsetVersion=9.4&locale=en

Reference

- http://documentation.sas.com/?docsetId=lefunctionsref&docsetTarget=p0s9ilagexmjl8n1u7e1t1jfnzlk.htm&docsetVersion=9.4&locale=en

- https://algs4.cs.princeton.edu/lectures/54RegularExpressions.pdf

Thank you!
Any questions are welcome.