Regular Expression Examples in R

Eamonn

19 November, 2016

```
  # example of regular expression use

      d <- c("0011070009_CFFP", "0011070001-M1_XY1", "0011070001-M2_XY1",
             "0011070002-M1_XY1", "0011070002-M2_XY1", "0011070003-M1_XY1",
             "0011070003-M2_XY1", "0011070005_NPC", "0011070013_CPPS1", "0011070017_CPPS2")

      # parsing a variable in which characters - and _ appear

      (tmp <- gsub("(.*)\\_.*", "\\1", d )) # remove all after first occurance of _
```

```
 [1] "0011070009"    "0011070001-M1" "0011070001-M2" "0011070002-M1"
 [5] "0011070002-M2" "0011070003-M1" "0011070003-M2" "0011070005"
 [9] "0011070013"    "0011070017"
```

```
      gsub("(.*)\\-.*", "\\1", tmp)         # remove all after first occurance of -
```

```
 [1] "0011070009" "0011070001" "0011070001" "0011070002" "0011070002"
 [6] "0011070003" "0011070003" "0011070005" "0011070013" "0011070017"
```

```
      gsub(".*\\-","",d)                    # remove before and including -
```

```
 [1] "0011070009_CFFP"  "M1_XY1"            "M2_XY1"
 [4] "M1_XY1"           "M2_XY1"            "M1_XY1"
 [7] "M2_XY1"           "0011070005_NPC"    "0011070013_CPPS1"
[10] "0011070017_CPPS2"
```

```
      gsub("\\-.*","", d)                   # remove all text after -
```

```
 [1] "0011070009_CFFP"  "0011070001"        "0011070001"
 [4] "0011070002"       "0011070002"        "0011070003"
 [7] "0011070003"       "0011070005_NPC"    "0011070013_CPPS1"
[10] "0011070017_CPPS2"
```

```
      gsub("^.*7", "", d)                   # remove all text before and including 7
```

```
 [1] "0009_CFFP"    "0001-M1_XY1" "0001-M2_XY1" "0002-M1_XY1"
 [5] "0002-M2_XY1" "0003-M1_XY1" "0003-M2_XY1" "0005_NPC"
 [9] "0013_CPPS1"   "_CPPS2"
```

```
      gsub('.{2}$', '', d)                  # remove last two characters
```

```
 [1] "0011070009_CF"   "0011070001-M1_X" "0011070001-M2_X"
 [4] "0011070002-M1_X" "0011070002-M2_X" "0011070003-M1_X"
 [7] "0011070003-M2_X" "0011070005_N"    "0011070013_CPP"
[10] "0011070017_CPP"
```

```r
sub('.*(?=.{2}$)', '', d , perl=T)    # extract last two characters
```

```
 [1] "FP" "Y1" "Y1" "Y1" "Y1" "Y1" "Y1" "PC" "S1" "S2"
```

```r
sub('.*(?=.{1}$)', '', d , perl=T)    # extract last character
```

```
 [1] "P" "1" "1" "1" "1" "1" "1" "C" "1" "2"
```

```r
sub("[_][^_]*", "", d)                # remove underscore
```

```
 [1] "0011070009"    "0011070001-M1" "0011070001-M2" "0011070002-M1"
 [5] "0011070002-M2" "0011070003-M1" "0011070003-M2" "0011070005"
 [9] "0011070013"    "0011070017"
```

```r
gsub("001107000", "", d)              # remove these characters
```

```
 [1] "9_CFFP"            "1-M1_XY1"          "1-M2_XY1"
 [4] "2-M1_XY1"          "2-M2_XY1"          "3-M1_XY1"
 [7] "3-M2_XY1"          "5_NPC"             "0011070013_CPPS1"
[10] "0011070017_CPPS2"
```

```r
d[grepl("CFFP", d) ]                  # pull out using grep
```

```
[1] "0011070009_CFFP"
```

```r
# select characters before first occurance of 3
gsub("(.*?)(3.*)", "\\1",d)
```

```
 [1] "0011070009_CFFP"    "0011070001-M1_XY1" "0011070001-M2_XY1"
 [4] "0011070002-M1_XY1" "0011070002-M2_XY1" "001107000"
 [7] "001107000"          "0011070005_NPC"    "001107001"
[10] "0011070017_CPPS2"
```

```r
sub('_([^_]*.)$', '',  d )            # remove after and including _
```

```
 [1] "0011070009"    "0011070001-M1" "0011070001-M2" "0011070002-M1"
 [5] "0011070002-M2" "0011070003-M1" "0011070003-M2" "0011070005"
 [9] "0011070013"    "0011070017"
```

```r
x <- gsub("001107000","0070 10 0", d) # introduce spaces
gsub("[\\ \\ ]", "",                  # select all text between " " and " "
     regmatches(x,
                gregexpr("\\ .*?\\ ", x)))
```

```
 [1] "10"             "10"             "10"             "10"
 [5] "10"             "10"             "10"             "10"
 [9] "character(0)" "character(0)"
```

```r
gsub("(.*?)( .*)", "\\1", x)          # select chars before 1st space
```

```
 [1] "0070"             "0070"             "0070"            
 [4] "0070"             "0070"             "0070"            
 [7] "0070"             "0070"             "0011070013_CPPS1"
[10] "0011070017_CPPS2"
```

```r
gsub("(.*?)( .*)", "\\2", x)          # select chars after 1st space
```

```
 [1] " 10 09_CFFP"      " 10 01-M1_XY1"    " 10 01-M2_XY1"   
 [4] " 10 02-M1_XY1"    " 10 02-M2_XY1"    " 10 03-M1_XY1"   
 [7] " 10 03-M2_XY1"    " 10 05_NPC"       "0011070013_CPPS1"
[10] "0011070017_CPPS2"
```

```r
(x <- gsub("^.*?M","", d))             # extract characters after M
```

```
 [1] "0011070009_CFFP"  "1_XY1"            "2_XY1"           
 [4] "1_XY1"            "2_XY1"            "1_XY1"           
 [7] "2_XY1"            "0011070005_NPC"   "0011070013_CPPS1"
[10] "0011070017_CPPS2"
```

```r
 gsub("(.*?)(P.*)", "\\1", x)         # then select all before first occurance of P
```

```
 [1] "0011070009_CFF" "1_XY1"          "2_XY1"         
 [4] "1_XY1"          "2_XY1"          "1_XY1"         
 [7] "2_XY1"          "0011070005_N"   "0011070013_C"  
[10] "0011070017_C"  
```

```r
# Extract all before first occurance of M and 000
gsub("(.*?)(M.*)", "\\1", d )
```

```
 [1] "0011070009_CFFP"  "0011070001-"      "0011070001-"     
 [4] "0011070002-"      "0011070002-"      "0011070003-"     
 [7] "0011070003-"      "0011070005_NPC"   "0011070013_CPPS1"
[10] "0011070017_CPPS2"
```

```r
gsub("(.*?)(000.*)", "\\1", d )
```

```
 [1] "001107"           "001107"           "001107"          
 [4] "001107"           "001107"           "001107"          
 [7] "001107"           "001107"           "0011070013_CPPS1"
[10] "0011070017_CPPS2"
```

```r
x <- gsub("001107000","0070 1000", d) # introduce a space
gsub( " .*$", "", x )                 # remove everything after the occurance of the blank space
```

```
 [1] "0070"             "0070"             "0070"            
 [4] "0070"             "0070"             "0070"            
 [7] "0070"             "0070"             "0011070013_CPPS1"
[10] "0011070017_CPPS2"
```

```r
gsub("[[:space:]]", "", x)             # remove space(s?)
```

```
 [1] "007010009_CFFP"   "007010001-M1_XY1" "007010001-M2_XY1"
 [4] "007010002-M1_XY1" "007010002-M2_XY1" "007010003-M1_XY1"
 [7] "007010003-M2_XY1" "007010005_NPC"    "0011070013_CPPS1"
[10] "0011070017_CPPS2"
```

```r
# using grep and ifelse to create variables

ifelse(grepl("M1",  d ),1,
       ifelse(grepl("PC",  d ), 0, 2))
```

```
 [1] 2 1 2 1 2 1 2 0 2 2
```

```r
ifelse(grepl("NPC",  d),"NPC",
       ifelse(grepl("CFFP",  d), "CFFP", "Clinical"))
```

```
 [1] "CFFP"     "Clinical" "Clinical" "Clinical" "Clinical" "Clinical"
 [7] "Clinical" "NPC"      "Clinical" "Clinical"
```

```r
x <- gsub("001107000","0070 1000", d) # introduce a space
stringr::word(x, 1)                    # extract first word
```

```
 [1] "0070"             "0070"             "0070"
 [4] "0070"             "0070"             "0070"
 [7] "0070"             "0070"             "0011070013_CPPS1"
[10] "0011070017_CPPS2"
```

```r
# replacing values in variable
d<- as.character(d)
plyr::mapvalues(d, from = c("0011070013_CPPS1", "0011070001-M2_XY1"),
                to = c("HEY", "WHAT"))
```

```
 [1] "0011070009_CFFP"   "0011070001-M1_XY1" "WHAT"
 [4] "0011070002-M1_XY1" "0011070002-M2_XY1" "0011070003-M1_XY1"
 [7] "0011070003-M2_XY1" "0011070005_NPC"    "HEY"
[10] "0011070017_CPPS2"
```

```r
# extract all strings in alphanumeric variable
# http://stackoverflow.com/questions/17215789/extract-a-substring-in-r-according-to-a-pattern
gsub("[0-9]", "", d)
```

```
 [1] "_CFFP" "-M_XY" "-M_XY" "-M_XY" "-M_XY" "-M_XY" "-M_XY" "_NPC"
 [9] "_CPPS" "_CPPS"
```

CONCLUSION

REFERENCES

COMPUTING ENVIRONMENT

```
R version 3.2.2 (2015-08-14)
Platform: x86_64-w64-mingw32/x64 (64-bit)
Running under: Windows 8 x64 (build 9200)

locale:
[1] LC_COLLATE=English_United Kingdom.1252
[2] LC_CTYPE=English_United Kingdom.1252
[3] LC_MONETARY=English_United Kingdom.1252
[4] LC_NUMERIC=C
[5] LC_TIME=English_United Kingdom.1252

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods
[7] base

other attached packages:
[1] stringr_1.1.0 plyr_1.8.4    knitr_1.15

loaded via a namespace (and not attached):
 [1] magrittr_1.5    assertthat_0.1  tools_3.2.2     htmltools_0.3.5
 [5] yaml_2.1.14     tibble_1.2      Rcpp_0.12.8     stringi_1.1.2
 [9] rmarkdown_1.1   digest_0.6.10   evaluate_0.10

[1] "C:/Users\\User\\Documents\\GIT\\Regular-expressions"
```

This took 0.46 seconds to execute.