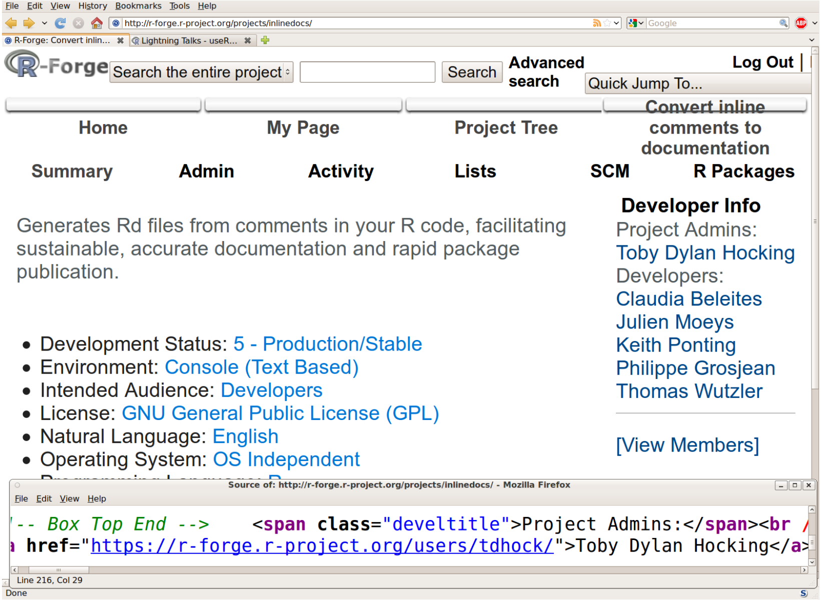
Fast, named capture regular expressions in R 2.14

Toby Dylan Hocking
toby.hocking@inria.fr
`http://cbio.ensmp.fr/~thocking/`

16 August 2011

Example: extract developer and project data from HTML


File Edit View History Bookmarks Tools Help
http://r-forge.r-project.org/projects/inlinedocs/
Google
R-Forge: Convert inlin...
Lightning Talks - useR...
R-Forge
Search the entire project
Search
Advanced search
Log Out
Quick Jump To...
Home
My Page
Project Tree
Convert inline comments to documentation
Summary
Admin
Activity
Lists
SCM
R Packages
Source of: http://r-forge.r-project.org/projects/inlinedocs/ - Mozilla Firefox
File Edit View Help
Line 216, Col 29
Done

Generates Rd files from comments in your R code, facilitating sustainable, accurate documentation and rapid package publication.

- Development Status: 5 - Production/Stable
- Environment: Console (Text Based)
- Intended Audience: Developers
- License: GNU General Public License (GPL)
- Natural Language: English
- Operating System: OS Independent
- Programming Language: R

Developer Info

Project Admins:
Toby Dylan Hocking
Developers:
Claudia Beleites
Julien Moeys
Keith Ponting
Philippe Grosjean
Thomas Wutzler

[View Members]

```
'-- Box Top End -->     <span class="develtitle">Project Admins:</span><br /
a href="https://r-forge.r-project.org/users/tdhock/">Toby Dylan Hocking</a>
```

How to extract user ids and names from HTML?

Data:

```
<a href="https://r-forge.r-project.org/users/tdhock/">
Toby Dylan Hocking </a>
<br /><span class="develtitle">Developers:</span><br />
<a href="https://r-forge.r-project.org/users/kmpont/">
Keith Ponting </a><br />
...
```

Want: table of extracted information.

id	name
tdhock	Toby Dylan Hocking
kmpont	Keith Ponting
...	

Solution: extract data using capturing regular expressions

```
<a href="https://r-forge.r-project.org/users/tdhock/">
Toby Dylan Hocking </a>
```

Capturing regular expression:

```
<a href="https://r-forge.r-project.org/users/([^/]+)/">
([^<]+) </a>
```

Named capture regular expression:

```
<a href="https://r-forge.r-project.org/users/(?<id>[^/]+)/'
(?<name>[^<]+) </a>
```

	R 2.13	R 2.13	R 2.14
	gregexpr()	str_match_all	gregexpr()
whole match	✓	✓	✓
capture		✓	✓
fast C code	✓		✓
named capture			✓

Introduction: regular expressions in R 2.13 give you the position and length of the entire match, not groups!

```
> u <- "http://r-forge.r-project.org/projects/inlinedocs"
> html <- paste(readLines(u),collapse="\n")
> pattern <-
+   paste(’<a href="https://r-forge.r-project.org/users/’,
+         ’([ˆ/]+)/">’, # capture group for user id
+         ’([ˆ<]+)’,    # capture group for user name
+         ’</a>’,sep="")
> gregexpr(pattern,html)[[1]]

[1] 14241 14372 14455 14531 14608 14693
attr(,"match.length")
[1] 76 77 70 71 79 77

> named.p <-
+   paste(’<a href="https://r-forge.r-project.org/users/’,
+         ’(?<id>[ˆ/]+)/">’, # named capture group
+         ’(?<name>[ˆ<]+)’,  # named capture group
+         ’</a>’,sep="")
```

Perl-Compatible Regular Expressions in R 2.14

```
> gregexpr(pattern,html,perl=TRUE)[[1]]

[1] 14241 14372 14455 14531 14608 14693
attr(,"match.length")
[1] 76 77 70 71 79 77
attr(,"capture.start")

[1,] 14286 14295
[2,] 14417 14429
[3,] 14500 14509
[4,] 14576 14585
[5,] 14653 14666
[6,] 14738 14752
attr(,"capture.length")

[1,]  6 18
[2,]  9 16
[3,]  6 12
[4,]  6 13
```

Capture names can be used to identify groups

```
> gregexpr(named.p,html,perl=TRUE)[[1]]

[1] 14241 14372 14455 14531 14608 14693
attr(,"match.length")
[1] 76 77 70 71 79 77
attr(,"capture.start")
        id  name
[1,] 14286 14295
[2,] 14417 14429
[3,] 14500 14509
[4,] 14576 14585
[5,] 14653 14666
[6,] 14738 14752
attr(,"capture.length")
     id name
[1,]  6   18
[2,]  9   16
[3,]  6   12
[4,]  6   13
```

stringr::str_match_all extracts groups using R code

```
> str_match_all(html,pattern)[[1]]

     [,1]
[1,] "<a href=\"https://r-forge.r-project.org/users/tdhock/
[2,] "<a href=\"https://r-forge.r-project.org/users/cbeleit
[3,] "<a href=\"https://r-forge.r-project.org/users/jmoeys/
[4,] "<a href=\"https://r-forge.r-project.org/users/kmpont/
[5,] "<a href=\"https://r-forge.r-project.org/users/phgrosj
[6,] "<a href=\"https://r-forge.r-project.org/users/tomasch
     [,2]          [,3]
[1,] "tdhock"      "Toby Dylan Hocking"
[2,] "cbeleites"   "Claudia Beleites"
[3,] "jmoeys"      "Julien Moeys"
[4,] "kmpont"      "Keith Ponting"
[5,] "phgrosjean"  "Philippe Grosjean"
[6,] "tomaschwutz" "Thomas Wutzler"
```

A function based on the new C code in R 2.14

```
> str_match_all_perl(html,
+   named.p)[[1]]

[1,] "<a href=\"https://r-forge.r-project.org/users/tdhock/
[2,] "<a href=\"https://r-forge.r-project.org/users/cbeleit
[3,] "<a href=\"https://r-forge.r-project.org/users/jmoeys/
[4,] "<a href=\"https://r-forge.r-project.org/users/kmpont/
[5,] "<a href=\"https://r-forge.r-project.org/users/phgrosj
[6,] "<a href=\"https://r-forge.r-project.org/users/tomasch
     id             name
[1,] "tdhock"       "Toby Dylan Hocking"
[2,] "cbeleites"    "Claudia Beleites"
[3,] "jmoeys"       "Julien Moeys"
[4,] "kmpont"       "Keith Ponting"
[5,] "phgrosjean"   "Philippe Grosjean"
[6,] "tomaschwutz" "Thomas Wutzler"
```

The new group parsing in C is 10x faster!

```
> system.time(replicate(1000,{
+   str_match_all(html,pattern)
+ }))

   user  system elapsed
  6.290   0.020   6.315
> system.time(replicate(1000,{
+   str_match_all_perl(html,pattern)
+ }))

   user  system elapsed
  0.460   0.010   0.472
```

New group extraction is 10x faster than existing methods for extracting the first substring!

Text to extract:

```
<a href="https://r-forge.r-project.org/users/tdhock/">
Toby Dylan Hocking</a>
</ul>Registered:&nbsp;2009-07-29 14:37
```

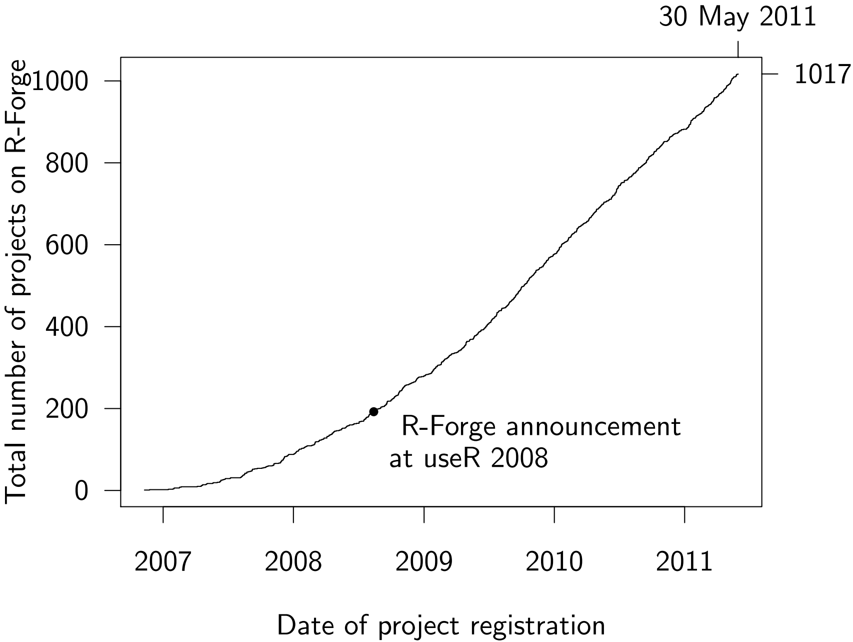
```
>
time.method("users/","[^/]+")
```

```
              seconds result
stringr         3.252 tdhock
gsub            0.761 tdhock
lookbehind      0.806 tdhock
R.2.14          0.078 tdhock
```

```
>
time.method("Registered:&nbsp;","[^<]+")
```

```
              seconds                  result
stringr         3.312 2009-07-29 14:37\t\t
gsub            0.802 2009-07-29 14:37\t\t
lookbehind      0.726 2009-07-29 14:37\t\t
R.2.14          0.072 2009-07-29 14:37\t\t
```

Efficient algorithms crucial for processing more data



Extracted developer and project data shows collaboration frequency in R-Forge projects

Project	Developers
ctv	25
rmetrics	22
phyloc	22
phylobase	16
phylohelper	13
mlr	12
genabel	12
yuima	11
rsiena	11
flr	10
distr	10
blotter	10
sedar	9
diseasemapping	9
.	.
.	.

Developers	Number of projects
25	1
22	2
16	1
13	1
12	2
11	2
10	3
9	2
8	4
7	7
6	20
5	34
4	54
3	114
2	254
1	513

Use regular expressions for fast and easy text processing!

Example to match:

```
<a href="https://r-forge.r-project.org/users/tdhock/">
Toby Dylan Hocking </a>
```

Named capture regular expression:

```
<a href="https://r-forge.r-project.org/users/(?<id>[^/]+)/"
(?<name>[^<]+) </a>
```

Available R functions:

	R 2.13	R 2.13	R 2.14
	gregexpr()	str_match_all()	gregexpr()
whole match	✓	✓	✓
groups		✓	✓
fast C code	✓		✓
named groups			✓

Conclusion: faster, easier text processing in R 2.14

- Before the 2.14 release, you can download and compile `ftp://ftp.stat.math.ethz.ch/Software/R/R-devel.tar.gz` to get access to the new `gregexpr()`.
- After: `str_match_all_perl()` function in the `stringr` package?
- Slides and Sweave source available on my web page: `http://cbio.ensmp.fr/~thocking/`
- Questions? Contact me directly: toby.hocking@inria.fr