Web Scraping with R

Xiao Nan @road2stat

6th China R Beijing

Outline

- Overview
- Toolkit
- Exception Handling
- Parallelization
- Outro

Part I

Overview

Two Types of Scrapers / Crawlers

The REAL ones and the ...

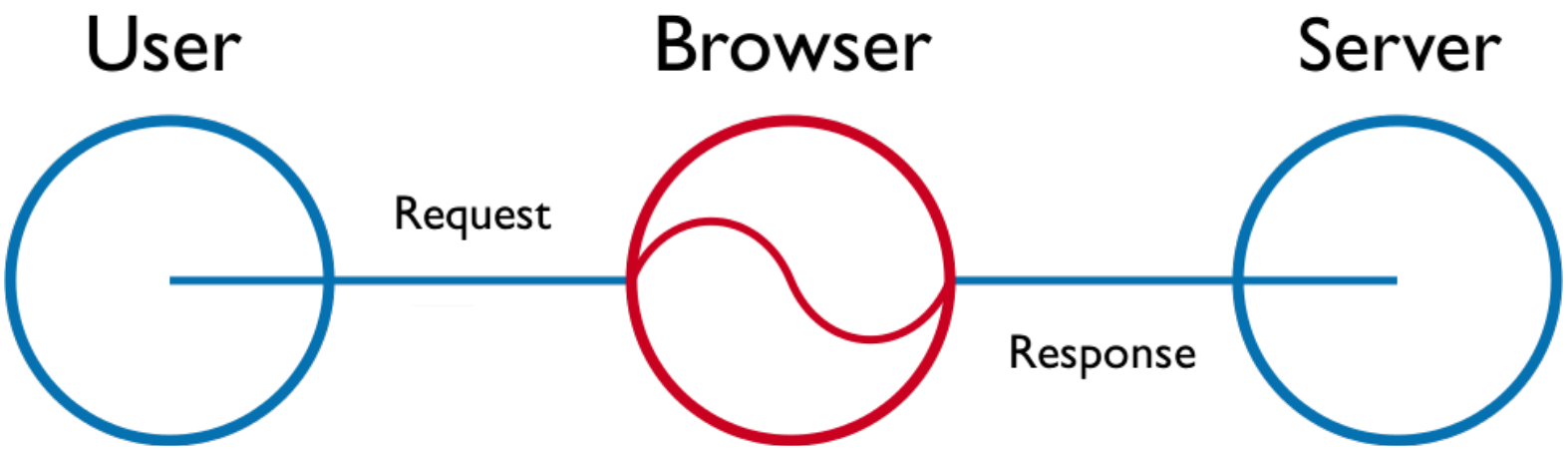

Fake ones?


I’m not a fake!
I can crawl the web, too!

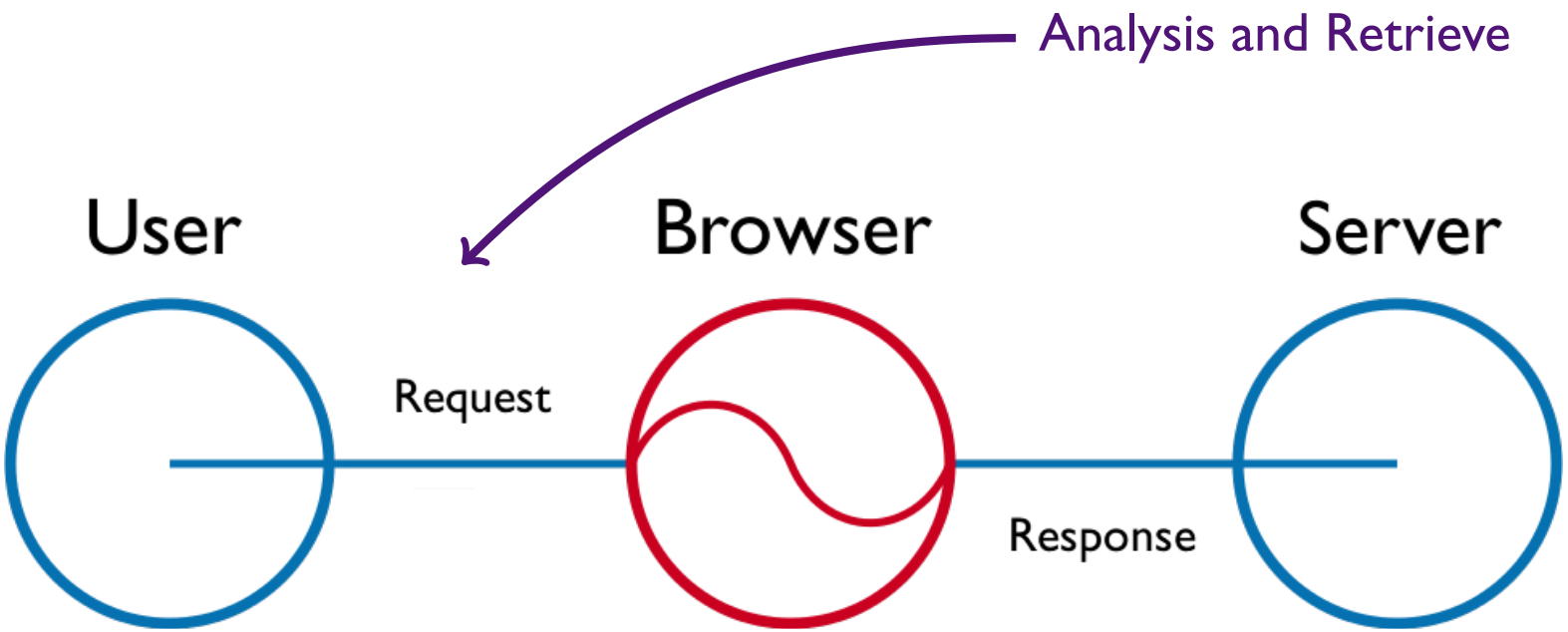
Two Types of Scrapers

- General-purpose Crawlers
- Focused Crawlers (our focus today)

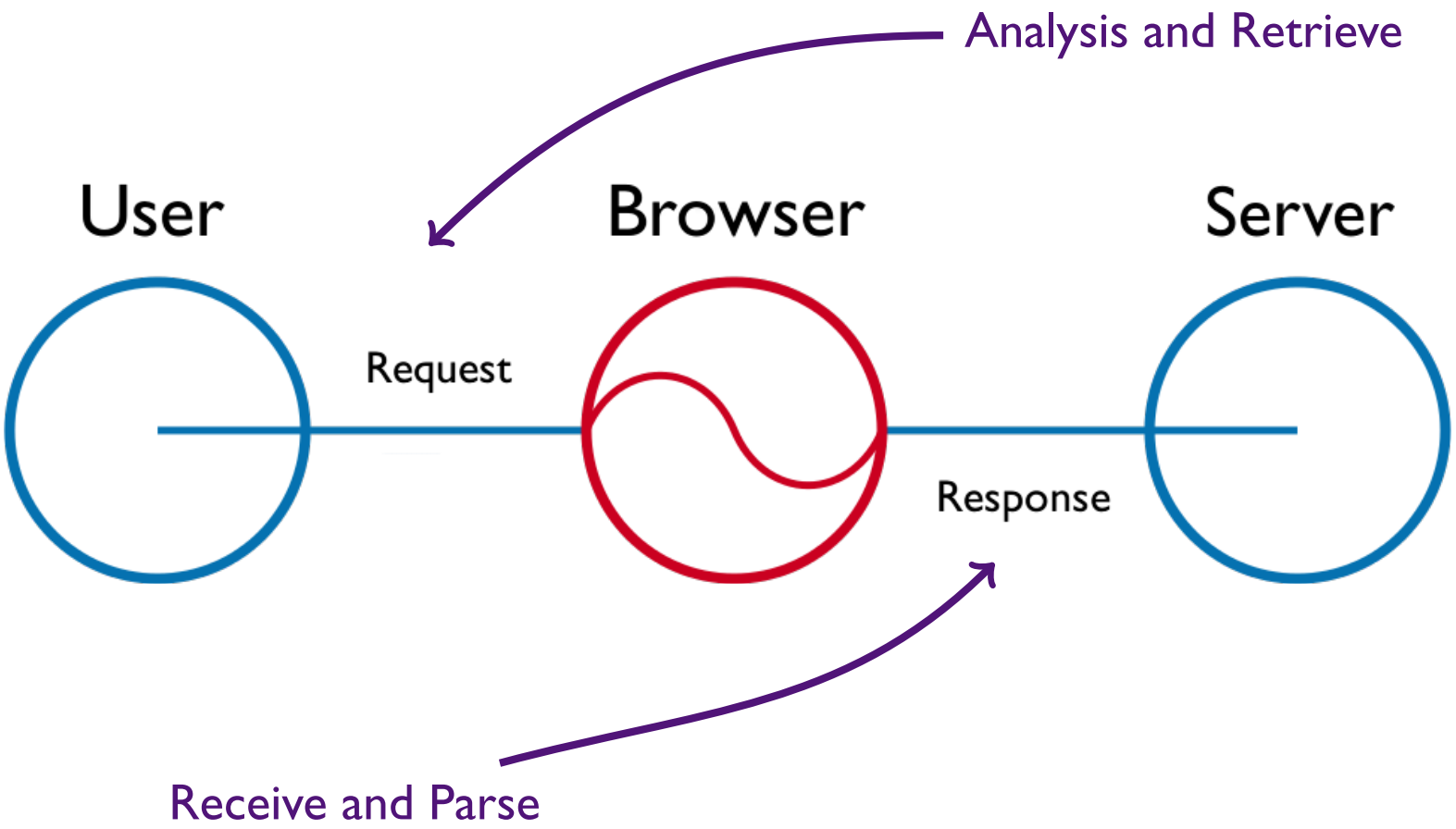
Browser Revisited



Browser Revisited



Browser Revisited



Comparing to Other Languages

Pros & Cons

Pros

- Lightweight
- Easy to implement
- Easy to debug
- Seamless modeling integration: less I/O

Comparing to Other Languages

Pros & Cons

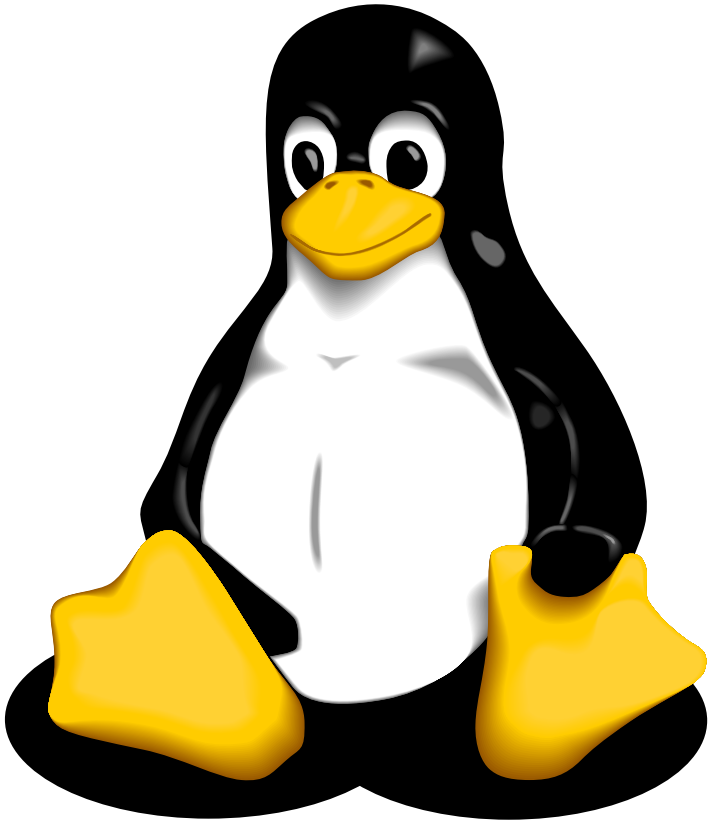
Pros

- Lightweight
- Easy to implement
- Easy to debug
- Seamless modeling integration: less I/O

Cons

- Fewer libraries (than Python & Ruby)
- Multi-Process Parallelization: forking is deficient ...

Choosing A Better Platform



Choosing A Better Platform

Why Linux?

Choosing A Better Platform

Why Linux?

- Network performance & mem. management $\rightarrow$ Faster

Choosing A Better Platform

Why Linux?

- Network performance & mem. management $\rightarrow$ Faster
- Better parallelization support $\rightarrow$ Faster

Choosing A Better Platform

Why Linux?

- Network performance & mem. management $\rightarrow$ Faster
- Better parallelization support $\rightarrow$ Faster
- Unified encoding & locale $\rightarrow$ Faster (for coders)

Choosing A Better Platform

Why Linux?

- Network performance & mem. management $\rightarrow$ Faster
- Better parallelization support $\rightarrow$ Faster
- Unified encoding & locale $\rightarrow$ Faster (for coders)
- More recent third party libs $\rightarrow$ Faster (less bugs)

Part 2

Toolkit

Retrieve & Parse

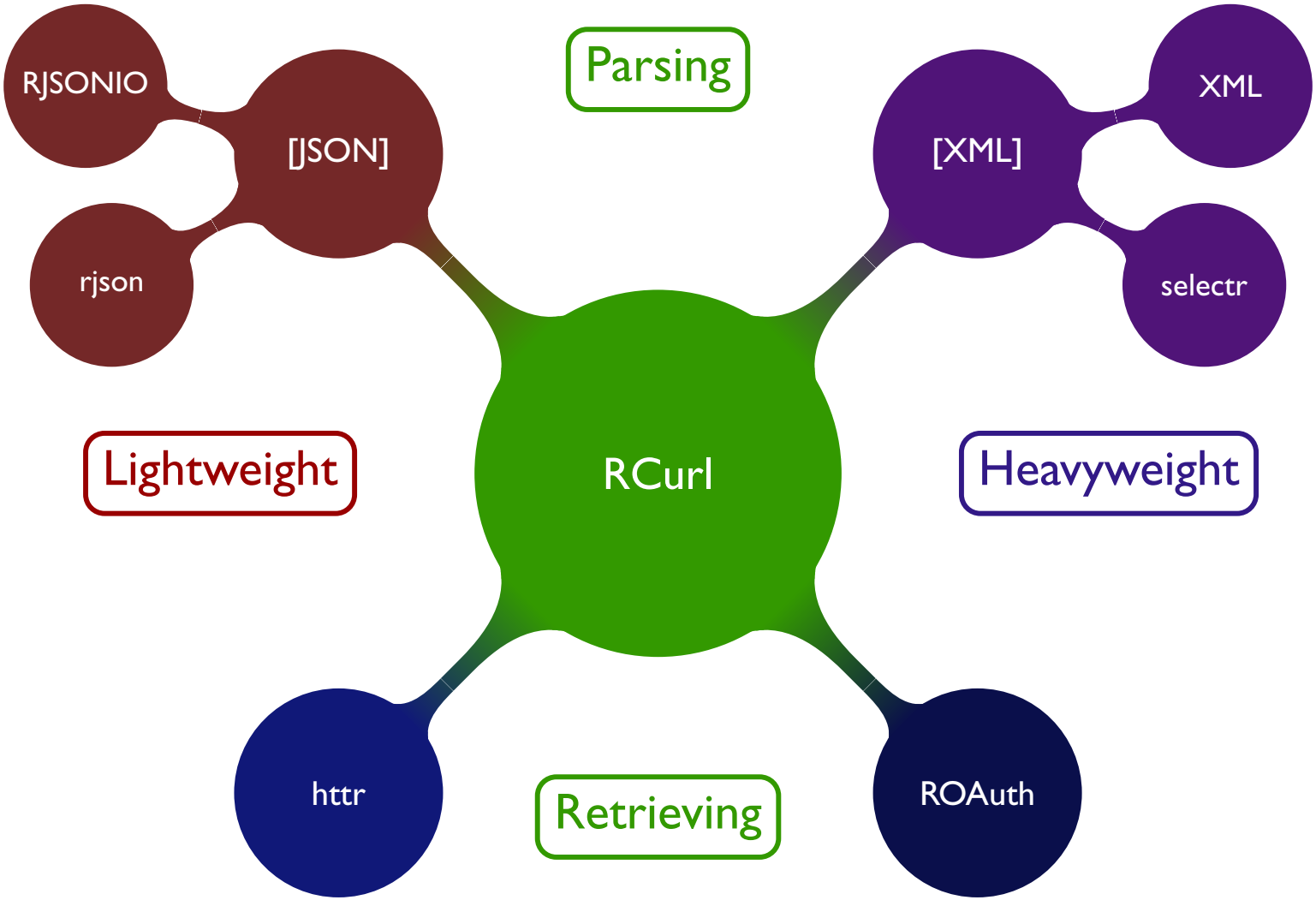
Toolkit

Available R Packages

Pkg. Name	Retrieve?	Parse?	Must-Know?
RCurl	Yes	No	Yes
XML	Limited	Yes	Yes
rjson	No	Yes	Yes
RJSONIO	No	Yes	Optional
httr	Yes	Yes	Optional
selectr	No	Yes	Optional
ROAuth	No	No	Optional

Toolkit

Available R Packages



Toolkit

Available R Packages

- RCurl → Header Configuration

  R 不务正业之 RCurl: `http://cos.name/cn/topic/17816`

Toolkit

Available R Packages

- RCurl → Header Configuration

  R 不务正业之 RCurl: `http://cos.name/cn/topic/17816`

- XML → XPath, 3 Critical Functions

  `http://www.road2stat.com/cn/r/rxml.html`

Toolkit

Available R Packages

- RCurl → Header Configuration

  R 不务正业之 RCurl: `http://cos.name/cn/topic/17816`

- XML → XPath, 3 Critical Functions

  `http://www.road2stat.com/cn/r/rxml.html`

- rjson → Officially Listed / RJSONIO → by D.T.L.

Toolkit

Available R Packages

- RCurl → Header Configuration

  R 不务正业之 RCurl: `http://cos.name/cn/topic/17816`

- XML → XPath, 3 Critical Functions

  `http://www.road2stat.com/cn/r/rxml.html`

- rjson → Officially Listed / RJSONIO → by D.T.L.
- httr → Simplification version RCurl + XML + rjson
  Not Recommended for not discreet enough:

  `http://randyzwitch.com/r-error-message-fun/`

Toolkit

Available R Packages

- RCurl → Header Configuration

  R 不务正业之 RCurl: `http://cos.name/cn/topic/17816`
- XML → XPath, 3 Critical Functions

  `http://www.road2stat.com/cn/r/rxml.html`
- rjson → Officially Listed / RJSONIO → by D.T.L.
- httr → Simplification version RCurl + XML + rjson
  Not Recommended for not discreet enough:

  `http://randyzwitch.com/r-error-message-fun/`
- ROAuth → Useful for APIs. see RWeibo of @lijian001

Toolkit

Available R Packages

- RCurl → Header Configuration

  R 不务正业之 RCurl: `http://cos.name/cn/topic/17816`
- XML → XPath, 3 Critical Functions

  `http://www.road2stat.com/cn/r/rxml.html`
- rjson → Officially Listed / RJSONIO → by D.T.L.
- httr → Simplification version RCurl + XML + rjson
  Not Recommended for not discreet enough:

  `http://randyzwitch.com/r-error-message-fun/`
- ROAuth → Useful for APIs. see RWeibo of @lijian001
- selectr → Translate CSS Selectors to XPath Expressions

Toolkit

Front-End and Miscellaneous

- Chrome Developer Tools / FireBug →
  Analyzing AJAX Requests: `http://cos.name/cn/topic/107729`

Toolkit

Front-End and Miscellaneous

- Chrome Developer Tools / FireBug $\rightarrow$ Analyzing AJAX Requests: `http://cos.name/cn/topic/107729`
- JSONView $\rightarrow$ Output Formatted JSON

Toolkit

Front-End and Miscellaneous

- Chrome Developer Tools / FireBug → Analyzing AJAX Requests: `http://cos.name/cn/topic/107729`
- JSONView → Output Formatted JSON
- Visual Event → Bounded event on DOM elements

Toolkit

Front-End and Miscellaneous

- Chrome Developer Tools / FireBug → Analyzing AJAX Requests: `http://cos.name/cn/topic/107729`
- JSONView → Output Formatted JSON
- Visual Event → Bounded event on DOM elements
- tcpdump + Wireshark → Packet Capture & Protocol Analysis

Part 3

Exception Handling

More than 70%

Exception Handling

Coding Strategy

- Dirty HTML & XML: Preprocess with `htmltidy`

Exception Handling

Coding Strategy

- Dirty HTML & XML: Preprocess with `htmltidy`
- Build-in Condition/Error Handler Function: `XML::xmlStructuredStop`

Exception Handling

Coding Strategy

- Dirty HTML & XML: Preprocess with `htmltidy`
- Build-in Condition/Error Handler Function: `XML::xmlStructuredStop`
- Coding Strategy: Interative until fault-tolerant.

Exception Handling

FAQ on COS BBS

- Cookie Operation $\rightarrow$ `http://cos.name/cn/topic/108806`
- Referer Validation $\rightarrow$ `http://cos.name/cn/topic/109407`
- Session Validation $\rightarrow$ `http://cos.name/cn/topic/107802`
- Encoding Errors $\rightarrow$ Identify the Problem Source

  Google: 编码 `site:cos.name/cn/`

Exception Handling

The Various Data Source

- Choose Official API first: NCBI with rOpenSci

Exception Handling

The Various Data Source

- Choose Official API first: NCBI with rOpenSci
- Restricted API usage: Private API key

Consumer keys of official Twitter clients

Gist Detail
Revisions 5
Stars 275
Forks 128

Download Gist

Clone this gist
/rhenium/3878505

Embed this gist
<script src="https:/

Link to this gist
https://gist.github

gistfile1.md Markdown

Twitter公式クライアントのコンシューマキー

Twitter for iPhone

```
Consumer key: IQKbtAYlXLripLGPWd0HUA
Consumer secret: GgDYlkSvaPxGxC4X8liwpUoqKwwr3lCADbz8A7ADU
```

Twitter for Android

```
Consumer key: 3nVuSoBZnx6U4vzUxf5w
Consumer secret: Bcs59EFbbsdF6Sl9Ng71smgStWEGwXXKSjYvPVt7qys
```

Twitter for Google TV

```
Consumer key: iAtYJ4HpUVfIUoNnif1DA
Consumer secret: 172fOpzuZoYzNYaU3mMYvE8m8MEyLbztOdbrUolU
```

Twitter for iPad

```
Consumer key: CjulERsDeqhhjSme66ECg
Consumer secret: IQWdVyqFxghAtURHGeGiWAsmCAGmdW3WmbEx6Hck
```

Twitter for Mac

Exception Handling

The Various Data Source

- Choose Official API first (NCBI with rOpenSci)
- Restricted API usage: Private API key
- SSL and SSL Decryption: Trusted MITM

`http://www.webos-internals.org/wiki/Decrypt_SSL_(trusted_man-in-the-middle_technique)`

Part 4

Parallelization

Parallelization

The best solution is?

A Conventional Way: `RCurl::getURIAsynchronous()`

Parallelization

The best solution is?

A Conventional Way: `RCurl::getURIAsynchronous()`

- Native. Extremely easy to use.

Parallelization

The best solution is?

A Conventional Way: `RCurl::getURIAsynchronous()`

- Native. Extremely easy to use.
- Pitfalls: Have to control the process number by hand.
  This seems weird!

Parallelization

The best solution is?

A Better Way: doMC + foreach

Parallelization

The best solution is?

A Better Way: doMC + foreach

- Full control / Easy to migrate / Natural to code:

```
require(doMC)
registerDoMC(20)
x <- foreach(i = 1:1e+5, ...) %dopar% {
  xxx <- getURL(urls[i])
}
```

Parallelization

The best solution is?

A Better Way: doMC + foreach

- Full control / Easy to migrate / Natural to code:

```
require(doMC)
registerDoMC(20)
x <- foreach(i = 1:1e+5, ...) %dopar% {
  xxx <- getURL(urls[i])
}
```

- Single machine, `registerDoMC(20)`, 10 min, 1e+5 pages.

Parallelization

The best solution is?

A Better Way: doMC + foreach

- Full control / Easy to migrate / Natural to code:

```
require(doMC)
registerDoMC(20)
x <- foreach(i = 1:1e+5, ...) %dopar% {
  xxx <- getURL(urls[i])
}
```

- Single machine, `registerDoMC(20)`, 10 min, 1e+5 pages.
- Pitfalls: (Almost) Linux only.

Parallelization

More Pitfalls

- Requires high-perf storage $\rightarrow$
  Redis (rredis) or MongoDB (RMongo)?
- Memory leak (RCurl & XML) $\rightarrow$ Avoid long exec. time
- Intensive testing before run $\rightarrow$ Minimize errors

Part 5

Outro

Remarks

Web Crawler Ethics

- Web Crawler Ethics

Remarks

Web Crawler Ethics

- Web Crawler Ethics
- Honor `robots.txt`

Remarks

Web Crawler Ethics

- Web Crawler Ethics
- Honor `robots.txt`
- A Balanced Crawling Rate

Remarks

Web Crawler Ethics

- Web Crawler Ethics
- Honor `robots.txt`
- A Balanced Crawling Rate
- Spammer Shame

Remarks

Web Crawler Ethics

- Web Crawler Ethics
- Honor `robots.txt`
- A Balanced Crawling Rate
- Spammer Shame
- *With great power comes great responsibility.*

Further Reading

1. XML & JSON Specification (esp. XPath)
2. RCurl & XML Documentation
3. Web Data Mining (Chapter 8) by Bing Liu
4. XML and Web Technologies for Data Sciences with R by Duncan Temple Lang, et al. (Due Sep. 2013)
5. Curl.jl `https://github.com/forio/Curl.jl`

http://cos.name/cn/

Q & A