Web Data Collection with R

HTML Forms POST Case Study

Peter Meißner / 2016-02-29 – 2016-03-04 / ECPR WSMT

Posting HTML Forms

GET and POST within Developer Tools

Posting HTML Forms

Teaser

- ADCR author blog post on `http://www.r-datacollection.com/` battle against each other for the readability price

a first glance

```
# ADCR: page 236
attr_inspector <- function(parsed_html, xpath){
  x <- html_nodes(parsed_html, xpath=xpath)
  x <- html_attrs(x)
  x <- lapply(x, function(x) as.data.frame(t(x)) )
  do.call(plyr::rbind.fill, x)
}
```

```
library(httr)
library(rvest)
```

```
## Loading required package: xml2
```

```
library(stringr)
```

a first glance

```
url  <- "http://read-able.com/"
html <- read_html(url)
attr_inspector( html, "//form")
```

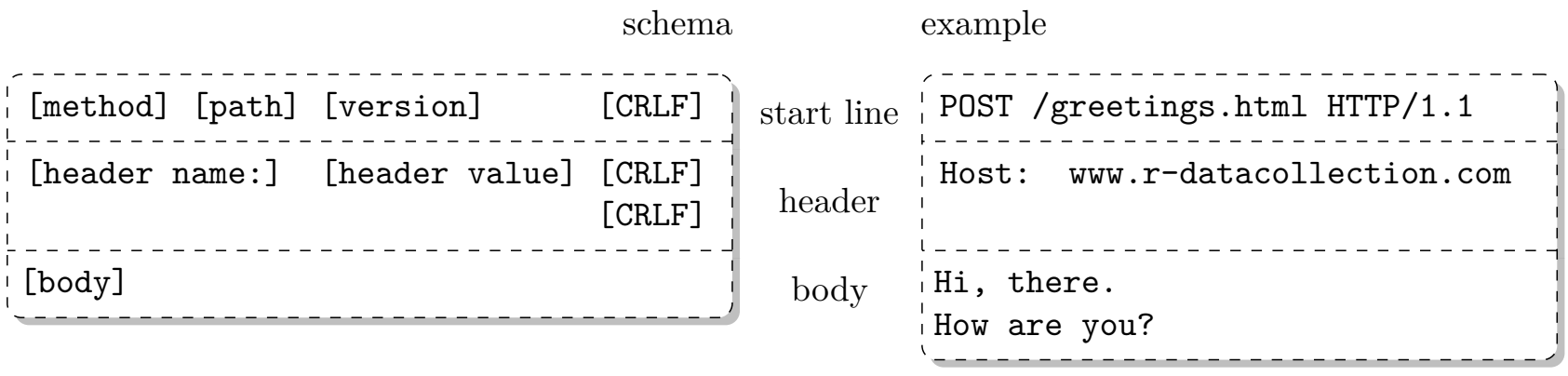
```
##   method    action
## 1    get check.php
## 2   post check.php
```

```
attr_inspector( html, "//form[2]//input|//textarea|//select
```

```
##             id         name rows cols
## 1 directInput directInput   10   60
```

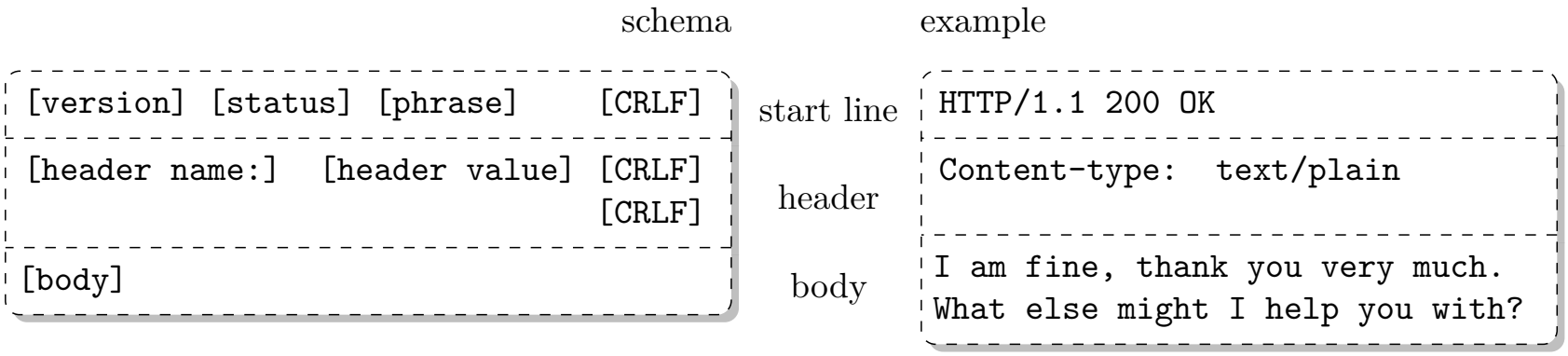
HTTP messages

- but where goes our data?



HTTP messages

- but where goes our data?



getting the texts

```
dominic <- read_html("http://www.r-datacollection.com/blog/
dominic <- html_nodes(dominic, xpath="//p")
dominic <- html_text(dominic)
dominic <- str_c(dominic, collapse="\n")
```

getting the texts

```
peter <- read_html("http://www.r-datacollection.com/blog/In
peter <- html_nodes(peter, xpath="//p")
peter <- html_text(peter)
peter <- str_c(peter, collapse="\n")
```

getting the texts

```
simon <- read_html("http://www.r-datacollection.com/blog/Pr
simon <- html_nodes(simon, xpath="//p")
simon <- html_text(simon)
simon <- str_c(simon, collapse="\n")
```

getting the texts

```
christian <- read_html("http://www.r-datacollection.com/blo
christian <- html_nodes(christian, xpath="//p")
christian <- html_text(christian)
christian <- str_c(christian, collapse="\n")
```

posting texts

```
force <- F # redo or not
if ( !file.exists("dominic.html") | force==T){
  resp_d <- POST("http://read-able.com/check.php",
              body=list(directInput=dominic),
              encode="form")
  writeBin(content(resp_d, "raw"),
           con="dominic.html" , useBytes=T)
}
```

posting texts

```
if ( !file.exists("peter.html")  | force==T){
  resp_p <- POST("http://read-able.com/check.php",
             body=list(directInput=peter),
             encode="form")
  writeBin(content(resp_p, "raw"),
           con="peter.html" , useBytes=T)
}
```

posting texts

```
if ( !file.exists("simon.html") | force==T ){
  resp_s <- POST("http://read-able.com/check.php",
             body=list(directInput=simon),
             encode="form")
  writeBin(content(resp_s, "raw"),
           con="simon.html" , useBytes=T)
}
```

posting texts

```
if ( !file.exists("christian.html")  | force==T){
  resp_c <- POST("http://read-able.com/check.php",
             body=list(directInput=christian),
             encode="form")
  writeBin(content(resp_c, "raw"),
           con="christian.html" , useBytes=T)
}
```

the verdict

```
verdict <- function(file){
  read_html(file) %>%
    html_table() %>%
    magrittr::extract2(1) %>%
    magrittr::extract(2,)
}
```

the verdict

```
verdict("simon.html")
```

```
##                                X1  X2 X3
## 2 Flesch Kincaid Grade Level 7.4 NA
```

```
verdict("dominic.html")
```

```
##                                X1  X2 X3
## 2 Flesch Kincaid Grade Level 8.1 NA
```

```
verdict("peter.html")
```

```
##                                X1   X2 X3
## 2 Flesch Kincaid Grade Level 10.6 NA
```

```
verdict("christian.html")
```

```
##                                X1 X2 X3
## 2 Flesch Kincaid Grade Level 11 NA
```

GET and POST within Developer Tools

GET and POST within Developer Tools

Libve Clicking