Web Data Collection with R

JSON/API Case Study

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

The old wikipedia pageviews API - stats.grok.se

The old wikipedia pageviews API - stats.grok.se

packages

```r
library(rvest)
library(jsonlite)
library(lubridate)
library(dplyr)
```

http://stats.grok.se/

A first try

```r
url  <- "http://stats.grok.se/json/en/latest90/Influenza"
json <- html_text(read_html(url))
json
```

```
## [1] "{\"daily_views\": {\"2015-12-17\": 2100, \"2015-12-
```

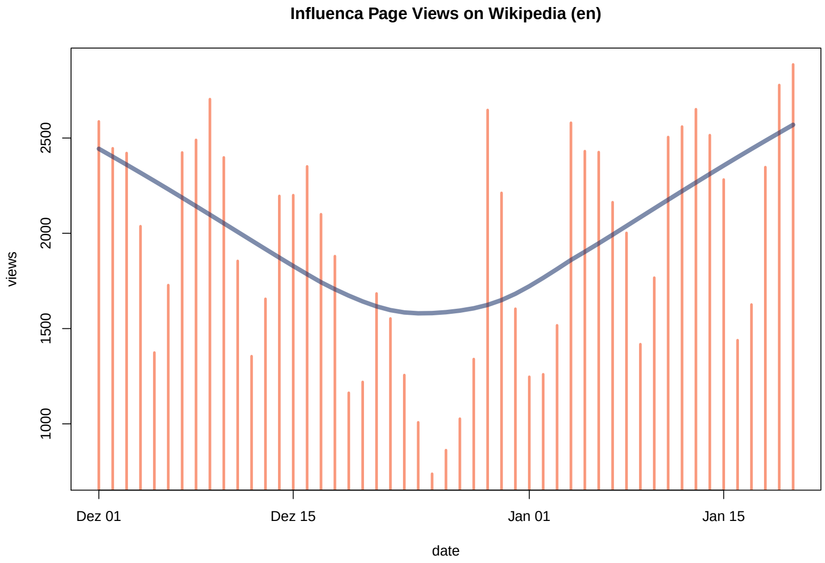
A first try

```
## {"daily_views": {"2015-12-17": 2100,
##   "2015-12-07": 2424,
##   "2015-12-06": 1728,
##   "2016-01-04": 2580,
##   "2016-01-05": 2431,
##   "2015-12-09": 2704,
##   "2015-12-08": 2490,
##   "2016-01-01": 1247,
##   "2016-01-02": 1260,
##   "2016-01-03": 1517,
##   "2015-12-03": 2421,
##   "2015-12-02": 2446,
##   "2015-12-01": 2587,
##
## ...
```

Getting JSON in

```
data  <- fromJSON(json)
date  <- as.Date(names(data$daily_views))
views <- unlist(data$daily_views)

plot( date, views,
      ylim = c(0, 3000),
      type = "h",
      col  = "#F54B1A90",
      lwd=3,
      main="Influenca Page Views on Wikipedia (en)")

lines(lowess(views ~ date), col = "#1B346C90", lwd=5)
```

Getting JSON in



Getting serious

```r
url_pt1  <- "http://stats.grok.se/json/en/"
url_pt2  <-
  paste0(
    rep(2014:2015, each=12),
    str_pad(1:12, width=2, side="left", "0")
  )
url_pt3  <- "/Influenza"
URL <- paste0(url_pt1, url_pt2, url_pt3)
```

Getting serious

```
# downloadig the data
JSON <- list()
for( i in seq_along(URL) ){
  fname <- basename(dirname(URL[i]))
  if( !file.exists(fname) ){
    download.file(URL[i], fname )
    Sys.sleep(1)
  }
  JSON[i] <- readLines(fname, warn = FALSE)
}
```

Getting serious

```r
# parsing the JSON
JSON_parsed <- lapply(JSON, fromJSON)

json <- JSON_parsed[[1]]

date <-
  json$daily_views %>%
  names() %>%
  ymd()

views <-
  json$daily_views %>%
  unlist()

views[1:3]
```

```
## 2014-01-15 2014-01-14 2014-01-17
##       6009       5951       5219
```

Getting serious

```r
# putting it in a function
page_views_to_df <- function(json){
  date  <- json$daily_views %>% names() %>% ymd()
  views <- json$daily_views %>% unlist()
  df <- data.frame(date, views)[!is.na(date),]
  rownames(df) <- NULL
  return(df)
}
```

Getting serious

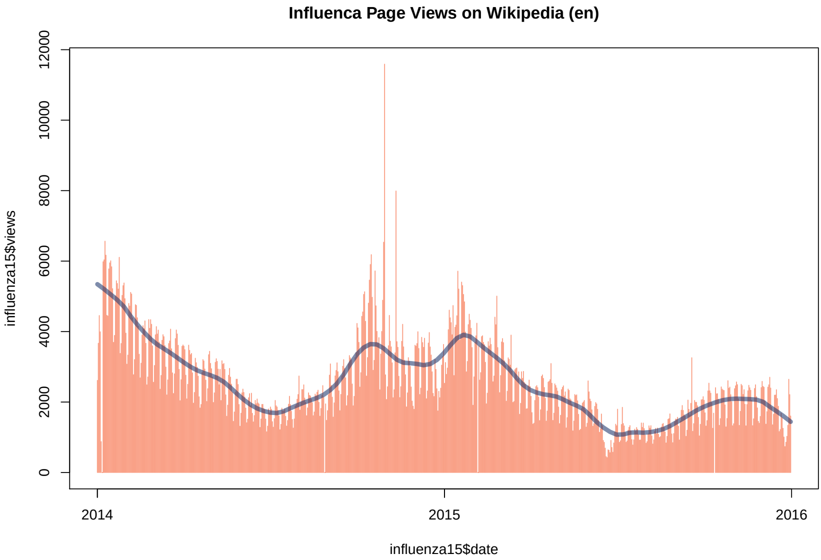
```r
influenza15 <-
  JSON_parsed %>%
  lapply(page_views_to_df) %>%
  do.call(rbind, .)
```

```
## Warning: 3 failed to parse.

## Warning: 1 failed to parse.


## Warning: 1 failed to parse.


## Warning: 1 failed to parse.


## Warning: 1 failed to parse.

## Warning: 3 failed to parse.

## Warning: 1 failed to parse.


## Warning: 1 failed to parse.
```

Getting serious

```
plot( influenza15$date, influenza15$views,
      type = "h",
      col  = "#F54B1A90",
      lwd=1,
      main="Influenca Page Views on Wikipedia (en)")

lowess(influenza15$views ~ influenza15$date, f=0.08) %>%
  lines(col = "#1B346C90", lwd=5)
```

Getting serious


Influenca Page Views on Wikipedia (en)
influenza15$views
influenza15$date