Webscraping NY Times in R using Rvest and RNYTIMES

Loren Collingwood

To harvest text, the webscraping skillset is a must. This vignette introduces webscraping in R using the package Rvest and NY Times API.

Step 1

Download and install the requisite packages. Then load them into R.

```
    # Install R package if you have not done so #
    #install.packages("rvest")

    # Load package into current R session #
    library(rvest)
```

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

```
    #install.packges("devtools") # May need to install Rtools on windows or
    # Xcode (+ command line tools) on MAC OSx.
    # devtools::install_github("omegahat/RNYTimes")
    library(RNYTimes)

    #install.packages("jsonlite")
    library(jsonlite)
```

Step 2

Set up NY Times developer API article search app and get your own api key. You will need to set up an account at NYTimes. Then set up search term, start, and end date, as below:

```
#####################
# Muslim Ban Search #
#####################

term <- "muslim+ban" # Need to use + to string together separate words
begin_date <- "20170101" # Begin Date
end_date <- "20180101" # End Date

# Get an Article Search NY-Times KEY

# Have to set up an APP api and from there you can copy/paste your key
NYTIMES_KEY <- "HVAgAcTcESrWWOhIMnWb9R2L6FBAj1QP" # NYT Password need to get from website
```

Step 3

Extract individual article links:

```
############################
# Extract the Links First #
############################

# Set string to send to fromJson
baseurl <- paste0("http://api.nytimes.com/svc/search/v2/articlesearch.json?q=",term,
                  "&begin_date=",begin_date,"&end_date=",end_date,
                  "&facet_filter=true&api-key=",NYTIMES_KEY, sep="")

# Grab first Page; tag API via fromJSON
initialQuery <- jsonlite::fromJSON(baseurl)

# Figure out Number of pages to loop over
#maxPages <- round((initialQuery$response$meta$hits[1] / 10)-1)

pages <- list()

for(i in 1:9){
    nytSearch <- jsonlite::fromJSON(paste0(baseurl, "&page=", i),
                                    flatten = TRUE) %>% data.frame()
    message("Retrieving page ", i)
    pages[[i]] <- nytSearch  # use i+1 here, and start with 0:maxPages but will throw error at end
    Sys.sleep(1)
}
```

```
## Retrieving page 1
## Retrieving page 2
## Retrieving page 3
## Retrieving page 4
## Retrieving page 5
## Retrieving page 6
## Retrieving page 7
## Retrieving page 8
## Retrieving page 9
```

```
# Combine Pages using rbind_pages() function #
allNYTSearch <- rbind_pages(pages)

# Extract links #
links <- allNYTSearch$response.docs.web_url
```

Step 4

Loop over links, capture text

```r
# Loop over Links, bringing in rvest package #
# Use SelectorGadget on googlechrome to sort out the text to scrape (in this case: div p) #

articles <- list()

# Muslim Ban Search #
for (i in 1:length(links)) {

    link <- read_html(links[i])

    articles[[i]] <- paste( html_text(html_nodes(link, "div p")), collapse = " ")

    grep("muslim ban", tolower(articles[[i]]))

}

# Put into Vector #
articles_df <- unlist(articles)

# Put into Dataframe
nyt_muslim_ban_2017 <- data.frame(allNYTSearch, text = articles_df, stringsAsFactors = F)

# Save as .RData file # This can be loaded into R vis load() function #
save(nyt_muslim_ban_2017, file="nyt_muslim_ban_2017.RData")
```