Final Project - PSYC789W - R Programming

Eric Dunford

Introduction

The project I proposed in my initial proposal was a little more involved than I had expected. The project sought to create a web-scraper, collect data, and then map that data onto an interactive time-series map. After diving further into the demands of such a project, I realized that it might be best to take a step back and just focus on the scraping part of the project.

Below you will find a web-scraping function that pulls from the BBC. The goal was to build a function that generated a data frame of all the top stories with regard to some subject. The user should be capable of running queries within the function and to view the stories relevance by the degree of social media activity the story received on Facebook. The output of the function generates an object as a data frame, orders that data frame by the stories relevancy, and formats the dates as Dates (via the `lubridate` function).

All in all, this is just a small step toward the larger project that I initially outlined; however, I believe it to be sufficient for this course, given the parameters outlined in the syllabus.

bbcStoryScrape()

As noted above, the following function scrapes relevant news topics given a query provided by the user. The function then returns a data frame of the BBCs most recent stories regarding that query.

```
bbcStoryScrape <- function(query){
  require(dplyr)
  require(rvest)
  require(rjson)
  require(XML)
  require(lubridate)
  require(plyr) #I know: having the two siblings together is a 'no no', but I need `ldply()`. So I'll l
  url <- paste0("http://www.bbc.co.uk/search?q=",query)
  links <- htmlParse(url) %>% html_nodes(".search-results article") %>% html_nodes("a") %>% html_attr("
  links <- gsub("programmes",NA,links)
  links <- gsub("music",NA,links)
  links <- gsub("blogs",NA,links)
  links <- links[!is.na(links)]
  bbcScraper <- function(url){
    if(!is.character(url)){
      return("Issue: Non-character value detected")
      } else{
        raw.data <- html(url)
        if(is.null(html_node(raw.data,".date"))){
          date <- html_node(raw.data,"td span") %>% html_text() %>% strtrim(.,width=35) %>% gsub("Last
          } else{
            date <- html_node(raw.data,".date") %>% html_text()
            }}
    if(is.null(html_node(raw.data,".story-header")) & is.null(html_node(raw.data,".headlinestory b")) &
      title <- html_node(raw.data,".sh") %>% html_text()
      } else {if(is.null(html_node(raw.data,".story-header")) & is.null(html_node(raw.data,".headlinest
        title <- html_node(raw.data,".story-body__h1") %>% html_text()
        } else{if(!is.null(html_node(raw.data,".story-header")) & is.null(html_node(raw.data,".headline
```

```
          title <- html_node(raw.data,".story-header") %>% html_text()
          } else{
            title <- html_node(raw.data,".headlinestory b") %>% html_text()
            }
          }
        }
    if(!is.null(title) & !is.null(data)){
      c(title,date)
      }
    }
  data <- ldply(links,bbcScraper)
  colnames(data) <- c("Headline","Date_Published")
  Relevance <- function(url){
    queryUrl = paste0('http://graph.facebook.com/fql?q=','select share_count,comment_count,like_count,
    lookUp <- URLencode(queryUrl)
    rd <- readLines(lookUp, warn="F")
    data <- fromJSON(rd)
    output <- data.frame(Shares=data$data[[1]]$share_count,No.of.Comments=data$data[[1]]$comment_count,
    return(output)
    }
  SM_data <- ldply(links,Relevance)
  output <- cbind(data,SM_data)
  output$link <- links
  #Arrange the output in descending order
  output <- arrange(output,desc(Total))
  #Date formating
  output$Date_Published <- output$Date_Published %>% as.Date(.,"%d %B %Y")
  output
  }
```

Testing the function out.

```
bbcStoryScrape("Saudi Arabia")
```

```
##                                                 Headline Date_Published
## 1          Saudi Arabia's King Abdullah bin Abdulaziz dies     2015-01-23
## 2    Saudi King Abdullah death: Cameron to pay his respects     2015-01-23
## 3        Saudi Arabia's new King Salman promises continuity     2015-01-23
## 4                                  Obituary: King Abdullah     2015-01-23
## 5                                    Saudi Arabia profile      2015-01-23
## 6                             Saudi Arabia oil: What next?     2015-01-23
## 7               Saudi: Turbulent times for new King Salman     2015-01-23
## 8  Saudi Arabia: Lashings, hawks and friends in high places     2015-01-23
## 9                                    Saudi Arabia profile      2015-01-23
## 10         Saudi Arabia's King Abdullah leaves mixed legacy     2015-01-23
##    Shares No.of.Comments No.of.Likes Total
## 1    4765           4420       10904 20089
## 2     310            385         452  1147
## 3     434            230         343  1007
## 4     116             53         217   386
## 5      61             21          22   104
## 6      40              2           0    42
## 7      16              1           5    22
```

```
## 8       7              5           5    17
## 9       6              0           0     6
## 10      3              0           0     3
##                                                      link
## 1  http://www.bbc.co.uk/news/world-middle-east-30945324
## 2                 http://www.bbc.co.uk/news/uk-30946159
## 3  http://www.bbc.co.uk/news/world-middle-east-30950731
## 4  http://www.bbc.co.uk/news/world-middle-east-10214554
## 5  http://www.bbc.co.uk/news/world-middle-east-14703523
## 6           http://www.bbc.co.uk/news/business-30950263
## 7  http://www.bbc.co.uk/news/world-middle-east-30949483
## 8               http://www.bbc.co.uk/newsbeat/30948582
## 9  http://www.bbc.co.uk/news/world-middle-east-14703480
## 10 http://www.bbc.co.uk/news/world-middle-east-30950433
```

Persisting Issues

It is important to note that the function has difficulty dealing with older stories (i.e. stories that hit the web around the turn of the century). This is due to the structure of the BBC website during that time period, which was much different 15 years ago than it was today. This issue, though outstanding, can still be resolved. Rather, it requires the addition of more conditional elements in the function to deal with the alternative structure.

Also, the function doesn't discriminate by language; thus, it's capable of pulling news stories in other languages, which can present date formatting issues (and general readability issues). I'm still working my way through that.

Conclusion

I hope you enjoyed the function. Though still fragile (given different search terms, it is still possible to break the function), I think it is a good first step in the right direction. Within political science, the data generating process can be quite arduous – since we primarily rely on observational data. Thus, learning alternative processes to the data generating process can be useful down the line.

I look forward to your feedback and appreciate all your help during the course of the class.