Web scraping

Venkatesh Mummidi (2016Z19)

Raj Kumar R (2016Z21)

23 January 2017

Objective:

To do web scraping and get the headlines, content of the body from all the links given to us

Programming Language used: **R**

**Process:**

* The csv file contains the all the urls in each row. R reads the content as factor , so we are changing it to charachter.

Read all the given urls

library(xlsx)

library(rvest)

et <- read.xlsx("ET\_Links\_Sample.xlsx",sheetIndex = 1,header = F)  
et[,1]=as.character(et[,1])  
str(et)

## 'data.frame': 100 obs. of 1 variable:  
## $ X1: chr "http://economictimes.indiatimes.com/archivelist/year-2015,month-8,starttime-42242.cms" "http://economictimes.indiatimes.com/archivelist/year-2015,month-8,starttime-42243.cms" "http://economictimes.indiatimes.com/archivelist/year-2015,month-8,starttime-42244.cms" "http://economictimes.indiatimes.com/archivelist/year-2015,month-8,starttime-42245.cms" ...

dim(et)

## [1] 100 1

* Initializations are done

title=c()  
body=c()  
date=c()  
n=1

* Read the urls provided one by one by a for loop

for(i in 1:nrow(et){  
 url=read\_html(et[i,1])

* The URL contains lot of links in itself (500 links in each url) and they are stored as content in the webpage. All the links are read as hrefs

#get hrefs of all the links in provided url  
 hrefs <- url%>%html\_nodes("ul.content li a")%>%html\_attr("href")

*#ul.content is the class and li a is the css tag, where we can fetch the links.*

date.et <- url%>%html\_nodes("td.contentbox5 b")%>%html\_text()

*#Extract date of each article. Date is associated with td.contentbox5 class and b tag.*   
 for(j in 1:length(hrefs){

*#Title and body are extracted for each of the href*  
   
 appended.url=paste("http://economictimes.indiatimes.com",hrefs[j],sep = '')  
 link=read\_html(appended.url)

* While reading the links, the complete address of the website was not there so we append the web address in front of the content. Ex: link extracted from href is *“/news/politics-and-nation/quota-row-curfew-imposed-in-gujarats-mehsana-district/articleshow/48674445.cms".* So we need to append “[*http://economictimes.indiatimes.com*](http://economictimes.indiatimes.com)*”* infront of every link.

#Each href has title and content which are stored as title and body. Title can be obtained from h1.title tag and body can be obtained from div. Normal tag.

title[n] <- link%>%html\_nodes("h1.title")%>%html\_text()   
   
 body[n] <- link%>%html\_nodes("div.Normal")%>%html\_text()  
   
 date[n] <- date.et[2]  
 n=n+1  
 }  
}

date[1]

## [1] "26 Aug, 2015"

title[1]

## [1] "Quota row: Curfew imposed in Gujarat's Mehsana district"

body[1]

## [1] "AHMEDABAD: After widespread incidents of violence across the state of Gujarat, curfew was imposed in Mehsana city and the towns of Unjha and Visnagar in the district, the collector of Mehsana said here today. \"Curfew has been imposed in Mehsana city as well as Unjha and Visnagar towns after incidents of violence in Mehsana district,\" Mehsana District Collector Lochan Sehra told PTI. Mehsana is around 90 km from here. The district administration also called in paramilitary forces to control the situation, the collector said. When asked about the number of incidents of violence in Mehsana district, the collector said that around two to three incidents of violence occurred today. Mob blocked roads in the towns of Unjha and Visnagar, after which curfew was imposed to control the situation, the collector said. The collector said that no casualty has been reported so far. "

#Combine date,title and body and write to a excel file

final=as.data.frame(cbind(date,title,body))

# Once all the contents of the link are read, they are written into date, title and body respectively. loop reiterates and the contents of the next link are read and the contents in it are read and written until the last link. Finally combine they as dataframe and write to output file.

write.xlsx(final,"final.xlsx")

As there are total of 100\*500=50,000 articles, we ran our code to fetch first 30 articles.