**Scraping bibliometric data from google scholar using R**

Overview: Describe the steps to scrape data from html and java rendered web pages.

Data source: available from Google Scholar, e.g. citations, subject areas. Top 100 researches.

Target audience: R users with a basic knowledge of html and java rendered websites.

Main R packages: splashr, rvest, scholar

**Outline**

Data Scraping

Terms of use

HTML vs Java rendered websites

What is bibliometric data?

Data Scraping HTML

Scraping lists of authors and citations by university (Top 100)

Data Scraping JAVA

Scraping lists of authors IDs by university (Top 100)

Data analysis

Main areas of research

Co-citation Network

```{r eval=FALSE, include=FALSE}

# Search a list of uottawa authors

ottawa <- "https://scholar.google.ca/citations?view\_op=view\_org&org=5757600927927532557&hl=en&oi=io"

toronto <- "https://scholar.google.ca/citations?view\_op=view\_org&org=8515235176732148308&hl=en&oi=io"

page <- httr::GET(ottawa)

str(page)

page\_content <- httr::content(page, "text")

write.table(x = page\_content,

col.names = F,

row.names = F,

quote = F,

file = "test\_1.html")

)

library (rvest)

webpage <- xml2::read\_html(ottawa)

#Using CSS selectors to scrap the rankings section

data\_html <- html\_nodes(webpage,'#gs\_hdr\_tsi')

#Converting the ranking data to text

data <- html\_text(data\_html)

#Loading both the required libraries

library(rvest)

library(V8)

#URL with js-rendered content to be scraped

link <- 'https://food.list.co.uk/place/22191-brewhemia-edinburgh/'

#Read the html page content and extract all javascript codes that are inside a list

link %>%

html\_nodes('li') %>%

html\_nodes('script') %>%

html\_text()

# Create a new v8 context

ct <- v8()

#parse the html content from the js output and print it as text

read\_html(ct$eval(gsub('document.write','',link))) %>%

html\_text()

```

moher <- scholar::get\_citation\_history("QhVtC1wAAAAJ")

scholar::get\_profile("QhVtC1wAAAAJ")

```{r eval=FALSE, include=FALSE}

# Search a list of uottawa authors

webpage <- xml2::read\_html(test[10])

)

#Using CSS selectors to scrap the rankings section

data\_html <- html\_nodes(webpage,'.gs\_ai\_name, .gs\_ai\_cby')

data\_html\_2 <- html\_nodes(data\_html, '.gs\_ai\_int')

#Converting the ranking data to text

data <- html\_text(data\_html)

```{r include=TRUE, eval=FALSE}

test <- paste0(url\_ins, ins\_id)

```

page <- httr::GET(test[10])

str(page)

```

page\_content <- httr::content(page, "text")

write.table(x = page\_content,

col.names = F,

row.names = F,

quote = F,

file = "test\_1.html")

)

webpage <- xml2::read\_html(ottawa)