

Actividad #2 – Curso Data Science - Stage Barcelona
Datos Elegantes + Análisis de Datos con Web Scrapping

GRUPO #4

Integrantes:

- **Richard Croce**
- **Kevin Moscoso**
- **Roberto Ibarra**

Contenido

Control de Cambios.....	3
(I) Pregunta 1.....	4
(II) Pregunta 2.....	7
(III) Repositorio en GITHUB	9

Control de Cambios

Fecha	Versión	Editado por	Comentarios
20-11-2023	1.0	Grupo 04	Creación Documento Actividad #2
21-11-2023	2.0	Grupo 04	Trabajo de Preguntas 1.5 en adelante.

(I) Pregunta 1

Queremos programar un programa de tipo web scrapping con el que podamos obtener una página web, mediante su URL, y poder analizar su contenido HTML con tal de extraer datos e información específica.

1. Pregunta #1:

Descargar la página web de la URL indicada, y almacenarlo en un formato de R apto para ser tratado.

```
#Inicio ejercicio Actividad #2
#Instalar el paquete httr para descargar paginas web
#Instalar paquete XML
install.packages("httr")
install.packages("XML")
install.packages("dplyr")
install.packages("stringr")

library(XML)
library(dplyr)

#Pregunta 1.1 agregando
url <- "https://www.mediawiki.org/wiki/Mediawiki"
web_page <- httr::GET(url)

> #Pregunta 1.1 agregando
> url <- "https://www.mediawiki.org/wiki/Mediawiki"
> web_page <- httr::GET(url)
> |
```

Environment	History	Connections	Git	Tutorial
R 174 MiB Global Environment				
Data				
web_page	List of 10			
values				
url	"https://www.mediawiki.org/wiki/Mediawiki"			

2. Pregunta #2

El título de la página es = Mediawiki

```
> xml_title
[[1]]
<title>Mediawiki</title>

attr(,"class")
[1] "XMLNodeSet"
> print(xml_title[1])
[[1]]
<title>Mediawiki</title>

> |
```

Analizar el contenido de la web, buscando el título de la página (que en HTML se etiqueta como “title”).

3. Pregunta #3

Analizar el contenido de la web, buscando todos los enlaces (que en HTML se etiquetan como “a”), buscando el texto del enlace, así como la URL.

Empleamos un dataframe de nombre df.

main.R x df x	
Filter	
HREF	TEXTO
1 #bodyContent	Jump to content
2 /wiki/MediaWiki	Main page
3 /wiki/Download	Get MediaWiki
4 /wiki/Special:MyLanguage/Category:Extensions	Get extensions
5 //techblog.wikimedia.org/	Tech blog
6 /wiki/Special:MyLanguage/How_to_contribute	Contribute
7 /wiki/Special:MyLanguage/Help:Contents	User help
8 /wiki/Special:MyLanguage/Manual:FAQ	FAQ
9 /wiki/Special:MyLanguage/Manual:Contents	Technical manual
10 /wiki/Project:Support_desk	Support desk
11 /wiki/Special:MyLanguage/Communication	Communication
12 https://developer.wikimedia.org/	Developer portal
13 /wiki/Development_statistics	Code statistics
14 /wiki/Project:Help	Community portal
15 /wiki/Special:RecentChanges	Recent changes
16 /wiki/Special:LanguageStats	Translate content
17 /wiki/Special:Random	Random page
18 /wiki/Project:Village_Pump	Village pump
19 /wiki/Project:Sandbox	Sandbox

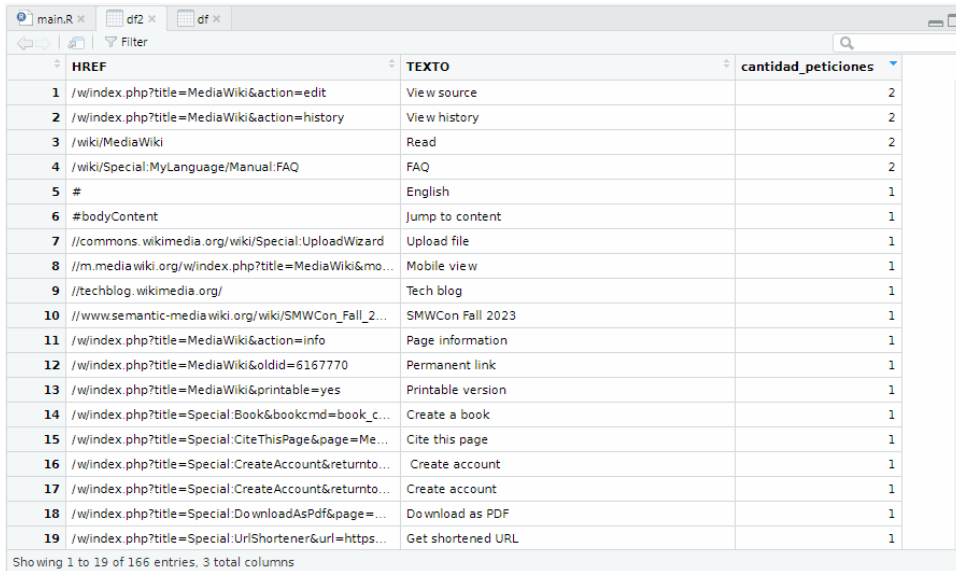
Showing 1 to 19 of 170 entries, 2 total columns

4. Pregunta #4

Generar una tabla con cada enlace encontrado, indicando el texto que acompaña el enlace, y el número de veces que aparece un enlace con ese mismo objetivo.

Empleamos funciones o métodos de la clase dataframe, para poder agrupar y sumarizar los valores de HREF.

```
#Pregunta 1.4 ver la cantidad de URLs que han sido llamado mas de una vez y agruparlo con su texto
df2 <- df %>% group_by(HREF,TEXTO) %>% summarise(cantidad_peticiones = n())
```



	HREF	TEXTO	cantidad_peticiones
1	/w/index.php?title=MediaWiki&action=edit	View source	2
2	/w/index.php?title=MediaWiki&action=history	View history	2
3	/wiki/MediaWiki	Read	2
4	/wiki/Special:MyLanguage/Manual:FAQ	FAQ	2
5	#	English	1
6	#bodyContent	Jump to content	1
7	//commons.wikimedia.org/wiki/Special:UploadWizard	Upload file	1
8	//m.media.wiki.org/w/index.php?title=MediaWiki&mo...	Mobile view	1
9	//techblog.wikimedia.org/	Tech blog	1
10	//www.semantic-media.wiki.org/wiki/SMWCon_Fall_2...	SMWCon Fall 2023	1
11	/w/index.php?title=MediaWiki&action=info	Page information	1
12	/w/index.php?title=MediaWiki&oldid=6167770	Permanent link	1
13	/w/index.php?title=MediaWiki&printable=yes	Printable version	1
14	/w/index.php?title=Special:Book&bookcmd=book_c...	Create a book	1
15	/w/index.php?title=Special:CiteThisPage&page=Me...	Cite this page	1
16	/w/index.php?title=Special:CreateAccount&returnto...	Create account	1
17	/w/index.php?title=Special:CreateAccount&returnto...	Create account	1
18	/w/index.php?title=Special:DownloadAsPdf&page=...	Download as PDF	1
19	/w/index.php?title=Special:UrlShortener&url=https...	Get shortened URL	1

Showing 1 to 19 of 166 entries. 3 total columns

5. Pregunta #5

Para cada enlace, seguirlo e indicar si está activo (podemos usar el código de status HTTP al hacer una petición a esa URL).

Hicimos uso de de la función case_when para normalizar los valores de HREF del dataframe2 (df2), obtener si son valores relativos o absolutos y finalmente ejecutar un for al dataframe para iterar constantemente y obtener el http status por cada validación.

HREF	TEXTO	cantidad_peticiones	url_type	responseCode
1 https://www.media.wiki.org/wiki/MediaWiki#	English	1	relativo	200
2 https://www.media.wiki.org/wiki/MediaWiki#bodyCo...	Jump to content	1	relativo	200
3 https://www.media.wiki.org/wiki/MediaWiki/commons...	Upload file	1	relativo	404
4 https://www.media.wiki.org/wiki/MediaWiki/m.media...	Mobile view	1	relativo	404
5 https://www.media.wiki.org/wiki/MediaWiki/techblog...	Tech blog	1	relativo	404
6 https://www.media.wiki.org/wiki/MediaWiki/www.sem...	SMWCon Fall 2023	1	relativo	404
7 https://www.media.wiki.org/wiki/MediaWiki/w/index.p...	View source	2	relativo	200
8 https://www.media.wiki.org/wiki/MediaWiki/w/index.p...	View history	2	relativo	404
9 https://www.media.wiki.org/wiki/MediaWiki/w/index.p...	Page information	1	relativo	200
10 https://www.media.wiki.org/wiki/MediaWiki/w/index.p...	Permanent link	1	relativo	200
11 https://www.media.wiki.org/wiki/MediaWiki/w/index.p...	Printable version	1	relativo	404
12 https://www.media.wiki.org/wiki/MediaWiki/w/index.p...	Create a book	1	relativo	404
13 https://www.media.wiki.org/wiki/MediaWiki/w/index.p...	Cite this page	1	relativo	404
14 https://www.media.wiki.org/wiki/MediaWiki/w/index.p...	Create account	1	relativo	404
15 https://www.media.wiki.org/wiki/MediaWiki/w/index.p...	Create account	1	relativo	404
16 https://www.media.wiki.org/wiki/MediaWiki/w/index.p...	Download as PDF	1	relativo	404
17 https://www.media.wiki.org/wiki/MediaWiki/w/index.p...	Get shortened URL	1	relativo	404
18 https://www.media.wiki.org/wiki/MediaWiki/w/index.p...	Log in	1	relativo	404
19 https://www.media.wiki.org/wiki/MediaWiki/w/index.p...	Log in	1	relativo	404
20 https://www.media.wiki.org/wiki/MediaWiki/wiki/Cate...	Languages pages	1	relativo	404
21 https://www.media.wiki.org/wiki/MediaWiki/wiki/Deve...	Code statistics	1	relativo	404
22 https://www.media.wiki.org/wiki/MediaWiki/wiki/Do w...	Get MediaWiki	1	relativo	404
23 https://www.media.wiki.org/wiki/MediaWiki/wiki/File...		1	relativo	404
24 https://www.media.wiki.org/wiki/MediaWiki/wiki/Help...		1	relativo	404
25 https://www.media.wiki.org/wiki/MediaWiki/wiki/Help...		1	relativo	404
26 https://www.media.wiki.org/wiki/MediaWiki/wiki/Help...	learn more	1	relativo	404
27 https://www.media.wiki.org/wiki/MediaWiki/wiki/Help...		1	relativo	404
28 https://www.media.wiki.org/wiki/MediaWiki/wiki/How...		1	relativo	404
29 https://www.media.wiki.org/wiki/MediaWiki/wiki/Medi...		1	relativo	404

Showing 1 to 29 of 166 entries. 5 total columns

```
#Pregunta 1.5 de la lista anterior se tiene que iterar y obtener los valores del URL

url_test <- httr::GET(url)
url_test$status_code

## NORMALIZANDO la información de HREF, convirtiendo una URL Relativa a absoluta (tipo1)
df2 <- df2 %>%
# mutate(HREF = ifelse(stringr::str_starts(HREF, "^/", negate = FALSE),
# paste0(url, gsub("/", "/", HREF)),
# HREF), url_type = "relativo")

## NORMALIZANDO la información de HREF, convirtiendo una URL Relativa a absoluta (tipo2)
df2 <- df2 %>%
mutate(
url_type = case_when(
stringr::str_starts(HREF, "^/", negate = FALSE) ~ "relativo",
stringr::str_starts(HREF, "^\\w+", negate = FALSE) ~ "relativo",
grepl("^#", HREF) ~ "relativo",
TRUE ~ "absoluto"
),
HREF = case_when(
stringr::str_starts(HREF, "^/", negate = FALSE) ~ paste0(url, gsub("/", "/", HREF)),
stringr::str_starts(HREF, "^\\w+", negate = FALSE) ~ paste0(url, HREF),
grepl("^#", HREF) ~ paste0(url, HREF),
TRUE ~ HREF
)
)

# ...

# Agregar la columna responseCode a df2
df2$responseCode <- NA

# Iterar sobre las filas de df2
for (i in 1:nrow(df2)) {
row <- df2[i,]
url_local <- row$HREF

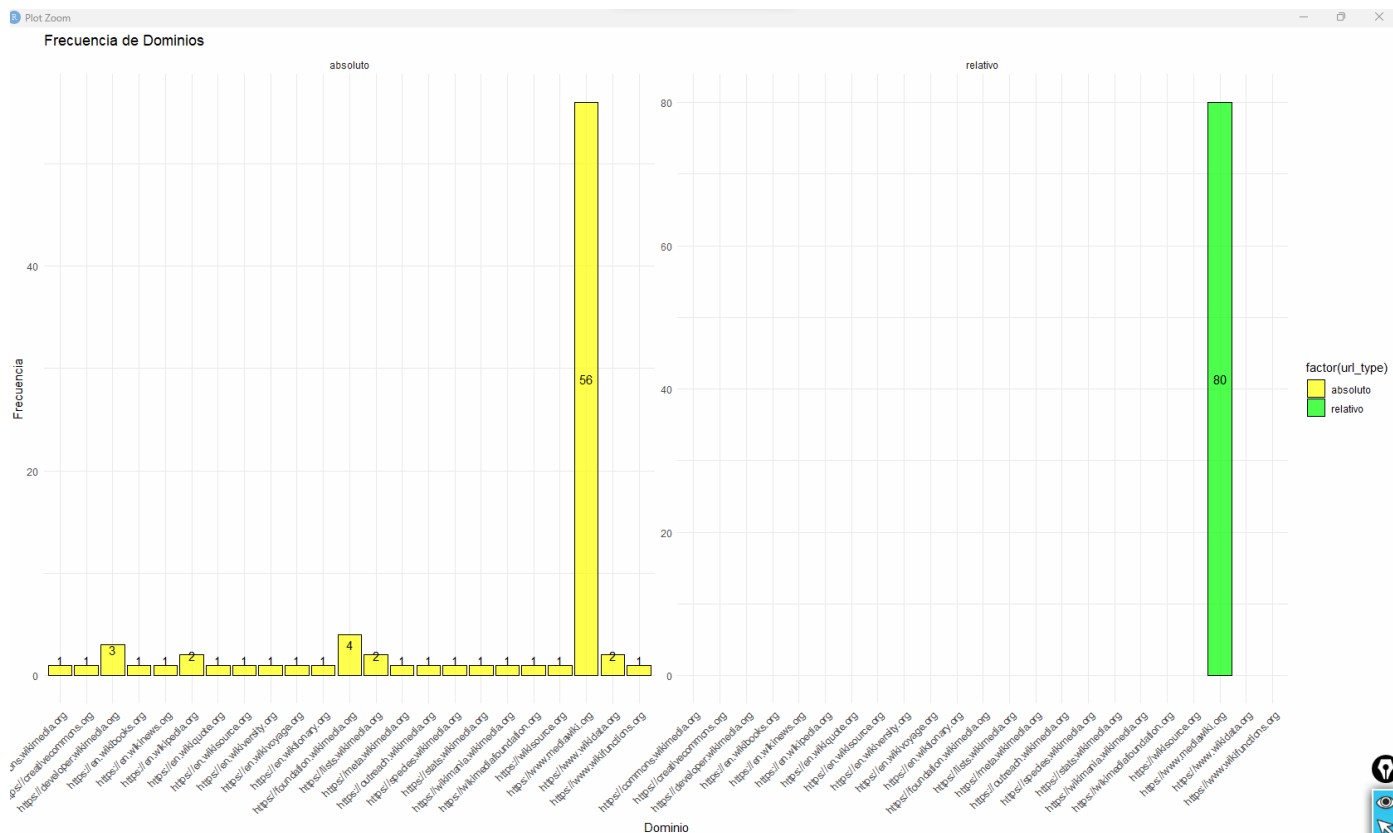
# Realizar la solicitud HTTP y almacenar el código de respuesta
if (!is.na(url_local)) {
if (startswith(url_local, "https")) {
response <- httr::GET(url_local)
df2[i, "responseCode"] <- response$status_code
}
}
```

(II) Pregunta 2

Elaborar, usando las librerías de gráficos base y qplot (ggplot2), una infografía sobre los datos obtenidos. Tal infografía será una reunión de gráficos donde se muestren los siguientes detalles:

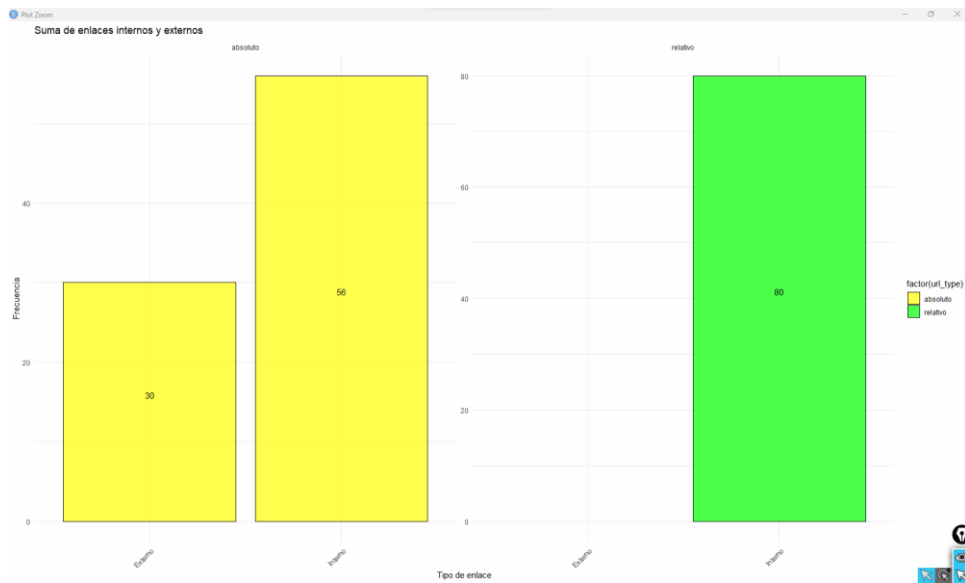
1. Pregunta #1:

Un histograma con la frecuencia de aparición de los enlaces, pero separado por URLs absolutas (con “http...”) y URLs relativas.



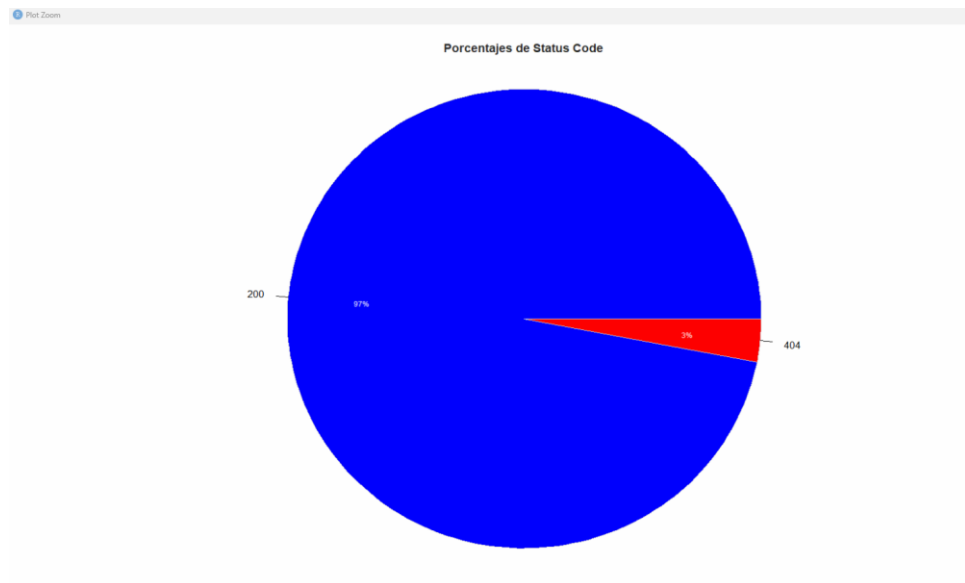
2. Pregunta #2:

Un gráfico de barras indicando la suma de enlaces que apuntan a otros dominios o servicios (distinto a <https://www.mediawiki.org> en el caso de ejemplo) vs. la suma de los otros enlaces.



3. Pregunta #3:

Un gráfico de tarta (pie chart) indicando los porcentajes de Status de nuestro análisis.



(III) Repositorio en GITHUB

URL: <https://github.com/vanisheriii/barcelona-data-science-2>

< Fin del documento >