iGraph_URLs_SEO.R

Ana Laguna

4 de diciembre de 2017

Visualizing your site's Internal Linking

https://data-seo.com/2015/07/07/r-tools-seo-part-1/

https://stackoverflow.com/questions/47641684/labels-on-only-root-and-terminal-vertices-in-igraph-r

http://kateto.net/netscix2016

```
library(igraph)
```

```
##
## Attaching package: 'igraph'
```

```
## The following objects are masked from 'package:stats':
##
##     decompose, spectrum
```

```
## The following object is masked from 'package:base':
##
##     union
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:igraph':
##
##     as_data_frame, groups, union
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(sqldf)
```

```
## Loading required package: gsubfn
```

```
## Loading required package: proto
```

```
## Loading required package: RSQLite
```

```
file_csv <- "/Users/alaguna/Desktop/Ana/DATOS externos/all_inlinks.csv"
# import semRush
## skip first line
DF <- read.csv("/Users/alaguna/Desktop/Ana/DATOS
externos/all_inlinks.csv", header=TRUE, sep = ";", stringsAsFactors = F )

head(DF)
```

```
##    Type
## 1 AHREF
## 2 AHREF
## 3 AHREF
## 4 AHREF
## 5 AHREF
## 6 AHREF
##
Source
## 1  https://www.bbva.es/productos/ficha-planes/bbva-plan-mercado-
monetario-ppi/7661
## 2  https://www.bbva.es/productos/ficha-planes/bbva-plan-mercado-
monetario-ppi/7661
## 3  https://www.bbva.es/productos/ficha-planes/bbva-plan-mercado-
monetario-ppi/7661
## 4  https://www.bbva.es/productos/ficha-planes/bbva-plan-mercado-
monetario-ppi/7661
## 5 https://www.bbva.es/bancapersonal/ahorroeinversion/planes-de-
pensiones/index.jsp
## 6 https://www.bbva.es/bancapersonal/ahorroeinversion/planes-de-
pensiones/index.jsp
##
Destination
## 1 https://www.bbva.es/productos/ficha-planes/bbva-plan-mercado-
monetario-ppi/7661
## 2 https://www.bbva.es/productos/ficha-planes/bbva-plan-mercado-
monetario-ppi/7661
## 3 https://www.bbva.es/productos/ficha-planes/bbva-plan-mercado-
monetario-ppi/7661
## 4 https://www.bbva.es/productos/ficha-planes/bbva-plan-mercado-
monetario-ppi/7661
## 5 https://www.bbva.es/productos/ficha-planes/bbva-plan-mercado-
monetario-ppi/7661
## 6 https://www.bbva.es/productos/ficha-planes/bbva-plan-mercado-
monetario-ppi/7661
##    Size Alt_Text                            Anchor Status_Code Status
## 1 87026                                    Espanol         200     OK
## 2 87026                                    Catalan         200     OK
## 3 87026                                    English         200     OK
## 4 87026          BBVA Plan Mercado Monetario, PPI          200     OK
## 5 87026                                  Informate         200     OK
## 6 87026                                  Informate         200     OK
```

```
##   Follow
## 1   true
## 2   true
## 3   true
## 4   true
## 5   true
## 6   true
```

```
dim(DF)
```

```
## [1] 17832     9
```

```
dim(unique(DF))
```

```
## [1] 16463     9
```

```
file_outlinks         <- '/Users/alaguna/Desktop/Ana/DATOS
externos/all_inlinks.csv'
website_url          <- 'https://www.bbva.es'

## we keep only link
#DF <- DF[DF$Type=="HREF",]
DF <- select(DF,Source,Destination)
DF <- as.data.frame(sapply(DF,gsub,pattern=website_url,replacement=""))
DF <- as.data.frame(sapply(DF,gsub,pattern="\"",replacement=""))
## delete subdomain
DF <- subset(DF, !grepl("^http", DF$Source))
DF <- subset(DF, !grepl("^http", DF$Destination))
## adapt colnames and rownames
colnames(DF) <- c("From","To")
rownames(DF) <- NULL

DF <- unique(DF)
# generate graph with data.frame
graphObject = graph.data.frame(DF, directed = TRUE)
# to run pagerank we need a simple, undirected graph
graphObject = simplify(as.undirected(graphObject))


head(DF)
```

```
##
From
## 1           /productos/ficha-planes/bbva-plan-mercado-monetario-
ppi/7661
## 5          /bancapersonal/ahorroeinversion/planes-de-
pensiones/index.jsp
## 7           /bancaprivada/ahorroeinversion/planes-de-
pensiones/index.jsp
## 9           /particulares/ahorro-inversion/planes-de-
pensiones/index.jsp
```

```
## 11           /autonomos/ahorro-e-inversion/planes-de-
pensiones/index.jsp
## 15 /general/banca-online/movimientos-previstos-gastos-
ingresos/index.jsp
##
To
## 1           /productos/ficha-planes/bbva-plan-mercado-monetario-
ppi/7661
## 5           /productos/ficha-planes/bbva-plan-mercado-monetario-
ppi/7661
## 7           /productos/ficha-planes/bbva-plan-mercado-monetario-
ppi/7661
## 9           /productos/ficha-planes/bbva-plan-mercado-monetario-
ppi/7661
## 11          /productos/ficha-planes/bbva-plan-mercado-monetario-
ppi/7661
## 15 /general/banca-online/movimientos-previstos-gastos-
ingresos/index.jsp
```

```
V(graphObject)$name = V(graphObject)


plot(get.edgelist(graphObject))
```

```
#edgest
plot(E(graphObject))
```

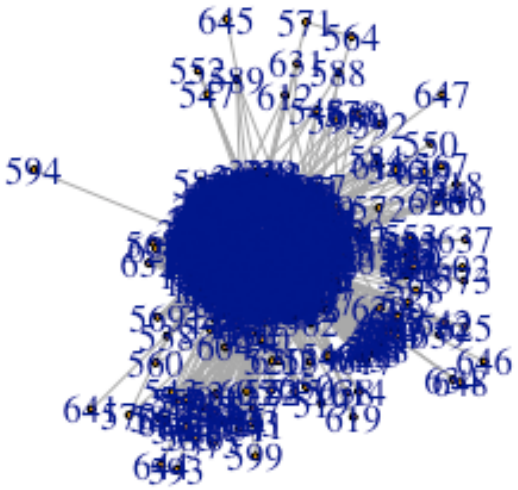
```
E(graphObject)$name
```

```
## NULL
```

```
#vertices
plot(V(graphObject))
```

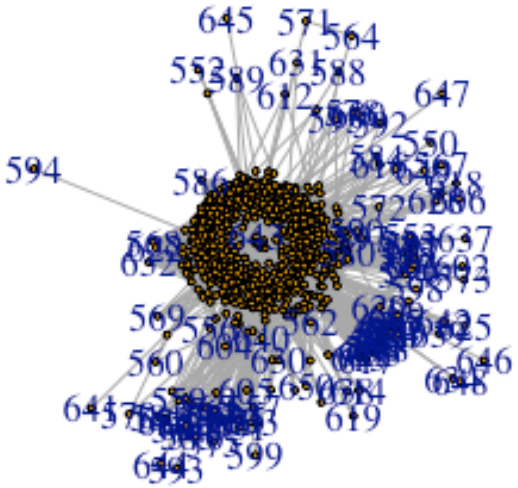
```
e = get.edgelist(graphObject)

# Root vertices are in first column but not in second column
root = setdiff(e[,1],e[,2])

# Terminal vertices are in second column but not in first column
terminal = setdiff(e[,2], e[,1])

# Vertices to remove are not in root or terminal vertices
remove = setdiff(unique(c(e)), c( terminal))

# Remove names of intermediate vertices
```

```r
V(graphObject)$name[V(graphObject)$name %in% remove] = ""

g = set.vertex.attribute(graphObject, "name", value=c(1,rep("",
length(2:11)),12:21))

## Warning in vattrs[[name]][index] <- value: número de items para para
## sustituir no es un múltiplo de la longitud del reemplazo

## FIJO UN NUMERO DE VERTICES
# summary(as.integer(terminal))
#    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#   387.0   484.5   554.0   544.2   604.8   650.0


set.seed(2)
plot(graphObject,
     layout=layout.fruchterman.reingold,
     vertex.size        = 3,          # Smaller nodes
     vertex.label       =   V(graphObject)     # Set the labels)
     )
```

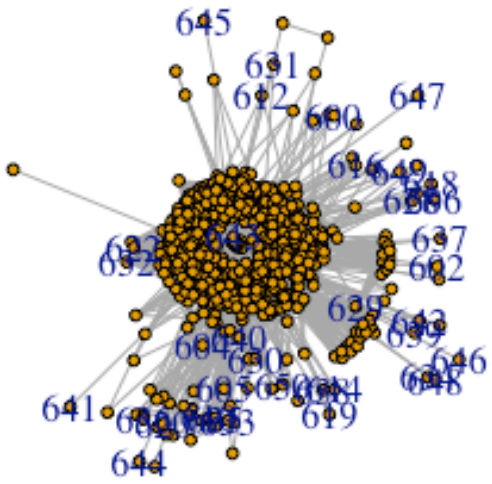



```r
set.seed(2)
plot(graphObject,
```

```
     layout=layout.fruchterman.reingold,
     vertex.size        = 3,          # Smaller nodes
     vertex.label       = ifelse(V(graphObject) >= 550 ,  V(graphObject),
"")      # Set the labels)
     )
```




```
set.seed(2)
plot(graphObject,
     layout=layout.fruchterman.reingold,
     vertex.size        = 5,          # Smaller nodes
     vertex.label       = ifelse(V(graphObject) >= 600 ,
V(graphObject)$name, "")      # Set the labels)
     )
```

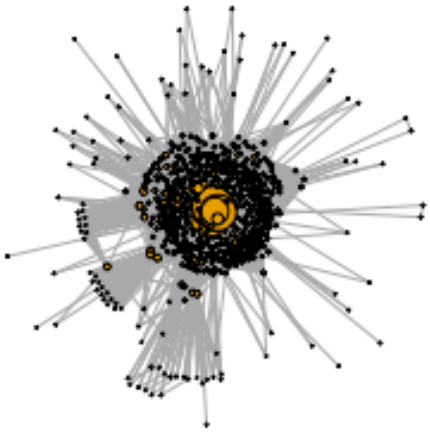



```
dataset_far_nodes <- ifelse(V(graphObject) >= 544 ,  V(graphObject)$name,
"")
clean_dataset_far_nodes <- unique(dataset_far_nodes)
```

PageRank

```
map <- function(x, range = c(0,1), from.range=NA) {
  if(any(is.na(from.range))) from.range <- range(x, na.rm=TRUE)

  ## check if all values are the same
  if(!diff(from.range)) return(
    matrix(mean(range), ncol=ncol(x), nrow=nrow(x),
           dimnames = dimnames(x)))

  ## map to [0,1]
  x <- (x-from.range[1])
  x <- x/diff(from.range)
  ## handle single values
  if(diff(from.range) == 0) x <- 0

  ## map from [0,1] to [range]
  if (range[1]>range[2]) x <- 1-x
  x <- x*(abs(diff(range))) + min(range)
```

```
  x[x<min(range) | x>max(range)] <- NA

  x
}

# calculate pagerank
pr <- page.rank(graphObject,directed=TRUE)
# print graph with size node linked with pagerank
plot(graphObject,
     layout=layout.fruchterman.reingold,
     vertex.size        = map(pr$vector, c(1,20)),
     vertex.label       = NA,
     vertex.label.color = "black",
     edge.arrow.size=.2
)
```



It is very easy to notice that pagerank is badly distributed and internnal linking is unbalanced.