Web Scraping Project

Iñigo Portillo, Arantxa Urdangarin

May 27, 2017

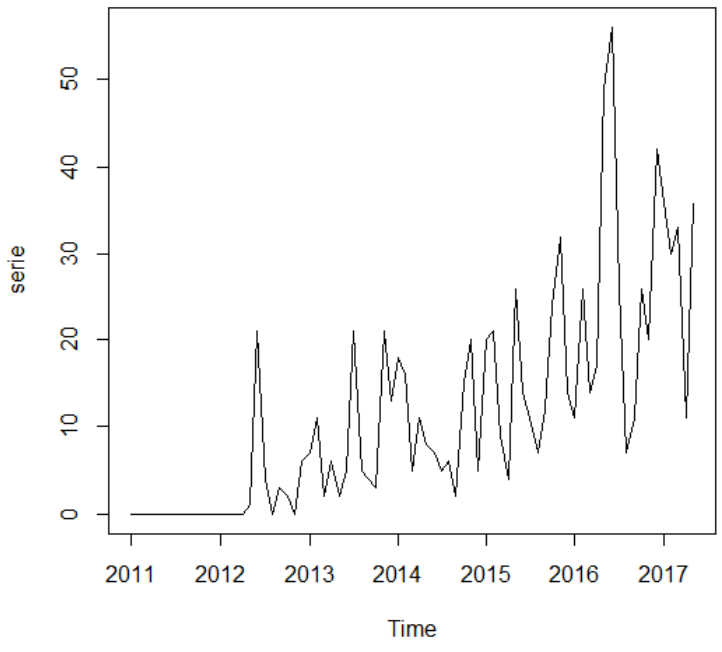
The goal of this work is to analize the temporal structure and the titles of the news related with the residue incinerator of Zubieta (Gipuzkoa) which construction started some days ago. During 2011-2017 have been a lot of controversy in this point between politics and the people who live in Gipuzkoa. Most people and doctors think that the residue incineraton is not good for the health of the people who lives near it, also, they think that is not good for the environment. Those people, tried to implement some different ways to recycle the residues, giving some ideas and implementing them in some cities. However, this methods to collect residuals hadn't been successful. Some days ago, the Parliament of Gipuzkoa approbed the construction of the incinerator.

First of all, we took a newspaper called *Naiz*, then we found every notice related to the residual incinerator using the search engine. Once we saw the structure of those pages we *Web-scrapped* the information we needed. In our case, *dates* and *titles* of every new. To finish, we plotted a time serie where the quantity of news in each month were grouped. Also, using the *twitteR* package we drew a wordcloud, to see the most common words on those titles.

To scrap the dates, firstly we used a loop to extract the dates of every new in every page. Then, we converted those dates from character to date format. To finish we grouped every date in months and we plotted the time serie.

To extract the titles of the news we followed the same procedure. Using a loop we extracted the titles of every page. Once we got the titles we defined the forbidden words, those we want to remove from titles once we started constructing the *wordcloud*. Then using the *twitteR* package, we contructed the *wordcloud*.

These are the results obtained:



We can observe that news related to the residual incinerator have been growing up significantly since 2012. In june of 2016, we can see that had been a lot of news. The reason could be that in that month had been a lot of public protest against the project.

We obtain the following results from the analysis of the words of the titles.

We can see in the *wordcloud* that the most common words of the titles are *Gipuzkoa*, followed by the name of the place where the incinerator will be constructed *zubieta, donostia*, some associations against the project such as *ghk, gurasos*, the names of politic parties that are in favour and against the proyect *bildu, pnv, pse* and some keywords *contra, residuos, erraustegiaren, diputación....* We can conclude that the news about the movement of the people agains the proyect have had more impact.

Finally, we have had some difficulties when we wanted to select those news talking about the incinerator. We found them using the search engine of the webpage instead of using *R*.

Appendix

```
############## Assessment-Web Scraping Project #############
########## Iñigo Portillo and Arantxa Urdangarin ##########

# Analysis of dates
install.packages("rvest")
require(rvest)
vectori <- list()
for (i in 1:18){
urlweb <- paste0("http://www.naiz.eus/es/actualidad/busqueda/page/",
i,"?order_criteria=date&per_page=50&query=incineradora&scope=all&search=")
webpage <- read_html(urlweb)
TAB <- webpage %>%
html_nodes(".date") %>%
html_text()
vectori[[i]] <- as.vector(TAB)
}

dates <- 1:(17*50+length(vectori[[18]]))
class(dates) <- "Date"
for(i in 1:length(vectori)){
for(j in 1:length(vectori[[i]])){
dates[length(vectori[[i]])*(i-1)+j] <- as.Date(vectori[[i]][j], "%d/%m/%Y")
}
}

serie <- c()
for(i in 2011:2017){
for(j in 1:12){
valor <- 0
for(k in 1:length(dates)){
if(as.numeric(format(dates[k], "%Y"))==i & as.numeric(format(dates[k], "%m"))==j){
valor <- valor+1
}
}
serie[12*(i-2011)+j] <- valor
}
}

serie <- serie[-(78:84)]
(serie=ts(serie,start=2011,freq=12))
plot(serie)


# Analysis of words
if (!require(twitteR)) install.packages("twitteR")
if (!require(ROAuth)) install.packages("ROAuth")
if (!require(httr)) install.packages("httr")
if (!require(tm)) install.packages("tm")


require(tm)
require(ROAuth)
require(httr)
require(twitteR)
vectortit <- list()
for (i in 1:18){
urlweb <- paste0("http://www.naiz.eus/es/actualidad/busqueda/page/",i,
"?order_criteria=date&per_page=50&query=incineradora&scope=all&search=")
webpage <- read_html(urlweb)
```

```
TAB2 <- webpage %>%
html_nodes(".content-text a") %>%
html_text()
vectortit[[i]] <- as.vector(TAB2)
}

stopwords <- c("cuatro", "nos", "eta", "dugu", "die", "los", "<<el nos", "sin", "que", "egin",
"las", "buruz", "una", "ditu", "ala", "≪no", "del", "28an", "egingo", "sobre",
"una","más","por","≪No","con","para","dute","sus","dos","las","<<el","tras",
"<<la","hay","pese","tres","san","como","dituzte","dos","antes","los")
tweets.text.corpus <- Corpus(VectorSource(vectortit))
tweets.text.corpus <- tm_map(tweets.text.corpus, function(x) removeWords(x, stopwords))


if (! require(wordcloud) ) install.packages("wordcloud")
require("wordcloud")

wordcloud(tweets.text.corpus, min.freq = 3, scale=c(7,0.5),
colors=brewer.pal(8, "Dark2"),random.color= TRUE, random.order = FALSE,
max.words = 150)
```