Project 2 – Automated Data Collection

Code

Soufiane Fadel (5314E)

Introduction

In this project, we are collecting data from various online sources in order to build a collection of 5 text corpora, each one consisting of documents written in a different languages:

- *English* : collected from the United Kingdom's Government website.
- *French* : collected from Wikipedia.
- *Spanish* : collected from twitter.
- *Italian* : collected from the PDF document 'Giovannino Guareschi's Tutto don Camillo'.
- *Germany* : collected from the n-tv German free-to-air television news channel.

Our Goal is to automate the task of scrapping this text corpora in one peace of code in order to produce one final dataset that consist of all of the observations (text) placed in rows andeach row associated with a specific language code ("Eng", "Fra", "Esp", "Ita", "Ger").

To do so we need to proceed according to workflow with following parts:

Loading R libraries

Hide

Hide

```
library(bitops)
library(RCurl)
library(XML)
library(stringr)
library(stringi)
library(rvest)
library(magrittr)
library(xml2)
library(curl)
library(twitteR)
library(purrr)
library(tidytext)
library(dplyr)
library(tidyr)
library(lubridate)
library(scales)
library(broom)
library(pdftools)
```

English Text

We are scrapping all the UK Government press releases from "News and Communications" page (https://www.gov.uk/) published in 2018

(from Jan 1, 2018 to Dec 31, 2018). The first step is to capture the URLs of all the required press releases and then download the press releases to a local folder in order extract the main text of each press release and save it to a string.

Back to top of notebook

Codes of English text

Hide
Hide

```
ENG_TEXT <- function() {

signatures = system.file("CurlSSL", cainfo = "cacert.pem", package = "RCurl")
all_sub_links_articles <- character()    # initialize
sub_link_page <- 'search/news-and-communications?public_timestamp%5Bfrom%5D=01%2F01%2F2018&public_timestamp%
5Bto%5D=01%2F01%2F2019&order=updated-newest'
count <- 0
while( length(sub_link_page)>0  ){
  full_link_page <- str_c("https://www.gov.uk/", sub_link_page)
  html_page <- getURL(full_link_page, cainfo = signatures)
  html_page_tree <- htmlParse(html_page)
  sub_links_of_articles = xpathSApply(html_page_tree, "//li[@class='gem-c-document-list__item  ']//a", xmlGe
tAttr, "href")
  all_sub_links_articles <- c(all_sub_links_articles,sub_links_of_articles)
  sub_link_page <- xpathSApply(html_page_tree,  "//li[@class= 'gem-c-pagination__item gem-c-pagination__item
--next']//a", xmlGetAttr,"href")
  count <- count +1
}

# Download all press releases in a the Folder 'Press_Releases_Eng' that you need to create before.
dir.create("Press_Releases_Eng")
for(i in 1:length(all_sub_links_articles)){
    url <- str_c("https://www.gov.uk", all_sub_links_articles[i])     # visit URL i
    tmp <- getURL(url, cainfo = signatures)              # get the HTML at URL i
    write(tmp, str_c("Press_Releases_Eng/", i, ".html"))     # write HTML to Press_Releases/i.html
}

Eng_text <- character()

for(i in 1:length(list.files("Press_Releases_Eng")) ){
  tmp <- readLines(str_c("Press_Releases_Eng/", i, ".html"))
  tmp <- str_c(tmp, collapse = "")
  tmp <- htmlParse(tmp)
  release <- xpathSApply(tmp, "//div[@class='govspeak']", xmlValue)
  Eng_text <- c(Eng_text,release)
}


return(c(count,Eng_text))
}

Eng <- ENG_TEXT()
```

Back to top of notebook

Testing the English text Result

Hide
Hide

```
number_of_articles <- length(Eng[2:length(Eng)])
number_of_pages <- Eng[1]
sample_of_Eng_text <-  Eng[3:4]

cat("> the total number of articles is : ", number_of_articles, '\n','\n','\n')
```

```
> the total number of articles is :  9183
```

Hide

Hide

```
cat("> the total number of pages containing articles is : ",number_of_pages,'\n','\n','\n')
```

```
> the total number of pages containing articles is :  460
```

Hide

Hide

```
cat("> Here is some samples from the English text : ", '\n','\n',sample_of_Eng_text)
```

```
> Here is some samples from the English text :

 Government delivers on manifesto commitment to protect millions from unjustified price rises as energy pric
e cap comes into force today  new cap will mean 11 million loyal energy customers on “default” tariffs pay a
fair price for their energy  part of the government’s commitment to tackle fuel poverty and protect consumer
sAround 11 million households who have stayed loyal to energy suppliers on poor value energy tariffs will pa
y a fair price from today (1 January 2019) thanks to the government’s price cap.The cap will bring down the
amount consumers have been overpaying to energy companies, including the Big Six, by £1 billion a year, star
ting this winter when households are typically using more energy to heat and light their homes. It will rema
in in place until at least 2020, while energy suppliers and industry continue to work with the energy regula
tor Ofgem and the government to build an energy market that works better for all consumers.Prime Minister Th
eresa May said:  Our energy price cap will cut bills for millions of families and people across the UK who h
ave been ripped off by energy companies for far too long. From today, money will go straight back into the p
ockets of loyal consumers, including the elderly and those on lower incomes who feel the pinch more acutely.
But work to tackle this issue doesn’t stop there. We’re working with regulators and industry to ensure that
consumers are not unfairly overcharged in the future – whether on their phone bills or their insurance premi
ums.Energy and Clean Growth Minister Claire Perry said:  Today marks the end of unjustified price rises on e
nergy bills as this government delivers on time on its promise to protect millions of households from poor v
alue deals, especially the vulnerable.  For too long, suppliers have failed to pass on any savings to their
customers, who deserve to pay a fair price for their gas and electricity. Switching supplier is still the be
st way to find a better deal, but that doesn’t mean customers should be punished for their loyalty.  Bill pa
yers can now be confident that any change to the price cap will be a fair representation of the actual costs
of energy, rather than suppliers passing on inefficiencies to their customers or as excess profits.Following
a consultation, Ofgem set the price cap level at £1,137 per year for a typical dual fuel customer paying by
direct debit. The amount customers will pay depends on how much energy they actually use, as the price cap s
ets a limit on how much suppliers can charge per unit of gas and electricity not on overall energy bills.The
cap will also protect around one million households who receive the Warm Home Discount currently protected b
y Ofgem’s safeguard tariff. Ofgem already caps energy prices through its safeguard tariff for 4 million hous
eholds on pre-payment meters.Ofgem will review the level of the cap every 6 months taking into account any c
hanges to the actual costs of providing gas and electricity to energy customers. The first review will take
place in early February coming into effect on 1 April 2019.Dermot Nolan, chief executive at Ofgem, said:  Un
der the cap, Ofgem will protect consumers from being overcharged and ensure they pay a fair price to heat an
d light their homes. Consumers can have confidence that any rise in prices in the future will only be down t
o genuine increases in energy costs rather than supplier profiteering while falls in energy costs will alway
s be passed on to them.  Households who are protected by the cap will be able to save even more money by sho
pping around for a better deal. In the meantime Ofgem will continue with reforms which aim to deliver a smar
ter, more competitive energy market which, combined with protection for those who need it, works for all con
sumers.The Domestic Gas and Electricity Act, which passed Parliamentary scrutiny and became law on 19 July 2
018, put in place a requirement on Ofgem to cap standard variable and default energy tariffs after the Compe
tition and Markets Authority (CMA) found consumers had been overpaying the Big Six an average of £1.4 billio
n a year.While the temporary cap is in place, energy suppliers and industry will continue to work with Ofgem
and government to build an energy market that works better for all consumers, ensuring they get the best ser
vice for a fair price so that everyone reaps the benefits of the move to a smarter, more digital economy.Oth
er measures designed to deliver the government’s objective of clean, affordable and innovative energy, while
tackling fuel poverty, as part of our modern Industrial Strategy include:the rollout of smart meters  initia
tives to promote smarter and faster switching  a joint review with Ofgem on the future of the retail marketN
otes to editors    Ofgem already caps energy prices through its safeguard tariff for 4 million households on
pre-payment meters. This cap was extended to a further one million vulnerable households in receipt of the W
arm Home Discount, a £140 discount on winter energy bills, early in 2018. These Warm Home Discount household
s will be transferred on 31 December to the new price cap for direct debit customers, which has been designe
d for those on standard, rather than prepayment, meters.        The new cap will protect households in every
region of Great Britain who are on standard variable or default fixed-term tariffs. Around 60% of households
currently pay for their energy on these tariffs.        The cap will remain in place until at least 2020 and
can be extended until end of 2023 if the conditions for effective competition are not in place.        For m
ore information on the energy price cap, including the pre-payment meter cap, visit Ofgem’s Energy price cap
webpages.   The Patent Cooperation Treaty (PCT) fee structure changed on 1 January 2019 and are as follows:T
ransmittal fee:£75Search fee:£1,576International fee:£1,037 for the first 30 sheets  £12 for each sheet over
30Restoration for restoration of priority£150Reductions for e-filing£156: electronic filing (not being in ch
aracter coded format)  £234: electronic filing (being in character coded format)Fees for preparation of prio
rity document£20PCT formsPCT Request Form (RO/101)  PCT Request Form (RO/101) - completed example  PCT Deman
d Form (IPEA/401)  PCT Power of Attorney (POA)Further informationprotecting your patent abroad  protecting y
our inventions abroad: PCT frequently asked questions
```

Back to top of notebook

French Text

Here we are scrapping the Actrices françaises page on (French Wikipedia):
https://fr.wikipedia.org/wiki/Cat%C3%A9gorie:Actrice_fran%C3%A7aise . We will identify and capture the URLs which yields French actresses whose family name (or last name) starts with an “L” an “M” and then we will extract the main text of each entry and save it to a

string.

Back to top of notebook

Codes of French text

Hide

Hide

```
Fr_TEXT <- function(){

signatures = system.file("CurlSSL", cainfo = "cacert.pem", package = "RCurl")



url_1 <- 'https://fr.wikipedia.org/w/index.php?title=Cat%C3%A9gorie:Actrice_fran%C3%A7aise&from=L'
html_page <- getURL(url_1, cainfo = signatures)
html_page_tree <- htmlParse(html_page)
sub_links_of_actress_1 = xpathSApply(html_page_tree, "//div[@id='mw-pages']//div[@class='mw-category-group']
//ul//li//a", xmlGetAttr, "href")

url_2 <- 'https://fr.wikipedia.org/w/index.php?title=Cat%C3%A9gorie:Actrice_fran%C3%A7aise&subcatfrom=L&file
from=L&pagefrom=Lesache%2C+Bernadette%0ABernadette+Le+Sach%C3%A9#mw-pages'
html_page <- getURL(url_2, cainfo = signatures)
html_page_tree <- htmlParse(html_page)
sub_links_of_actress_2 = xpathSApply(html_page_tree,'/html/body/div[3]/div[3]/div[4]/div[2]/div[2]/div[2]/di
v/div[1]/ul/li/a', xmlGetAttr, "href")

url_3 <- 'https://fr.wikipedia.org/w/index.php?title=Cat%C3%A9gorie:Actrice_fran%C3%A7aise&from=M'
html_page <- getURL(url_3, cainfo = signatures)
html_page_tree <- htmlParse(html_page)
sub_links_of_actress_3 = xpathSApply(html_page_tree, "//div[@id='mw-pages']//div[@class='mw-category-group']
//ul//li//a", xmlGetAttr, "href")

url_4 <- 'https://fr.wikipedia.org/w/index.php?title=Cat%C3%A9gorie:Actrice_fran%C3%A7aise&pagefrom=Meurisse
%2C+Nina%0ANina+Meurisse&subcatfrom=M&filefrom=M#mw-pages'
html_page <- getURL(url_4, cainfo = signatures)
html_page_tree <- htmlParse(html_page)
sub_links_of_actress_4 = xpathSApply(html_page_tree,'/html/body/div[3]/div[3]/div[4]/div[2]/div[2]/div[2]/di
v/div[1]/ul/li/a', xmlGetAttr, "href")

sub_links_of_actress <- c(sub_links_of_actress_1,sub_links_of_actress_2,sub_links_of_actress_3,sub_links_of_
actress_4)



# Download all press releases
dir.create("Press_Releases_Fr")
for(i in 1:length(sub_links_of_actress)){
    url <- str_c("https://fr.wikipedia.org", sub_links_of_actress[[i]])     # visit URL i
    tmp <- getURL(url, cainfo = signatures)              # get the HTML at URL i
    write(tmp, str_c("Press_Releases_Fr/", i, ".html"))     # write HTML to Press_Releases/i.html
}


Fr_text <- character()
for(i in 1:length(list.files("Press_Releases_Fr")) ){
  tmp <- readLines(str_c("Press_Releases_Fr/", i, ".html"))
  tmp <- str_c(tmp, collapse = "")
  tmp <- htmlParse(tmp)
  release <- xpathSApply(tmp, "//div[@class='mw-parser-output']//p", xmlValue)
  release <- str_c(release, collapse = " ")
  Fr_text <- c(Fr_text,release)
}

return(Fr_text)
}


Fr <- Fr_TEXT()
```

Back to top of notebook

Testing the French text Result

Hide

Hide

```
number_of_actresses <- length(Fr)
sample_of_Fr_text <-  Fr[2:4]

cat("> the number of French actresses is : ", number_of_actresses, '\n','\n','\n')
```

```
> the number of French actresses is :  619
```

Hide

Hide

```
cat("> Here is a sample of the French text : ",'\n',sample_of_Fr_text, fill = 2)
```

```
> Here is a sample of the French text :


modifier  Barbara Laage est une actrice française, née le 30 juillet 1920 à Menthon-Saint-Bernard (Haute-Sav
oie) et morte le 19 mai 1988 à Deauville (Calvados). Elle fait ses débuts à l'écran en 1942 avec le film Sig
né illisible. Pressentie pour le premier rôle dans La Dame de Shanghaï[2], Orson Welles confiera finalement
l'interprétation à sa femme, Rita Hayworth. Barbara Laage décroche son premier rôle international avec le pe
rsonnage d’Eugenia Taris dans L'Indomptée (1948) aux côtés de Van Heflin, et accède à la célébrité avec son
interprétation de Lizzie McKay dans l'adaptation cinématographique de la pièce de Jean-Paul Sartre, La Putai
n respectueuse (1952). Surtout à l'aise dans le cinéma américain (ainsi aux côtés de Kirk Douglas dans Un ac
te d'amour d’Anatole Litvak, ou partenaire de Gene Kelly dans La Route joyeuse), elle se produit en France e
ntre autres pour Alex Joffé (Les Assassins du dimanche, 1956). En 1960, elle apparaît aux côtés de Paul Newm
an dans Paris Blues de Martin Ritt, ainsi que dans deux productions allemandes : Les Mille-et-une Nuits avec
Karl Lieffen et Georg Jacoby, et Une nuit à Monte Carlo avec Eddie Constantine. Dans Domicile conjugal de Fr
ançois Truffaut (1970), elle interprète une scène d'anthologie que Woody Allen adaptera dans Annie Hall : al
ors que Claude Jade parle dans les escaliers avec sa voisine de sa vie conjugale avec Jean-Pierre Léaud, ce
dernier échange simultanément sur le même sujet au café d’en-face avec une collègue de bureau (jouée par Bar
bara Laage) ; au montage, le son et les images des deux dialogues sont intervertis.  Au cours des années 197
0, elle interprète encore deux seconds rôles : Défense de savoir aux côtés de Jean-Louis Trintignant et Proj
ection privée face à Jane Birkin. Sa dernière apparition à l'écran vient en 1976 avec Une place forte de Guy
Jorré. Elle s'établit avec sa sœur Christiane à Deauville en 1985, où réside un de ses filleuls[3]. C'est là
qu'elle meurt trois ans plus tard à 67 ans, quelques jours après s'être étouffée au cours d'un repas[3]. Ses
cendres se trouvent au crématorium de Trouville-sur-Mer[3].
Vous pouvez partager vos connaissances en l’améliorant (comment ?) selon les conventions filmographiques. mo
difier  Lou de Laâge, née le 27 avril 1990 à Bordeaux[1], est une actrice française. Née d'un père journalis
te à Sud Ouest (1958-2018) et d'une mère peintre, elle passe sa jeunesse entre Bordeaux et Montendre (Charen
te-Maritime). Après un bac littéraire, elle intègre l'école de théâtre parisienne Claude Mathieu pour deveni
r actrice[1]. Elle fait ses débuts au cinéma en 2011 dans la comédie J'aime regarder les filles. Elle joue l
a jeune fille dont le héros, incarné par Pierre Niney, tombe amoureux. La même année, elle est la tête d'aff
iche du film à petit budget, Nino (une adolescence imaginaire de Nino Ferrer), écrit et réalisé par Thomas B
ardinet. En 2013, elle fait partie de la distribution de projets exposés immédiatement : le drame hippique J
appeloup, réalisé par Christian Duguay, écrit et interprété par Guillaume Canet. Elle y a aussi pour partena
ires Marina Hands et Daniel Auteuil. Sa performance lui vaut le César 2014 du meilleur espoir féminin. Elle
fait aussi partie du casting choral de la comédie dramatique Des gens qui s'embrassent, réalisée par Danièle
Thompson. En 2014, elle partage l'affiche du drame intimiste Respire, second long-métrage de Mélanie Laurent
, avec Joséphine Japy. Cette fois, elle est nommée au César 2015 du meilleur espoir féminin. En 2015, elle s
'investit dans deux projets franco-italiens : elle partage l'affiche du drame indépendant Le Tournoi avec Mi
chelangelo Passaniti, puis seconde Juliette Binoche, actrice principale de L'Attente, co-écrit et réalisé pa
r Piero Messina. L'année suivante, elle incarne l'héroïne du drame historique Les Innocentes, réalisé par An
ne Fontaine. Elle remporte cette fois le Prix Romy-Schneider 2016. Elle reste dans le registre historique po
ur le drame portugais Le Cahier noir, réalisé par Valeria Sarmiento. L'année 2019 est marquée par la sortie
de la satire Blanche comme neige, nouveau projet d'Anne Fontaine. Elle y a pour principale partenaire Isabel
le Huppert. En juin de la même année elle fait partie du jury de Sandrine Bonnaire lors du 33e Festival du f
ilm de Cabourg. Sur les autres projets Wikimedia :
modifier  Lââm, née le 1er septembre 1971 dans le 12e arrondissement de Paris, est une chanteuse française d
e pop et de RnB.  Elle a vendu 4 millions de disques[1]. Son style musical oscille entre la variété français
e et un style plus « urbain ». Née le 1er septembre 1971 dans le 12e arrondissement de Paris dans une famill
e d'origine tunisienne, Lââm, née Lamia Naoui (épouse Suber), vit une enfance difficile, en raison de problè
mes familiaux, et elle est placée dans un foyer à Bourges. Très jeune, elle est déjà passionnée par la musiq
ue et, dès sa majorité, décide d'aller chanter dans le métro pour exercer à tout prix cette passion. Un prod
ucteur la remarque ainsi, est séduit par sa façon d'interpréter les chansons : sa carrière peut enfin débute
```

r. En 1990, elle entre aux ACP La Manufacture Chanson pour une formation de 2 ans au metier de chanteur. Elle publie en 1998 son premier single Chanter pour ceux qui sont loin de chez eux, une reprise de la célèbre chanson de Michel Berger. Le public est séduit par la nouvelle version de la chanson et le disque s'écoule à plus d'un million d'exemplaires (disque de diamant), se classant no 2 au Top 50 durant neuf semaines consécutives. Dans la suite de ce succès, Lââm sort son premier album, Persévérance, qui est agrémenté de la reprise de Michel Berger, mais aussi de singles comme Jamais loin de toi (no 4 au Top), Assez (no 51 au Top), Les enfants de l'an 2000 (no 3 au Top) ou encore Face à face (no 83 au Top 100). Certains de ces singles se transforment rapidement en tubes et l'album est sacré disque de platine avec plus de 500 000 exemplaires vendus[2]. Dans sa première année de carrière reconnue (1999), elle produit la bande son d'un dessin animé pour France télévisions et le programme de divertissement jeunesse Les Minikeums intitulé Tom et Sheenah. En mars 2001, elle est l'une des nombreuses interprètes du titre Que serais-je demain ? en tant que membre du collectif féminin Les Voix de l'espoir créé par Princess Erika[3]. La même année, elle sort son nouvel album intitulé Une vie ne suffit pas et décroche très rapidement un disque d'or, avec plus de 100 000 exemplaires vendus. Deux tubes se détachent de l'album, Que l'amour nous garde (no 20 au Top) et De ton indifférence (no 29 au Top). Parallèlement, Lââm joue en 2002 dans la comédie musicale Cindy de Luc Plamondon et Romano Musumarra. Sorti le 26 février 2002, l'album tiré du spectacle se classe no 24 en France et no 17 en Belgique francophone[4]. De ce disque est notamment extrait en février 2002 un duo entre Lââm et Frank Sherbourne, Un monde à nous. Ce titre se classe no 5 en France et en Belgique francophone[5]. Un autre duo avec Jay Je l'aime en secret sorti en octobre 2002 est également classé. Luc Plamondon, producteur du spectacle, reconnaît plus tard que cette comédie musicale est un échec commercial, malgré un disque d'or pour l'album[6] avec plus de 100 000 exemplaires vendus[7]. En 2004, la chanteuse sort un album éponyme plus diversifié que les précédents : dans cet opus, on passe de la variété au rock, du R'n'B au rap, avec plusieurs duos avec des artistes reconnus (Jean-Jacques Goldman, Lisa Stansfield, Princess Aniès…). La sortie de l'album est accompagnée par deux singles écrits et composés par Jean-Jacques Goldman : Tu es d'un chemin et On pardonne. Les deux titres sont proposés aux radios mais ces dernières les jugent trop lents et refusent de les passer. La promotion n'est donc pas excellente et le disque ne s'écoule qu'à 25 000 exemplaires. Lââm entame cependant une tournée française en passant par Paris, au Zèbre de Belleville, pour plusieurs soirs. En août 2005, elle publie un nouveau single inédit, Petite Sœur. Cette chanson au rythme très R'n'B se classe directement à la 5e place des charts, et s'écoule à plus de 250 000 exemplaires. Ce titre replace Lââm sur le devant de la scène. En septembre de la même année, son album éponyme est réédité sous le nom de Pour être libre, contenant trois inédits : Petite Sœur, Pour être libre (no 17 au Top) et Elle est toujours là. L'album est alors certifié disque d'or, moins d'un an après, pour plus de 75 000 exemplaires vendus. En 2006, la chanteuse est nominée en tant qu'Artiste francophone de l'année aux NRJ Music Awards et aux Victoires de la musique. Elle organise ensuite une grande tournée d'été en France et en Belgique. Elle se lance alors dans la réalisation d'un nouvel album pour la fin de l'année 2006. Ce disque est précédé par la sortie du single Le Sang chaud en duo avec la rappeuse Princess Aniès, qui se vend à plus de 50 000 exemplaires, se classant no 7 au Top 50. Malgré le succès du single, l'album Le Sang chaud ne s'écoule qu'à 15 000 exemplaires, un score très décevant pour la chanteuse. Lââm est néanmoins élue « Voix de l'année » par le public de la chaîne Filles TV ; c'est le premier prix que Lââm reçoit dans sa carrière. En mars 2007, elle publie un nouveau single, Rien ne dure, qui est envoyé aux radios qui le boudent totalement. Le disque n'est donc pas commercialisé. En mai, elle sort un nouveau single, Relève-toi, qui parle des femmes battues. Lââm sillonne alors les routes de France au cours d'une tournée durant l'été et l'automne. Au printemps 2008, elle passe également par le Sentier des Halles, à Paris, où elle présente un spectacle de reprises des chanteurs qu'elle admire (comme Jean-Jacques Goldman, Léo Ferré…). La même année, Lââm interprète également la bande originale de High School Musical 2 avec la chanson Savoir qui je suis, mais qui ne rencontre aucun succès en France. Malgré ces échecs commerciaux, elle continue à donner des concerts en France, en Belgique, en Suisse et au Canada. En décembre 2008, elle sort le single Ta voix en duo avec Jennifer Paige. Ce titre permet d'annoncer son Best of On a tous quelque chose de Lââm, qui contient ses plus grands succès, ainsi que trois inédits, dont ce duo[8]. Il est publié le 26 janvier 2009. En 2011, Lâam publie son nouvel et dernier album, Au cœur des hommes. L'album ne trouve pas son public et est un échec commercial. Sa maison de disques met un terme à son contrat. Depuis, la chanteuse continue la scène mais ne réalise plus d'album. En février 2012, elle participe à la soirée au profit de l'association « Angèle en rêve », dont elle est la marraine. En novembre, elle sort le single Révolution, une collaboration avec le DJ Kastilla dont les revenus sont reversés au Téléthon[9]. Lââm est une fidèle des concerts des Enfoirés de 2000 à 2012 puis de 2014 à 2015. Elle est la marraine de l'association « Regart's » qui a pour objectif d'aider les jeunes, les enfants et personnes en difficulté à mieux s'intégrer dans la vie de leur quartier et leur vie sociale. En 2013, elle participe au single Notre liberté au profit des Restos du cœur belges[10] et au single Je reprends ma route de l'association Les voix de l'enfant[11], et rejoint le collectif d'artistes « Les grandes voix des comédies musicales chantent pour les enfants hospitalisés »[12] aux côtés notamment d'anciens chanteurs de comédies musicales pour le single Un faux départ[13]. Le 7 avril 2015, elle est présente au dîner d'état organisé au palais de l'Élysée en l'honneur du président de la Tunisie Béji Caïd Essebsi et publie le single Tu me manques[14]. En fin de cette année, elle collabore sur le titre OH! B.I.G. du rapper The Notorious B.I.G.[15]. En 2016, elle est invitée avec Kijahman sur le titre Fo'Wyner de Joss Project[16],[17]. Elle devient chroniqueuse dans Touche pas à mon poste ! le 6 septembre 2016 sur C8, mais ne participera qu'à une seule émission. En mars 2017, elle déclarera avoir été humiliée par Cyril Hanouna. En 2017, elle apparaît en featuring sur le titre Caribbean King du rappeur Stan Apocalptik[18]. Le 15 mai 2018, elle participe à The Island : Célébrités sur M6[19],[20],[21]. En 2019, elle sera à l'affiche de la tournée Born in 90, qui réunit plusieurs chanteurs des années 1990 comme Larusso, Ménélik, Zouk machine, Allan Théo, World Apart, Indra, Benny B ou encore Génération Boys Band[22],[23]. Pendant l'été 2017, elle se fait remarquer par une série de tweets jugeant « raciste » le titre du roman d'Agatha Christie Dix petits nègres, visant en particulier une adaptation télévisée récente sur TF1. Estimant que le « mot nègre ne devrait plus exister », elle se fait critiquer sur les réseaux sociaux[24]. Au même moment, elle tourne la page de la musique, annonçant qu'elle ne sortirait plus de disque à l'avenir[25]. En 2000, elle est apparue dans le clip de Dje-an Mon Caractère, clip dans lequel est aussi apparue Carole Fredericks. Elle apparaît également dans son propre r

```
ôle dans l'épisode L'Arnaque de Sous le soleil diffusé le samedi 12 janvier 2008. Elle est aussi apparue dan
s une parodie sur le milieu de la musique aux côtés de Michal de Star academy 3. Elle apparaît dans l'émissi
on Un dîner presque parfait sur M6 diffusée dans la semaine du 5 mars 2012. Elle participe au cinquième prim
e de Star Academy le 3 janvier 2013 sur NRJ 12. Elle apparaît dans l'émission Un dîner presque parfait sur M
6 diffusée dans la semaine du 11 mai 2013. En mai 2013, Lââm apparaît dans un spot publicitaire pour la marq
ue RasoirClub[27]. C'est la première fois que la chanteuse participe à un spot publicitaire. En mai 2018, Lâ
âm participe à The Island spécial célébrités sur M6. Sur les autres projets Wikimedia :
```

Back to top of notebook

Spanish Text

We are scrapping 700 tweets (total) from :

- @realmadrid (200 tweets)
- @FCBarcelona (200 tweets)
- @LaLiga (100 tweets)
- @PaulinaRubio (100 tweets)
- @Armada_esp (100 tweets)

FC Barcelona

Back to top of notebook

Codes of Spanish text

Hide

Hide

```
SPAN_TEXT <- function() {
key <- "NR90Z1GOKWYNePoEwhX6SzTKL"
secret <- "8fmkbthBRwFvjdJPnHOrkxiYbdYsn7c1EtrG3xcy4Qlo7aulhv"
accessToken <- "948925352-lbmVCizm3AFLxLa1NbNMyzk6xAOOK3Ryh1ZY88mF"
accessSecret <- "4exQ4m58BjvcHw1xH2JNlbQsHGas1X9gmRpyPgMFILejp"
options(httr_oauth_cache=TRUE)
setup_twitter_oauth(key,secret, accessToken, accessSecret)

Realmadridtweets_brut <- userTimeline("realmadrid", n = 250)
Fcbtweets_brut <- userTimeline("FCBarcelona", n = 250)
LaLiga_brut <- userTimeline("LaLiga", n = 150)
PaulinaRubio_brut <- userTimeline("PaulinaRubio", n = 150)
Armada_esp_brut <- userTimeline("Armada_esp", n = 150)

df_1 <- tbl_df(map_df(Realmadridtweets_brut, as.data.frame))['text']
df_2 <- tbl_df(map_df(Fcbtweets_brut, as.data.frame))['text']
df_3 <- tbl_df(map_df(LaLiga_brut, as.data.frame))['text']
df_4 <- tbl_df(map_df(PaulinaRubio_brut, as.data.frame))['text']
df_5 <- tbl_df(map_df(Armada_esp_brut, as.data.frame))['text']


my_data <- c(df_1$text[1:200],df_2$text[1:200],df_3$text[1:100],df_4$text[1:100],df_5$text[1:100])


return(my_data)
}

Esp <- SPAN_TEXT()
```

```
[1] "Using direct authentication"
```

Back to top of notebook

Testing the Spanish text Result

Hide

Hide

```
number_of_tweets <- length(Esp)


cat("> the total number of Tweets is : ", number_of_tweets, '\n','\n','\n')
```

```
> the total number of Tweets is :  700
```

Hide

Hide

```
cat("> Here is some Spanich Tweets  : ",'\n', Esp[1:20], fill = 2)
```

```
> Here is some Spanich Tweets  :


□□ ¡El equipo se concentró para el encuentro ante el @ClubBrugge! #RMUCL https://t.co/FIUyu5H7ce
□ ¡ESTÁS DENTRO! □
De los campos de entrenamiento a la sala de cine, de la piscina al gimnasio.
Os enseñamos… https://t.co/GUKi0LidLW
□ ¡Nuestros 19 convocados para el partido contra el @ClubBrugge!
#RMUCL | #HalaMadrid https://t.co/5j1eYnKgqX
□□ ¡Así fue el día de ensueño de los 5 Coderistas y sus acompañantes animando al @RealMadrid en el derbi! ¿Q
uieres… https://t.co/o7GNlKNTai
□□✨ ¡Último entrenamiento antes del primer partido de @LigadeCampeones en el Santiago Bernabéu!
#RMUCL | #RMCity https://t.co/d3JCW3TK44
□ ¡¡FELIZ CUMPLEAÑOS #RMCity!! □

□□□ La mejor ciudad deportiva del mundo cumple hoy 14 años.
¡Descubre las instala… https://t.co/AakadOQysx
□ #Zidane: "Mañana es una oportunidad para devolver lo que se merece la afición y también para nosotros porq
ue nece… https://t.co/sj2M0mZzXX
□ @hazardeden10: "Tenemos la obligación de ganar al @ClubBrugge y vamos a dar el máximo"
#RMCity | #RMTV https://t.co/TKTco5HBpq
□□ ¡@hazardeden10  y #Zidane atienden a los medios en la rueda de prensa previa al encuentro □ Brujas! #RMU
CL https://t.co/ZjCXtPvUp2
□□ DIRECTO: ¡Último entrenamiento antes de recibir al @ClubBrugge! #RMCity https://t.co/bom0bu3gXy
El @RealMadrid ha recibido 9 veces como local a equipos belgas en @LigadeCampeones. ¿Sabrías decir cuántas t
erminar… https://t.co/elbBQpg2Aj
□□ ¡El equipo ya prepara el partido de Champions frente al @ClubBrugge!
#RMUCL | #HalaMadrid https://t.co/Plm8GifcLj
□ □ ¡El behind the scenes del #RMDerbi en el Metropolitano!
#RMLiga | #HalaMadrid https://t.co/mg7zGmugbo
□□ El equipo blanco empató en su visita al @Atleti y es el único invicto de la LaLiga.

□ Galería y crónica haciend… https://t.co/VAW0h5MeVf
□□ #Zidane: "Creo que merecíamos mucho más. Los jugadores han hecho un gran partido."
#RMTV | #RMDerbi https://t.co/008tPgvw4N
□ ¡Estas han sido las palabras de nuestro capitán @SergioRamos a pie de campo!
#RMDerbi | #HalaMadrid https://t.co/RgV6hWw1yI
□ FP: @Atleti 0-0 @RealMadrid
#Emirates | #HalaMadrid https://t.co/oxGABBYk5j
88' □ | 0-0 | Tercer y último cambio. Entra Luka Jović en lugar de @Benzema.
#RMDerbi | #RMLiga https://t.co/kVWCwnNrdn
77' □ | 0-0 | Segundo cambio en las filas del @realmadrid. Entra  @jamesdrodriguez sustituyendo a @hazarded
en10.… https://t.co/ijTveo006i
68' □ | 0-0 | Primer cambio del encuentro. Entra al campo @lukamodric10 en lugar de @fedeevalverde.
#RMDerbi |… https://t.co/tjbmy6NCQv
```

Back to top of notebook

Italien Text

We are scrapping the Italien text from Giovannino Guareschi's Tutto don Camillo (I racconti del Mondo piccolo) – Volume 1 di 5 (PDF) 1 page per row.

Tutto don Camillo

Back to top of notebook

Codes of Italien text

Hide

Hide

```
ITA_TEXT <- function() {

download.file("http://www.flyemail.com/public/libri/1%20-%20Guareschi%20Giovannino%20-%20Tutto%20Don%20Camillo%20%20Volume.pdf", "./tutto.pdf")
text <- pdf_text("./tutto.pdf")

return(text)
}

ITA <- ITA_TEXT()
```

```
trying URL 'http://www.flyemail.com/public/libri/1%20-%20Guareschi%20Giovannino%20-%20Tutto%20Don%20Camillo%20%20Volume.pdf'
Content type 'application/pdf' length 1990948 bytes (1.9 MB)
==================================================
downloaded 1.9 MB
```

Back to top of notebook

Testing the Italien text Result

Hide

Hide

```
number_of_pages <- length(ITA)


cat("> the total number of pages in the book is : ", number_of_pages, '\n','\n','\n')
```

```
> the total number of pages in the book is :  848
```

Hide

Hide

```
cat("> Here is some pages from the Italian text  : ",'\n', ITA[1:5], fill = 2)
```

```
> Here is some pages from the Italian text  :



Giovannino Guareschi
Tutto don Camillo
        MONDO PICCOLO
              Volume 1 di 5
Racconti dal 1 al 84 e tre racconti nel prologo

   Qui, con tre storie e una citazione, si spiega il
              mondo di "Mondo piccolo'
     Io da giovane facevo il cronista in un giornale e andavo
in giro tutto il giorno in bicicletta per trovare dei fatti da
raccontare.
     Poi conobbi una ragazza, e allora passavo le giornate
pensando a come sì sarebbe comportata quella ragazza se io
fossi diventato imperatore del Messico o se fossi morto. E,
alla sera, riempivo la mia pagina inventando i fatti di crona-
ca, e questi fatti piacevano parecchio alla gente perché era-
no molto più verosimili di quelli veri.
     Io, nel mio vocabolario, avrò sì e no duecento parole, e
son le stesse che usavo per raccontare l'avventura del vec-
chio travolto da un ciclista o quella della massaia che, sbuc-
ciando le patate, ci rimetteva un polpastrello.
     Quindi niente letteratura o altra mercanzia del genere:
in questo libro io sono quel cronista di giornale e mi limito a
raccontare dei fatti di cronaca. Roba inventata e perciò tan-
to verosimile che mi è successo un sacco di volte dì scrivere
una storia e di vederla, dopo un paio dì mesi, ripetersi nella
realtà. E non c'è niente di straordinario, è semplice questio-
ne di ragionamento: uno considera il tempo, la stagione, la

moda e il momento psicologico e conclude che, stando così
le cose, in un ambiente x possono verificarsi questa e que-
st'altra vicenda. (…)
      L'ambiente è un pezzo della pianura padana: e qui biso-
gna precisare che, per me, il Po comincia a Piacenza.
      Il fatto che da Piacenza in su sia sempre lo stesso fiu-
me, non significa niente: anche la Via Emilia, da Piacenza a
Milano, è in fondo la stessa strada; però la Via Emilia è
quella che va da Piacenza a Rimini.
      Non si può fare un paragone tra un fiume e una strada
perché le strade appartengono alla storia e i fiumi alla geo-
grafia.
      E con questo?
      La storia non la fanno gli uomini: gli uomini subiscono
la storia come subiscono la geografia. E la storia, del resto,
è in funzione della geografia.
      Gli uomini cercano di correggere la geografia bucando
le montagne e deviando i fiumi e, così facendo, si illudono dì
dare un corso diverso alla storia, ma non modificano un bel
niente, perché, un bel giorno, tutto andrà a catafascio. E le
```

```
acque ingoieranno i ponti, e romperanno le dighe, e riempi-
ranno le miniere; crolleranno le case e i palazzi e le cata-
pecchie, e l'erba crescerà sulle macerie e tutto ritornerà ter-
ra. E i superstiti dovranno lottare a colpi di sasso con le be-
stie, e ricomincerà la storia.
      La solita storia.

      Poi, dopo tremila anni, scopriranno, sepolto sotto qua-
ranta metri di fango, un rubinetto dell'acqua potabile e un
tornio della Breda di Sesto San Giovanni e diranno: «Guar-
da che roba!».
      E si daranno da fare per organizzare le stesse stupidag-
gini dei lontani antenati. Perché gli uomini sono delle di-
sgraziate creature condannate al progresso, il quale pro-
gresso porta irrimediabilmente a sostituire il vecchio Padre-
terno con le nuovissime formule chimiche. E così, alla fine,
il vecchio Padreterno si secca, sposta di un decimo di milli-
metro l'ultima falange del mignolo della mano sinistra e tut-
to il mondo va all'aria.
      Dunque il Po comincia a Piacenza, e fa benissimo per-
ché è l'unico fiume rispettabile che esista in Italia: e i fiumi
che si rispettano si sviluppano in pianura, perché l'acqua è
roba fatta per rimanere orizzontale, e soltanto quando è per-
fettamente orizzontale l'acqua conserva tutta la sua naturale
dignità. Le cascate del Niagara sono fenomeni da baracco-
ne, come gli uomini che camminano sulle mani.
      Il Po comincia a Piacenza, e a Piacenza comincia an-
che il Mondo piccolo delle mie storie, il quale Mondo picco-
lo è situato in quella fetta di pianura che sta fra il Po e l'Ap-
pennino.
      «… Il cielo è spesso d'un bell'azzurro, come ovunque in
Italia, salvo nella stagione men buona, in cui si levano fittis-
sime nebbie. (…) Il suolo è la più parte gentile, arenoso e
fresco, alquanto forte a monte e talora schiettamente argillo-
```

Back to top of notebook

German Text

We are scrapping 500 press releases from n-tv German free-to-air television news channel. The first step is to capture the URLs of all the required press releases and then download the press releases to a local folder in order extract the main text of each press release and save it to a string.

Back to top of notebook

Codes of German text

Hide

Hide

```
GER_TEXT <- function() {

signatures = system.file("CurlSSL", cainfo = "cacert.pem", package = "RCurl")
all_links_articles <- character()    # initialize
link_page <- 'https://www.n-tv.de/thema/trends/archiv-0'
while( length(all_links_articles) < 501  ){
   html_page <- getURL(link_page, cainfo = signatures)
  html_page_tree <- htmlParse(html_page)
  links_of_articles = xpathSApply(html_page_tree, "//div/section/section/article/figure/a", xmlGetAttr, "hre
f")
  all_links_articles <- c(all_links_articles,links_of_articles)
  link_page <- xpathSApply(html_page_tree,  "//div[@class= 'paging']//div//a[@class= 'paging__next1 icon ico
n__arrow'] ", xmlGetAttr,"href")
}

# Download all press releases in a the Folder 'Press_Releases_Eng' that you need to create before.
dir.create("Press_Releases_german")
for(i in 1:length(all_links_articles)){
  url <- all_links_articles[i]
  tmp <- getURL(url, cainfo = signatures)              # get the HTML at URL i
  write(tmp, str_c("Press_Releases_german/", i, ".html"))     # write HTML to Press_Releases/i.html
}

German_text <- character()
for(i in 1:length(list.files("Press_Releases_german")) ){
  tmp <- readLines(str_c("Press_Releases_german/", i, ".html"))
  tmp <- str_c(tmp, collapse = "")
  tmp <- htmlParse(tmp)
  release <- xpathSApply(tmp, "//div[@class='article__text']", xmlValue)
  release <- str_c(release, collapse = " ")
  German_text <- c(German_text,release)
}



return(German_text[1:500])
}

GER <- GER_TEXT()
```

```
'Press_Releases_german' already exists
```

Back to top of notebook

Testing the German text Result

Hide

Hide

```
number_of_articles <- length(GER)
sample_of_GER_text <-  GER[4]


cat("> the total number of articles is : ", number_of_articles, '\n','\n','\n')
```

```
> the total number of articles is :  500
```

Hide

Hide

```
cat("> Here is some samples from the German text : ", '\n','\n',sample_of_GER_text[[1]])
```

```
> Here is some samples from the German text :

              In Deutschland sinkt die Nachfrage nach RohÃ¶l. Im vergangenen Jahr kauft die Bundesrepublik so wenig des fossilen Rohstoffs ein, wie noch nie seit der Wiedervereinigung. Damit setzt sich ein Trend fort, der nach der Jahrtausendwende begonnen hat.              Deutschland hat im vergangenen Jahr so wenig RohÃ¶l eingefÃ¼hrt wie nie seit der Wiedervereinigung. Gut 84,8 Millionen Tonnen des wichtigen Rohstoffs, der zum Beispiel zu HeizÃ¶l und Sprit verarbeitet wird, wurden 2018 aus dem Ausland eingekauft, wie das Statistische Bundesamt mitteilte.       Wirtschaft  29.07.19    Welterste 9000-Meter-FÃ¶rderung China erschlieÃt gewaltige ErdÃ¶l-Mengen     Damit sank die importierte Menge im dritten Jahr in Folge: 2016 waren es noch rund 91,8 Millionen Tonnen, 2017 knapp 90 Millionen Tonnen. Auf dem Niveau von 2017 lag die importierte RohÃ¶lmenge auch 1991, dem ersten vollen Jahr nach der deutschen Wiedervereinigung, wie die Wiesbadener Statistiker erlÃ¤uterten. Am meisten Ã l wurde in dem betrachteten Zeitraum im Jahr 2005 eingefÃ¼hrt: 114,5 Millionen Tonnen.Der mit Abstand wichtigste Lieferant des Rohstoffs fÃ¼r Deutschland ist Russland. Mit rund 29,2 Millionen Tonnen bezog Deutschland 2018 gut ein Drittel seines Ã ls von dort. Aus Norwegen kamen knapp neun Prozent (rund 7,6 Mio Tonnen), aus Libyen rund 8,6 Prozent (rund 7,3 Mio Tonnen) der gesamten Importmenge.Saudi-Arabien, das nach Angaben der Internationalen Energie-Agentur (IEA) der weltweit grÃ¶Ã te ErdÃ¶lexporteur ist, hat Zahlen der Wiesbadener BehÃ¶rde zufolge als Ã llieferant fÃ¼r Deutschland stark an Bedeutung verloren. Im vergangenen Jahr bezog Europas grÃ¶Ã te Volkswirtschaft von dort gut 1,4 Millionen Tonnen RohÃ¶l. Das waren gerade einmal 1,7 Prozent der gesamten RohÃ¶leinfuhren.Die Bedeutung des ErdÃ¶ls nimmt fÃ¼r Deutschland seit Jahren ab, wie die Statistiker betonten. So ging der Verbrauch von HeizÃ¶l von 2005 bis 2016 zurÃ¼ck, auch die Verwendung von Benzin. Der Verbrauch von Diesel dagegen stieg in dem Zeitraum an. Gemessen wird der Verbrauch auch in Petajoule. Die privaten Haushalte in Deutschland verbrauchten 2016 laut Statistik 455 Petajoule fÃ¼rs Heizen mit Ã l und 1315 Petajoule fÃ¼rs Autofahren mit Benzin und Diesel.
```

Back to top of notebook

The Final dataset of 5 text corpora

Hide

Hide

```
MAIN_DATASET <- function() {

Eng <- ENG_TEXT()
Fr <- Fr_TEXT()
Esp <- SPAN_TEXT()
ITA <- ITA_TEXT()
GER <- GER_TEXT()


x_1 <- data.frame("language_code" = "ENG", "text" = Eng[2:length(Eng)])
x_2 <- data.frame("language_code" = "Fr", "text" = Fr)
x_3 <- data.frame("language_code" = "Esp", "text" = Esp)
x_4 <- data.frame("language_code" = "ITA", "text" = ITA)
x_5 <- data.frame("language_code" = "GER", "text" = GER)


df_3 <- rbind(x_1, x_2,x_3,x_5)

write.csv(df_3, file = "Final_data_set.csv")

df_4 <- df_3[sample(nrow(df_3)),]
head(df_4,n=10)


}

MAIN_DATASET()
```

```
'Press_Releases_Eng' already exists'Press_Releases_Fr' already exists
```

```
[1] "Using direct authentication"
```

```
trying URL 'http://www.flyemail.com/public/libri/1%20-%20Guareschi%20Giovannino%20-%20Tutto%20Don%20Camillo%20%20Volume.pdf'
Content type 'application/pdf' length 1990948 bytes (1.9 MB)
==================================================
downloaded 1.9 MB

'Press_Releases_german' already exists
```

	language_code <fctr>
7176	ENG
847	ENG
7904	ENG
1686	ENG
8562	ENG
6430	ENG
2866	ENG
7445	ENG
4620	ENG
2687	ENG

1-10 of 10 rows | 1-2 of 2 columns

Hide

Hide

```
df_4 <- df_3[sample(nrow(df_3)),]
head(df_4,n=10)
```

	language_code <fctr>
10721	GER
6850	ENG
10847	GER
5682	ENG
862	ENG
9464	Fr
1782	ENG
9714	Fr
9271	Fr
6242	ENG

1-10 of 10 rows | 1-2 of 2 columns

Back to top of notebook