TUTORIAL: NUVEM DE PALAVRAS NO R

Carla Cristina Passos Cruz

SUMÁRIO

OBSERVAÇÕES GERAIS:

O código na íntegra também está disponível em:
https://github.com/carlapassos/Tutorial_Wordcloud_R

Este tutorial e código pode ser usado para outros tipos de documento e texto, com devidas adaptações

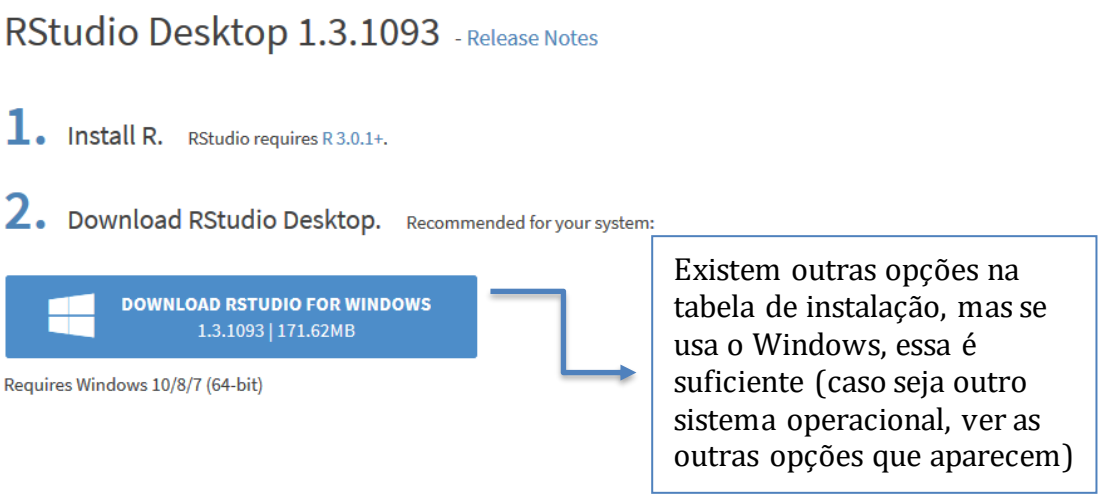
PARTE I – INSTALAÇÃO DOS PROGRAMAS NECESSÁRIOS

PASSO 1: Baixar e instalar o *software* R no computador

Link: https://cran.r-project.org/bin/windows/base/R-4.0.3-win.exe

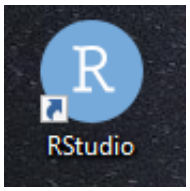
PASSO 2: Após instalar o R, baixar e instalar o RStudio (o RStudio não funciona sem o R instalado, por isso a ordem de instalação é importante)

Site: https://rstudio.com/products/rstudio/download/#download



PARTE II – ABRINDO O RSTUDIO, INSTALAÇÃO E ATIVAÇÃO DE PACOTES

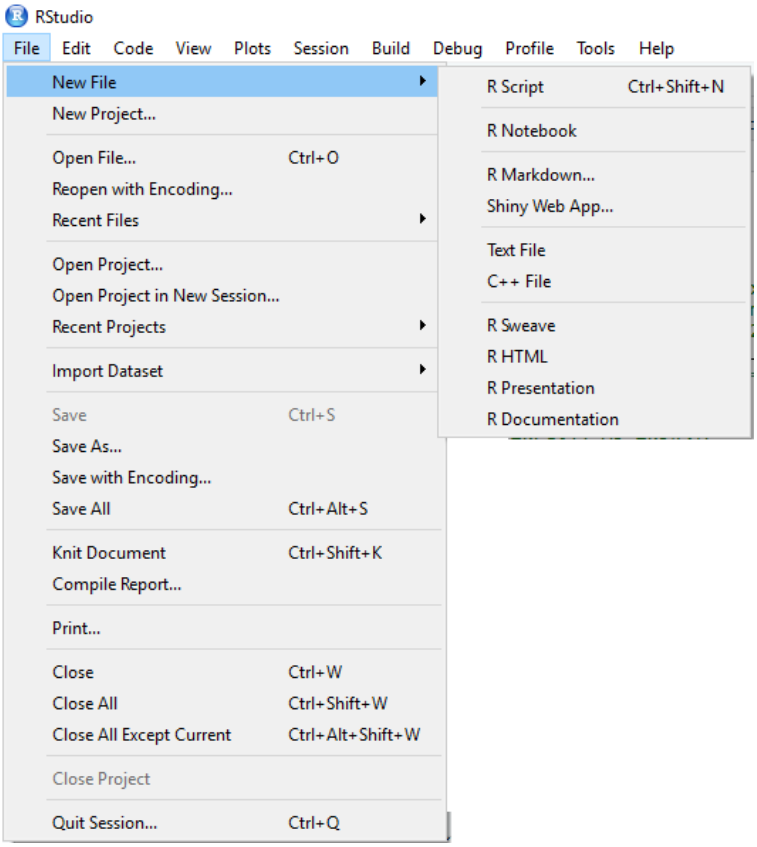
PASSO 1: Após a instalação dos dois programas, abrir o RStudio



OBSERVAÇÃO:

É necessária que a instalação seja nesta ordem, caso contrário, ocorrerá erro na execução do programa.

PASSO 2: Ao abrir o RStudio ir em: File -> New File -> R Script



PASSO 3: Abrirá uma janela em branco, para que os comandos sejam digitados. Logo, os códigos (programação) serão diferentes, dependendo do trabalho que deseja executar. Como nesta dinâmica o objetivo é gerar a Nuvem de Palavras (*wordcloud*), será necessário instalar pacotes que auxiliam na execução do trabalho a se realizar. Copiar e colar no RStudio os pacotes abaixo:

```
#INSTALAÇÃO
install.packages("stringr") #pacote para strings
install.packages("stringi") #outro pacote para strings
install.packages("tm") #pacote para a mineração de texto
install.packages("stopwords") #pacote para a limpeza das stopwords
install.packages("wordcloud") #pacote para a nuvem de palavras
install.packages("textreadr") #pacote que lê textos em pdf

#LEITURA
library(stringr)
library(stringi)
library(tm)
library(stopwords)
library(wordcloud)
library(textreadr)
```

Observações:

a) Os termos entre aspas na parte install.packages, precisam ficar igual ao que está sendo exibido, ou seja, escrito da cor verde. Caso não esteja, trocar as “” copiadas, apagando-as e colocando-as novamente;

b) Após a instalação e ativação dos pacotes, caso queira usar o código novamente, colocar o símbolo “#” antes de install.packages, conforme mostrado a seguir. Não é obrigatório, mas evita que o programa instale os pacotes novamente e de forma desnecessária, pois uma vez instalado, basta somente ativar (ler) através do comando **library**;

```
#INSTALAÇÃO
#install.packages("stringr") #pacote para strings
##install.packages("stringi") #outro pacote para strings
#install.packages("tm") #pacote para a mineração de texto
#install.packages("stopwords") #pacote para a limpeza das stopwords
#install.packages("wordcloud") #pacote para a nuvem de palavras
#install.packages("textreadr") #pacote que lê textos em pdf
```

c) No caso das *stopwords*, como é um tutorial simples, serão utilizadas as fornecidas pelo software. No entanto, para trabalhos mais complexos, há a necessidade de complementação em outras fontes de pesquisa.

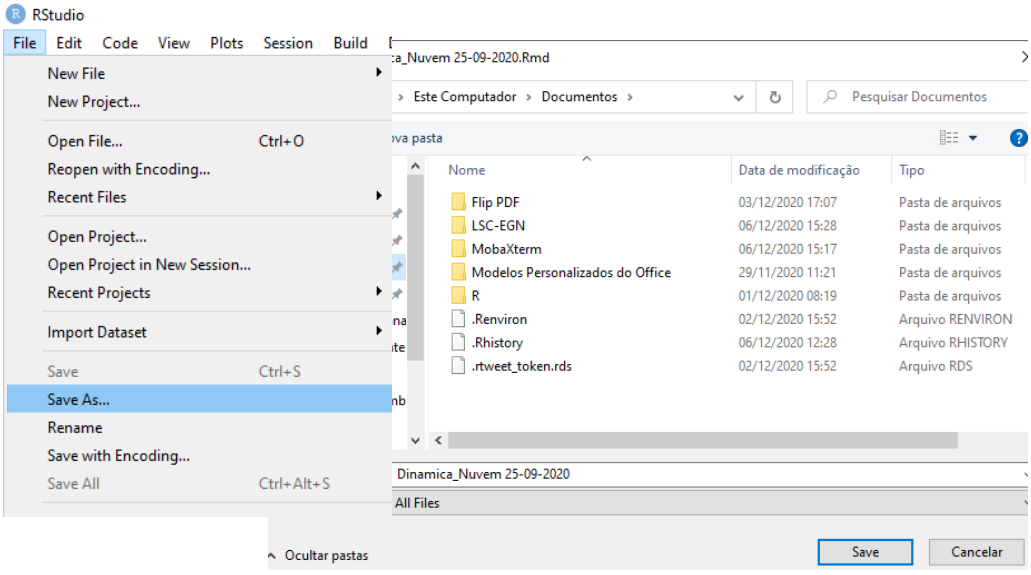
PARTE III – APLICAÇÃO: NUVEM DE PALAVRAS

PASSO 1: Realizada as etapas acima, agora a nuvem de palavras pode começar a ser construída. A Nuvem de palavras pode usar palavras de qualquer fonte (artigo, site, facebook, twitter, etc.) e/ou arquivo (pdf, word, excel, etc.). Como é um exemplo, será utilizado o resumo de uma dissertação do Programa de Pós-Graduação em Estudos Marítimos da Escola de Guerra Naval (PPGEM/EGN), enviada junto com o tutorial, mas também pode ser encontrada em:

http://www.redebim.dphdm.mar.mil.br/vinculos/00001b/00001b45.pdf

PASSO 2: O arquivo deve ser salvo no mesmo lugar do arquivo do RStudio. Para isso, salva-se o arquivo .R em um local de fácil acesso e o arquivo em .pdf no mesmo lugar, pois precisaremos saber o "caminho" de onde ele se encontra:

Salvando o arquivo .R: file -> save as -> escolher o local e nome do arquivo -> save



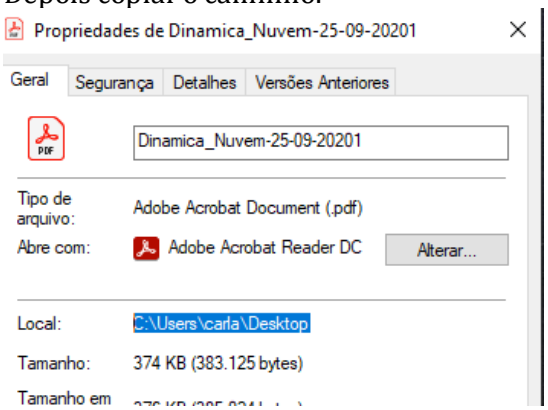
PASSO 3: Lendo o arquivo (OBS.: Copiar e colar o código abaixo no RStudio. Caso o que está escrito entre aspas não ficar verde, trocar as aspas):

```
#Leitura do resumo da Dissertacao 1
doc1 <- read_pdf("C:/Users/carla/Documents/LSC-EGN/dissertacaoegn.pdf")
doc1
```

Este é o caminho onde está salvo o arquivo. Por isso é importante saber onde salvou ou, se possível, salvar na mesma pasta do arquivo R. Para ler o caminho do arquivo (no meu caso o arquivo está salvo na área de trabalho:

Botão direito do mouse sobre o arquivo -> propriedades

Depois copiar o caminho:



Copiar e colar o caminho no R e fazer ajustes. Ficará assim:

```
De: "C:\Users\carla\Desptop"

Para:"C:/Users/carla/Desktop/dissertacaoegn.pdf"
```

Nome do arquivo

PASSO 4: Em seguida, será mostrado os primeiros elementos contidos no arquivo. Como queremos apenas o item RESUMO, escreveremos um código no qual o programa procurará onde começa o mesmo. De forma a facilitar o trabalho, a linha onde começa e termina o resumo, (linhas 115 e 127, respectivamente) encontram-se descritas no código (copiar e colar ou escrever no arquivo R):

```
doc1$text == "RESUMO" #verificando a partir de que linha está o resumo

resumo1 <- doc1$text[115:127] #armazenando o resumo da dissertacao1
resumo1 <- paste(resumo1,collapse = " ") #armazenando o resumo
resumo1

##################ESTA  PARTE É O RESULTADO A SER MOSTRADO############

## [1] "A presente dissertação repousa na hipótese de que, a despeito da existên
cia de dois subcomplexos de segurança sul-americanos e de suas diferenças, as aç
ões da política externa brasileira juntamente com as dos organismos regionais po
dem mitigar ou impedir a ocorrência de conflitos armados na região e, até mesmo,
consolidar tais subcomplexos, transformando o subcontinente em uma comunidade de
segurança. O enfoque metodológico da pesquisa proposta decorre do seu referencia
l teórico, baseado no conceito de segurança proposto por Barry Buzan, Ole Wæver
e Jaap de Wilde e, sob uma abordagem analítica, subsidiada por pesquisas bibliog
ráfica de documentos oficiais do Brasil, de outros Estados sul-americanos e de O
rganizações Regionais. Este trabalho se justifica por duas razões principais. Pr
imeiramente, porque envolve as relações entre Estados em uma região na qual o Br
asil está geográfica, econômica e politicamente inserido. Ademais, porque os con
flitos, armados ou não, afetarão diretamente a política de defesa e os demais in
teresses do Brasil, uma vez que ocorrerão dentro de seu entorno estratégico."
```

PASSO 5: Agora é feito o processo de limpeza e tratamento do texto (conhecido como pré-processamento na Mineração de Texto):

```
#Deixando em linha única (necessário para texto em pdf)
corpus1 <- gsub(pattern = "\\W",replace = " ", resumo1)
#Convertendo de UTF-8 para ASCII*
corpus1 <- iconv(corpus1,"UTF-8",to="ASCII//TRANSLIT")
#Removendo caracteres especiais
corpus1 <- gsub(pattern="\\b[A-z]\\b{1}",replace= " ",corpus1)
#Conversão do formato de matriz para corpus (formato para text mining)
corpus1 <- VectorSource(corpus1)
corpus1 <- VCorpus(corpus1)
#Convertendo as letras de maiúsculo para minúsculo
corpus1 <- tm_map(corpus1,tolower)
#Remove pontuação
corpus1 <- tm_map(corpus1,removePunctuation)
```

```
#Remove as stopwords
corpus1 <- tm_map(corpus1,removeWords,stopwords("pt"))
#Remove espaços extras
corpus1 <- tm_map(corpus1,stripWhitespace)
#Verificar resultado
inspect(corpus1)

###########RESULTADO FINAL#####################################################
[1]  presente dissertacao repousa hipotese despeito existencia dois subcomplexos
seguranca sul americanos diferencas acoes politica externa brasileira juntamente
organismos regionais podem mitigar impedir ocorrencia conflitos armados regiao a
te consolidar tais subcomplexos transformando subcontinente comunidade seguranca
enfoque metodologico pesquisa proposta decorre referencial teorico baseado conce
ito seguranca proposto barry buzan ole waver jaap wilde sob abordagem analitica
subsidiada pesquisas bibliografica documentos oficiais brasil outros estados sul
americanos organizacoes regionais trabalho justifica duas razoes principais prim
eiramente porque envolve relacoes estados regiao brasil geografica economica pol
iticamente inserido ademais porque conflitos armados nao afetarao diretamente po
litica defesa demais interesses brasil vez ocorrerao dentro entorno estratégico
```

Obs: *Importante verificar resultado, pois dependendo da configuração do computador, o passo *#Convertendo de UTF-8 para ASCII** não é necessário.

PASSO 6: Há palavras que precisam unificadas mas o programa não o faz. Unindo as palavras compostas (obs.: este passo é opcional, ou seja, dependerá das palavras contidas no texto usado):

```
for (j in seq(corpus1)){
  corpus1[[j]] = gsub("sul americanos", "sul_americanos", corpus1[[j]])

  corpus1[[j]] = gsub("politica externa", "politica_externa", corpus1[[j]])
  corpus1[[j]] = gsub("referencial teorico", "referencial_teorico", corpus1[[j]])
  corpus1[[j]] = gsub("documentos oficiais", "documentos_oficiais", corpus1[[j]])
  corpus1[[j]] = gsub("conflitos armados", "conflitos_armados", corpus1[[j]])
}
```

PASSO 7: Criando a matriz termo-documento, que possui as frequências das palavras para a nuvem de palavras e verificando as frequências dos termos:

```
t1 <- corpus1
m1   <- tm_map(t1,PlainTextDocument)
mtd1 <- DocumentTermMatrix(m1) #termos nas linhas e documento na coluna
Terms(mtd1) #exibe os termos (caso queira verificar. Não é necessário)


#Verificando a frequencia com que os termos aparecem
freqt1<- colSums(as.matrix(mtd1))
freqt1

#Verificando os termos e suas frequencias
R1 <- order(freqt1, decreasing = T) #Recebe os indices dos termos de menor e mai
or freq.
```

```
#Verificando os termos e suas frequencias
R1 <- order(freqt1, decreasing = T) #Recebe os indices dos termos de menor e maior freq;
```

PASSO 8: Criando a matriz de frequências para a nuvem de palavras e a nuvem de palavras:

```
#Criando uma matriz de frequências para a nuvem de palavras
mfreq1 <- data.frame(names(freqt1[R1]),as.integer(freqt1[R1]))
names(mfreq1) <- c("termos","freqabs")

#Salvando a matriz de frequências em um arquivo excel/csv
write.csv2(mfreq1,"C:/Users/carla/Desktop/freqtexto1.csv")
```

Salvar preferencialmente no mesmo local do arquivo .R e do arquivo em pdf utilizados

```
#NUVEM DE PALAVRAS

#Abrindo/lendo o arquivo .csv salvo
freqt1 <- read.csv2(,"C:/Users/carla/Desktop/freqtexto1.csv",sep=";")

#Passos necessários para que a nuvem de palavras leia as frequências e termos
freqg1 <- as.data.frame(freqt1)
nomes1 <- as.character(freqg1[,2])
valor1 <- as.integer(freqg1[,3])

#Exibindo a Nuvem de Palavras
wordcloud(nomes1, valor1, min.freq = 1, scale = c(2,.0), random.order = F,
colors = brewer.pal(8,"Dark2))
```

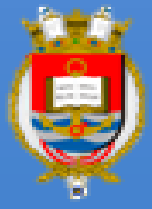
Escola de Guerra Naval

Avenida Pasteur, 480 - Rio de Janeiro – RJ – Brasil

CEP.: 22290-240

✉ arranjos.lsc@gmail.com

https://www.linkedin.com/company/arranjos-metodologicos-lsc



Escola de Guerra Naval

Laboratório de Simulações e Cenários

Subgrupo Arranjos Metodológicos