TEXT MINING no R

Grimaldo Oliveira

O que é Text Mining

1. Processo de obtenção de informações importantes de um texto.
2. É um campo interdisciplinar que se baseia na recuperação de informações, extração de dados, aprendizado de máquina, estatísticas e linguística computacional.
3. Tarefas típicas de mineração de texto incluem categorização e agrupamento de texto, extração de conceito/entidade, produção de taxonomias granulares, análise de sentimentos.

Sites Importantes

TM CRAN

https://cran.r-project.org/web/packages/tm/tm.pdf

Text Mining with R

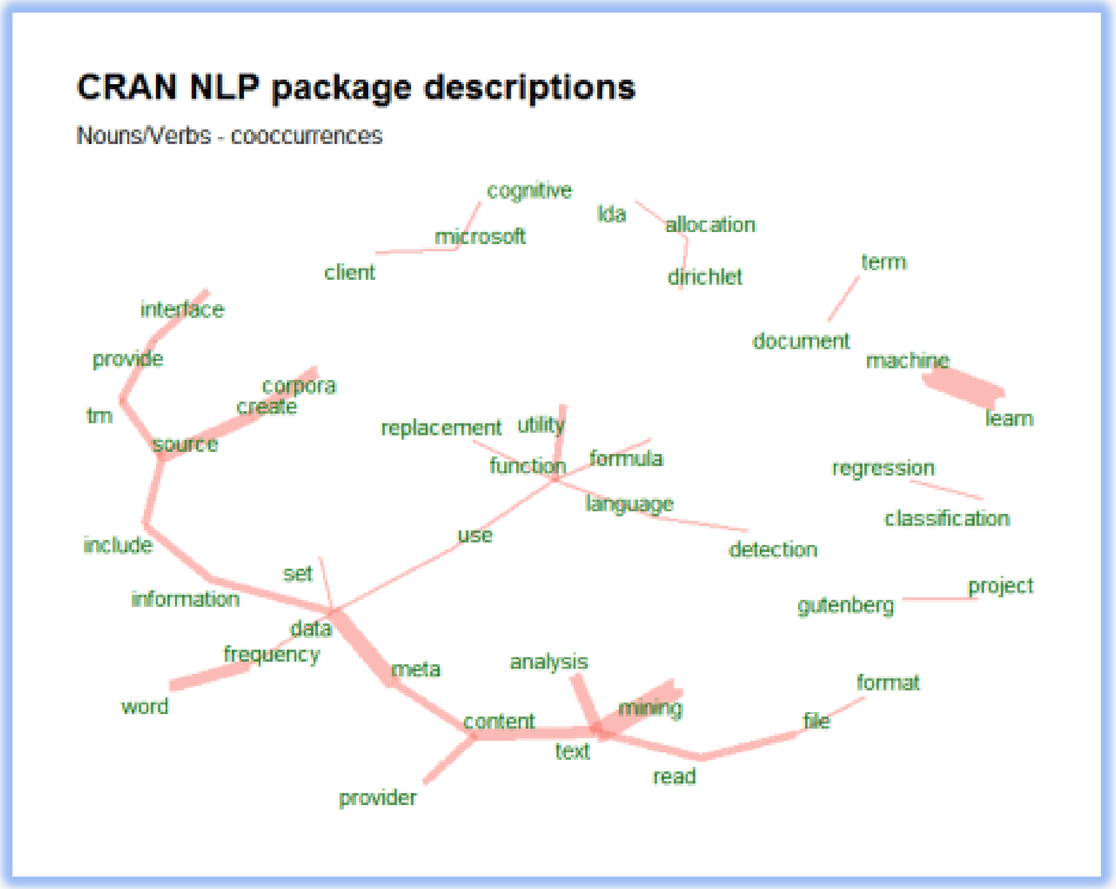
https://www.tidytextmining.com/

R Studio Cloud

https://rstudio.cloud/

Exemplo

Este é um exemplo de uma Mineração de Texto com análise dos seus termos



Fonte: https://www.r-bloggers.com/text-mining-with-r-upcoming-courses-in-belgium/

1. Nuvem de Palavras
2. Corpus
3. Tokenização
4. Tokens
5. Frequência de Palavras
6. Stopwords

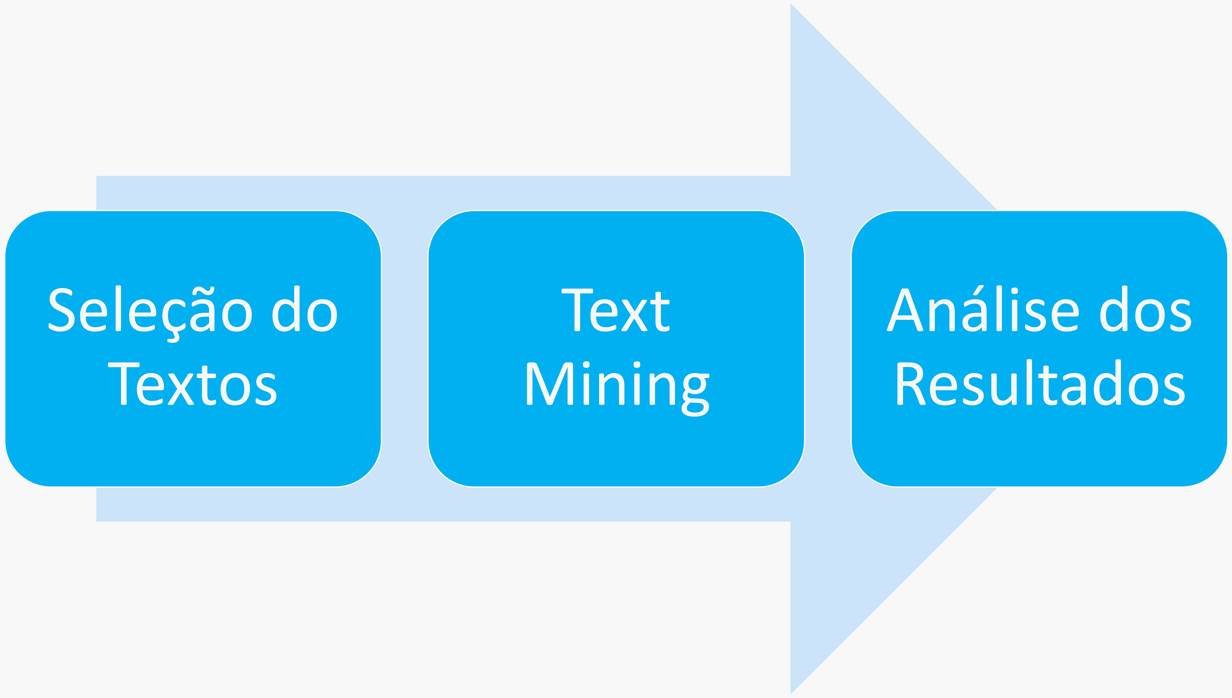
Termos que serão comuns

Vamos começar!

Processo de ajustes nos textos

Processo de Text Mining

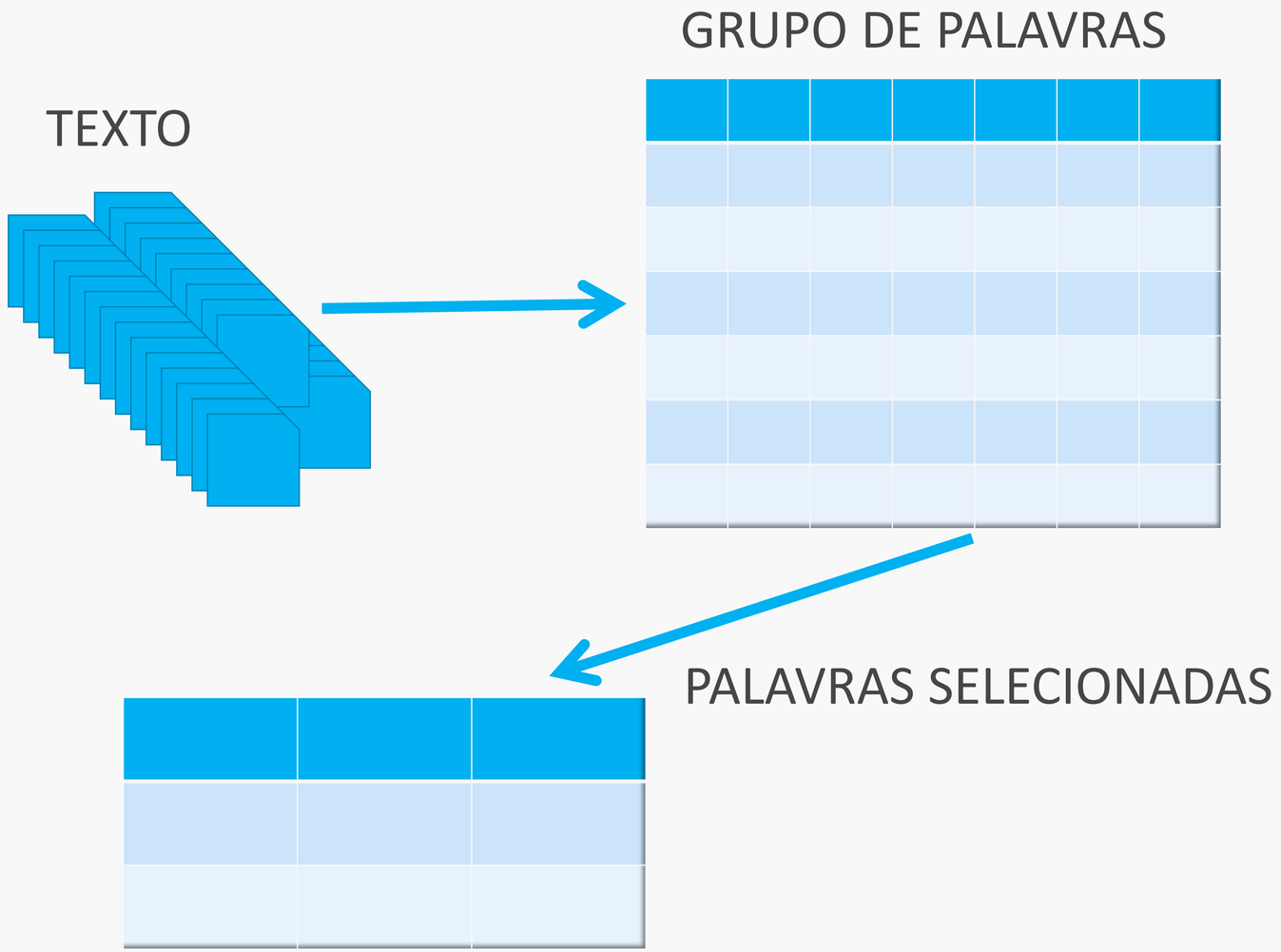
A construção de análises em texto, deriva de atenção na coleta, preparação e análise das informações, deveremos ter um planejamento para processar os textos.



Vamos começar!

Separando o que importa

Teremos que ter em mente, que os textos são geralmente compostos de diversos caracteres que não desejamos analisar e palavras que gostaríamos de excluir: preposições, pronomes, pontos, etc.

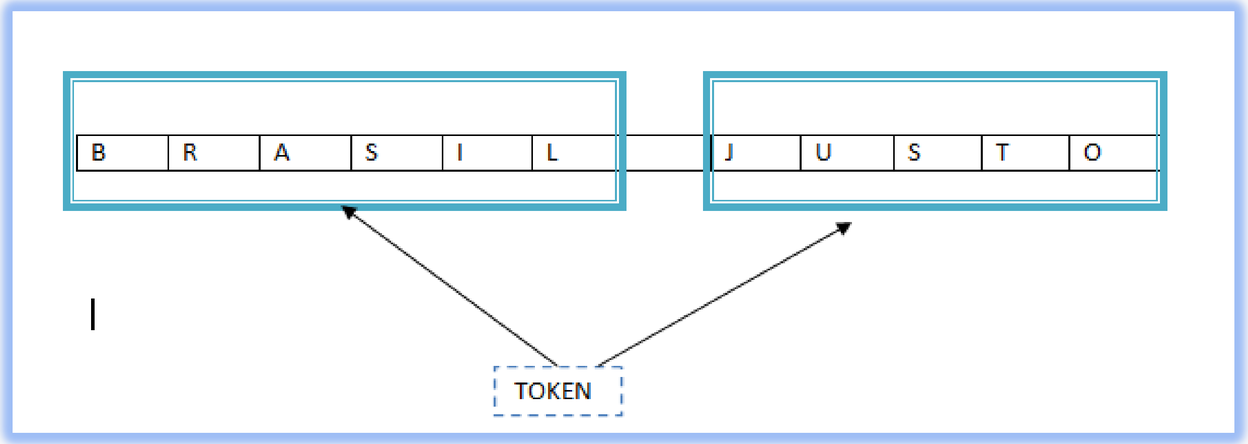


Vamos começar!

Entendendo os termos

Token

Um token é uma unidade significativa de texto, como uma palavra, que estamos interessados em usar para análise, e tokenização é o processo de dividir o texto em tokens.

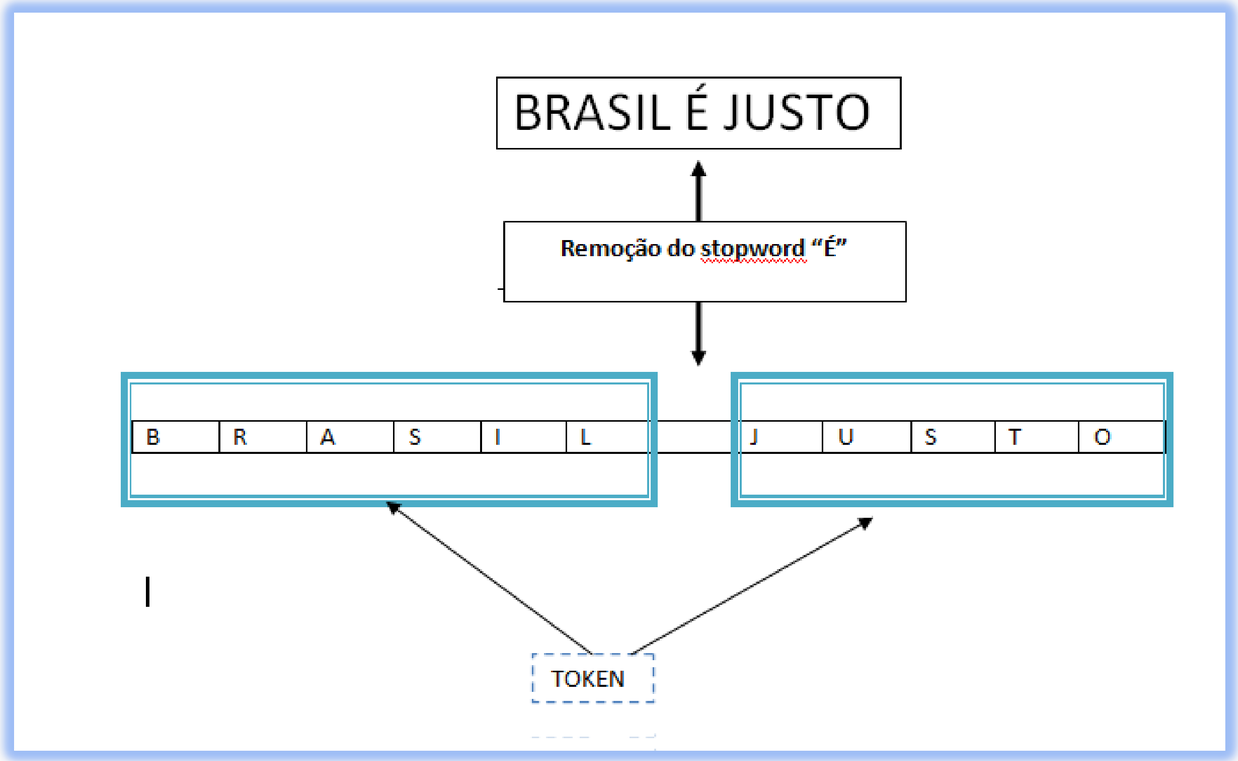


Vamos começar!

Entendendo sobre exclusão de termos

Stopwords

É a identificação do que pode ser desconsiderado, é a forma de excluir tudo que não desejamos ser considerado conhecimento nos textos. Veremos que nos textos há muitos tokens que não tem valor semântico algum, estes tokens são chamados de stopwords.

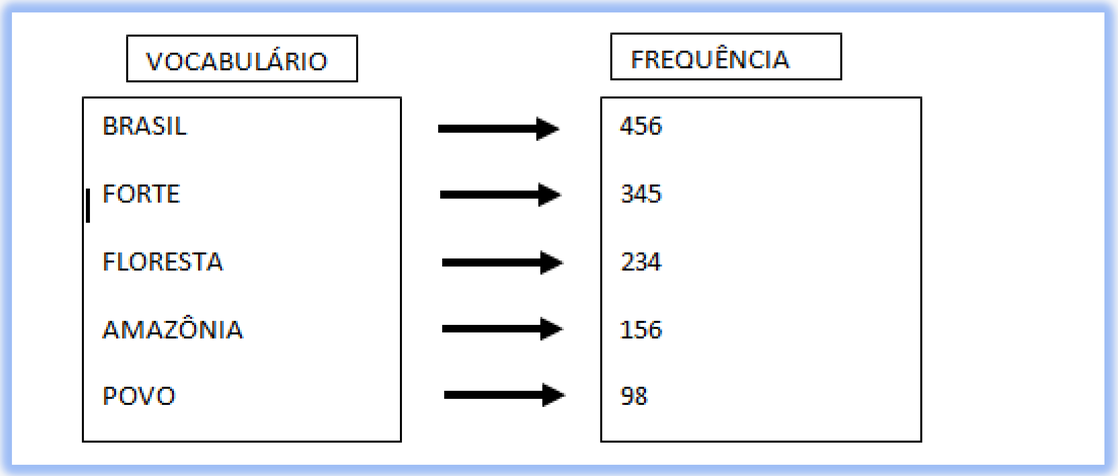


Vamos começar!

Entendendo sobre as frequências dos termos

Frequência de Palavras

É a identificação com que frequência uma palavra/termo ocorre em determinado texto.

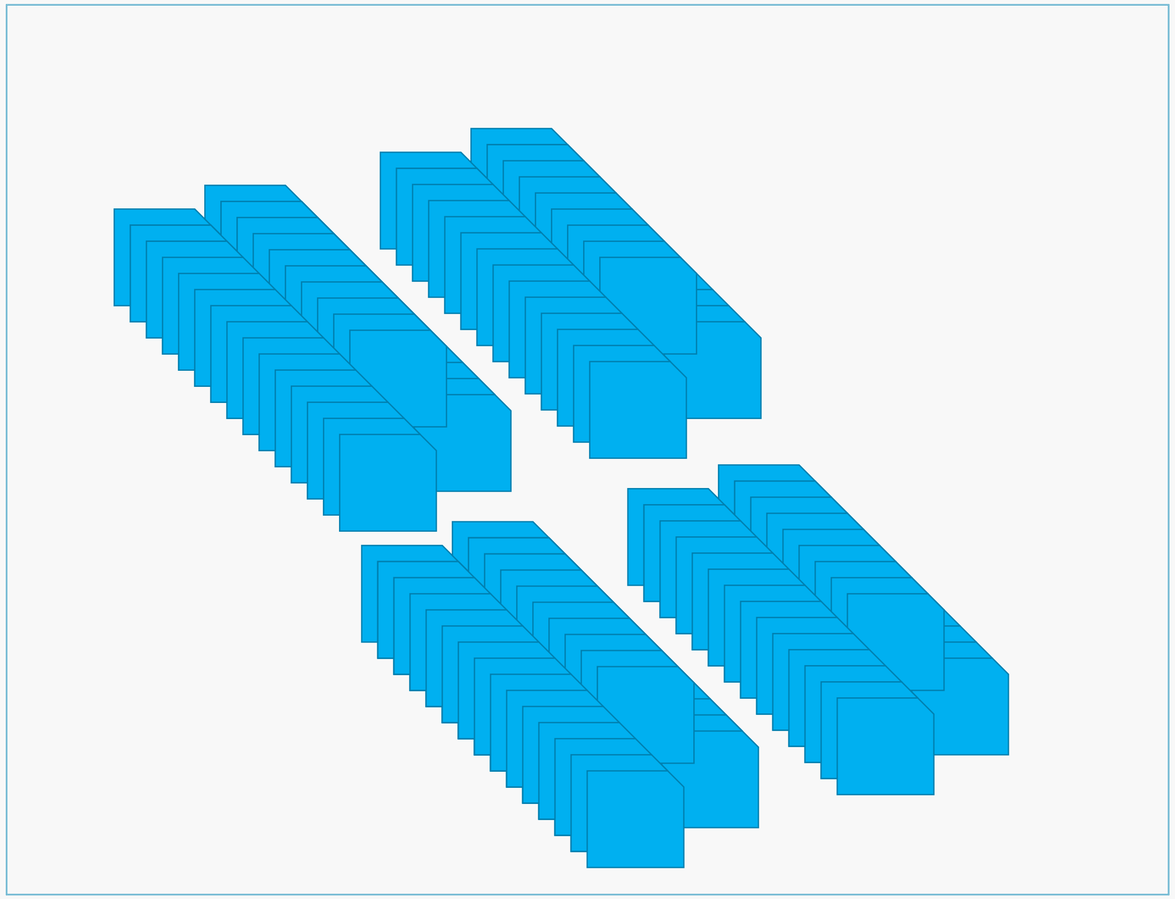


Vamos começar!

Entendendo sobre coleção de textos

Corpus

É uma coleção de textos, é o conjunto completo dos textos que queremos analisar

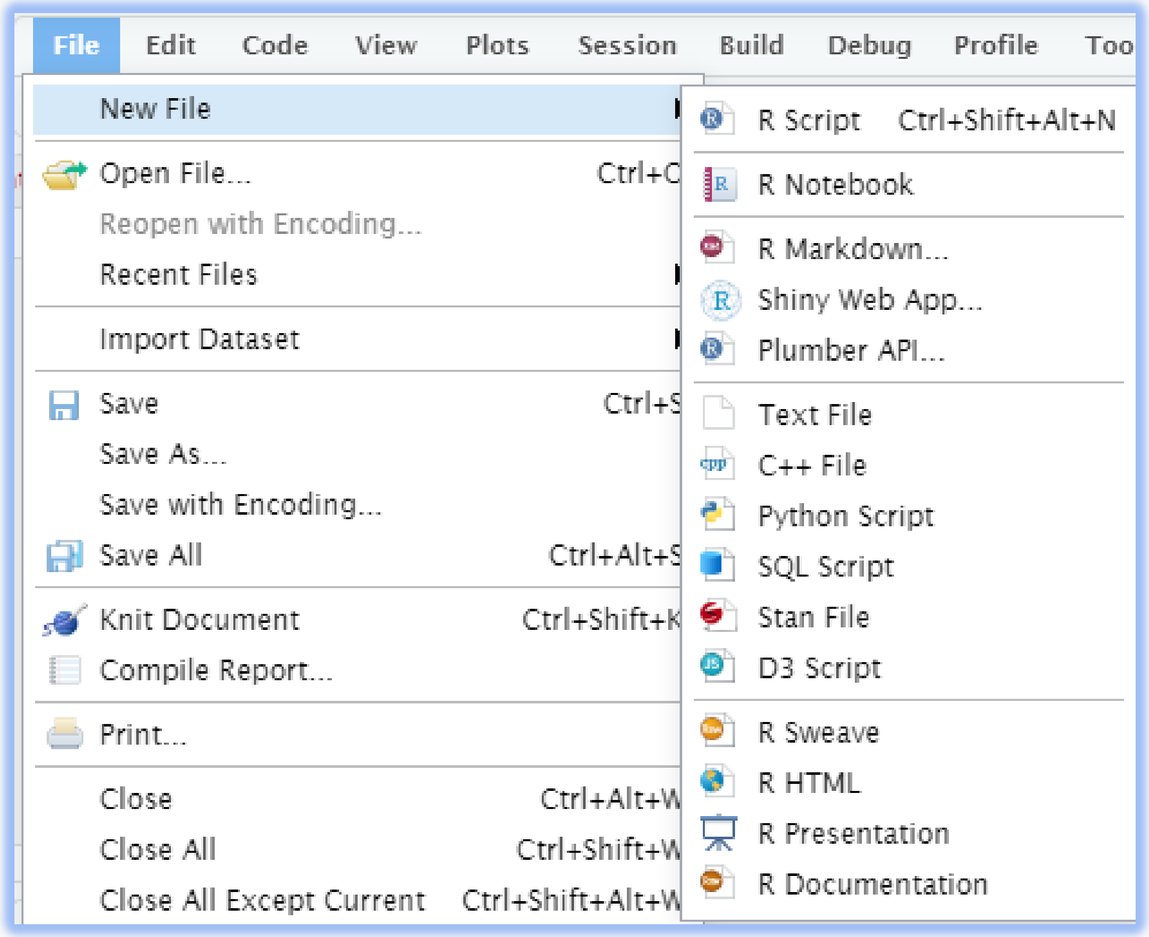


Vamos começar!

Crie um arquivo no formato .Rmd.

Vamos utilizar o R Studio Cloud, mas você pode instalar o Rstudio na sua máquina, sem problemas.

File > New File > R Script



Vamos começar!

Criando o primeiro script para leitura de textos

Iniciando na leitura de Textos

Vamos carregar o pacote **tm.**

Install.packages(tm)

```
1  install.packages(tm)
2  library(tm)
3  # testando a identificação de arquivos
4  docs <- c("meu primeiro texto", "entrada de dados")
5  VCorpus(VectorSource(docs))
```

```
> docs <- c("meu primeiro texto", "entrada de dados")
> VCorpus(VectorSource(docs))
<<VCorpus>>
Metadata:  corpus specific: 0, document level (indexed): 0
Content:  documents: 2
```

TEXT MINING no R

Grimaldo Oliveira

Análise de Texto

Poemas

Leitura arquivos físicos poemas

Criando o segundo script para leitura de textos, agora em arquivos físicos.

1. Escolha dos arquivos

Poemas sobre a vida

- **Rafael Monteiro**
- **Vinicius de Moraes**

```
Os ventos que às vezes
Levam para longe o que amamos
São os mesmos
Que trazem algo mais para ser amado

Nós não podemos chorar pelo
Que nos foi tirado
Nós não iremos... / Nós não iremos...
Nós amaremos o que nos foi dado
Pois tudo que é realmente nosso, não irá embora.
```

```
É claro que a vida é boa
E a alegria, a única indizível emoção
É claro que te acho linda
Em ti bendigo o amor das coisas simples
É claro que te amo
E tenho tudo para ser feliz

Mas acontece que eu sou triste
```

Leitura arquivos físicos poemas

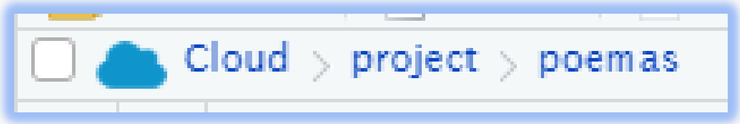
Criando o segundo script para leitura de textos, agora em arquivos físicos.

1. Carregando arquivos

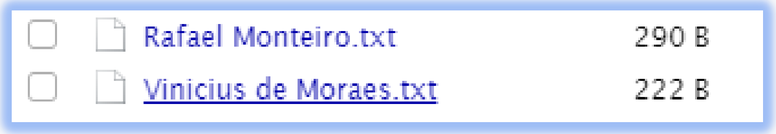
Poemas sobre a vida

- **Rafael Monteiro**
- **Vinicius de Moraes**

2. Criando diretório poemas



3. Upload arquivos



4. Executar o segundo script

```
#Carregando dois poemas de Rafael Monteiro e Vinicius de Moraes
setwd("/cloud/project/poemas")
arquivotxt <- c("/cloud/project/poemas")
textos <- VCorpus(DirSource(arquivotxt, encoding = "UTF-8"),readerC
VCorpus(VectorSource(textos))


# inspeciojnando os arquivos
inspect(textos[1:2])

# identificando os textos
meta(textos[[1]], "id")
meta(textos[[2]], "id")
```

Leitura arquivos físicos poemas

Trabalhando com o segundo script para leitura de textos, utilizando transformações no texto

5. Retirando espaços em branco
6. Transformando tudo em minúscula
7. Colocando retirada *stopwords*
8. Acrescentando mais palavras *stopwords*
9. *Redução de radical da palavra*

```
#####################################
#        TRANSFORMACOES             #
#####################################

# Retirando os espaços em branco
textostrans <- tm_map(textos, stripWhitespace)
inspect(textostrans[[1]])

# Transformando todos os caracteres para minusculo, para uniformizar
textostrans  <- tm_map(textostrans, content_transformer(tolower))
inspect(textostrans[[1]])
```

```
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 152

 claro   vida  boa
  alegria,  única indizível emoção
 claro   acho linda
 ti bendigo  amor  coisas simples
 claro   amo
  tudo  ser feliz
```

TEXT MINING no R

Grimaldo Oliveira

Análise de Texto Regiões Economicas

Leitura arquivos físicos Regiões

Trabalhando com o terceiro script para leitura de textos, utilizando transformações no texto

1. Avaliando a esparcividade
2. Criando a Matriz Termo-Frequencia

```
1  #Carregando dados sobre economia das reiões do Brasil
2  library(tm)
3  setwd("/cloud/project/poemas")
4  arquivotxt <- c("/cloud/project/economia")
5  textos <- VCorpus(DirSource(arquivotxt, encoding = "UTF-8"),readerControl = li
6  VCorpus(VectorSource(textos))
7
8
9  # inspeciojnando os arquivos
10 inspect(textos[1:5])
11
12 # identificando os textos
```

```
<<DocumentTermMatrix (documents: 5, terms: 326)>>
Non-/sparse entries: 393/1237
Sparsity           : 76%
Maximal term length: 17
Weighting          : term frequency (tf)
Sample             :
                    Terms
Docs                 criação estado grande grandes industrial maior produção
  regiao centro-oeste.txt       2      2      0       1          0     0        0
  regiao nordeste.txt           0      0      2       1          1     2        4
  regiao norte.txt              0      0      1       1          3     0        1
  regiao sudeste.txt            0      2      2       0          0     4        1
  região sul.txt                2      1      2       1          0     0        2
                    Terms
```

TEXT MINING no R

Grimaldo Oliveira

Análise de Texto

Discurso

Presidentes

Leitura arquivos Discurso Presidentes

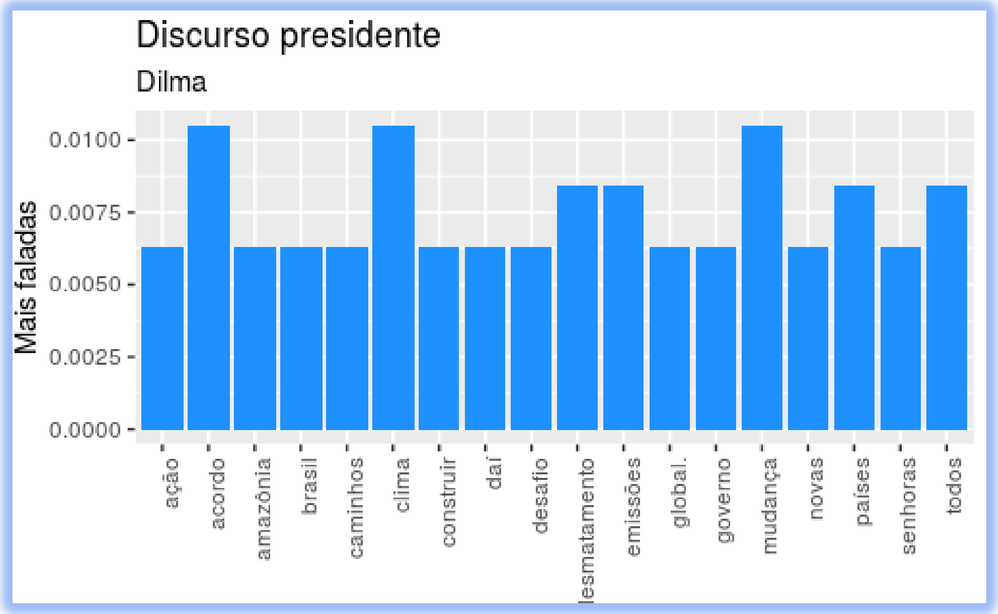
Trabalhando com o quarto script para leitura de textos, utilizando transformações no texto

1. Avaliando as palavras mais ditas

```
#Carregando dados sobre economia das reiões do Brasil
library(tm)
setwd("/cloud/project/poemas")
arquivotxt <- c("/cloud/project/economia")
textos <- VCorpus(DirSource(arquivotxt, encoding = "UTF-8"),readerControl = li
VCorpus(VectorSource(textos))


# inspeciojnando os arquivos
inspect(textos[1:5])

# identificando os textos
```



TEXT MINING no R

Grimaldo Oliveira

Análise de Texto

Discurso

Presidentes

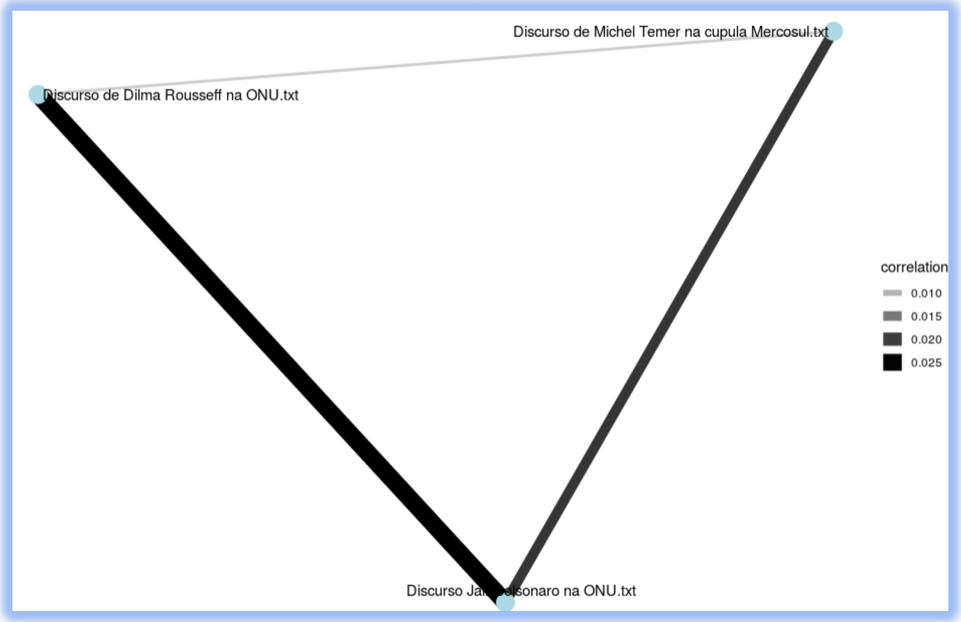
Parte II

Leitura arquivos Discurso Presidentes

Trabalhando com o quinto script para leitura de textos, utilizando transformações no texto

1. Avaliando correlação entre os textos

```
#Gráfico com as correlação. correlação >=0.8 ou <=-0.8
library(ggraph)
library(igraph)
grupo_correlacao %>%
  filter(correlation > .004) %>%
  graph_from_data_frame() %>%
  ggraph(layout = "fr") +
  geom_edge_link(aes(alpha = correlation, width = correlation)) +
  geom_node_point(size = 6, color = "lightblue") +
  geom_node_text(aes(label = name), repel = TRUE) +
  theme_void()
```



TEXT MINING no R

Grimaldo Oliveira

Análise de Texto

Opinião

Saque Emergencial

Análise de Sentimento

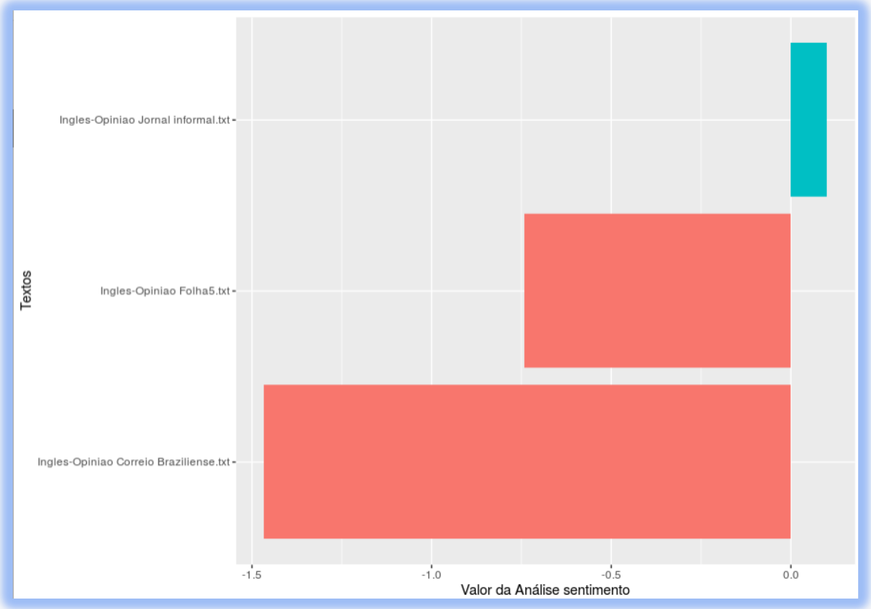
Leitura arquivos Opinião Saque Emergencial

Trabalhando com o sexto script para leitura de textos, utilizando transformações no texto

1. Realizando análise de sentimento
2. Tradução do arquivo para inglês

```
#Preparacao da analise de sentimento, calculando valores.
#Estamos usando o léxico Afinn, pois oferece os valores por sentir
grupo_sentimento <- word_grupo %>%
  inner_join(get_sentiments("afinn"), by = "word") %>%
  group_by(id) %>%
  summarize(value = sum(value * n) / sum(n))


#gráfico de sentimento, opiniao da Folha tem menos negatividade qu
grupo_sentimento %>%
  mutate(recodifica = reorder(id, value)) %>%
  ggplot(aes(recodifica, value, fill = value > 0)) +
```



TEXT MINING no R

Grimaldo Oliveira

Análise de Texto

Resenha Capitã Marvel

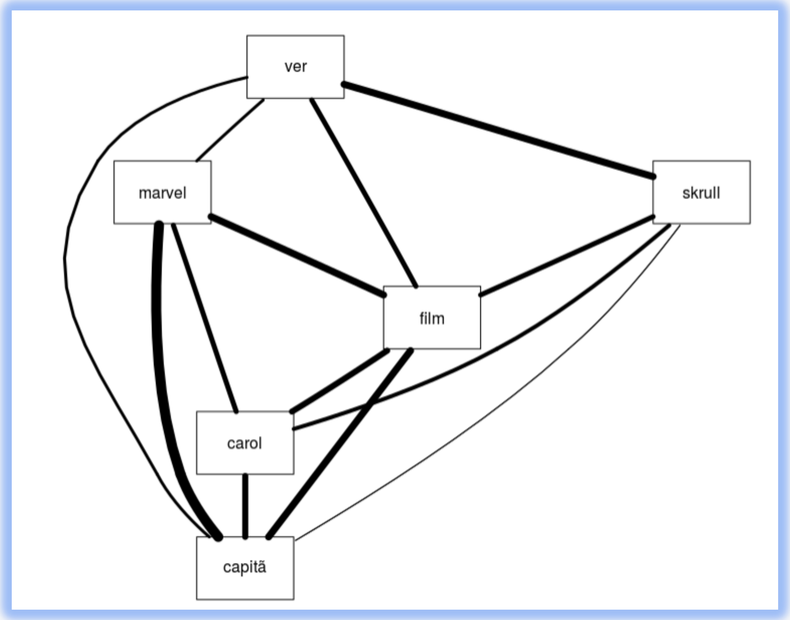
Leitura arquivos Resenha Capitã Marvel

Trabalhando com o sétimo script para leitura de textos, utilizando transformações no texto

1. Realizando análise de texto
2. Avaliando associação de palavras

```
####################################
#Analisando resenha do filme Capitã Marvel
####################################
#lendo o arquvio de dados
setwd("/cloud/project/resenha")
arquivo <- readLines("capita_marvel.txt",16)

#Preparando o Corpus do Arquivo
library(tm)
resenha <- Corpus(VectorSource(arquivo))
resenha
```



TEXT MINING no R

Grimaldo Oliveira

Análise de Texto

Webscrapping

Leitura arquivo site ogol.com.br

Trabalhando com o oitavo script para leitura de texto, utilizando transformações no texto

1. Realizando análise de texto - webscrapping
5. *Realizando estudo das palavras*

```
[48] "retorno   gávea  outra copa "
[49] " flamengo   condição financeira nenhuma  receber zic
[50] " principais preocupações  diretoria flamenguista entr
[51] "ainda assim zico     copa  mundo  voltou  defender  s
[52] " jogo acabou empatado   careca abriu  placar   brasil
[53] "paguei  desobedecer     coração mandava    ir   copa
[54] "últimos anos  idolatria  japão"
[55] " zico  lampejos  grandes momentos  semifinal  campeon
[56] "zico jogou  fla    maior artilheiro  história  clube
[57] " tornou ídolo  kashima antlers ainda  capaz  agraciar
[58] "  importante  mostrar       fazer  futebol  profissã
[59] "anos  tarde comandou  seleção japonesa  país conquist
[60] " técnico  ídolo   turquia comandando  título turco  t
```

PRÁTICA

ENVIE AO PROFESSOR

PREPARE UM ESTUDO
COM AS CARACTERÍSTICAS
APRESENTADAS AQUI NESTE CURSO.
VAMOS TENTAR, FAÇA UMA ANÁLISE
COMPLETA NOS TEXTOS NA WEB

Ciência de dados e mercado de trabalho