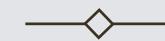




Crédit photographique: Laura Kapfe, Unsplash

# R ET RSTUDIO POUR L'ANALYSE DE DONNÉES TEXTUELLES



Pascal Brissette  
Automne 2023

# Pourquoi analyser des textes avec un langage de programmation?

1. Un langage de programmation offre beaucoup plus de *liberté* qu'un logiciel même très bien fait.
2. Un logiciel, surtout s'il est très puissant (exemple: Excel), utilise beaucoup de *ressources* pour simplement fonctionner.
3. Avec la programmation, vous conservez les *traces* de chaque opération de chacune des parties de votre analyse. Cela est utile pour vous et pour ceux qui voudraient *reproduire* votre expérience.
4. La programmation contribue à formaliser une analyse.
5. La programmation est un appel à la *créativité* et à la *collaboration*.
6. Les ordinateurs et les LLM ne lisent pas des textes comme l'être humain, mais ils sont excellents pour repérer des régularités, des motifs, des corrélations, des cas aberrants. Leur mode de lecture est complémentaire à celui de l'être humain.



Crédit photographique: Janko Ferlic, Unsplash

# Qu'est-ce que R?

---

- R est un langage de programmation complet et un logiciel gratuit distribué sous licence GNU (utilisation, partage, modification autorisés);
- créé en 1993 par Ross Ihaka et Robert Gentleman à l'Université d'Auckland (NZ), inspiré du langage S-PLUS.
- développé et distribué par des statisticiens pour l'analyse de données (intègre de nombreuses fonctions statistiques);
- la première version officielle du langage R a été publiée en février 2000.

# Pourquoi R en particulier?

## ATOOTS

- R est *gratuit*, son code est *ouvert* et son développement est soutenu par une très large *communauté scientifique* (abondante documentation en anglais et en français);
- langage interprété (non compilé): *exécution facile* ;
- il comporte plus de **25 000 extensions** et peut faire beaucoup plus que des calculs statistiques;
- RStudio (Posit), son environnement de développement, est *convivial, attrayant et bien intégré*;
- il peut produire des *graphiques* et documents variés (Word, Power Point, pdf, etc.) et de très *grande qualité*.

## LIMITES

- Il faut consacrer un certain nombre d'heures à l'apprentissage de la syntaxe;
- la plupart de ses opérations se font « en mémoire », ce qui veut dire que l'exécution de certaines tâches peuvent être impossibles ou plus lentes qu'avec d'autres langages (Julia, Go, C#, C++).

# R STUDIO



RStudio

Project: (None)

Environment History Connect

Global Environment

Environment is empty

Files Plots Packages Help

New Folder Delete Rename

ladirec\_class > corpus\_fr > code

Name

20210914\_PB\_0prretraitement\_corpus.R

```
1 # Prétraitement du corpus - Nettoyage et filtrage
2 setwd('~/McGill University/Ladirec_Group - Documents/ladirec_class')
3 libs <- c("tidyverse", "tidytext")
4 lapply(libs, require, character.only = TRUE)
5 rm(list = ls())
6 gc()

# ===== Importation des données =====
7 recits_faim <- read_tsv('data/CRIEM_recitsfaim_supercat')
8 colnames(recits_faim)[6] <- "text"
9 colnames(recits_faim)[27] <- "doc_id"
10 recits_faim <- recits_faim %>% select(doc_id, text, even)
11
12 # ===== Traitements =====
13 # Création d'un data frame rassemblant les documents dont le texte est null
14 recits_faim_null <- filter(recits_faim, text == "null")
15
16 # Remplacer la cellule vide ("null") de la colonne text
17 for (i in 1:nrow(recits_faim)) {
18   if (recits_faim$text[i] == "null") {
19     recits_faim$text[i] <- NA
20   }
21 }
22
23 # Sauvegarde du data frame
24 write_csv(recits_faim, "recits_faim.csv")
25 write_csv(recits_faim_null, "recits_faim_null.csv")

(R Script)
```

Console Terminal Jobs

R 4.1.0 · ~/McGill University/Ladirec\_Group - Documents/ladirec\_class/corpus

```
> lapply(libs, require, character.only = TRUE)
Loading required package: tidyverse
— Attaching packages — tidyverse 1.3.1 —
✓ ggplot2 3.3.5    ✓ purrr  0.3.4
✓ tibble  3.1.4    ✓ dplyr  1.0.7
✓ tidyverse 1.1.3   ✓ stringr 1.4.0
✓ readr   2.0.1    ✓ forcats 0.5.1
— Conflicts — tidyverse_conflicts() —
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()   masks stats::lag()
Loading required package: tidytext
[[1]]
[1] TRUE

[[2]]
[1] TRUE
```

# Installer RStudio sur son ordinateur

TÉLÉCHARGER ET INSTALLER LA DERNIÈRE VERSION DU LANGAGE R



The screenshot shows the CRAN (Comprehensive R Archive Network) website. On the left, there's a sidebar with links like 'CRAN Mirrors', 'What's new?', 'Search', and 'CRAN Team'. The main content area has a large 'R' logo at the top. It discusses precompiled binary distributions for Windows and Mac users. Below that, it lists 'Source Code for all Platforms' and provides links for 'Download R for Linux (Debian/Fedora/Redhat/Ubuntu)', 'Download R for macOS', and 'Download R for Windows'. At the bottom, there's a section titled 'Questions About R' and a 'Supporting CRAN' section.

<https://cran.r-project.org/>

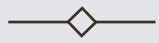
TÉLÉCHARGER ET INSTALLER LA DERNIÈRE VERSION DE RSTUDIO

The screenshot shows the posit.co website. It features a search bar and navigation links for 'Finance', 'Programmation', 'LLM', 'Charts', 'Ville et littérature', 'Textes et articles...', and 'Admin de vie'. The main content is about the 'RStudio IDE', described as the most popular coding environment for R. It highlights its use by millions of people weekly and its features like a console, syntax-highlighting editor, and tools for plotting, history, debugging, and managing your workspace. Two blue buttons at the bottom are labeled 'DOWNLOAD RSTUDIO' and 'DOWNLOAD RSTUDIO SERVER'. Below the main content, there's a section for 'RStudio Desktop'.

<https://posit.co/downloads/>

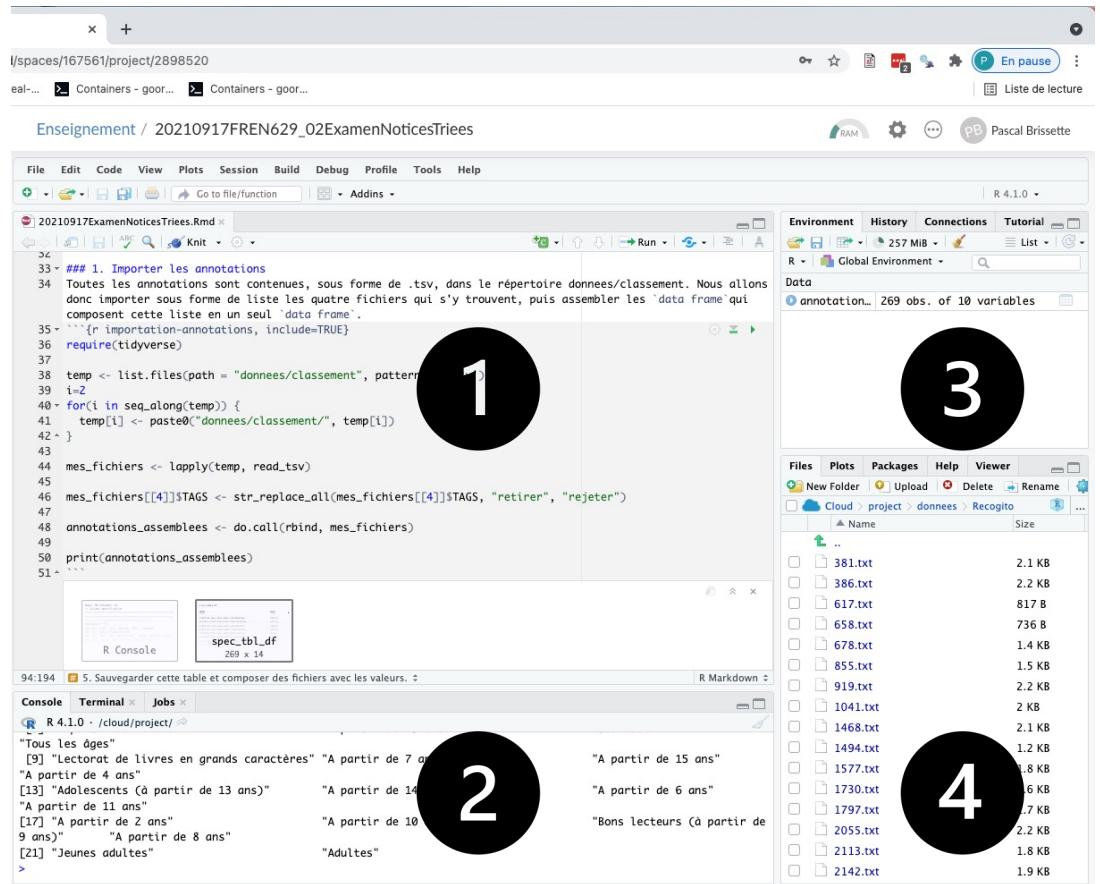


# Environnement de développement RStudio



Par défaut

1. Éditeur de script
2. Console (exécution des commandes)
3. Objets en mémoire
4. Dossiers/graphiques/librairies/aide



# À vous de jouer!

- Ouvrez RStudio
- Dans L'espace de la console, écrivez  $4 + 4$  et enfoncez la touche « Return ».
- Voilà! Vous avez fait vos premiers pas dans R!
- Ci-contre, la commande `getwd()` retourne le chemin du répertoire de travail (*working directory*)

The screenshot shows the RStudio interface with the following details:

- Console Tab:** Displays the R session output:

```
> 4 + 4
[1] 8
>
>
> print('Hello world!')
[1] "Hello world!"
```
- Environment Tab:** Shows the Global Environment, which is currently empty.
- Files Tab:** Lists the contents of the current working directory:

Nom	Type	Lo...	Taille	Valeur
.gitignore	Texte	44 B		
.RData	Données	3,5 KB		
.Rhistory	Texte	4,3 KB		
Atelier_1_vecteur.qmd	Document	15,4 KB		
Atelier_2_liste.qmd	Document	6,1 KB		
Atelier_3_matrice.qmd	Document	7 KB		
Atelier_4_tableau.qmd	Document	7,2 KB		
Atelier_5_filtrer_tableau.qmd	Document	35,8 KB		

# Quatre commandes essentielles

getwd()

install.packages()

library()

The screenshot shows the RStudio interface with the following components:

- Code Editor:** Displays an R script titled "Untitled1" containing the following code:

```
1 ##### Quatre commandes essentielles #####
2
3 getwd()
4
5 setwd("~/Users/pascalbrissette/Desktop")
6
7 install.packages("proustr")
8
9 library(proustr)
10
```
- Console:** Shows the output of the R session:

```
R version 4.1.0 (2021-05-18) -- "Camp Pontanezen"
Copyright (C) 2021 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin17.0 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[Workspace loaded from ~/RData]

> getwd()
[1] "/Users/pascalbrissette"
> setwd("~/Users/pascalbrissette/Desktop")
> getwd()
[1] "/Users/pascalbrissette/Desktop"
> install.packages("proustr")
trying URL 'https://cran.rstudio.com/bin/macosx/contrib/4.1/proustr_0.4.0.tgz'
Content type 'application/x-gzip' length 2979706 bytes (2.8 MB)
=====
downloaded 2.8 MB

The downloaded binary packages are in
  /var/folders/3r/dbg36kd94kdd3dpfx8h0dqym000gn/T//RtmpQVC5dn/downloaded_packages
> library(proustr)
> |
```
- Environment:** Shows the Global Environment table with the message "Environment is empty".
- Packages:** Shows the Packages tab with a list of installed packages:

Name	Description	Version
proto	Prototype Object-Based Programming	1.0.0
<b>proustr</b>	Tools for Natural Language Processing in French	0.4.0
proxy	Distance and Similarity Measures	0.4-26
PRROC	Precision-Recall and ROC Curves for Weighted and Unweighted Data	1.3.1
ps	List, Query, Manipulate System Processes	1.6.0
psych	Procedures for Psychological, Psychometric, and Personality Research	2.1.6
psychotools	Psychometric Modeling Infrastructure	0.6-1
psychotree	Recursive Partitioning Based on Psychometric Models	0.15-4
purrr	Functional Programming Tools	0.3.4
qap	Heuristics for the Quadratic Assignment Problem (QAP)	0.1-1
qdap	Bridging the Gap Between Qualitative Data and Quantitative Analysis	2.4.3
qdapDictionar...	Dictionaries and Word Lists for the 'qdap' Package	1.0.7
qdapRegex	Regular Expression Removal, Extraction, and Replacement Tools	0.7.2
qdapTools	Tools for the 'qdap' Package	1.3.5
qpdf	Split, Combine and Compress PDF Files	1.1

# Vous avez des questions? Consultez la communauté R!

## STACK OVERFLOW

A screenshot of a Stack Overflow question titled "Mutate multiple / consecutive columns (with dplyr or base R)". The question asks how to create "waves" of variables representing mean values for variables 1-10, 11-20, ..., 91-100. It includes sample code using `matrix` and `dplyr::mutate` to generate a data frame with 10 rows and 100 columns. The post has been asked 5 years, 9 months ago and viewed 5k times.

```
mat <- matrix(runif(1000, 1, 10), ncol = 100)
df <- data.frame(mat)
dplyr::mutate(df, ...)
```

The response section shows a user's attempt to use `dplyr` to achieve the same result:

```
df <- df %>%
  mutate(wave_1 = (X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9 + X10) / 10,
        wave_2 = (X11 + X12 + X13 + X14 + X15 + X16 + X17 + X18 + X19 + X20) / 10,
        ...,
        wave_10 = (X91 + X92 + X93 + X94 + X95 + X96 + X97 + X98 + X99 + X100) / 10)
```

Comments invite others to share their approaches.

## R-BLOGGERS

A screenshot of an R-bloggers tutorial titled "Tutorials for learning R". The page features a sidebar with navigation links like HOME, ABOUT, RSS, ADD YOUR BLOG!, LEARN R, R JOBS, and CONTACT US. The main content area displays a post by Tal Galili from December 10, 2015, with a graph showing data points. The post discusses various R packages and tools for data analysis. Below the post are social sharing buttons for Facebook and Twitter, and a sidebar for "Most viewed posts (weekly)".

Basic R : Read so many CSV files  
The quest for faster(er?) row-oriented workflows  
5 Ways to Subset a Data Frame in R  
How to confuse your shareholders by bad data visualization  
How to write the first for loop in R  
Date Format Capture d'écran  
Function With Special Talent from 'caret'

# Ressources



CRAN Project -  
<https://cran.r-project.org/>



RStudio  
<https://www.rstudio.com/>



RStudio Cloud  
<https://rstudio.cloud/>



Stack Overflow  
<https://stackoverflow.com/questions/tagged/r>



R-Bloggers <https://www.r-bloggers.com/>