MANIPULATION DE STRING avec R

Léonard Boisson

15/11/2020

Figure 1: Légende

I - INTRODUCTION

Définition

Le package stringr , développé par Hadley Wickham, a été conçu pour agir comme une simple 'enveloppe' permettant de rendre les fonctionnalités de R applicables aux chaînes de caractères plus cohérentes, simples et faciles à utiliser

Historique

Stringr a été construit à partir du package stringi, qui lui utilise la librairie C/C++ de la ICU (International Components for Unicode), fournissant des implémentations rapides et robustes couvrant pratiquement toutes les manipulations de chaînes de caractères imaginables.

Cette particularité permet au package stringr d'offrir des fonctions qui gèrent convenablement les valeurs manquantes NA ainsi que les caractères de longueur nulle, en plus d'assurer une cohérence au niveau des noms de fonction et d'argument.

Finalement, toutes les fonctionnalités de stringr retournent des structures de données en sortie qui correspondent à celles reçues en entrée par les autres fonctions du package. Cette dernière caractéristique simplifie de beaucoup l'utilisation du résultat d'une fonction comme argument en entrée d'une autre fonction.

Pourquoi ?

Deux raison principales :

- Car nous souhaitons travaillez sur des textes et plus précisement faire du *text mining* sur ces données.
- Car nous souhaitons *nettoyer/récupérer* des infos empriosnnées dans des chaînes de caractères.

Chargement du package

```
library('stringr')
```

II - FONCTIONS DE BASE

Petit aide mémoire visuelles de quelques fonctions expliquées plus bas

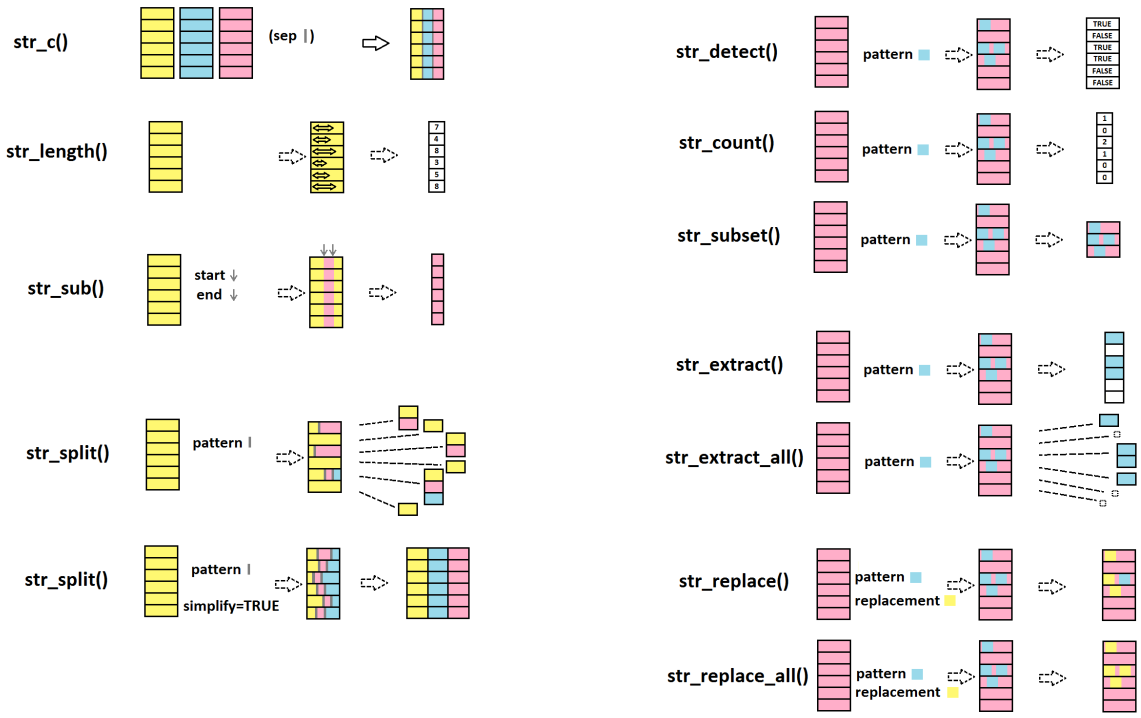

Figure 2: Légende

On va considérer le tableau suivant pour tous les exemples qui suivront :

```
library(readxl)
```

```
d <- read_excel("C:/Users/leona/netflix_titles.xlsx", col_types = c("text",
    "text", "text", "text", "text", "text",
    "text", "text", "text", "text", "text",
    "text"))
```

```
d
```

```
## # A tibble: 15 x 12
##    show_id type  title director cast  country date_added release_year rating
##    <chr>   <chr> <chr> <chr>    <chr> <chr>   <chr>      <chr>        <chr>
##  1 811456~ Movie Norm~ Richard~ Alan~ United~ September~ 2019         TV-PG
##  2 801174~ Movie Jand~ <NA>     Jand~ United~ September~ 2016         TV-MA
##  3 702344~ TV S~ Tran~ <NA>     Pete~ United~ September~ 2013         TV-Y7~
##  4 800586~ TV S~ Tran~ <NA>     Will~ United~ September~ 2016         TV-Y7
##  5 801259~ Movie #rea~ Fernand~ Nest~ United~ September~ 2017         TV-14
##  6 801638~ TV S~ Apac~ <NA>     Albe~ Spain   September~ 2016         TV-MA
##  7 703049~ Movie Auto~ Gabe Ib~ Anto~ Bulgar~ September~ 2014         R
##  8 801640~ Movie Fabr~ Rodrigo~ Fabr~ Chile   September~ 2017         TV-MA
##  9 801179~ TV S~ Fire~ <NA>     <NA>  United~ September~ 2017         TV-MA
## 10 703049~ Movie Good~ Henrik ~ Jame~ United~ September~ 2014         R
## 11 801697~ Movie Joaq~ JosÃ© M~ Joaq~ <NA>    September~ 2017         TV-MA
## 12 702992~ Movie Kidn~ Daniel ~ Jim ~ Nether~ September~ 2015         R
## 13 801824~ Movie Kris~ <NA>     Dama~ <NA>    September~ 2009         TV-Y7
## 14 801824~ Movie Kris~ Munjal ~ Dama~ <NA>    September~ 2013         TV-Y7
## 15 801825~ Movie Kris~ Munjal ~ Dama~ <NA>    September~ 2016         TV-Y
## # ... with 3 more variables: duration <chr>, listed_in <chr>, description <chr>
```

Movies and TV Shows listings on Netflix

1. Concaténer des chaînes de caractères avec str_c

- Par défaut, str_c concatène en ajoutant un espace entre les différentes chaînes. Il est possible de spécifier un autre séparateur avec son argument sep.

```
str_c(d$title, d$director, sep = " - ")
```

```
##  [1] "Norm of the North: King Sized Adventure - Richard Finn, Tim Maltby"
##  [2] NA
##  [3] NA
##  [4] NA
##  [5] "#realityhigh - Fernando Lebrija"
##  [6] NA
##  [7] "Automata - Gabe IbÃ¡Ã±ez"
##  [8] "Fabrizio Copano: Solo pienso en mi - Rodrigo Toro, Francisco Schultz"
##  [9] NA
## [10] "Good People - Henrik Ruben Genz"
## [11] "JoaquÃn Reyes: Una y no mÃ¡s - JosÃ© Miguel Contreras"
## [12] "Kidnapping Mr. Heineken - Daniel Alfredson"
## [13] NA
## [14] "Krish Trish and Baltiboy: Battle of Wits - Munjal Shroff, Tilak Shetty"
## [15] "Krish Trish and Baltiboy: Best Friends Forever - Munjal Shroff, Tilak Shetty"
```

- Si on veut concaténer les différents éléments d'un vecteur entre eux, il faut ajouter l'argument collapse qui renvoi le séparateur.

```
str_c(d$type, collapse = " - ")
```

```
## [1] "Movie - Movie - TV Show - TV Show - Movie - TV Show - Movie - Movie - TV Show - Movie - Movie -
```

2. Convertir en majuscule et minuscule

a) str_to_lower : convertit en minuscule

```
str_to_lower(d$cast)
```

```
##  [1] "alan marriott, andrew toth, brian dobson, cole howard, jennifer cameron, jonathan holmes, lee
##  [2] "jandino asporaat"
##  [3] "peter cullen, sumalee montano, frank welker, jeffrey combs, kevin michael richardson, tania gu
##  [4] "will friedle, darren criss, constance zimmer, khary payton, mitchell whitfield, stuart allan,
##  [5] "nesta cooper, kate walsh, john michael higgins, keith powers, alicia sanz, jake borelli, kid i
##  [6] "alberto ammann, eloy azorãn, verã³nica echegui, lucãa jimã©nez, claudia traisac"
##  [7] "antonio banderas, dylan mcdermott, melanie griffith, birgitte hjort sã¸rensen, robert forster,
##  [8] "fabrizio copano"
##  [9] NA
## [10] "james franco, kate hudson, tom wilkinson, omar sy, sam spruell, anna friel, thomas arnold, oli
## [11] "joaquãn reyes"
## [12] "jim sturgess, sam worthington, ryan kwanten, anthony hopkins, mark van eeuwen, thomas cocquere
## [13] "damandeep singh baggan, smita malhotra, baba sehgal"
## [14] "damandeep singh baggan, smita malhotra, baba sehgal, deepak chachra"
## [15] "damandeep singh baggan, smita malhotra, deepak chachra"
```

***b) str_to_upper : convertit en majuscule**

```
str_to_upper(d$cast)
```

```
##  [1] "ALAN MARRIOTT, ANDREW TOTH, BRIAN DOBSON, COLE HOWARD, JENNIFER CAMERON, JONATHAN HOLMES, LEE
##  [2] "JANDINO ASPORAAT"
##  [3] "PETER CULLEN, SUMALEE MONTANO, FRANK WELKER, JEFFREY COMBS, KEVIN MICHAEL RICHARDSON, TANIA GU
##  [4] "WILL FRIEDLE, DARREN CRISS, CONSTANCE ZIMMER, KHARY PAYTON, MITCHELL WHITFIELD, STUART ALLAN,
##  [5] "NESTA COOPER, KATE WALSH, JOHN MICHAEL HIGGINS, KEITH POWERS, ALICIA SANZ, JAKE BORELLI, KID I
##  [6] "ALBERTO AMMANN, ELOY AZORÃN, VERÃ³NICA ECHEGUI, LUCÃA JIMÃ©NEZ, CLAUDIA TRAISAC"
##  [7] "ANTONIO BANDERAS, DYLAN MCDERMOTT, MELANIE GRIFFITH, BIRGITTE HJORT SÃ¸RENSEN, ROBERT FORSTER,
##  [8] "FABRIZIO COPANO"
##  [9] NA
## [10] "JAMES FRANCO, KATE HUDSON, TOM WILKINSON, OMAR SY, SAM SPRUELL, ANNA FRIEL, THOMAS ARNOLD, OLI
## [11] "JOAQUÃN REYES"
## [12] "JIM STURGESS, SAM WORTHINGTON, RYAN KWANTEN, ANTHONY HOPKINS, MARK VAN EEUWEN, THOMAS COCQUERE
## [13] "DAMANDEEP SINGH BAGGAN, SMITA MALHOTRA, BABA SEHGAL"
## [14] "DAMANDEEP SINGH BAGGAN, SMITA MALHOTRA, BABA SEHGAL, DEEPAK CHACHRA"
## [15] "DAMANDEEP SINGH BAGGAN, SMITA MALHOTRA, DEEPAK CHACHRA"
```

***c) str_to_title : capitalise les éléments d'un vecteurde chaînes de caractères**

```
str_to_title(d$description)
```

```
##  [1] "Before Planning An Awesome Wedding For His Grandfather, A Polar Bear King Must Take Back A Sto
##  [2] "Jandino Asporaat Riffs On The Challenges Of Raising Kids And Serenades The Audience With A Rou
```

```
##  [3] "With The Help Of Three Human Allies, The Autobots Once Again Protect Earth From The Onslaught
##  [4] "When A Prison Ship Crash Unleashes Hundreds Of Decepticons On Earth, Bumblebee Leads A New Aut
##  [5] "When Nerdy High Schooler Dani Finally Attracts The Interest Of Her Longtime Crush, She Lands I
##  [6] "A Young Journalist Is Forced Into A Life Of Crime To Save His Father And Family In This Series
##  [7] "In A Dystopian Future, An Insurance Adjuster For A Tech Company Investigates A Robot Killed Fo
##  [8] "Fabrizio Copano Takes Audience Participation To The Next Level In This Stand-Up Set While Refl
##  [9] "As California's 2016 Fire Season Rages, Brave Backcountry Firefighters Race To Put Out The Fla
## [10] "A Struggling Couple Can't Believe Their Luck When They Find A Stash Of Money In The Apartment
## [11] "Comedian And Celebrity Impersonator Joaquãn Reyes Decides To Be His Zesty Self For A Night Of
## [12] "When Beer Magnate Alfred \"Freddy\" Heineken Is Kidnapped In 1983, His Abductors Make The Larg
## [13] "A Team Of Minstrels, Including A Monkey, Cat And Donkey, Narrate Folktales From The Indian Reg
## [14] "An Artisan Is Cheated Of His Payment, A Lion Of His Throne And A Brother Of His Inheritance In
## [15] "A Cat, Monkey And Donkey Team Up To Narrate Folktales About Friendship From Northeast India, A
```

3. Trouver la longueur d'une chaîne de caractère avec str_lenght

- Renvoie le nombre de caractères de chaque éléments

```
str_length(d$description)
```

```
##  [1] 140 145 140 126 148 137 149 149 147 134 150 121 131 140 139
```

4. Suppression des espace en début et fin

- Supprime les espaces en début et fin de chaîne de caractères

```
str_trim('       Le vent souffle sur les plaines de la Bretagne armoricaine       ')
```

```
## [1] "Le vent souffle sur les plaines de la Bretagne armoricaine"
```

5. Découper des chaînes de caractères avec str_split

- Découpe la chaîne de caractère en fonction d'un délimiteur.

```
str_split('Et un et deux et trois zéro', ' ')
```

```
## [[1]]
## [1] "Et"    "un"    "et"    "deux"  "et"    "trois" "zéro"
```

- Appliqué à un vecteur/colonne d'un tableau, str_split créé une ou plusieurs listes

```
str_split(d$cast, ",")
```

```
## [[1]]
##  [1] "Alan Marriott"     " Andrew Toth"      " Brian Dobson"
##  [4] " Cole Howard"      " Jennifer Cameron" " Jonathan Holmes"
##  [7] " Lee Tockar"       " Lisa Durupt"      " Maya Kay"
## [10] " Michael Dobson"
```

```
##
## [[2]]
## [1] "Jandino Asporaat"
##
## [[3]]
##  [1] "Peter Cullen"              " Sumalee Montano"
##  [3] " Frank Welker"             " Jeffrey Combs"
##  [5] " Kevin Michael Richardson" " Tania Gunadi"
##  [7] " Josh Keaton"              " Steve Blum"
##  [9] " Andy Pessoa"              " Ernie Hudson"
## [11] " Daran Norris"             " Will Friedle"
##
## [[4]]
## [1] "Will Friedle"        " Darren Criss"       " Constance Zimmer"
## [4] " Khary Payton"       " Mitchell Whitfield" " Stuart Allan"
## [7] " Ted McGinley"       " Peter Cullen"
##
## [[5]]
##  [1] "Nesta Cooper"          " Kate Walsh"           " John Michael Higgins"
##  [4] " Keith Powers"         " Alicia Sanz"          " Jake Borelli"
##  [7] " Kid Ink"              " Yousef Erakat"        " Rebekah Graf"
## [10] " Anne Winters"         " Peter Gilroy"         " Patrick Davis"
##
## [[6]]
## [1] "Alberto Ammann"     " Eloy AzorÃn"      " VerÃ³nica Echegui"
## [4] " LucÃa JimÃ©nez"   " Claudia Traisac"
##
## [[7]]
## [1] "Antonio Banderas"          " Dylan McDermott"
## [3] " Melanie Griffith"         " Birgitte Hjort SÃ¸rensen"
## [5] " Robert Forster"           " Christa Campbell"
## [7] " Tim McInnerny"            " Andy Nyman"
## [9] " David Ryall"
##
## [[8]]
## [1] "Fabrizio Copano"
##
## [[9]]
## [1] NA
##
## [[10]]
##  [1] "James Franco"      " Kate Hudson"      " Tom Wilkinson"
##  [4] " Omar Sy"          " Sam Spruell"      " Anna Friel"
##  [7] " Thomas Arnold"    " Oliver Dimsdale"  " Diana Hardcastle"
## [10] " Michael Jibson"   " Diarmaid Murtagh"
##
## [[11]]
## [1] "JoaquÃn Reyes"
##
## [[12]]
## [1] "Jim Sturgess"      " Sam Worthington"  " Ryan Kwanten"
## [4] " Anthony Hopkins"  " Mark van Eeuwen"  " Thomas Cocquerel"
## [7] " Jemima West"      " David Dencik"
##
```

```
## [[13]]
## [1] "Damandeep Singh Baggan" " Smita Malhotra"        " Baba Sehgal"
##
## [[14]]
## [1] "Damandeep Singh Baggan" " Smita Malhotra"        " Baba Sehgal"
## [4] " Deepak Chachra"
##
## [[15]]
## [1] "Damandeep Singh Baggan" " Smita Malhotra"        " Deepak Chachra"
```

- Appliqué à un tableau, str_plit peut aussi créer une matrice si on ajoute l'argument simplify = TRUE

```
str_split(d$date_added, ",", simplify = TRUE)
```

```
##       [,1]          [,2]
##  [1,] "September 9" " 2019"
##  [2,] "September 9" " 2016"
##  [3,] "September 8" " 2018"
##  [4,] "September 8" " 2018"
##  [5,] "September 8" " 2017"
##  [6,] "September 8" " 2017"
##  [7,] "September 8" " 2017"
##  [8,] "September 8" " 2017"
##  [9,] "September 8" " 2017"
## [10,] "September 8" " 2017"
## [11,] "September 8" " 2017"
## [12,] "September 8" " 2017"
## [13,] "September 8" " 2017"
## [14,] "September 8" " 2017"
## [15,] "September 8" " 2017"
```

6. Extraire des sous-chaînes par position avec str_sub

- Extrait des sous-chaînes par position en indiquant les positions des premier et dernier caractères

```
str_sub(d$type, 1, 4)
```

```
##  [1] "Movi" "Movi" "TV S" "TV S" "Movi" "TV S" "Movi" "Movi" "TV S" "Movi"
## [11] "Movi" "Movi" "Movi" "Movi" "Movi"
```

7. Détecter des motifs

a) str_detect

- Détecte la présence d'un ou plusieurs caractères parmi les éléments d'un vecteur, en renvoyant un vecteur de valeurs logiques

```
str_detect(d$type, 'Movie')
```

```
##  [1]  TRUE  TRUE FALSE FALSE  TRUE FALSE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE
## [13]  TRUE  TRUE  TRUE
```

b. str_count

- Renvoie le nombre de fois ou le caractère choisis est présent

```
str_count(d$title, 'o')
```

```
##  [1] 3 1 1 3 0 0 1 6 0 3 2 0 1 2 2
```

c. str_subset

- Renvoie seulement les éléments où la chaîne de caractère choisit est présente.

```
str_subset(d$type, 'TV')
```

```
## [1] "TV Show" "TV Show" "TV Show" "TV Show"
```

8. Extraire des caractères avec str_extract

- Permet d'extraire les valeurs correspondant à une chaîne de caractère en utilisant des expressions régulières en arguments.
- Str_extract extrait la première occurrence

```
str_extract(d$duration, '^\\d+')
```

```
##  [1] "90"  "94"  "1"   "1"   "99"  "1"   "110" "60"  "1"   "90"  "78"  "95"
## [13] "58"  "62"  "65"
```

```
# Isole les numéros
```

- Str_extract_all extrait l'ensemble des nombres présents

```
str_extract_all(d$date_added, "\\d+")
```

```
## [[1]]
## [1] "9"    "2019"
##
## [[2]]
## [1] "9"    "2016"
##
## [[3]]
## [1] "8"    "2018"
##
## [[4]]
## [1] "8"    "2018"
##
## [[5]]
## [1] "8"    "2017"
##
```

```
## [[6]]
## [1] "8"    "2017"
##
## [[7]]
## [1] "8"    "2017"
##
## [[8]]
## [1] "8"    "2017"
##
## [[9]]
## [1] "8"    "2017"
##
## [[10]]
## [1] "8"    "2017"
##
## [[11]]
## [1] "8"    "2017"
##
## [[12]]
## [1] "8"    "2017"
##
## [[13]]
## [1] "8"    "2017"
##
## [[14]]
## [1] "8"    "2017"
##
## [[15]]
## [1] "8"    "2017"
```

9. Remplacer des motifs avec str_replace

- Remplace une chaîne de caractère ou un motif par un autre

```
str_replace(d$date_added, 'September', '9,')
```

```
##  [1] "9, 9, 2019" "9, 9, 2016" "9, 8, 2018" "9, 8, 2018" "9, 8, 2017"
##  [6] "9, 8, 2017" "9, 8, 2017" "9, 8, 2017" "9, 8, 2017" "9, 8, 2017"
## [11] "9, 8, 2017" "9, 8, 2017" "9, 8, 2017" "9, 8, 2017" "9, 8, 2017"
```

- str_replace_all permet de spécifier plusieurs remplacements d'un coup

```
str_remove_all(d$type, c('Movie'='M', 'TV Show'='T'))
```

```
##  [1] "M" "M" "T" "T" "M" "T" "M" "M" "T" "M" "M" "M" "M" "M" "M"
```

10. Modificateur de motifs avec fixed

- On peut spécifier qu'un motif n'est pas une expression régulière mais une chaîne de caractère en lui appliquant la fonction fixed.

- Par exemple, si on veut compter le nombre de point, le paramétrage par défaut ne fonctionnera pas car dans une expression régulière, le point signifie "*n'importe quel caractère*".

```
# Paramétrage par défaut
str_count(d$description, '.')
```

```
##  [1] 140 145 140 126 148 137 149 149 147 134 150 121 131 140 139
```

```
# Paramétrage avec fixed

str_count(d$description, fixed('.'))
```

```
##  [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
```

III - LES EXPRESSIONS REGULIERES

Les ER sont utiles car les chaînes contiennent généralement des données non structurées ou semi-structurées, et les ER sont un langage concis pour décrire les motifs des chaînes. Lorsque vous regardez une ER pour la première fois, vous pensez qu'un chat a marché sur votre clavier, mais au fur et à mesure que votre compréhension s'améliore, les ER commencent à avoir un sens.

Très simplement, l'utilisation d'une expression régulière (ou regex) vous permet de parcourir une séquence de texte, afin d'en faire ressortir les motifs compatibles avec le pattern d'entrée. L'objectif ? Visualiser, modifier, ou encore supprimer...

Quelques liens pour mieux comprendre les RegEx :

- (https://data.hypotheses.org/959)
- (https://informatique-mia.inrae.fr/r4ciam/node/131)
- (https://thinkr.fr/r-les-expressions-regulieres/)

III. RESSOURCES

- (https://stringr.tidyverse.org/) Le site officiel de stringr
- (https://stringr.tidyverse.org/reference/index.html) Liste des fonctions et pages d'aide associées
- (https://stringr.tidyverse.org/articles/regular-expressions.html) Article dédié aux expressions régulières, en anglais
- (http://perso.ens-lyon.fr/lise.vaudor/Descriptoire/_book/intro.html#import-dans-r-de-donnees-textuelles) Cours plus avancé sur stringr
- (https://www.programmersought.com/article/7567693645/) Tutoriel détaillé sur le package stringr