Consolider sa reproductibilité avec le package prefixer

Bénédicte Garnier

2023-10-24

Résumé

Il est très simple d’ajouter, dans un script R, les noms des packages aux appels des fonctions. Sous forme de préfixes, ces indications permettent de repérer les packages utilisés et donc faciliter la reproductibilité de l’analyse. Nous prenons ici l’exemple d’un script générant un nuage de mots à partir du package de statistique textuelle R.temis.

Table des matières

**Fonctions utilisées dans la fiche**

| Packages | Fonctions |
| --- | --- |
| **prefixer** | prefixer |
| **R.temis** | import\_corpus |
| **R.temis** | build\_dtm |
| **R.temis** | frequent\_terms |
| **R.temis** | combine\_terms |
| **R.temis** | word\_cloud |

Pour illustrer ce propos, nous reprennons le script de la fiche de Coralie Cottet [*Générer des graphes de mot avec R.temis*](http://mthevenin.github.io/assistoolsms/R/assist/posts/nuage_de_mot/nuage_de_mot.html). Nous allons ajouter un prefixe à chaque fonction utilisée pour générer un nuage de mots.

**Documentation** :

* [**RStudio Addins**](http://docs.posit.co/ide/user/ide/guide/productivity/add-ins.html) sont des extensions pour l’environnement de développement intégré (IDE) RStudio
* [**prefixer: Shiny Gadget to interactively prefix function in a script**](http://rdrr.io/github/dreamRs/prefixer/man/prefixer.html)
* Séminaire R à l’Usage des Sciences Sociales, [*Esquisse, moins tu sais coder plus tu vas rigoler*](http://russ.site.ined.fr/fr/annee-2022-2023/esquisse/)

**Installation des packages** :

install.packages("remotes")  
install.packages("R.temis")  
install.packages("dyplyr")

**Fichiers à utilisés** :

Les [voeux du président Hollande pour les années 2013 à 2017](http://www.vie-publique.fr/discours-dans-lactualite/269998-les-voeux-des-presidents-de-la-republique-depuis-1974).

Chaque discours correspond à un fichier texte et ces 5 discours sont placés dans un dossier unique appelé *dossier\_de\_textes*

# Le script d’origine

## Appel des packages  
library(R.temis)  
library(tidyverse)  
  
## Statistiques textuelles  
## Création du tableau lexical associé au corpus des discours  
  
# Importation des textes avec la fonction `import\_corpus`  
corpusv <- import\_corpus("datas/dossier\_de\_textes", format="txt", language ="fr")  
  
# Création du tableau lexical associé   
# On ne garde pas les mots-outils et on garde aussi les mots de 1 seule lettre  
  
dtm <-build\_dtm(corpusv, remove\_stopwords = T, min\_length = 1)  
dtm  
  
# On obtient un \*tableau lexical\* de 5 unités de texte (les voeux) et 1575 mots différents ...   
  
  
## Afficher les occurrences des mots du corpus  
# Calcul des occurrences des mots   
frequent\_terms(dtm)   
  
# On remarque que les mots `tous`, `tout` et `toutes` sont très fréquents et qu ils sont employés dans des contextes similaires. On souhaite alors les associer à un terme unique : `tous.tes` (opération de `lemmatisation`)  
  
#### Personalisation du lexique  
  
# Création du dictionnaire/lexique associé au tableau lexical   
dic <-dictionary(dtm)   
  
# On crée une nouvelle table `dic2` en créant une première colonne appelée `word` avec l'intitulé des lignes du tableau et en remplaçant le contenu de la colonne cette première colonne ... C'est un lemmatiseur !  
dic2 = dic %>%  
 rownames\_to\_column(var="word") %>%   
 mutate(Term = word)  
  
row.names(dic2) <- dic2$word  
  
# On peut alors remplacer les mots de la colonne Term par l'intitulé souhaité. Ici tout.tes  
dic2$Term[dic2$word == "toutes"] <- "tous.tes"  
dic2$Term[dic2$word == "tout"] <- "tous.tes"  
dic2$Term[dic2$word == "tous"] <- "tous.tes"  
  
# Lemmatisation à l'aide du lemmatiseur personnalisé  
dtmlem <-combine\_terms(dtm, dic2)  
  
# Ensemble de mots à retirer  
mots\_a\_retirer <- c("a", "plus")  
  
# Suppression de mots dans le tableau lexical  
dtm2<-dtmlem[, !colnames(dtmlem) %in% mots\_a\_retirer]  
  
# Nouvelles fréquences des mots associés au corpus  
frequent\_terms(dtm2)  
  
  
### Affichage du nuage de mots  
  
cloud<-word\_cloud(dtm2, color= 'black', min.freq=1,n =50)   
title(main = "Mots les plus fréquents dans les discours de F. Hollande entre 2013 et 2017")

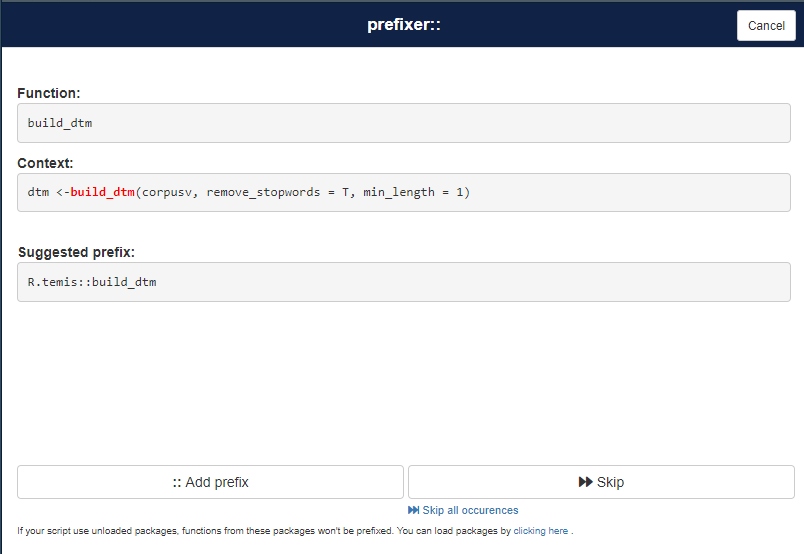
# Ajouter le nom des packages aux fonctions avec prefixer

## Installer le package *prefixer* à partir de GitHub

L’addin prefix est disponible sur la forge logicielle de *dreamRs*

install.packages("remotes")  
install\_github("dreamRs/prefixer")

Celui-ci permet d’ouvrir une interface qui propose pour chaque fonction du script de lui ajouter un préfixe composé du nom de la fonction suivi de ::



## Prefixer le script

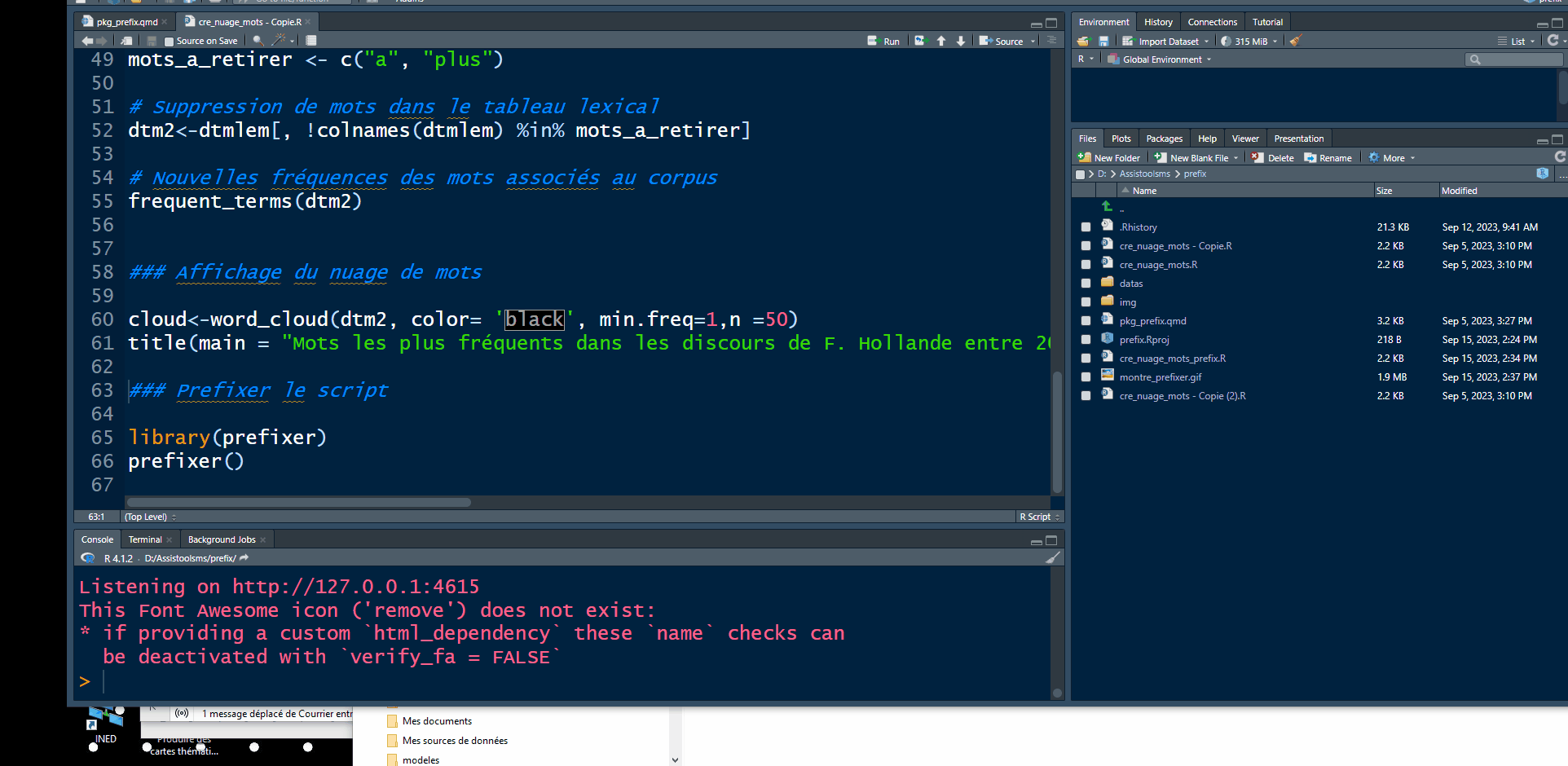
A la fin du script R, on charge le package prefixer puis on ouvre l’interface avec la commande prefixer()

## Appel du package  
library(prefixer)  
## Ouverture de l'interface RShiny  
prefixer()

Dans le cas de notre script:

#cloud<-word\_cloud(dtm2, color= 'black', min.freq=1,n =50)   
#title(main = "Mots les plus fréquents dans les discours de F. Hollande entre 2013 et 2017")  
  
library(prefixer)  
prefixer()

## Utiliser la fenêtre shiny permettant de choisir le préfixe



# Résultat: le script est enrichi

## Statistiques textuelles  
## Création du tableau lexical associé au corpus des discours  
  
# Importation des textes avec la fonction `import\_corpus`  
corpusv <- R.temis::import\_corpus("datas/dossier\_de\_textes", format="txt", language ="fr")  
  
# Création du tableau lexical associé   
# On ne garde pas les mots-outils et on garde aussi les mots de 1 seule lettre  
  
dtm <-R.temis::build\_dtm(corpusv, remove\_stopwords = T, min\_length = 1)  
dtm  
  
# On obtient un \*tableau lexical\* de 5 unités de texte (les voeux) et 1575 mots différents ...   
  
  
## Afficher les occurrences des mots du corpus  
# Calcul des occurrences des mots   
R.temis::frequent\_terms(dtm)   
  
  
# On remarque que les mots `tous`, `tout` et `toutes` sont très fréquents et qu ils sont employés dans des contextes similaires. On souhaite alors les associer à un terme unique : `tous.tes` (opération de `lemmatisation`)  
  
#### Personalisation du lexique  
  
# Création du dictionnaire/lexique associé au tableau lexical   
dic <-R.temis::dictionary(dtm)   
  
# On crée une nouvelle table `dic2` en créant une première colonne appelée `word` avec l'intitulé des lignes du tableau et en remplaçant le contenu de la colonne cette première colonne ... C'est un lemmatiseur !  
dic2 = dic %>%  
 tibble::rownames\_to\_column(var="word") %>%   
 dplyr::mutate(Term = stringr::word)  
  
row.names(dic2) <- dic2$word  
  
# On peut alors remplacer les mots de la colonne Term par l'intitulé souhaité. Ici tout.tes  
dic2$Term[dic2$word == "toutes"] <- "tous.tes"  
dic2$Term[dic2$word == "tout"] <- "tous.tes"  
dic2$Term[dic2$word == "tous"] <- "tous.tes"  
  
# Lemmatisation à l'aide du lemmatiseur personnalisé  
dtmlem <-R.temis::combine\_terms(dtm, dic2)  
  
# Ensemble de mots à retirer  
mots\_a\_retirer <- c("a", "plus")  
  
# Suppression de mots dans le tableau lexical  
dtm2<-dtmlem[, !colnames(dtmlem) %in% mots\_a\_retirer]  
  
# Nouvelles fréquences des mots associés au corpus  
R.temis::frequent\_terms(dtm2)  
  
  
### Affichage du nuage de mots  
  
cloud<-R.temis::word\_cloud(dtm2, color= 'black', min.freq=1,n =50)   
title(main = "Mots les plus fréquents dans les discours de F. Hollande entre 2013 et 2017")