Est-il possible d'améliorer encore la courbe d'apprentissage de R au delà de tidyverse? Les packages flow et chart.

Philippe Grosjean a and Guyliann Engels a

a Laboratoire d'Écologie numérique des Milieux aquatiques

A-Institut de Recherche Complexys

A-8 avenue du Champ de Mars, 7000 Mons, Belgique

philippe.grosjean@umons.ac.be

guyliann.engels@umons.ac.be

Mots clefs : écosystème R, tidyverse, tidyeval, courbe d'apprentissage.

R est un logiciel largement customisable et ouvert. L'écosystème R est représenté par plus de 12.000 packages additionnels sur CRAN, et encore autant répartis sur Bioconductor, les forges R, Github, ... [1] Ils offrent des possibilités immenses qui font le succès de ce logiciel. Mais cela ne facilite pas son apprentissage, et bien souvent, plusieurs options existent pour la même fonctionnalité. Par exemple, au moins trois systèmes graphiques principaux incompatibles co-existent: les graphiques de base de R, lattice et ggplot2.

Le "tidyverse" (https://www.tidyverse.org) est un ensemble de packages R cohérents entre eux, qui implémentent une approche différente de R de base, notamment pour des fonctions d'importation et de remaniement des données, de présentation graphique, etc. [2] Ce sous-écosystème de R est d'un abord plus facile pour les débutants [3], en particulier pour le remaniement des données avec dplyr et tidyr, ainsi que l'opérateur de "pipe" de magrittr, et pour les graphiques avec ggplot2.

Cependant, les fonctions dans tidyverse font massivement appel à l'évaluation non standard de leurs arguments, et cela introduit une difficulté supplémentaire lorsque du code doit être généralisé (passage d'un script R à une fonction; références non transparentes). Il faut alors ruser et utiliser des techniques particulières pour empaqueter et dépaqueter les expressions (les "quosures").

Nous présentons ici des fonctions nouvelles, regroupées, entre autres, dans les packages **flow** (https://github.com/SciViews/flow) et **chart** (https://github.com/SciViews/chart), visant à faciliter le flux d'analyse et sa généralisation d'une part, et à consolider les différents types de graphiques R d'autre part. **flow** propose un opérateur de pipe légèrement différent de celui de magrittr qui est capable de prendre en compte des variables internes au pipeline et à traiter les arguments qui doivent être évalués de manière non standard de façon transparente pour l'utilisateur. **chart** fourni un point d'accès unique pour les trois principaux moteurs graphiques de R, ainsi qu'une interface formule (optionnelle) étendue pour ggplot2. L'ensemble des nouveautés proposées visent un but unique: encore renforcer la cohérence des différentes fonctions de l'écosystème R et faciliter son apprentissage à différents niveaux, et en particulier, la transition encore délicate de l'utilisateur capable d'écrire un script R d'analyse simple vers un développeur capable de modulariser son code en fonctions, éventuellement incluses dans des packages.

Toutes ces améliorations sont inspirées de l'observation de nos étudiants qui apprennent R et le tidyverse. Elles sont incluses dans une version remaniée de nos cours de Science des Données à l'UMONS (voir http://biodatascience-course.sciviews.org).

Références

[1] Decan, A., Mens, T., Grosjean, Ph. (2018). An empirical comparison of dependency network evolution in seven software packaging ecosystems. Empirical Software Engineering Journal, https://doi.org/10.1007/

s10664-017-9589-y.

[2] Wickham, H., Grolemund, G. (2017). R for Data Science. 492pp. O'Reilly publ. Also available at: http://r4ds.had.co.nz.

[3] Robinson, D. (2017). Teach the tidyverse to beginners. Blog at http://varianceexplained.org/r/teach-tidyverse/.