Groupe de Travail « Analyse des données textuelles sous R »

N. Turenne [a]

[a] LISIS (UMR 1326), UPEM, INRA, CNRS
Cité Descartes, Batiment Bois de l'Etang 77430 Champs-sur-Marne
nturenne@u-pem.fr

Mots clefs : Statistique, Traitement du Langage Naturel, Données Textuelles, Apprentissage Automatique.

En extraction et gestion des connaissances, une partie de la recherche se préoccupe d'analyser des données non structurées et multivariées afin d'extraire des connaissances et de les restituer à des utilisateurs. Les corpus de textes d'origine variée (archives, web 1.0, web 2.0, objets connectés) ont une place prédominante dans ces données non structurées.

S'il existe des API pour analyser des données textuelles dans de nombreux langages (perl, python par exemple), R possède une bibliothèque d'algorithmes très riche pour aborder des sujets de linguistique quantitative. L'analyse statistique des données textuelles était peu pratiquée avant ces 5 dernières années, si ce n'est pour une analyse théorique des distributions [1]. Depuis lors, on trouve désormais des modules typiquement adaptés aux données textuelles pour se créer un corpus, pour découper un corpus et manipuler les segments textuels dans des représentations directement exploitables par des algorithmes d'analyse de données génériques (ou transversaux à tout type de problème et de données : comme une analyse factorielle par exemple)[2]. Cette concomitance de disposer d'un Atelier de Génie logiciel largement utilisé, d'une transformation des représentations qualitatif-quantitatif, de la présence récente de bibliothèques typiquement adaptées au texte tout en garantissant l'utilisation d'une vaste bibliothèque d'algorithmes génériques suscite un intérêt grandissant de R parmi les experts de traitement automatique du langage naturel [3][4][5][6][7].

En janvier 2016, un groupe de travail https://www.facebook.com/groups/rTextData/ a été créé pour dynamiser les activités et la promotion de R par ces experts. Il réunit 77 abonnés intéressés par l'usage de R pour l'analyse des corpus textuels. Le groupe a pour ambition de faire connaître les pratiques sur R mais aussi d'animer des activités autour de ces pratiques (formation flash à un module par exemple, création de synergies monde académique-monde entrepreneurial).

Références

[1] Baayen R.H. (2008) Analyzing Linguistic Data: A Practical Introduction to Statistics using R, 368 p. , Cambridge University Press.
[2] Turenne N (2016) Analyse de données textuelles sous R , 288 p., ISTE éditions.
[3] Jockers M. (2014) Text Analysis with R for Students of Literature, 194 p., Springer.
[4] Munzert S. , Rubba C., Meißner P. et Nyhuis D. (2015) Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining, 480 p., Wiley.
[5] Arnold T. et Tilton L. (2015) Humanities Data in R: Exploring Networks, Geospatial Data, Images, and Text , 211 p., Springer.
[6] Kumar A., Avinash P.A. (2016) Mastering Text Mining with R , 278 p., Packt Publishing.
[7] Marchette D.J. et Hohman E.L. (2016) Text Data Mining Using R, 304 p., Chapman and Hall/CRC.