R.temis, un paquet R d'analyse textuelle

Milan Bouchet-Valat & Bénédicte Garnier
(Ined)

http://rtemis.hypotheses.org

Artemis, sœur d'Apollon, était vénérée dans l'Antiquité pour sa dextérité dans la chasse. Elle parcourait la nuit les forêts, les collines et tous les espaces laissés en friche par les simples mortels.

Le projet R.TeMiS

Développement lancé en 2011 avec Gilles Bastin (IEP Grenoble), continué depuis avec Bénédicte Garnier

Limites des logiciels le plus souvent utilisés en SHS :

- Souvent propriétaires
- Fortement ancrés dans des contextes théoriques
- Codage manuel du corpus dans un format spécifique
- Isolement des méthodes statistiques et des logiciels *mainstream*

Risques :

- Enfermement dans un environnement du fait des coûts d'entrée élevés
- Surestimation des capacités de l'environnement d'un point de vue épistémologique et manque de réflexivité

Les avantages de R

L'environnement statistique R présente des qualités reconnues :

- sa **robustesse** (les procédures statistiques ont été éprouvées par des communautés d'utilisateurs avertis)
- la **transparence** de son code (l'utilisateur pouvant intervenir à chaque étape de l'analyse en modifiant celui-ci)
- sa **gratuité** (les possibilités de traitement — en termes de taille pour un corpus textuel — ne sont pas limitées par la licence acquise)
- son caractère **multi-plateforme**
- enfin sa nature **collaborative** qui permet des usages initialement non prévus grâce aux nombreux paquets disponibles

Le paquet RcmdrPlugin.temis

Environnement graphique via un greffon de RCommander

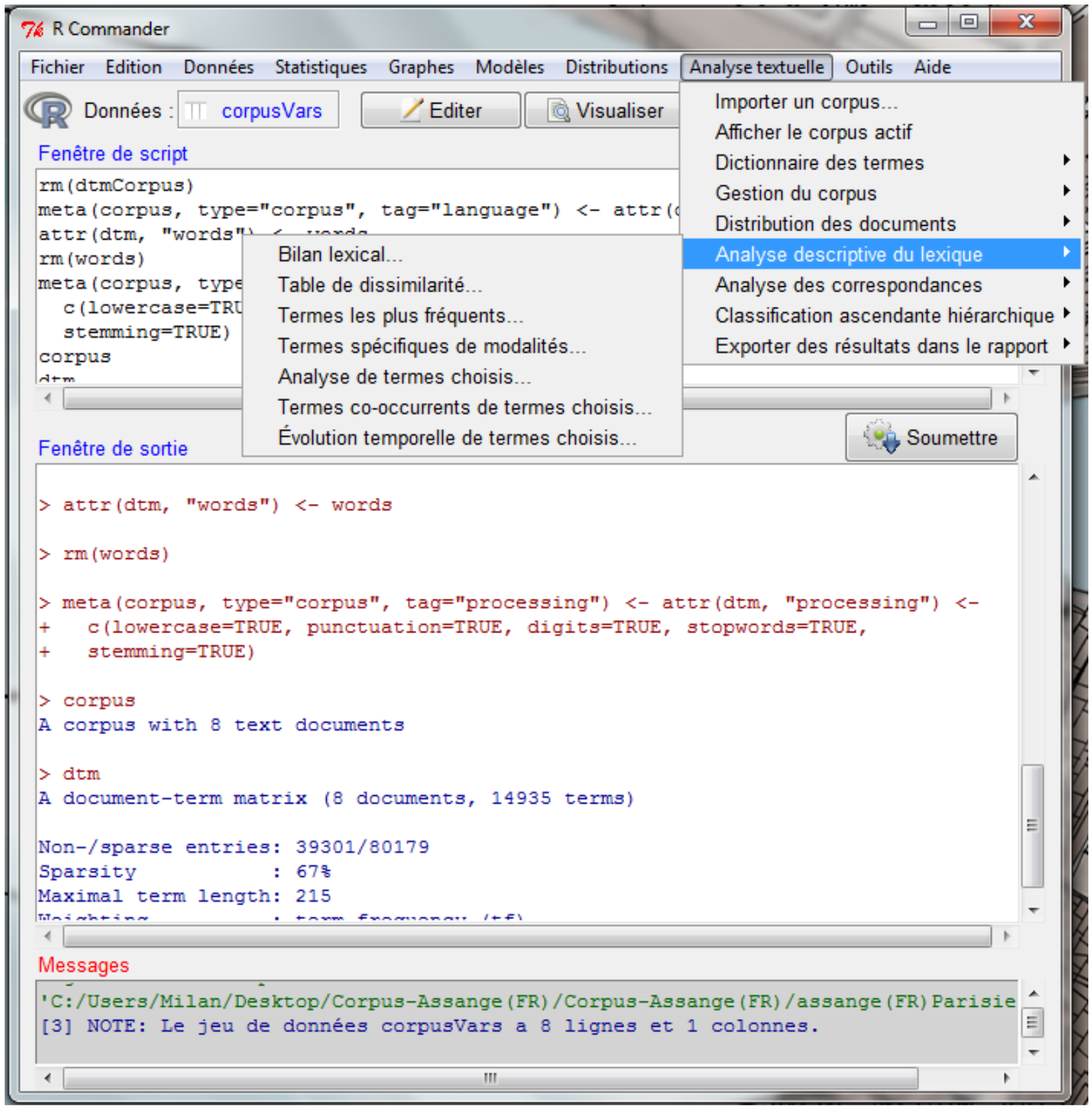
Avantages :

- Simplicité d'utilisation pour un public formé sur Alceste ou SPAD
- Génère du code R qui peut être relancé ou étendu manuellement

Limites :

- Impossible à utiliser en dehors de RCommander (RStudio…)
- Pas pensé principalement pour les utilisateurs qui écrivent directement leur code

Le paquet RcmdrPlugin.temis



Le paquet R.temis

Résultat du travail d'Antoine Chollet dans le cadre d'un stage ENSAI à l'Ined (juin-juillet 2018)

- Extraire les fonctions essentielles de RcmdrPlugin.temis
- Repenser leur organisation et leur fonctionnement pour s'adapter à une utilisation directe

Le paquet R.temis

Principe général : ne pas dupliquer les fonctions déjà présentes dans R

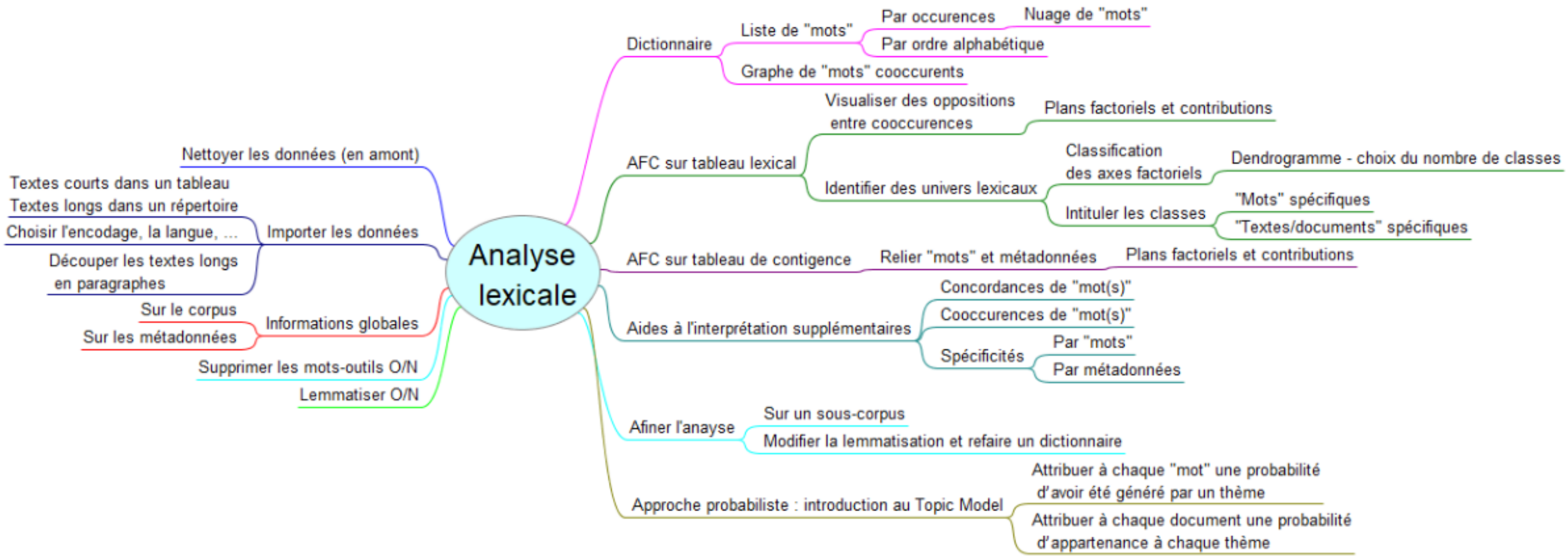
- Fonctions minimales
- Branchement aussi direct que possible sur les paquets existants et sur ceux développés dans le cadre du projet R.TeMiS
- Une partie du travail consiste en fait à documenter les manières les plus simples de réaliser des opérations génériques, mais appliquées à des corpus de textes

Fonctionnalités proposées

Méthodes classiques de l'école française :
Lebart & Salem, *Statistique textuelle*, 1994
Garnier & Guérin-Pace,
Appliquer les méthodes de la statistique textuelle, 2010

- Importation de corpus de différents formats
- Traitement des corpus : lemmatisation, découpage, extraction de sous-corpus
- Statistiques descriptives : tris à plat/croisés, graphiques
- Statistique lexicale : mesures d'occurrence et de cooccurrence de termes, spécificités, bilan lexical, nuage de mots
- Analyse de données textuelles : graphe de mots, classification ascendante hiérarchique et analyse factorielle des correspondances

Fonctionnalités proposées



Détails techniques

La gestion des corpus de textes est fondée sur le paquet tm (Feinerer, 2008 ; 2011 ; Feinerer, Hornik & Meyer, 2008)

- Interopérabilité avec d'autres paquets qui utilisent tm
- Interopérabilité indirecte avec quanteda

Paquets développés dans le cadre du projet R.TeMiS

- SnowballC pour la lemmatisation
- tm.plugin.* pour l'importation

Détails techniques

Recours à une série de paquets classiques pour les autres fonctions :

- questionr pour les tris croisés
- FactoMineR pour les AFC et la classification
- explor pour la visualisation des AFC
- igraph pour les graphes de mots
- wordcloud pour les nuages de mots
- extensibilité facile à volonté (ggplot2, maptools…)

Importation de corpus

- Quatre types de corpus peuvent être importés :
  - des fichiers de texte brut (au format .txt)
  - des fichiers au format CSV, où les lignes correspondent à des individus et les colonnes à des variables descriptives, plus une variable texte
  - des fichiers au format Alceste
  - des corpus de presse (ou Web)

L’importation automatique de corpus Web/presse

- Sources gérées : Factiva, Lexis-Nexis, Europresse
- Méta-données présentes : source, date et heure, auteur, section, zone géographique, thèmes, entreprises couvertes…
- Gain de temps : un clic -> 50/100 textes
- Limitation des erreurs de codage
- Traçabilité des opérations menées sur le corpus : toute modification des variables peut être réalisée en R et donc reproduite
- Extension facile lors de l’arrivée de nouvelles données
- Limitation des effets d’enfermement

Traitement du corpus

Traitements classiques :

- passage des termes en minuscule
- suppression de la ponctuation
- suppression des nombres
- suppression des mots vides (stopwords)
- lemmatisation automatique (paquet SnowballC : algorithme de Porter)
- lemmatisation manuelle ou fondée sur un dictionnaire (exemple : Lexique 3)

Traitement du corpus

- Possibilité de découper les textes en paragraphes (considérés chacun comme un « document »)
- Prise en compte des différents formats d'écriture de façon moins arbitraire qu'avec un découpage en segments de longueur uniforme
- S'applique aussi éventuellement aux entretiens
- Par défaut les fichiers tabulés sont découpés en autant de documents qu'ils comportent de lignes

Traitement du corpus

Des manipulations peuvent aussi être réalisées après l'importation :

- Choix d'un sous-ensemble de termes, ou élimination de certains termes
- Élimination de documents à partir de termes ou de variables
- Ces opérations peuvent s'appliquer à des paragraphes, et/ou aux résultats de l'analyse

Vocabulaire et fréquences

Bilan lexical :

- Longueur des documents, nombre de mots uniques (hapax), longueur des mots…

Fréquences du vocabulaire :

- Termes les plus fréquents
- Fréquence d’un terme particulier

Comparaisons possible entre documents ou groupes

Représentations graphiques

Nuage de mots :

- Représentation simple des termes les plus fréquents

Graphe de mots :

- Représentation des cooccurrences entre termes les plus fréquents sous forme d’un réseau

Spécificités et cooccurrences

Spécificités :

- Termes sur- ou sous-représentés dans un document ou une modalité
- Calcul à partir d'une loi hypergéométrique : tirage de termes au hasard dans l'ensemble du corpus

Cooccurrences :

- Termes spécifiques des documents qui contiennent un terme donné

Spécificités et cooccurrences

	Terme 1	Terme 2	Terme 3	Total
Doc 1	1	4	5	10
Doc 2	3	3	9	15
Doc 3	5	3	8	16
Doc 4	5	3	6	14
Doc 5	4	7	3	14
Total	18	20	31	69

Analyse des correspondances et classification

Deux méthodes :

- Analyse sur tableau lexical entier
- Analyse sur tableau lexical agrégé

Dans les deux cas, application d'une AFC classique avec FactoMineR puis visualisation interactive avec explor. Classification ascendante hiérarchique sur les résultats de l'AFC.

Analyse des correspondances sur tableau lexical entier

	Terme 1	Terme 2	Terme 3
Doc 1	1	4	5
Doc 2	3	3	9
Doc 3	5	3	8
Doc 4	5	3	6
Doc 5	4	7	3

A

B

	Terme 1	Terme 2	Terme 3
Doc 1	1	4	5
Doc 2	3	3	9
Doc 3	5	3	8
Doc 4	5	3	6
Doc 5	4	7	3
A	1+3+5	4+3+3	5+9+8
B	5+4	3+7	6+3

Modalités supplémentaires

Analyse des correspondances sur tableau lexical agrégé

	Terme 1	Terme 2	Terme 3
Doc 1	1	4	5
Doc 2	3	3	9
Doc 3	5	3	8
Doc 4	5	3	6
Doc 5	4	7	3

A

B

	Terme 1	Terme 2	Terme 3
A	1+3+5	4+3+3	5+9+8
B	5+4	3+7	6+3

Extensions

Parmi les nombreuses extensions possibles, on peut citer :

- Topic models : paquet topicmodels
- Classification selon la méthode Reinert/Alceste : paquet rainette

Références

http://rtemis.hypotheses.org

Bouchet-Valat & Bastin, « RcmdrPlugin.temis, a Graphical Integrated Text Mining Solution in R », *The R Journal*, 5 (1), 2013, p. 188-196.

Bastin & Bouchet-Valat, « Media Corpora, Text Mining and the Sociological Imagination. A Free Software Text Mining Approach to the Framing of Julian Assange in Three News Agencies Using R.TeMiS », *Bulletin de Méthodologie Sociologique*, 122 (1), 2014, p. 5-25.