R.TeMiS

A Text Mining Solution for R

Gilles Bastin & Milan Bouchet-Valat
Quanti-Lille 2012

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.

Pourquoi ?

Le sociologue (resp. politiste, anthropologue, historien...) souhaitant pratiquer l'analyse statistique de données textuelles est confronté à un troublant phénomène de double contrainte :

- du côté des données, la numérisation a conduit à un véritable « déluge » de textes disponibles à portée du clic de la souris (Hey & Trefethen, 2003).
- du côté des méthodes, l'offre logicielle a tout du « maquis » (Demazière & Brossaud, 2006), ce qui rend nécessaire un patient travail de défrichage et de choix préalable à toute analyse (Brugidou, et al., 2000; Jenny, 1996; 1997; Klein, 2001; Weitzman & Miles, 1995).

Pourquoi ?

- Les fonctionnalités usuelles de l'analyse de données textuelles sont pourtant en nombre limité (Lejeune, 2007)
- Elles tendent aussi à converger vers des traitements statistiques standard qui ne sont pas sensiblement différents de ceux appliqués à des données chiffrées.

Pourquoi ?

Malgré cela, les logiciels disponibles sont fortement ancrés dans les contextes théoriques dans lesquels ils ont été élaborés et fonctionnent souvent en vase clos du point de vue des routines statistiques ou des modalités de codage des corpus. Ce phénomène fait courir à l'utilisateur un double risque :

- celui de l'enfermement dans un environnement statistique déterminé (les coûts de constitution et de codage du corpus, comme ceux d'apprentissage du logiciel dissuadant la migration vers un autre environnement) ;
- celui d'une surestimation des capacités de cet environnement d'un point de vue épistémologique (l'indexation des opérations statistiques sur une doctrine théorique, une théorie de la langue, etc. pouvant faire croire à l'utilisateur qu'il ne fait qu'appliquer une méthode qu'il n'aurait pas à interroger) ;

Qu'est-ce que R.TeMiS ?

- R.TeMiS [R Text Mining Solution] est un environnement graphique de travail sous R permettant de créer, manipuler et analyser des corpus de textes.
- Il a été développé afin de minimiser les deux risques évoqués plus haut et de promouvoir une approche ouverte et réflexive des corpus de données textuelles.

Un environnement ouvert

- Dans R.TeMiS l'utilisateur peut appliquer sur le même corpus des opérations caractéristiques de la **statistique lexicale** (mesure d'occurrence et de cooccurrence de termes) et de **l'analyse de données textuelles** (classification hiérarchique ascendante et analyse factorielle des correspondances) ;
- Il peut le faire sans avoir à recoder ses données ;

Un environnement réflexif

Afin de limiter les effets de « boîte noire » et favoriser la réflexivité dans l'usage sociologique du logiciel (Demazière & Brossaud, 2006), R.TeMiS étend le champ d'intervention de l'utilisateur au-delà de l'analyse des termes qui caractérise les deux approches citées plus haut :

- Des fonctionnalités d'importation de corpus permettent de travailler différents formats de textes et de les structurer plus ou moins automatiquement, sans séparer cette phase de l'analyse (**« text mining »**) ;
- Des fonctionnalités d'analyse du corpus permettent de travailler à l'échelle du document et des variables contextuelles qui le caractérisent (date, source, auteur...) ;
- Enfin, l'utilisateur est placé au coeur de la réflexion. Grâce à l'utilisation de l'environnement statistique R les traitements disponibles et applicables au corpus sont très nombreux et ne s'arrêtent qu'aux limites de son imagination sociologique (Demazière, 2005) ;

R

L'environnement statistique R, dans lequel est développé R.TeMiS, 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** (un argument important notamment parce qu'à la différence de certains logiciels propriétaires, les possibilités de traitement — en termes de taille pour un corpus textuel — ne sont pas limitées par la licence acquise) ;
- son caractère **multi-plateforme** (R fonctionne aussi bien sous Mac que Windows et Linux) ;
- enfin sa nature **collaborative** qui permet d'envisager un développement communautaire du logiciel et son adaptation à des usages initialement non prévus ;
- Afin de faciliter l'usage de R.TeMiS aux néo-utilisateurs de R, le développement d'un **environnement graphique** a été privilégié. Celui-ci se présente donc comme un menu de l'application R Commander (Fox, 2005) ;

Quelques précisions

- L'architecture statistique de l'environnement R.TeMiS est fournie par le paquet **tm** développée par Ingo Feinerer (Feinerer, 2008; 2011; Feinerer, Hornik & Meyer, 2008).
- Celui-ci a été complété par d'autres paquets classiques de R comme **ca** pour la représentation des analyses factorielles des correspondances (Nenadic & Greenacre, 2007).
- Enfin des paquets spécifiques ont été développés pour faciliter l'usage de R.TeMiS dans le domaine des études sur les médias, par exemple pour la gestion des corpus constitués depuis la base de données d'articles de presse **Factiva**.

Quelques précisions

- Trois types de corpus peuvent être importés dans R.TeMiS :
  - des fichiers de texte brut (au format .txt) ;
  - des fichiers tabulés tirés d'enquêtes par questionnaire (au format .csv, .xls ou .ods, où les lignes correspondent à des individus et les colonnes à des variables descriptives et une variable texte) ;
  - des fichiers en .html ou .xml structurés par la base d'articles de presse Factiva.

Quelques précisions

- Lors de l'importation d'un corpus l'utilisateur peut décider de découper celui-ci en unités plus petites. L'unité choisie est le paragraphe. Si cette option est choisie chaque paragraphe sera considéré comme un document, ce qui peut permettre d'améliorer la qualité de l'analyse de données, notamment dans la perspective d'une classification ascendante hiérarchique ;
- Le choix d'une segmentation en paragraphes vise à prendre en compte les formats d'écriture médiatique de façon moins arbitraire qu'avec un découpage en segments de longueur uniforme (Jenny, 1999). Par défaut les fichiers tabulés sont découpés en autant de documents qu'ils comportent de lignes ;

Quelques précisions

- Lors de l'importation l'utilisateur définit le niveau de traitement lexical du corpus :
  - passage des termes en minuscule ;
  - suppression de la ponctuation ;
  - suppression des nombres ;
  - suppression des mots vides (stopwords) ;
  - lemmatisation (celle-ci est réalisée par le paquet Snowball qui utilise l'algorithme de Porter) ;

Installer R.TeMiS

- Télécharger et installer la dernière version de R (http://www.r-project.org/)
- Installer le paquet via l'installateur de paquets ou avec l'instruction

```
install.packages("RcmdrPlugin.temis", repos=c("http://R-Forge.R-project.org", getOption("repos")))
```

- Le charger par l'interface de gestion des paquets ou avec l'instruction

```
library(RcmdrPlugin.temis)
```

Le menu «Analyse textuelle» de R.TeMiS

- Importer un corpus
- Gestion du corpus
  - Afficher le corpus
  - Charger les variables
  - Sélectionner ou exclure des termes
  - Restreindre le corpus
- Analyse du corpus
  - Tri à plat
  - Tableau à double entrée
  - Evolution temporelle
  - Table de dissimilarité
- Analyse des occurrences
  - Cooccurrences
  - Termes typiques
  - Bilan du vocabulaire
  - Dictionnaire
  - Fréquence de termes particuliers
- Classification Ascendante Hiérarchique
  - Créer un dendrogramme
  - Créer des classes
- Analyse des correspondances
  - Réaliser une Analyse des Correspondances
  - Afficher l'Analyse des Correspondances
- Exporter les résultats

Le corpus de démonstration

- Dans ce qui suit les exemples sont tirés de l'analyse du corpus « Assange (fr) » ;
- Il est constitué de fichiers html téléchargés sur Factiva dans le cadre d'une enquête sur la médiatisation en France du personnage de Julian Assange, fondateur du site Wikileaks ;
- Ce corpus comporte tous les articles de Libération, Le Figaro, L'Humanité, Le Parisien et l'Agence France Presse contenant le terme « Assange » en texte intégral (dans le titre pour les dépêches AFP), quelque soit la date de publication ;
- Il se présente comme une archive zip contenant plusieurs fichiers html. Cette archives peut être téléchargée à l'adresse : http://mediacorpus.hypotheses.org/104. La décompresser puis lancer l'importation du corpus à partir du menu «Analyse Textuelle» de R Commander ;

Et maintenant, il n'y a plus qu'à faire preuve d'imagination...

Bibliographie

Brugidou, M., C. Escoffier, H. Folch, S. Lahlou, D. Le Roux, P. Morin-Andreani et G. Piat (2000). — « Les facteurs de choix et d'utilisation de logiciels d'Analyse de Données Textuelles », Actes des JADT 2000.

Brugidou, M. et D. Labbé (2000). — « Le vocabulaire syndical français à la lumière de l'analyse des données textuelles et de la statistique lexicale », Actes des Journées Internationales d'Analyse Statistique des Données Textuelles 2000.

Demazière, Didier (2005). — « Des logiciels d'analyse textuelle au service de l'imagination sociologique », Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique, 85 (1), pp. 5-9.

Demazière, Didier et Claire Brossaud (2006). — « Méthodes logicielles et réflexivité du sociologue », in Claire Brossaud, Patrick Trabal et Karl van Meter (Eds.), Analyses textuelles en sociologie. Logiciels, méthodes, usages, Rennes, Presses Universitaires de Rennes, pp. 11-22.

Feinerer, I. (2008). — « An introduction to text mining in R », R News, 8 (2), pp. 19-22.

Feinerer, I. (2011). — Introduction to the tm Package Text Mining in R.

Feinerer, I., K. Hornik et D. Meyer (2008). — « Text mining infrastructure in R », Journal of Statistical Software, 25 (5), pp. 1-54.

Fox, John (2005). — « The R Commander : A Basic-Statistics Graphical User Interface to R », Journal of Statistical Software, 14 (9).

Hey, T. et A. Trefethen (2003). — « The Data Deluge: An e-Science Perspective », in F. Berman, G. Fox et T. Hey (Eds.), Grid Computing: Making the Global Infrastructure a Reality, Wiley & Sons, pp. 809-824.

Jenny, Jacques (1996). — « Analyses de contenu et de discours dans la recherche sociologique française: pratiques micro-informatiques actuelles et potentielles », Current Sociology, 44 (3), pp. 279-290.

Jenny, Jacques (1997). — « Méthodes et pratiques formalisées d'analyse de contenu et de discours dans la recherche sociologique française contemporaine. Etat des lieux et essai de classification », Bulletin de Méthodologie Sociologique, 54, pp. 64-122.

Jenny, Jacques (1999). — « Pour engager un débat avec Max Reinert, à propos des fondements théoriques et des présupposés des logiciels d'analyse textuelle », Langage et société, 90 (1), pp. 73-85.

Klein, H. (2001). — « Overview of text analysis software », BMS. Bulletin de méthodologie sociologique (70), pp. 53-66.

Lejeune, C. (2007). — « Petite histoire des ressources logicielles au service de la sociologie qualitative », Humanités numériques (Tome 1).

Muller, Charles (1969). — « La statistique lexicale », Langue française, pp. 30-43.

Nenadic, O. et M. Greenacre (2007). — « Correspondence Analysis in R, with two- and three-dimensional graphics: The ca package », Journal of Statistical Software, 20 (3), pp. 1-13.

Weitzman, E.A. et M.B. Miles (1995). — Computer programs for qualitative data analysis, Thousand Oaks/London, Sage.