Exploiter les données

Analyse statistique et stylométrie avec *R*

Jean-Baptiste Camps & Simon Gabay

Univ. de Neuchâtel

Formation en philologie numérique :
encoder, exploiter, diffuser
12-16 février 2018

R : un langage et environnement d'analyse statistique

R est un langage et un environnement généraliste dédié à l'analyse statistique.

1. R est un logiciel gratuit et libre :
   R est distribué sous licence GNU - General Public License, la licence la plus répandue dans le logiciel libre.
2. R est un logiciel de référence :
   Le logiciel R connaît une popularité croissante, en particulier dans le monde de la recherche.
3. R est multiplateforme :
   L'installation est possible sous les systèmes Unix (Linux, Mac OS, etc.) ou Windows.

RStudio

- Pour rendre la manipulation de R un peu plus attrayante, nous allons utiliser un environnement de développement (un IDE - *Integrated Development Environment*) : RStudio
- RStudio permet d'appeler un grand nombre de fonctions depuis ses menus, intègre un débogueur, et permet, par son jeu de fenêtres, d'écrire et enregistrer des scripts, tout en visualisant des graphiques et en disposant d'une console R.

La stylométrie

définition

La stylométrie est l'étude et la mesure du style, souvent dans une perspective attributionniste.

Postulats

Chaque individu (et au-delà, chaque catégorie d'individus) emploie une langue démontrant des *propriétés particulières* et mesurables.

Stylométrie : quels données

Approche "sac de mot" (*bag of words*).

- Suppression des mots rares (lourdement liés au thème du texte ; suppression du bruit relatif aux thèmes évoqués) et accent sur les mots-outils, les 100/200/500 mots les plus fréquents (empiriquement, c'est ce qui caractérise le plus chaque individu et est le moins accessible à des changements conscients) ;
- conserver un maximum d'information grammaticale, graphique, ... voire travailler au niveau des séquences de n-caractères (*n-grams*) ;
- chercher les éléments les plus stables d'un texte à un autre du même auteur (quand cela est possible) ;
- pondérer pour éviter les biais dus à la longueur des textes, etc.

Les mots les plus fréquents

Pourquoi travailler sur les mots les plus fréquents (mots-vides, mots-outils) ?

Raisons statistiques

- Plus d'occurrences = plus de fiabilité ;
- éviter la survalorisation des hapax ;
- contourner la distribution parétienne.

Raisons philologiques et cognitives

- moins soumis aux variations de contenu, thèmes, niveau de langue, genre, versification,... ;
- usage inconscient des scripteurs (moins falsifiable, plus caractéristique d'un individu).

Pourquoi les mots fréquents ?

Préparez-vous à compter les / sur la diapositive suivante...

Idée empruntée à Mike Kestemont !

Il dit que la loi et le roi sont les meilleurs garants de la liberté civile.

Combien de / ?

5 ou 10 ?

Il dit que la loi et le roi sont les meilleurs garants de la liberté civile.

Programme du T.P.

1. un peu de statistique descriptive ;
2. traitement des données ;
3. partitionnement ;
4. analyse exploratoire par réduction de la dimensionnalité.