Analyse textuelle d'articles scientifiques évaluant l'impact des vers de terre sur l'environnement

Une approche *tidy* pour l'analyse textuelle d'articles scientifiques avec 

Antoine MALET - Stage M1, Parcours BIMS

Campus Agro Paris-Saclay - Unité Mathématiques et Informatique Appliqués

03/07/2024

Contexte scientifique:

Rôle écologique du vers de terre: deux mouvements de pensée opposés dans la littérature scientifique.

1. Europe : Importantes fonctions économiques et écosystémiques (productivité/richesse des sols).
2. Amérique du nord : Espèce invasive = dommages écosystémiques importants.

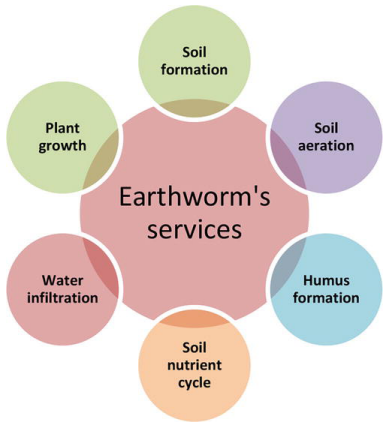

Figure 1: Les services écosystémiques rendus par les vers de terre[2].

[2] *Extrait de:* The Earthworms: Charles Darwin's Ecosystem Engineer, Kumar et al. (2023).

Objectifs:

Les méthodes statistiques de Text-mining permettent-elles d'identifier des groupes distincts d'articles scientifiques défendant une conception opposée du rôle écologique du vers de terre au sein du corpus de texte fourni pour l'analyse ?

Pour répondre à cette question de recherche, j'ai procédé en 2 étapes:

1. Constituer une base de données d'abstracts (courts résumés) d'articles scientifiques **issus de 4 métaanalyses** différentes du domaine par requêtage d'API et Web-scraping (récupération des données d'un site web grâce à un script) avec Python.
2. Analyser ces abstracts sous R, par une approche de Text-mining (analyse de données textuelles massives grâce à un script), pour tenter de résoudre cette controverse.

Base de données:

1. Ligne: 1 article / ligne.
2. La colonne "Abstract" est la plus importante, car elle contient les *textes à analyser*.

MA	Title	First author	Last author	Abstract	Date	DOI	URL
MA1	Influence of exotic earthworm invasion on soil organic matter, microbial biomass and denitrification potential in forest soils of the northeastern United States.	Amy E Burtelow	Peter M Groffman	Formerly glaciated regions of the northeastern United States have few native earthworm species and the region is dominated by exotic earthworms from Europe and Asia [...].	Sep 1998	10.1016/S0929-1393(98)00075-4	URL

Figure 2: Tableau présentant un exemple de la structure type du fichier CSV issu du Web-scraping et employé pour l'analyse. Le fichier originel comprend 168 lignes. Les données manquantes (non représentées ici), sont notées "N/A".

Scripts Python et R:

1. **Script Python:** Pour le Web-scraping.
   Principaux modules: habanero (Crossref) / ittertools / numpy / pandas / unidecode / ResearchGateScraper2 (module local, incluant re, time, parsel et playwright.sync_api).
2. **Script R:** Pour le Text-mining.
   Principaux modules: dplyr, ggplot2, ggraph, SnowballC, scales, tidytext.

Figure 3: Le code de Web-scraping a été implémenté en Python, le code de Text-mining en R (R Markdown).
Source de l'image: https://rstudio.github.io/reticulate/ (image employée à titre illustratif seulement).

Courbe de Zipf:

1. La loi de Zipf stipule que dans une collection de données ordonnées par fréquence décroissante, la fréquence d'un élément est inversement proportionnelle à son rang (*mots fréquents : en haut à gauche / mots rares: en bas à droite*).
2. La MA3 (*courbe grise*) contient davantage de mots très fréquents que les autres, dont la distribution de mots semble plus proche du modèle linéaire.

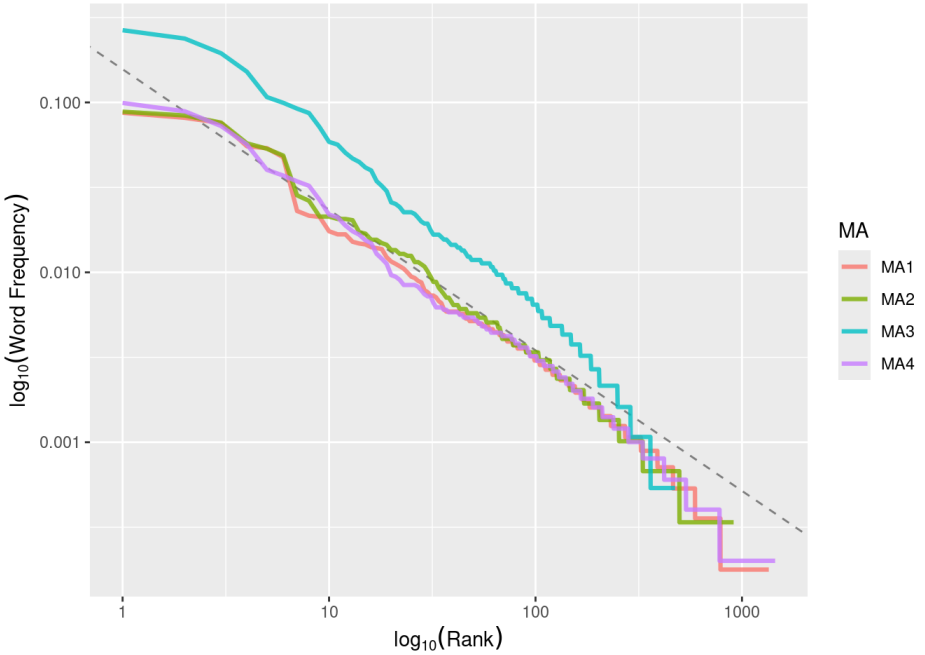

Figure 4: Allure de la courbe de Zipf pour les MA1, 2, 3 et 4.La droite en pointillés représente la valeur attendue selon un modèle linéaire sur l'intervalle [11,99].

Fréquences brutes:

1. Les trois racines les plus fréquentes dans les articles des quatres métaanalyses confondues sont *earthworm*, *soil* et *plant*.
2. Cela semble cohérent avec ce que nous savons des métaanalyses, qui traitent donc bel et bien, toutes trois, du rôle écologique du vers de terre.

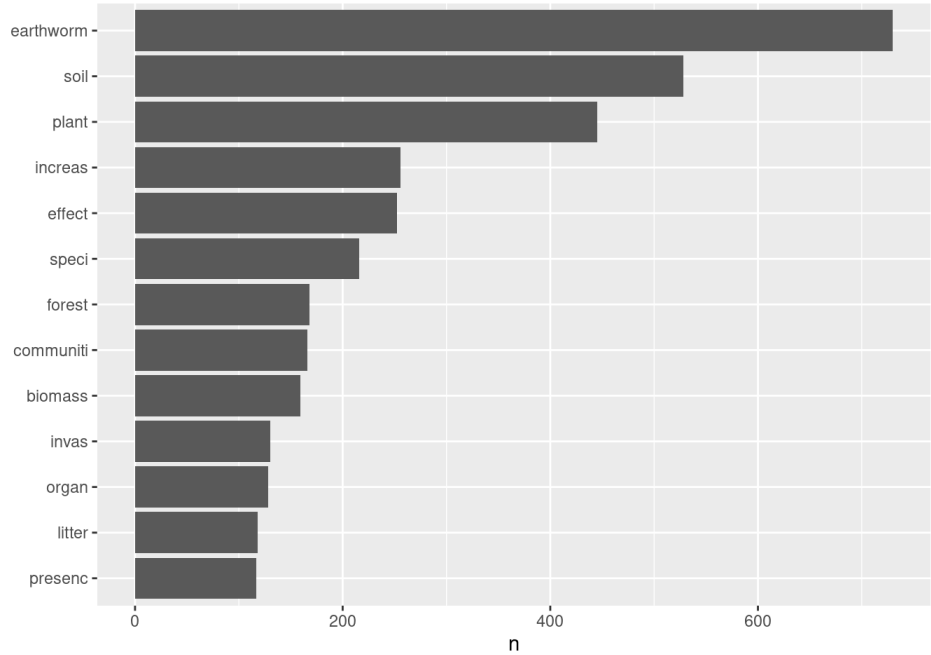

Figure 5: Diagramme à barre représentant les mots les plus fréquents dans l'ensemble des corpus étudiés. Les trois racines les plus fréquentes sont *earthworm*, *soil* et *plant*.

Comparaison de fréquences:

1. List item 1
2. List item 2

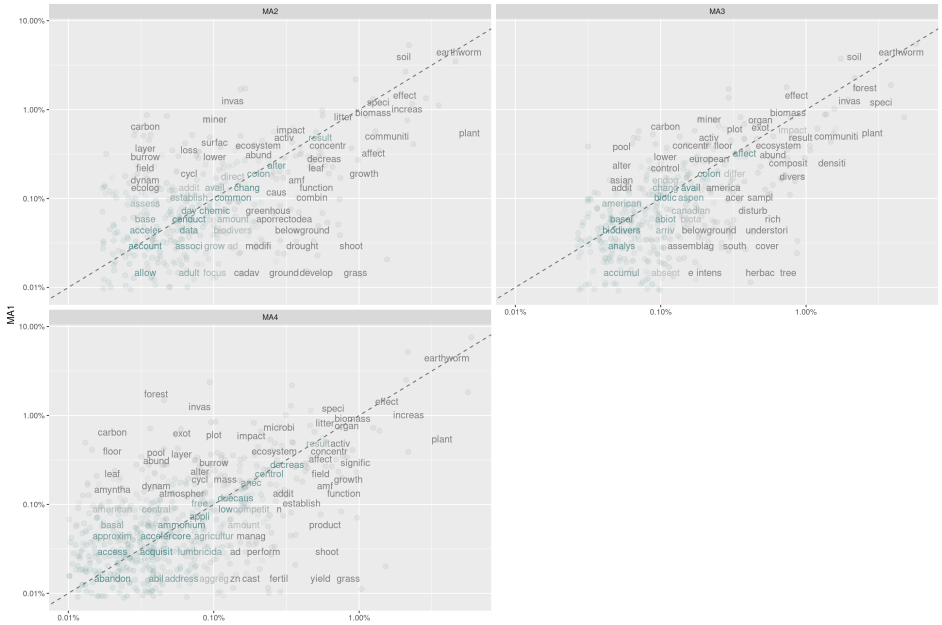

Figure 6: Log-log scatter plot montrant les corrélations du choix des mots entre les métaanalyses. Les mots sur la ligne grise en pointillés ont une fréquence similaire dans les deux MA comparées.

Approche TF-IDF:

Bla bla

1. List item 1
2. List item 2

- List item 1
- List item 2

Here is some rambling text

TF-IDF sur les bigrammes:

Bla bla

1. List item 1
2. List item 2

- List item 1
- List item 2

Here is some rambling text

Réseaux de bigrammes:

Bla bla

1. List item 1
2. List item 2

- List item 1
- List item 2

Here is some rambling text

Wordclouds:

Bla bla

1. List item 1
2. List item 2

- List item 1
- List item 2

Here is some rambling text

Mots simples:

Bla bla

1. List item 1
2. List item 2

- List item 1
- List item 2

Here is some rambling text

Bigrammes:

Bla bla

1. List item 1
2. List item 2

- List item 1
- List item 2

Here is some rambling text

Bla bla

1. List item 1
2. List item 2

Here is some rambling text

blabla