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 contextes géographiques qui s'opposent dans la littérature scientifique.

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

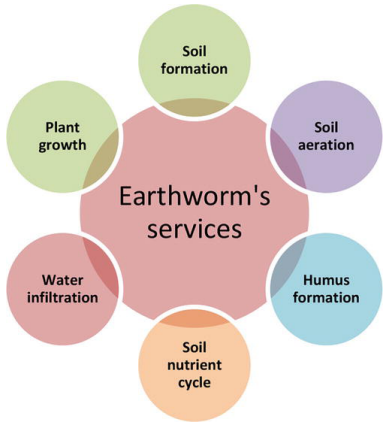

Figure 1: Les services écosystémiques rendus par les vers de terre[1].

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

Objectifs:

Les méthodes informatiques et statistiques de *Web-scraping* et de *Text-mining* permettent-elles de mettre en évidence cette opposition grâce à une approche automatisable ?

1. **Données:** Données textuelles (vers de terre), 4 métaanalyses et leur corpus[2] (*deux positives et deux négatives*).
2. **Objectif:** Tenter de retrouver cette opposition grâce à des méthodes statistiques automatisables.
3. **Méthodes:** *Web scraping* avec Python / *Text-mining* avec R.

[2] 1. "Soil chemistry turned upside down: a meta-analysis of invasive earthworm effects on soil chemical properties" 2. "Earthworms affect plant growth and resistance against herbivores: A meta-analysis" 3. "The unseen invaders: introduced earthworms as drivers of change in plant communities in North American forests (a meta-analysis)" 4. "Earthworms increase plant production: a meta-analysis"

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).[3]

[3] *Source de l'image:* https://rstudio.github.io/reticulate/ (image employée à titre illustratif seulement).

Stop words:

1. Mots-outil: **The, of, and, in, on, a, to, by.** Très haute fréquence (toutes MA confondues). Sémantiquement pauvres.
2. Approche en *"Stop words"*: Retirer les mots sémantiquement pauvres.

```
> freq_and_rank %>% filter(MA=='MA1')
# A tibble: 1,348 × 6
   MA    word          n Total word_frequency  rank
   <chr> <chr>     <int> <int>          <dbl> <int>
 1 MA1   the         488  5616         0.0869     1
 2 MA1   and         456  5616         0.0812     2
 3 MA1   of          425  5616         0.0757     3
 4 MA1   in          310  5616         0.0552     4
 5 MA1   earthworm   302  5616         0.0538     5
 6 MA1   soil        267  5616         0.0475     6
 7 MA1   to          129  5616         0.0230     7
 8 MA1   a           121  5616         0.0215     8
 9 MA1   forest      119  5616         0.0212     9
10 MA1   effect       98  5616         0.0175    10
# i 1,338 more rows
> freq_and_rank %>% filter(MA=='MA3')
# A tibble: 474 × 6
   MA    word          n Total word_frequency  rank
   <chr> <chr>     <int> <int>          <dbl> <int>
 1 MA3   the         495  1860          0.266     1
 2 MA3   of          442  1860          0.238     2
 3 MA3   and         362  1860          0.195     3
 4 MA3   in          282  1860          0.152     4
 5 MA3   earthworm   200  1860          0.108     5
 6 MA3   plant       185  1860          0.0995    6
 7 MA3   a           171  1860          0.0919    7
 8 MA3   soil        161  1860          0.0866    8
 9 MA3   to          133  1860          0.0715    9
10 MA3   increas     109  1860          0.0586   10
# i 464 more rows
```

```
> freq_and_rank %>% filter(MA=='MA2')
# A tibble: 911 × 6
   MA    word          n Total word_frequency  rank
   <chr> <chr>     <int> <int>          <dbl> <int>
 1 MA2   the         261  2959         0.0882     1
 2 MA2   of          248  2959         0.0838     2
 3 MA2   and         225  2959         0.0760     3
 4 MA2   plant       170  2959         0.0575     4
 5 MA2   in          158  2959         0.0534     5
 6 MA2   earthworm   144  2959         0.0487     6
 7 MA2   by           84  2959         0.0284     7
 8 MA2   a            78  2959         0.0264     8
 9 MA2   on           63  2959         0.0213     9
10 MA2   increas      63  2959         0.0213    10
# i 901 more rows
> freq_and_rank %>% filter(MA=='MA4')
# A tibble: 1,447 × 6
   MA    word          n Total word_frequency  rank
   <chr> <chr>     <int> <int>          <dbl> <int>
 1 MA4   the         495  4985         0.0993     1
 2 MA4   of          442  4985         0.0887     2
 3 MA4   and         362  4985         0.0726     3
 4 MA4   in          282  4985         0.0566     4
 5 MA4   earthworm   200  4985         0.0401     5
 6 MA4   plant       185  4985         0.0371     6
 7 MA4   a           171  4985         0.0343     7
 8 MA4   soil        161  4985         0.0323     8
 9 MA4   to          133  4985         0.0267     9
10 MA4   increas     109  4985         0.0219    10
# i 1,437 more rows
```

Figure 4: Capture d'écran issue de R montrant, pour chaque MA, les mots les plus fréquents rencontrés dans le corpus (rangs décroissants).

Fréquences brutes:

1. Racines les plus fréquentes: *earthworm*, *soil* et *plant*.
2. Thème général commun à toutes les MA: **Rôle écologique du vers de terre.**

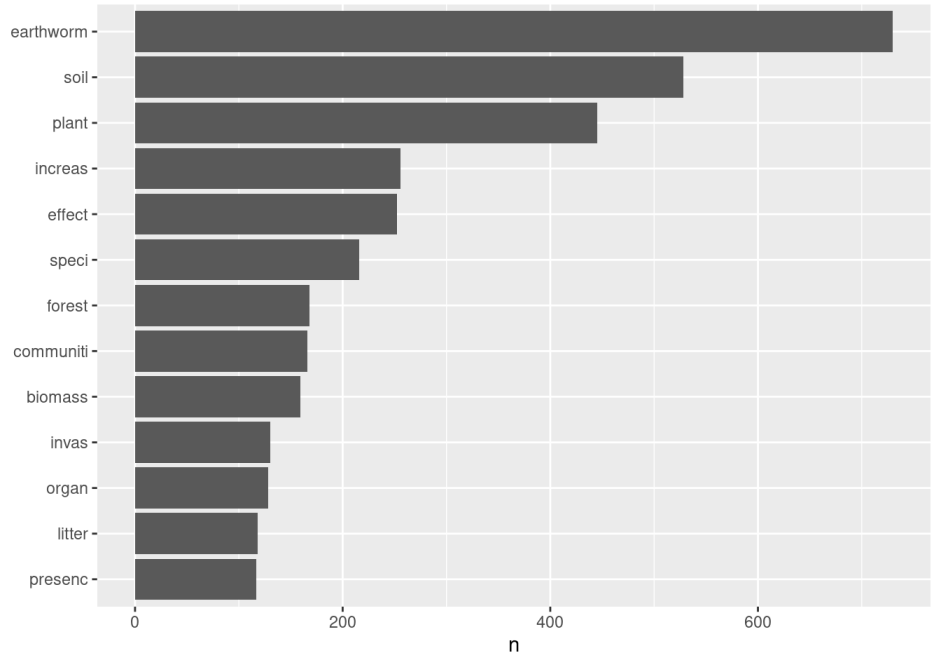

Figure 5: Diagramme à barre représentant les mots les plus fréquents dans l'ensemble des corpus étudiés. Les trois racines les plus fréquentes sont *earthworm*, *soil* et *plant*.

Comparaison de fréquences:

1. "earthworm" = très fréquent dans les 4 MA.
2. "soil" = très fréquent pour MA1, 2 et 3.
3. "plant", "community" et "growth" = MA2 > MA1.
4. "disturb", "density", "plant" = MA3 > MA1.
5. "forest", "invasive", "exotic" = MA1 > MA4.
6. "plant", "growth", "fertil" = MA4 > MA1.

Figure 6: Log-log scatter plot montrant les corrélations du choix des mots entre la MA1 et les 3 autres MA.

Approche TF-IDF (Term frequency x Inverse document frequency):

1. Thèmes majeurs MA1: "earthworm", "soil" et "forest".
2. Thèmes majeurs MA2: "aphid", "nematod" et "herbivor".
3. Thèmes majeurs MA3: "eastern", "canopy", "arthropod".
4. Thèmes majeurs MA4: "grain", "pb", "cu".

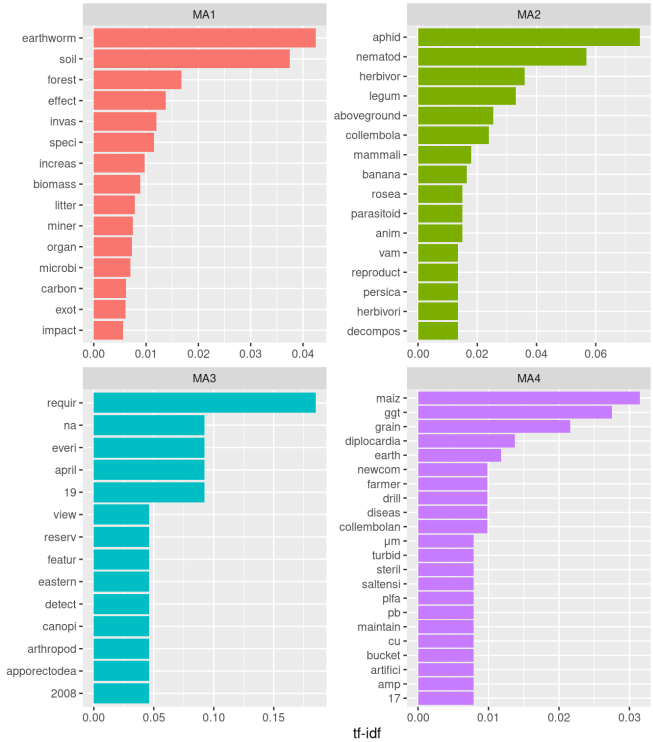

Figure 7: Ce graphe de tf-idf montre les mots thématiques plus fréquents dans une MA précise que dans l'ensemble du corpus.

TF-IDF sur les bigrammes:

1. Thèmes majeurs MA1: "mineral soil", "earthworm invasion", "exotic earthworms".
2. Thèmes majeurs MA2: "soil organisms", "plant responses", "plant mediated".
3. Thèmes majeurs MA3: "species richness", "earthworm invasion", "native earthworm".
4. Thèmes majeurs MA4: "soil organisms", "dry weight", "soil fertility".

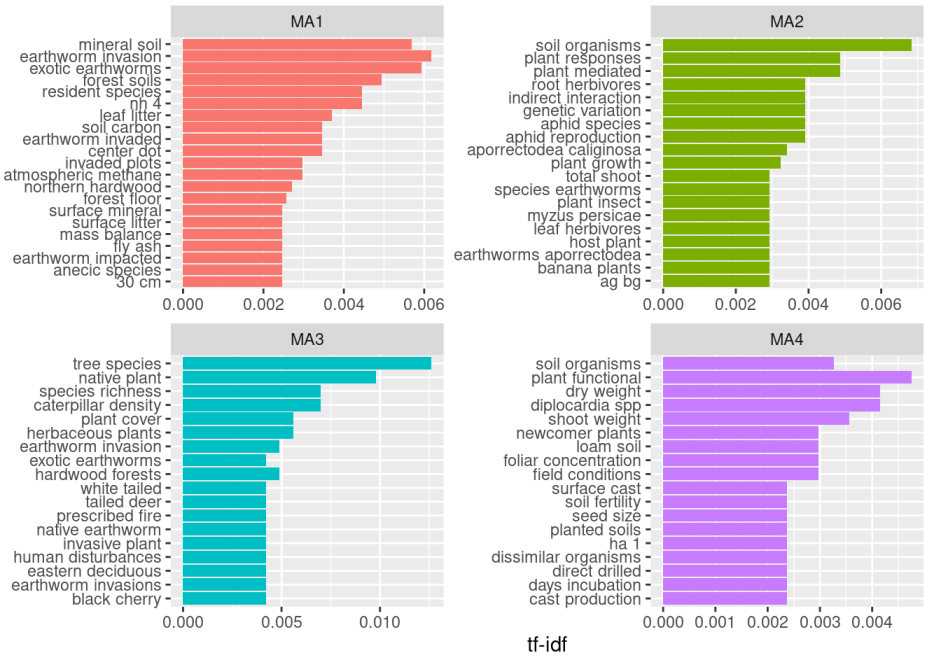

Figure 8: Ce graphe de tf-idf montre les bigrammes thématiques plus fréquents dans une MA précise que dans l'ensemble du corpus.

Réseaux de bigrammes:

1. Trois centres majeurs: **"plant"**, **"soil"** et **"earthworm"**.
2. Plant: *"plant community"*, *"plant growth"*, *"plant communities"*.
3. Soil: *"mineral soil"*, *"soil microbial"* / *"microbial biomass"* *(forte association)*, *"soil organisms"*.
4. Earthworm: *"earthworm invasion"*, *"earthworm species"*, *"earthworm activity"*.

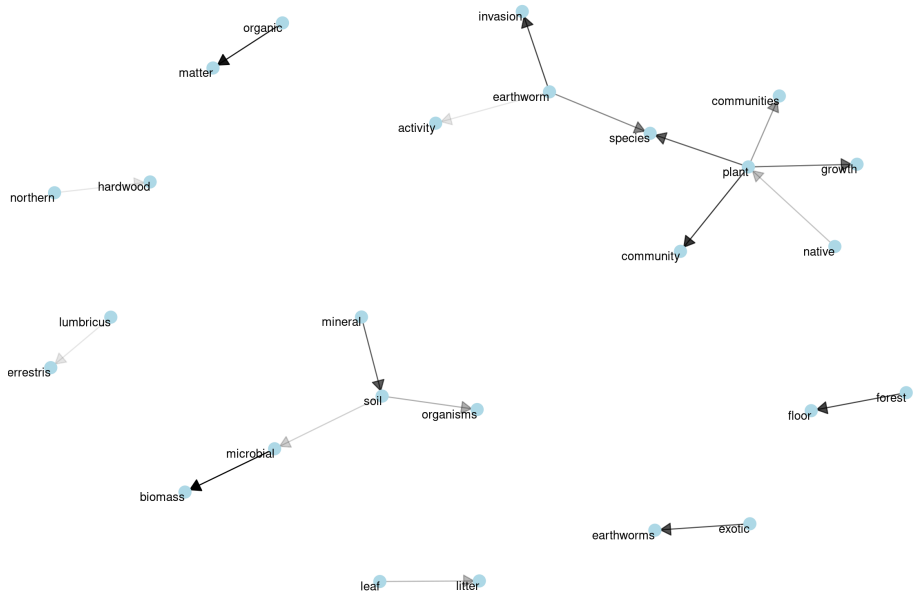

Figure 9: Réseau de bigrammes orienté montrant les différentes connections (flèches) existant entre les mots (noeuds) de l'ensemble du corpus.

Wordcloud:

1. Principaux contributeurs négatifs (-1 au score de sentiment): *"invasiv"*, *"invad"*, *"loss"*, *"negat"*.
2. Principaux contributeurs positifs (+1 au score de sentiment): *"effect"*, *"affect"*, *"enhance"*, *"product"*.

Figure 10: Nuage de mots réalisé sur les données des 4 MA confondues.
Sentiment négatif: Rouge.
Sentiment positif: Bleu.

Contribution:

1. MA1 et 3: Globalement négatives (*Figure 11, 1. et 3.*) **Principaux contributeurs:** "invas", "invad", "loss" (MA1) / "invas, "disturb", "burn" (MA3).
2. MA2 et 4: Globalement positives (*Figure 11, 2. et 4.*) **Principaux contributeurs:** "enhanc", "strong", "competitiv" (MA1) / "fertil", "enhanc", "competitiv" (MA3).

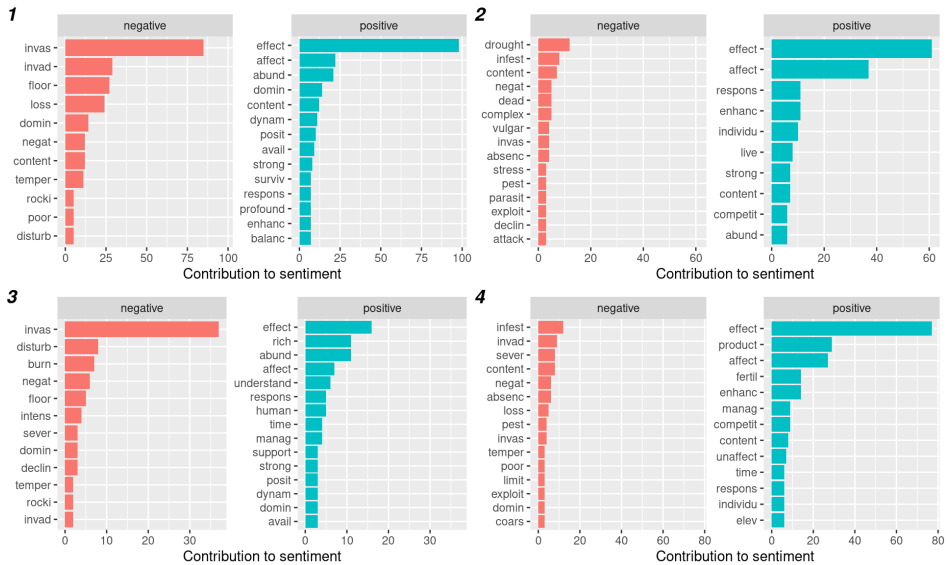

Figure 11: Diagramme à barres présentant, pour chacune des MA, les principaux contributeurs au score de sentiment global. **Sentiment négatif: Rouge. Sentiment positif: Bleu.**

Influence du mot "not" en n-1:

1. MA1: "benefit", "increase" = positif / "**not** benefit", "**not** increase" = négatif.
2. MA2 / 3: "affected" = négatif / "**not** affected" = positif.
3. MA4: "affected" = négatif / "**not** affected" = positif.
4. MA4: "responsible" / "**not** responsible" = Contexte dépendant.

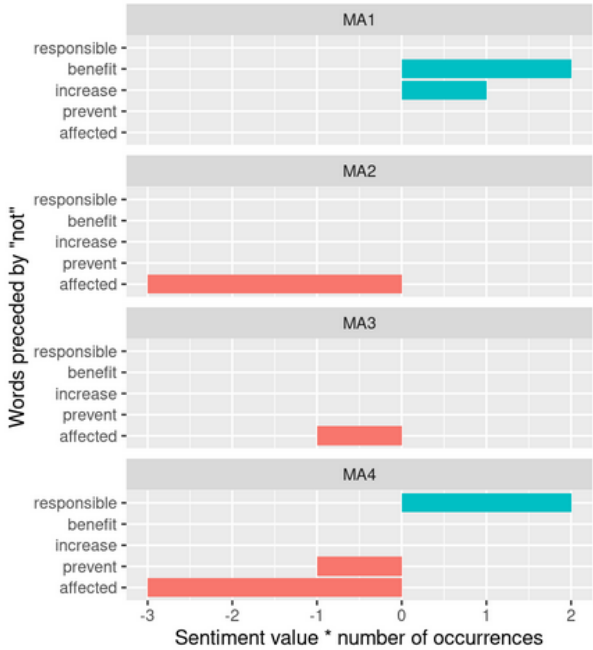

Figure 12: Diagramme à barres présentant, pour chacune des MA, les erreurs d'attribution de score de sentiment dûes à la présence du mot "not" en n-1.

- Collecte de données textuelles: *Web-scraping* avec Python.
- Analyse statistique: *Text-mining* avec R.
- MA1 et 3 globalement négatives.
- MA2 et 4 globalement positives.

- Pays d'origine des auteurs non trouvé (non inclus dans la base de données.)
- Difficulté à récupérer les abstracts par *Web-scraping*. Complétion manuelle des données.
- Approche de *topic modeling* prévue initialement, mais non abordée.