# Rapport Text Mining Earthworms

06/06/24

Premier bloc de code : importation des librairies.

```
library(dplyr)
library(ggplot2)
library(tidytext)
library(tm)
library(knitr)
library(SemNetCleaner)
library(textstem)
library(SnowballC)
library(gridExtra)
library(grid)
# WARNING : some libraries are not used but kept here at hand if the code were
# to be modified back (those libraries were used in
# previous versions of the script).
data("stop_words")
custom_stop_words <- tibble(word = c("plot","wild","slug","manipulation",
                                     "apple","significant",
                                     "significative","signific","free","led",
                                     "lead","warm","fresh","like"),
                                      lexicon = c("custom")) # remove words
#that does not have the same connotation in science compared to common language.
combined_stop_words <- bind_rows(stop_words, custom_stop_words) # didn't find
#the way to combine tibbles earlier, which is why some code chunks work with
#the two "stop words" dictionnaries separetly.
```

Deuxième bloc de code : définition des variables.

```
#-----------------------------
earthworms <- read.csv(file.path("~", "Documents", "GitHub",
                                 "EarthwormsMetanalysis",
                                 "ExtractionBiblio",
                                 "data_articles_metaanalyses.csv"))
abstracts<-c()
#----------------------
names(earthworms)
```

```
## [1] "X"              "MA"             "Title"          "First.author"
## [5] "Last.author"    "Abstract"       "OnResearchgate" "DOI"
## [9] "rg_URL"
```

```
n = nrow(earthworms)
```

Troisième bloc de code: ajout d'une colonne "aboutEarthworms".

```
earthworms <- earthworms %>% mutate(aboutEarthworms = rep(TRUE,n))
```

Quatrième bloc de code: Construction de la liste des abstracts.

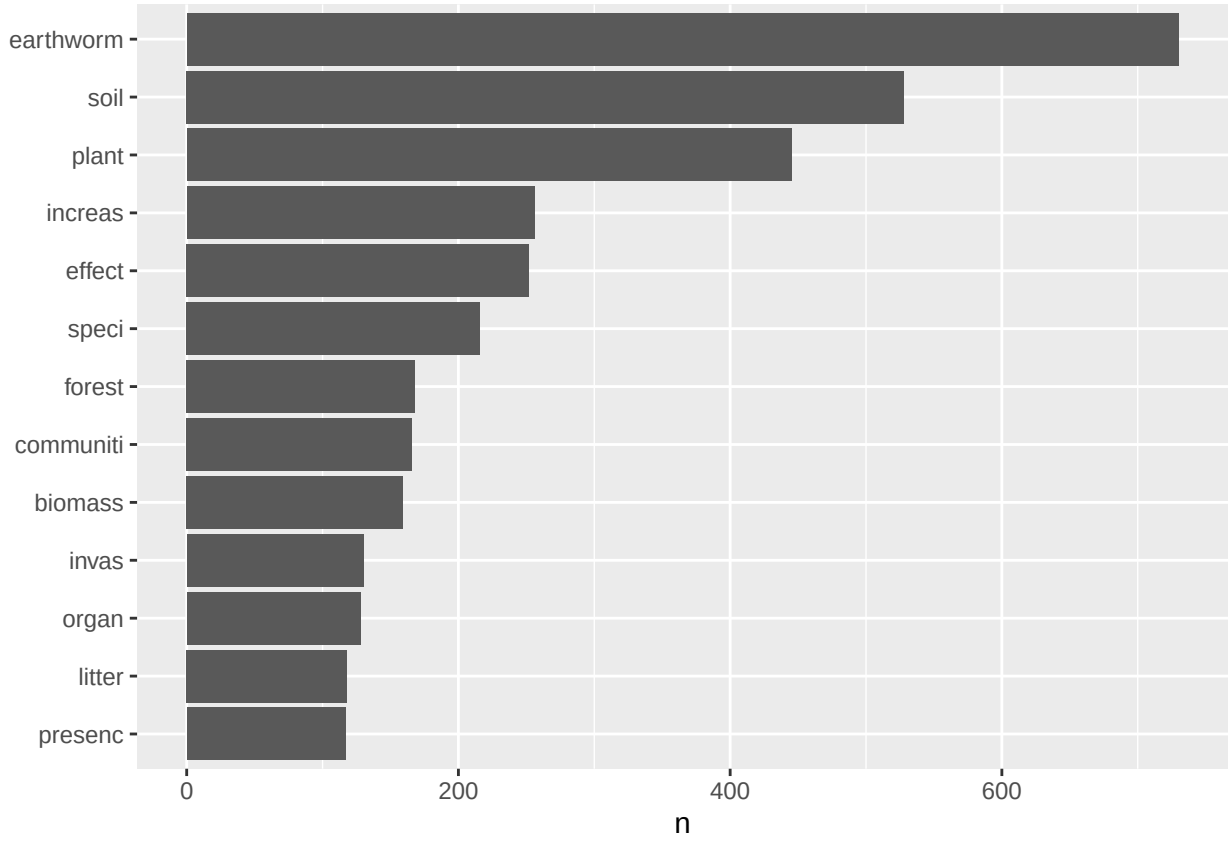
```
abstracts=earthworms$Abstract
```

Cinquième bloc de code : Construction d'un dataframe de tokens.

```
 # Unnest tokens for the current abstract
unnested_tokens <- tibble(abstracts) %>%
unnest_tokens(word, abstracts)
# Append the unnested tokens to the list
unnested_tokens<-unnested_tokens%>%anti_join(stop_words)
unnested_tokens<-unnested_tokens %>% rowwise() %>%
  mutate(word = wordStem(word, language = "en"))
```

Analyse du dataframe: trouver les mots les plus communs dans les quatre MA confondues.

```
ordre=unnested_tokens %>%
     count(word, sort = TRUE) %>%
     filter(n > 100) %>%
     mutate(word = reorder(word, n,decreasing=TRUE))

 ordre$word = factor(ordre$word,levels=rev(levels(ordre$word)))
 ggplot(ordre,aes(n, word)) +
     geom_col() +
     labs(y = NULL)
```




On peut voir que les trois mots les plus fréquents dans les articles des quatres métaanalyses confondues sont *earthworm*, *soil* et *plant*. On peut donc penser que les métaanalyses sont davantage portées sur le rôle écologique du ver de terre que sur l'étude des vers de terre à part entière.

MA1: "Soil chemistry turned upside down: a meta-analysis of invasive earthworm effects on soil chemical

properties" MA2: "Earthworms affect plant growth and resistance against herbivores: A meta-analysis" MA3: "The unseen invaders: introduced earthworms as drivers of change in plant communities in North American forests (a meta-analysis)" MA4: "Earthworms increase plant production: a meta-analysis"

## Création de subsets correspondant aux 4 MA individuellement, afin de pouvoir les comparer entre elles.

### Import subset MA1:

```
splitedData <- split(earthworms,earthworms$MA)

abstracts1<-c()
MA1 <- splitedData$MA1
nMA1 <- nrow(MA1)
abstracts1<-MA1$Abstract
unnested_tokens1 <- tibble(abstracts1) %>%
unnest_tokens(word, abstracts1)
unnested_tokens1<-unnested_tokens1 %>% anti_join(stop_words)
unnested_tokens1<-unnested_tokens1 %>% rowwise() %>%
  mutate(word = wordStem(word, language = "en"))
```

### Import subset MA2:

```
abstracts2<-c()
MA2 <- splitedData$MA2
nMA2 <- nrow(MA2)
abstracts2<-MA2$Abstract
unnested_tokens2 <- tibble(abstracts2) %>%
unnest_tokens(word, abstracts2)
unnested_tokens2<-unnested_tokens2 %>% anti_join(stop_words)
unnested_tokens2<-unnested_tokens2 %>% rowwise() %>%
  mutate(word = wordStem(word, language = "en"))
```

### Import subset MA3:

```
abstracts3<-c()
MA3 <- splitedData$MA3
nMA3 <- nrow(MA3)
abstracts3<-MA3$Abstract
unnested_tokens3 <- tibble(abstracts3) %>%
unnest_tokens(word, abstracts3)
unnested_tokens3<-unnested_tokens3 %>% anti_join(stop_words)
unnested_tokens3<-unnested_tokens3 %>% rowwise() %>%
  mutate(word = wordStem(word, language = "en"))
```

### Import subset MA4:

```
abstracts4<-c()
MA4 <- splitedData$MA4
nMA4 <- nrow(MA4)
abstracts4<-MA4$Abstract
```

```
unnested_tokens4 <- tibble(abstracts4) %>%
unnest_tokens(word, abstracts4)
unnested_tokens4<-unnested_tokens4 %>% anti_join(stop_words)
unnested_tokens4<-unnested_tokens4 %>% rowwise() %>%
  mutate(word = wordStem(word, language = "en"))
```

## Création des quatre plots:

### Plot MA1 :

```
ordre1=unnested_tokens1 %>%
     count(word, sort = TRUE) %>%
     filter(n > 25) %>%
     mutate(word = reorder(word, n,decreasing=TRUE))

 ordre1$word = factor(ordre1$word,levels=rev(levels(ordre1$word)))
 plot1<-ggplot(ordre1,aes(n, word)) +
     geom_col() +
     labs(y = NULL)
```

### Plot MA2 :

```
ordre2=unnested_tokens2 %>%
     count(word, sort = TRUE) %>%
     filter(n > 25) %>%
     mutate(word = reorder(word, n,decreasing=TRUE))

 ordre2$word = factor(ordre2$word,levels=rev(levels(ordre2$word)))
 plot2<-ggplot(ordre2,aes(n, word)) +
     geom_col() +
     labs(y = NULL)
```

### Plot MA3 :

```
ordre3=unnested_tokens3 %>%
     count(word, sort = TRUE) %>%
     filter(n > 25) %>%
     mutate(word = reorder(word, n,decreasing=TRUE))

 ordre3$word = factor(ordre3$word,levels=rev(levels(ordre3$word)))
 plot3<-ggplot(ordre3,aes(n, word)) +
     geom_col() +
     labs(y = NULL)
```

### Plot MA4:

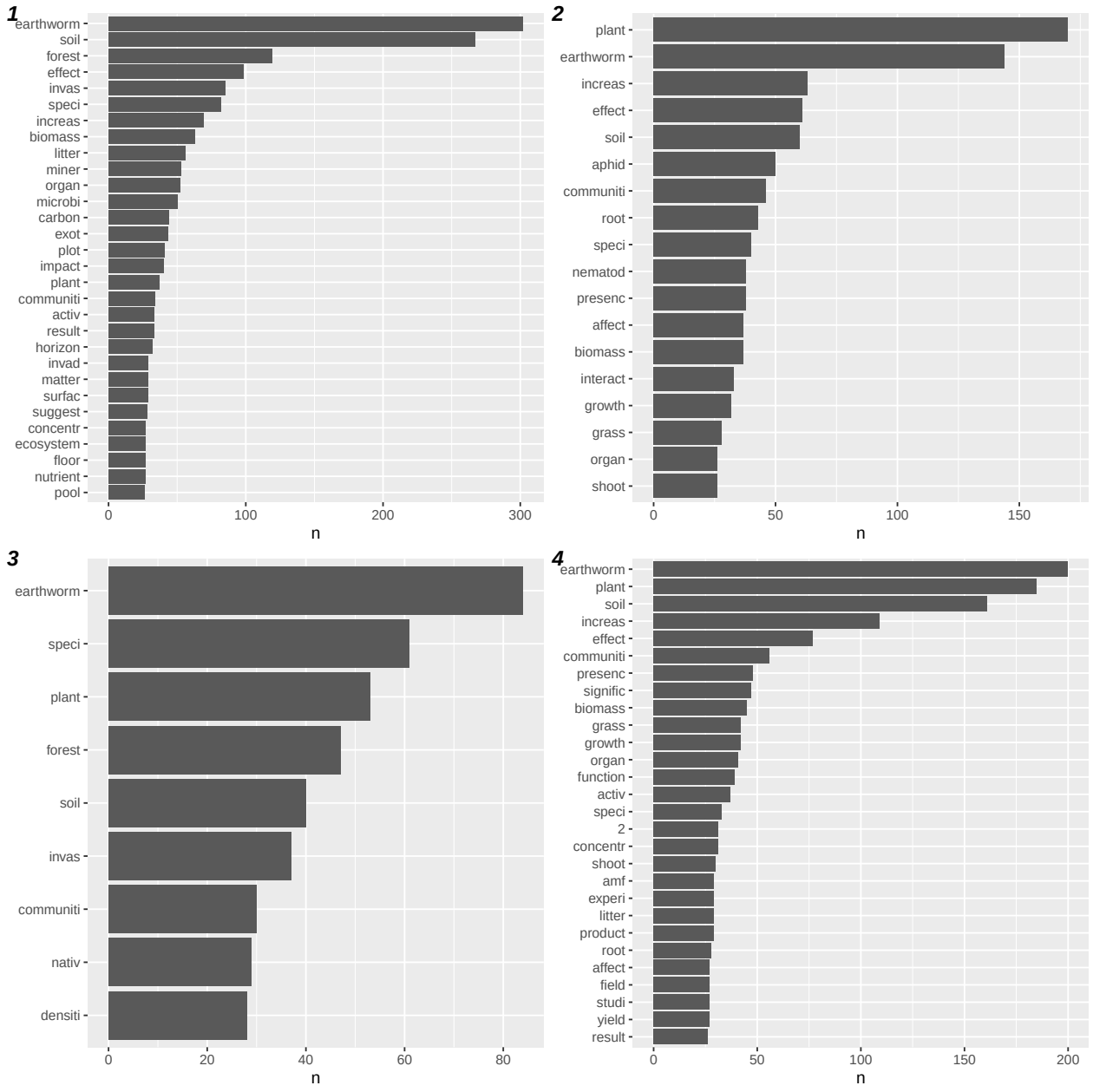
```
ordre4=unnested_tokens4 %>%
     count(word, sort = TRUE) %>%
     filter(n > 25) %>%
     mutate(word = reorder(word, n,decreasing=TRUE))
```

```
ordre4$word = factor(ordre4$word,levels=rev(levels(ordre4$word)))
plot4<-ggplot(ordre4,aes(n, word)) +
    geom_col() +
    labs(y = NULL)
```

## Comparaison des quatre MA:

Quels sont les mots les plus communs dans chacunes des MA?

```
library(egg)
ggarrange(plot1,plot2,plot3,plot4,widths = c(5,5),labels = c("1","2","3","4"))
```




On remarque que les mots les plus fréquents dans les articles des métaanalyses sont: - *earthworm*, *soil* et *forest* pour la MA1 (qui semble donc s'intéresser plus particulièrement à l'écosystème forestier). - *plant*, *earthworm* et

*increase* pour la MA2 (qui semble donc vouloir étudier précisément les effets des vers de terre sur les plantes). - *earthworm*, *specie*, et *plant* pour la MA3 (qui semble donc s'intéresser plus particulièrement à certaines espèces de vers de terre et leurs relations avec les plantes). - *earthworm*, *plant* et *soil* pour la MA4 (qui semble donc étudier l'influence des vers de terre sur les sols et les plantes). MA1: "Soil chemistry turned upside down: a meta-analysis of invasive earthworm effects on soil chemical properties" MA2: "Earthworms affect plant growth and resistance against herbivores: A meta-analysis" MA3: "The unseen invaders: introduced earthworms as drivers of change in plant communities in North American forests (a meta-analysis)" MA4: "Earthworms increase plant production: a meta-analysis" # DataFrame pour comparer les fréquences des 4 MA:

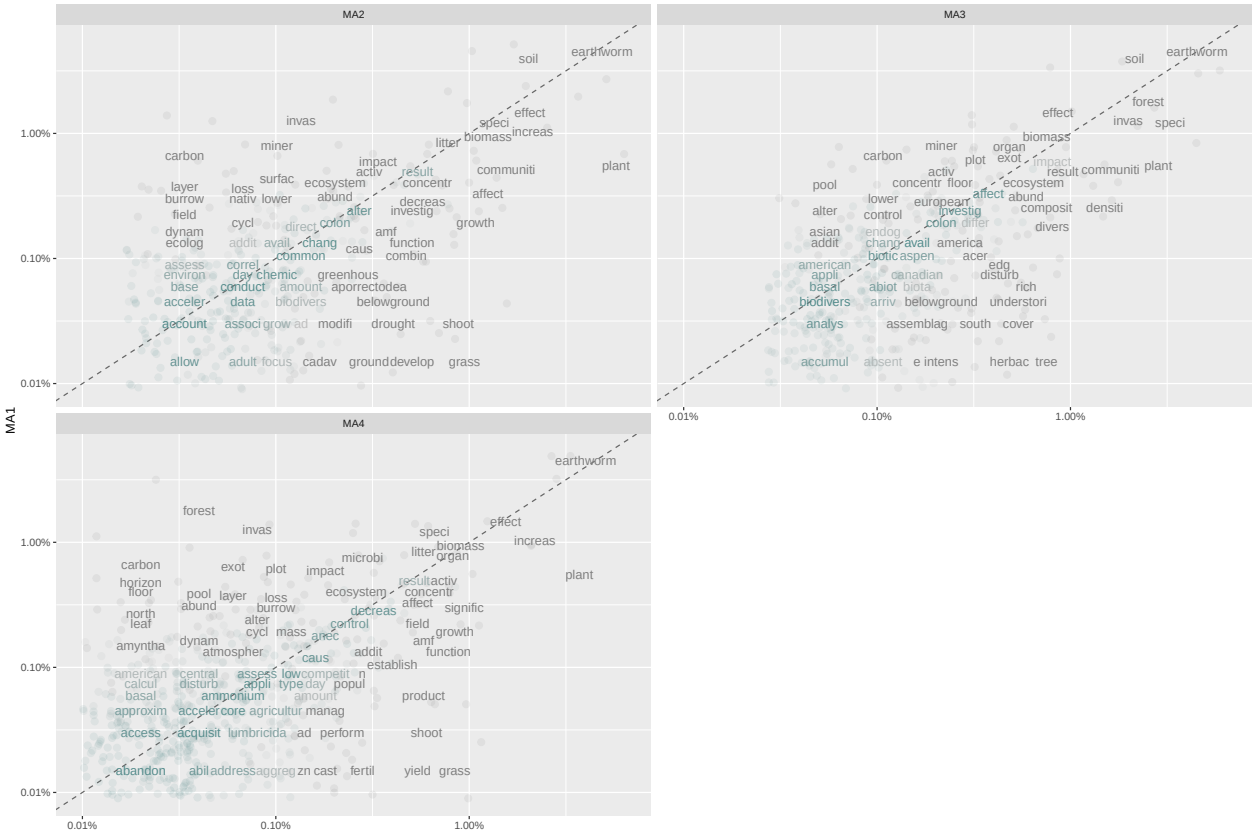
```
library(tidyr)

frequency1 <- bind_rows(mutate(unnested_tokens1, author = "MA1"),
                        mutate(unnested_tokens2, author = "MA2"),
                        mutate(unnested_tokens3, author = "MA3"),
                        mutate(unnested_tokens4, author = "MA4"))%>%
  mutate(word = stringr::str_extract(word, "[a-z']+")) %>%
  count(author, word) %>%
  group_by(author) %>%
  mutate(proportion = n / sum(n)) %>%
  select(-n) %>%
  pivot_wider(names_from = author, values_from = proportion) %>%
  pivot_longer(cols=c('MA2','MA3','MA4'),
               names_to = "author", values_to = "proportion")
```

```
frequency2 <- bind_rows(mutate(unnested_tokens1, author = "MA1"),
                        mutate(unnested_tokens2, author = "MA2"),
                        mutate(unnested_tokens3, author = "MA3"),
                        mutate(unnested_tokens4, author = "MA4"))%>%
  mutate(word = stringr::str_extract(word, "[a-z']+")) %>%
  count(author, word) %>%
  group_by(author) %>%
  mutate(proportion = n / sum(n)) %>%
  select(-n) %>%
  pivot_wider(names_from = author, values_from = proportion) %>%
  pivot_longer(cols=c('MA1','MA3','MA4'),
               names_to = "author", values_to = "proportion")
```

```
frequency3 <- bind_rows(mutate(unnested_tokens1, author = "MA1"),
                        mutate(unnested_tokens2, author = "MA2"),
                        mutate(unnested_tokens3, author = "MA3"),
                        mutate(unnested_tokens4, author = "MA4")) %>%
  mutate(word = stringr::str_extract(word, "[a-z']+")) %>%
  count(author, word) %>%
  group_by(author) %>%
  mutate(proportion = n / sum(n)) %>%
  select(-n) %>%
  pivot_wider(names_from = author, values_from = proportion) %>%
  pivot_longer(cols=c('MA1','MA2','MA4'),
               names_to = "author", values_to = "proportion")
```

## Graphe de comparaison des fréquences:

```
library(scales)

# expect a warning about rows with missing values being removed
ggplot(frequency1, aes(x = proportion, y = `MA1`,
                       color = abs(`MA1` - proportion))) +
  geom_abline(color = "gray40", lty = 2) +
  geom_jitter(alpha = 0.1, size = 2.5, width = 0.3, height = 0.3) +
  geom_text(aes(label = word), check_overlap = TRUE, vjust = 1.5) +
  scale_x_log10(labels = percent_format()) +
  scale_y_log10(labels = percent_format()) +
  scale_color_gradient(limits = c(0, 0.001),
                       low = "darkslategray4", high = "gray75") +
  facet_wrap(~author, ncol = 2) +
  theme(legend.position="none") +
  labs(y = "MA1", x = NULL)
```




## MA1 vs MA2 :

- Les mots proches de la ligne dans ces graphiques ont des fréquences similaires dans les deux ensembles de textes, soit, dans l'ordre croissant: "accelerated", "affecting","based", "dweling", "day", "change", "common", "availability", "negative", "abundance", "content", "increase", "activity", "concentration", **"litter", "biomass", "specie", "soil", "earthworm"**. On remarque que les mots les plus fréquents semblent être plus directement liés au sujet de l'article que les autres, qui paraissent plus générique au vocabulaire scientifique en écologie.

- Les mots **“invasion”, “carbon”, “layer” et “loss”** sont plus fréquents dans la MA1 que dans la MA2, tandis que les mots **“plant”, “root”, “shoot” et “grass”** sont plus fréquents dans la MA2 que de la MA1.
- Les mots **“specie”, “effect”, “soil” et “earthworm”** représentent entre 1 et 10% des mots totaux de chaque métaanalyse pour les métaanalyses 1 et 2 (échelle logarithmique).

### MA1 vs MA3 :

- Les mots proches de la ligne dans ces graphiques ont des fréquences similaires dans les deux ensembles de textes, soit, dans l’ordre croissant: “assessed”, “basal”, “abundant”, “abiotic”, “day”, “biotic”, “availability”, “compared”, “america”, “invasive”, “european”, “abundance”, “study”, “nutrient”, “increased”, “exotic”, “ecosystem”, “impact”, “litter”, “biomass”, “effect”, **“invasion”, “forest”, “specie”, “soil”, “earthworm”**. On remarque que certains mots dont la fréquence est proches pourraient peut-être faire partie du même groupe de mots (comme **“invasive european”**,**“nutrient increased”**, **“ecosystem impact”** ou encore **“litter biomass”**.)
- Les mots **“pool”, “carbon”, et “microbial** sont plus fréquents dans la MA1 que dans la MA3, tandis que les mots **“cover”, “tree”, “density” et “plant”** sont plus fréquents dans la MA3 que de la MA1.
- Les mots **“specie”, “forest,”soil” et “earthworm”** représentent entre 1 et 10% des mots totaux de chaque métaanalyse pour les métaanalyses 1 et 3 (échelle logarithmique). On peut remarquer une forte ressemblance avec les résultats obtenus entre les MA 1 et 2.

### MA1 vs MA4 :

- Les mots proches de la ligne dans ces graphiques ont des fréquences similaires dans les deux ensembles de textes, soit, dans l’ordre croissant: “access”,“acquisition”,“data”,“australia”, “amonium”, “applied”, “analysis”, “due”, “mass”, “direct”, “affect”, “system”, “native”, “ecosystem”, “density”, “experiment”, “organic”, “activity”, “litter”, “community”, “biomass”, “specie”, **effect, earthworm**. On remarque que certains mots dont la fréquence est proches pourraient peut-être faire partie du même groupe de mots (comme **acquisition data**, **direct effect**, **native ecosystem**, ou encore **litter community**).
- Les mots **“forest”,“invasion”, “carbon” et “floor”** sont plus fréquents dans la MA1 que dans la MA4, tandis que les mots **“invader”, “functional”, “grassland”, “grass” et “plant”** sont plus fréquents dans la MA4 que de la MA1.
- Les mots **“effect” et “earthworm”** représentent entre 1 et 10% des mots totaux de chaque métaanalyse pour les métaanalyses 1 et 4 (échelle logarithmique).

### Analyse globale:

- Dans cette configuration, la MA1 et la MA4 semblent être les plus similaires en termes de fréquences de mots (nuage de points davantage resserré autour de la droite.)
- Dans cette configuration, la MA1 et la MA2 semblent être les moins similaires en termes de fréquences de mots (nuage de points davantage dispersé autour de la droite.)

## Tests de corrélation:

Quel est le degré de corrélation entre les fréquences de mots de chaque métaanalyse ?

### MA2 VS MA1

```
cor.test(data = frequency1[frequency1$author == "MA2",],
         ~ proportion + `MA1`)
```

```
##
##  Pearson's product-moment correlation
##
## data:  proportion and MA1
## t = 16.521, df = 337, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.6055880 0.7238922
## sample estimates:
##       cor
## 0.6689551
```

### MA3 VS MA1

```
cor.test(data = frequency1[frequency1$author == "MA3",],
         ~ proportion + `MA1`)
```

```
##
##  Pearson's product-moment correlation
##
## data:  proportion and MA1
## t = 25.413, df = 369, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.7574403 0.8319765
## sample estimates:
##       cor
## 0.7977354
```

### MA4 VS MA1

```
cor.test(data = frequency1[frequency1$author == "MA4",],
         ~ proportion + `MA1`)
```

```
##
##  Pearson's product-moment correlation
##
## data:  proportion and MA1
## t = 28.579, df = 562, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.7337336 0.8013119
## sample estimates:
##       cor
## 0.7696699
```

### MA3 VS MA2

```
cor.test(data = frequency2[frequency2$author == "MA3",],
         ~ proportion + `MA2`)
```

```
##
##  Pearson's product-moment correlation
##
```

```
## data:  proportion and MA2
## t = 18.2, df = 224, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.7138335 0.8202087
## sample estimates:
##       cor
## 0.7723822
```

### MA4 VS MA2

```
cor.test(data = frequency2[frequency2$author == "MA4",],
         ~ proportion + `MA2`)
```

```
##
##  Pearson's product-moment correlation
##
## data:  proportion and MA2
## t = 31.618, df = 467, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.7944275 0.8524116
## sample estimates:
##       cor
## 0.8255868
```

### MA4 VS MA3

```
cor.test(data = frequency3[frequency3$author == "MA4",],
         ~ proportion + `MA3`)
```

```
##
##  Pearson's product-moment correlation
##
## data:  proportion and MA3
## t = 18.105, df = 306, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.6604871 0.7691044
## sample estimates:
##       cor
## 0.7191607
```

### Tableau récapitulatif :

| | **MA1** | **MA2** | **MA3** | **MA4** |
|---|---|---|---|---|
| **MA1** | - | 0.671 | 0.799 | 0.779 |
| **MA2** | | - | 0.793 | 0.825 |
| **MA3** | | | - | 0.740 |
| **MA4** | | | | - |

MA1: “Soil chemistry turned upside down: a meta-analysis of invasive earthworm effects on soil chemical

properties" MA2: "Earthworms affect plant growth and resistance against herbivores: A meta-analysis" MA3: "The unseen invaders: introduced earthworms as drivers of change in plant communities in North American forests (a meta-analysis)" MA4: "Earthworms increase plant production: a meta-analysis"

D'après le tableau ci-dessus, on peut voir que le choix de mots est le plus corrélé (correlation de Pearson) entre la MA2 et la MA4 ($r^2$=0.825), tandis que le choix de mots est le moins corrélé (corrélation de Pearson) entre la MA1 et la MA2 ($r^2$=0.671).

# Analyse de sentiment:

## Premiers graphiques.

On commence par construire des Tibbles contenant chaque mot exprimant un sentiment positif ou négatif (colonne *word*), puis on compte le nombre d'occurence de chaque mot (colonne *n*).

```
library(tidytext)
bing<-get_sentiments("bing")
bingStem <- bing %>% mutate(word = wordStem(word)) %>% unique()

senti1<-unnested_tokens1 %>%
  inner_join(bing) %>%
  count(word, sort = TRUE)

senti2<-unnested_tokens2 %>%
  inner_join(bing) %>%
  count(word, sort = TRUE)

senti3<-unnested_tokens3 %>%
  inner_join(bing) %>%
  count(word, sort = TRUE)

senti4<-unnested_tokens4 %>%
  inner_join(bing) %>%
  count(word, sort = TRUE)
```

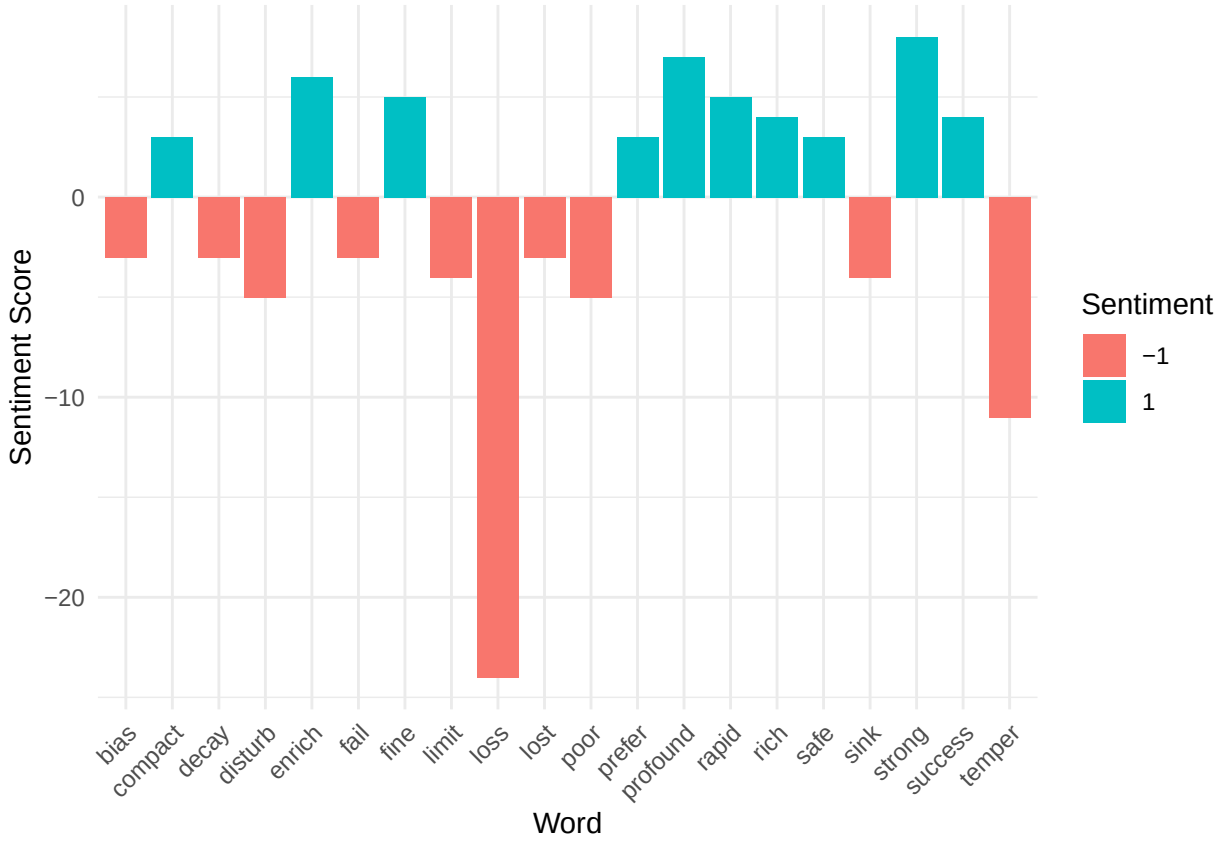
On ne conserve que les dix premières valeurs et les dix dernières valeurs, afin de pouvoir produire des graphiques plus lisibles ne conservant que les valeurs extrêmes.

```
library(tidyr)
MA1_senti <- senti1 %>%
  inner_join(get_sentiments("bing")) %>%
  anti_join(custom_stop_words, by = c("word")) %>%
  pivot_wider(names_from = sentiment, values_from = n, values_fill = 0) %>%
  mutate(sentiment = positive - negative)
sub1_MA1_senti<-MA1_senti %>%
    arrange(sentiment) %>%
    slice(1:10)
sub2_MA1_senti<-MA1_senti %>%
    arrange(desc(sentiment)) %>%
    slice(1:10)
sub_MA1_senti<-bind_rows(sub1_MA1_senti, sub2_MA1_senti)
```

On réalise un premier barplot de sentiment à titre d'exemple, en associant à chaque mot positif l'indice +1, et à chaque mot négatif l'indice -1. On représente ensuite cela graphiquement chaque mot en abscisse et la somme de ses indices (dont la valeur dépend de sa connotation et du nombre d'occurences) en ordonnées. On peut noter que certains mots neutres dans un contexte scientifique (comme "plot" ou "manipulation") ont été

rajoutés au dictionaire de “stop words”, pour éviter qu’ils soient considérés à tort comme des mots à valence négative.

```
ggplot(sub_MA1_senti, aes(x = word, y = sentiment,
                          fill = factor(sign(sentiment)))) +
  geom_bar(stat = "identity") +
  labs(x = "Word", y = "Sentiment Score", fill = "Sentiment") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

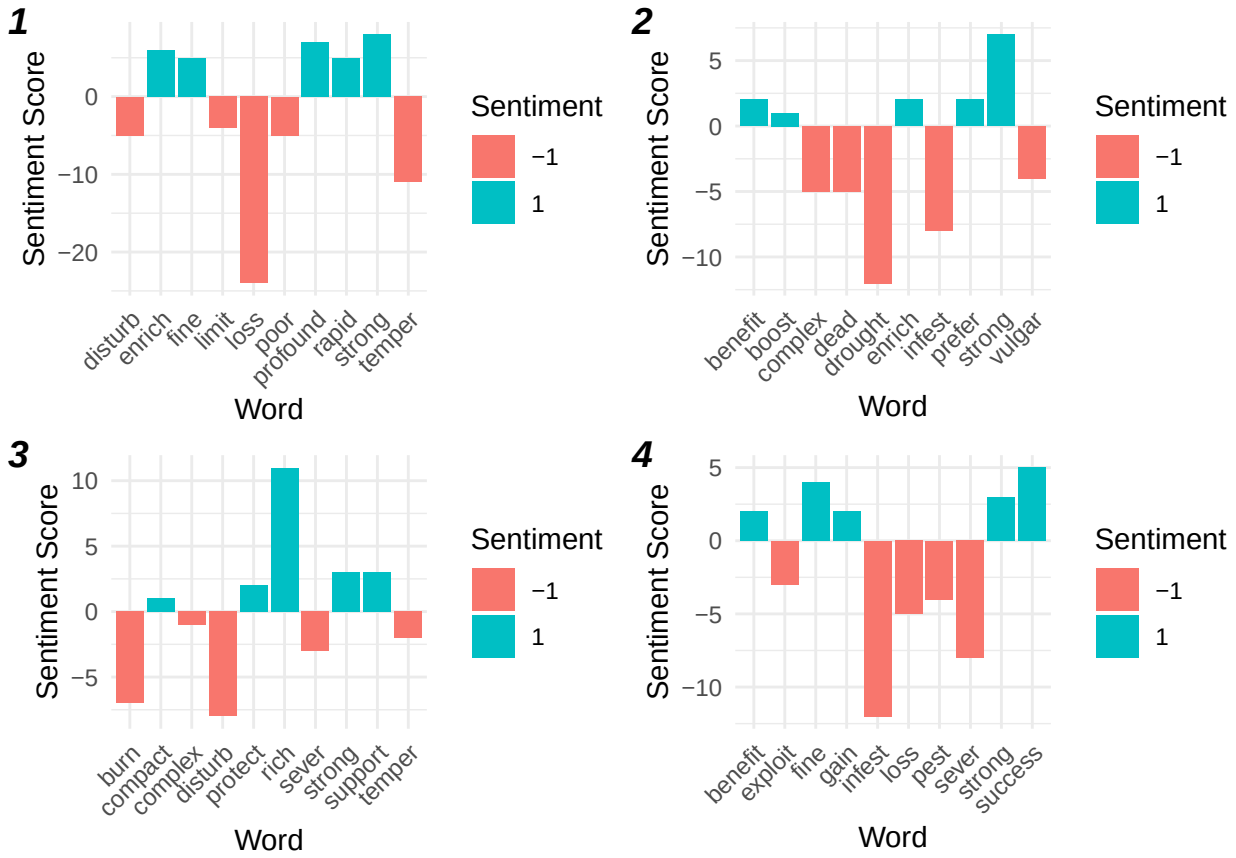



On répète le processus pour comparer les métaanalyses 1,2,3 et 4.

```
library(tidyr)
library(ggplot2)

plot_sentiment_subset <- function(senti) {
  MA_senti <- senti %>%
    inner_join(bingStem) %>%
    pivot_wider(names_from = sentiment, values_from = n, values_fill = 0) %>%
    mutate(sentiment = positive - negative)

  sub1_MA_senti <- MA_senti %>%
    anti_join(custom_stop_words, by = c("word")) %>%
    arrange(sentiment) %>%
    slice(1:5)

  sub2_MA_senti <- MA_senti %>%
    anti_join(custom_stop_words, by = c("word")) %>%
```

```
    arrange(desc(sentiment)) %>%
    slice(1:5)

  sub_MA_senti <- bind_rows(sub1_MA_senti, sub2_MA_senti)

  ggplot(sub_MA_senti, aes(x = word, y = sentiment,
                           fill = factor(sign(sentiment)))) +
    geom_bar(stat = "identity") +
    labs(x = "Word", y = "Sentiment Score", fill = "Sentiment") +
    theme_minimal() +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))
}

plot_senti1<-plot_sentiment_subset(senti1)
plot_senti2<-plot_sentiment_subset(senti2)
plot_senti3<-plot_sentiment_subset(senti3)
plot_senti4<-plot_sentiment_subset(senti4)
```

```
library(egg)
ggarrange(plot_senti1,plot_senti2,plot_senti3,plot_senti4,widths = c(2,2),
          labels = c("1","2","3","4"))
```




Tout d'abord, il est immportant de noter que tous les graphes n'ont pas les mêmes échelles ; les MA1 et 4 semblent avoir des échelles de sentiment comparables ([-20;10] et [-10;15]) et cela vaut aussi pour les MA2 et 3 ([-10;5] et [-5;10]).

Le terme *"strong"* est commun à toutes les métaanalyses. Cependant, sa valence peut grandement dépendre

du contexte de son utilisation. On peut constater que le mot positif *"rich/enrich"* est retrouvé dans la plupart de métaanalyses (1,2 et 3). Le mot *"benefit"* n'est retrouvé que dans les MA2 et 4. Le mot *"protect"* n'est retrouvé que dans la MA3.

On peut constater qu'il n'y a aucun mot négatif commun à tous les corpus. Le mot *"disturb"* n'est retrouvé que dans les MA 1 et 3. Le mot *"infest"* est seulement présent dans les MA2 et 4. Le mot *"loss"* n'est retrouvé que dans les MA1 et 4, ce qui semble indiquer que les effets négatifs relevés ne sont pas forcément liés à des pertes de caractéristiques. Le mot *"sever"* n'est retrouvé que dans les MA3 et 4, ce qui signifie peut-être qu'elles sont plus "quantitatives" que les deux autres. On peut enfin constater que le mot *"loss"* est le mot avec le score négatif le plus important des quatre graphiques, représenté dans la MA1.

MA1: "Soil chemistry turned upside down: a meta-analysis of invasive earthworm effects on soil chemical properties" MA2: "Earthworms affect plant growth and resistance against herbivores: A meta-analysis" MA3: "The unseen invaders: introduced earthworms as drivers of change in plant communities in North American forests (a meta-analysis)" MA4: "Earthworms increase plant production: a meta-analysis"

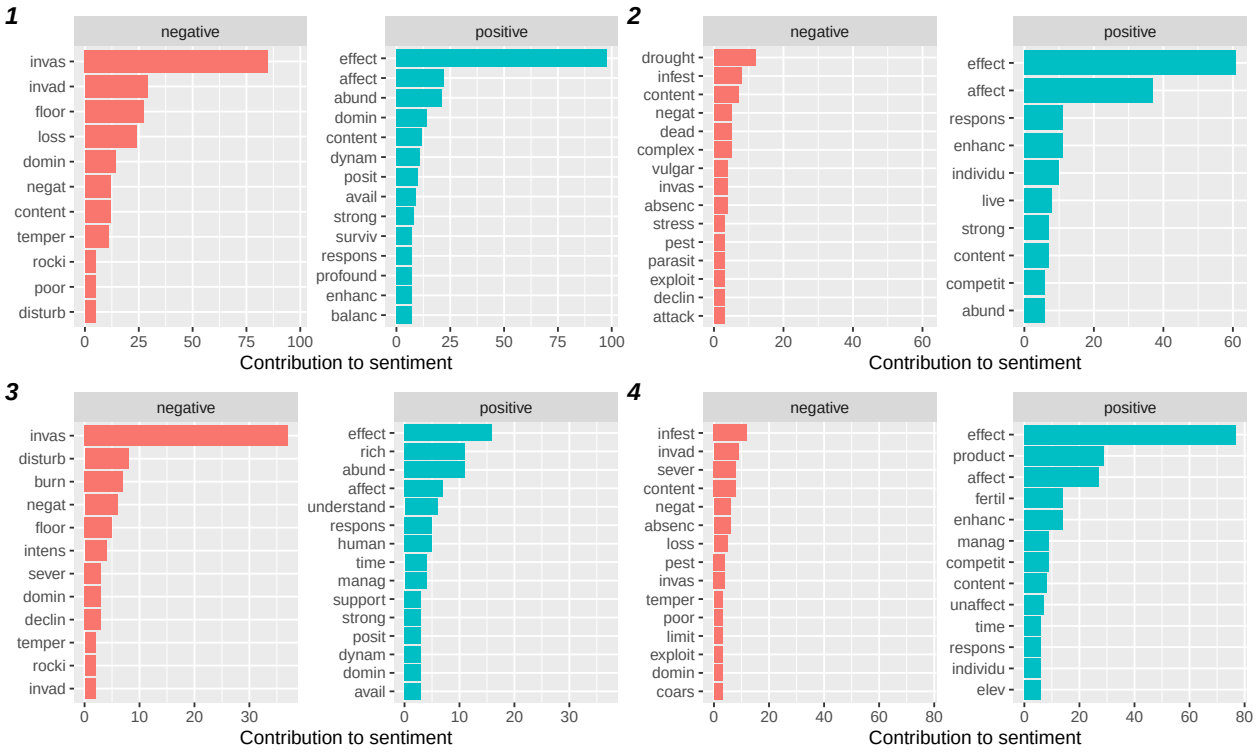
## Contribution de chaque terme au score final de sentiment.

```
bing_word_counts1 <- unnested_tokens1 %>%
  inner_join(bingStem) %>%
  count(word, sentiment, sort = TRUE) %>%
  mutate(MA="MA1") %>%
  ungroup()
bing_word_counts2 <- unnested_tokens2 %>%
  inner_join(bingStem) %>%
  count(word, sentiment, sort = TRUE) %>%
  mutate(MA="MA2") %>%
  ungroup()
bing_word_counts3 <- unnested_tokens3 %>%
  inner_join(bingStem) %>%
  count(word, sentiment, sort = TRUE) %>%
  mutate(MA="MA3") %>%
  ungroup()
bing_word_counts4 <- unnested_tokens4 %>%
  inner_join(bingStem) %>%
  count(word, sentiment, sort = TRUE) %>%
  mutate(MA="MA4") %>%
  ungroup()
```

```
contrib1<-bing_word_counts1 %>%
  anti_join(custom_stop_words, by = c("word")) %>%
  group_by(sentiment) %>%
  slice_max(n, n = 10) %>%
  ungroup() %>%
  mutate(word = reorder(word, n)) %>%
  ggplot(aes(n, word, fill = sentiment)) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~sentiment, scales = "free_y") +
  labs(x = "Contribution to sentiment",
       y = NULL)

contrib2<-bing_word_counts2 %>%
  anti_join(custom_stop_words, by = c("word")) %>%
  group_by(sentiment) %>%
  slice_max(n, n = 10) %>%
```

```
  ungroup() %>%
  mutate(word = reorder(word, n)) %>%
  ggplot(aes(n, word, fill = sentiment)) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~sentiment, scales = "free_y") +
  labs(x = "Contribution to sentiment",
       y = NULL)

contrib3<-bing_word_counts3 %>%
  anti_join(custom_stop_words, by = c("word")) %>%
  group_by(sentiment) %>%
  slice_max(n, n = 10) %>%
  ungroup() %>%
  mutate(word = reorder(word, n)) %>%
  ggplot(aes(n, word, fill = sentiment)) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~sentiment, scales = "free_y") +
  labs(x = "Contribution to sentiment",
       y = NULL)

contrib4<-bing_word_counts4 %>%
  anti_join(custom_stop_words, by = c("word")) %>%
  group_by(sentiment) %>%
  slice_max(n, n = 10) %>%
  ungroup() %>%
  mutate(word = reorder(word, n)) %>%
  ggplot(aes(n, word, fill = sentiment)) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~sentiment, scales = "free_y") +
  labs(x = "Contribution to sentiment",
       y = NULL)
```

```
library(egg)
ggarrange(contrib1,contrib2,contrib3,contrib4,widths = c(2,2),
          labels = c("1","2","3","4"))
```




```
bingStem %>% filter(word=="domin")
```

```
## # A tibble: 2 x 2
##   word  sentiment
##   <chr> <chr>
## 1 domin positive
## 2 domin negative
```

```
library(stringr)
bing %>% filter(str_detect(word, "^domin"))
```

```
## # A tibble: 5 x 2
##   word        sentiment
##   <chr>       <chr>
## 1 dominate    positive
## 2 dominated   positive
## 3 dominates   positive
## 4 domineer    negative
## 5 domineering negative
```

```
library(stringr)
unnested_tokens <- tibble(abstracts) %>%
unnest_tokens(word, abstracts)
unnested_tokens %>% filter(str_detect(word, "^domin")) %>% distinct() %>%
  print()
```

```
## # A tibble: 3 x 1
##   word
##   <chr>
## 1 dominated
## 2 dominant
## 3 dominating
```

```r
# Append the unnested tokens to the list
unnested_tokens<-unnested_tokens%>%anti_join(stop_words)
unnested_tokens<-unnested_tokens %>% rowwise() %>%
  mutate(word = wordStem(word, language = "en"))
```

Comme on peur le voir, dans le corpus de texte, *"domin"* a donc une valence **positive** et non négative. Dans la suite de cette analyse, les occurences négatives de ce mot seront donc négligées.

Tout d'abord, il faut noter que les graphiques ne sont pas à la même échelle en abscisse. Cela ne pose en théorie pas de problème pour interpréter les résultats au cas par cas, mais cela rentrera en jeux lors d'éventuelles comparaisons entre métaanalyses.

Les quatre mots positifs les plus importants en contribution sont: * *MA1:* "effect", affect", "abundant","dominate". * *MA2:* "effect", "affect", "response","enhance". * *MA3:* "effect", "rich", "abundant", "affect". * *MA4:* "effect", "product","affect","fertilize".

Les mots *"effect"* et *"affect"* sont présents dans toutes les métaanalyses. Leur sens précis, cependant, dépend du contexte. MA1: Les mots *"abundant"* et *"dominate"* évoquent l'écologie et les écosystèmes. La présence, dans le graphe, de mots comme *"content"*, *"dynamic"*, *"available"* ou encore *"survive"* laissent penser que le texte étudie la dynamique des ressources disponibles dans le sol, en lien avec des problématiques de survie. MA2: Les mots *"response"* et *"enhance"* semblent décrire des mécanismes physiologiques de réponse à l'environnement, ainsi qu'une augmentation, probablement de la croissance ou de la résistance de la plante, d'après le titre de la métaanalyse. Dans cet intervalle, il y a plus de mots positifs que de mots négatifs. MA3: Les mots *"rich"* et *"abundant"* semblent évoquer la richesse écosystémique, ainsi que la *"dynamique"*, probablement des écosystèmes, face à l'arrivée des vers de terre invasifs. MA4: Les mots *"product"* et *"fertilize"* font partie du champs lexical de l'agriculture. Avec *"enhance"*, ils semblent montrer un impact plutôt positif des vers de terre sur la production des plantes. Dans cet intervalle, il y a plus de mots positifs que de mots négatifs.

Les quatre mots négatifs les plus importants (en ignorant "content", "domin" et "floor") en contribution sont: * *MA1:* "invasive", "invade", "loss", "negative". * *MA2:* "drought", "infest", "negative","dead". * *MA3:* "invasive", "disturb", "burn","negative". * *MA4:* "infest", "invade", "sever","negative". MA2 et 4:On remarque la présence du mot *"infest"*, respectivement en deuxième et première positions pour les MA2 et 4. Son association à *"dead"* (MA2) et à *"loss"* et *"pest"* (MA4) traduit une vision probablement négative du vers de terre dans ces métaanalyse, avec probablement des dommages écosystémiques importants. MA1 et 4:Dans les MA1 et 4, la présence de *"invade"* et *"negative"*, présentant le vers de terre comme un problème, qui, sans forcément détruire totalement l'environnement natif, le **perturbe ("disturb")** (MA1) ou est perçu comme un **nuisible ("pest")** (MA4). MA1 et 3: Les mots *"invasive"* (1er positions dans les deux MA) et *"negative"* sont communs (dernière position dans les deux MA). La MA1 se focalise davantage sur l'**invasion** en tant que telle et la **perte**, et la MA3 sur la **perturbation** et la perte (**"burn"**). On constate donc que les thèses défendues semblent concordantes et négatives sur la question des vers de terre. La MA2 semble légèrement différente des autres dans son approche, avec des mots uniques qui lui sont propres tels que *"drought"* ou bien *"dead"*. D'autres mots comme *"stress"* ou *"parasite"*, au vu du titre de la métaanalyse, semblent décrire des effets négatifs sur les plantes.

Da façon simpliste, il semble donc possible, à partir de ces graphiques, de proposer une première catégorisation comme suit: - **MA1/MA3: Métaanalyses globalement défavorables aux vers de terre. Si certains bénéfices, pour la santé des sols notamment, semblent être reconnus, le vers de terre est perçu comme une espèce invasive perturbant les écosystèmes natifs. - MA2/MA4: Métaanalyses globalement favorables au vers de terre, mettant en avant les bienfaits physiologiques des vers de terre pour la plante et les avantages de ces interactions pour favoriser la productivité. Cependant, le vers de terre peut malgré tout rester nuisible, car c'est une espèce invasive qui risque de déséquilibrer l'écosystème natif.**

Cependant, il serait peut-être nécessaire d'étudier des unités textuelles plus larges, comme par exemple les phrases ou les n-grames, pour éviter au maximum le risque de contresens.

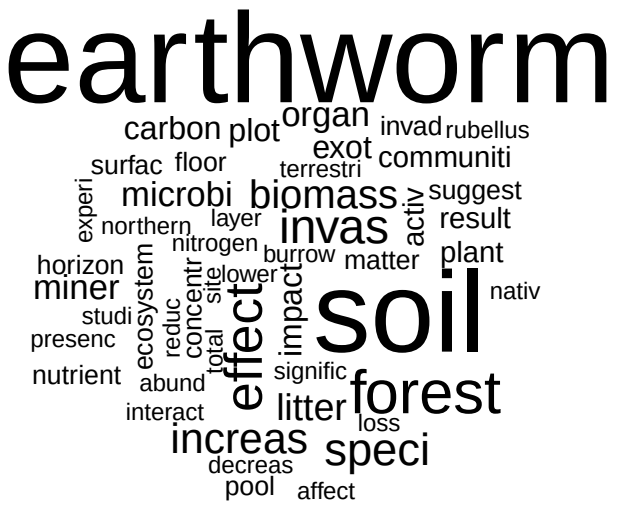
MA1: “Soil chemistry turned upside down: a meta-analysis of invasive earthworm effects on soil chemical properties” MA2: “Earthworms affect plant growth and resistance against herbivores: A meta-analysis” MA3: “The unseen invaders: introduced earthworms as drivers of change in plant communities in North American forests (a meta-analysis)” MA4: “Earthworms increase plant production: a meta-analysis”

## Représentation visuelle de la fréquence des mots: Wordclouds

### Wordcloud MA1: “Soil chemistry turned upside down: a meta-analysis of invasive earthworm effects on soil chemical properties”

Visuellement, on peut voir sur le nuage de mots que les termes prépondérants dans la MA1 sont : *“earthworm”,“soil”,“forest”,“increase”* et *“effect”*. On peut remarquer aussi que le mot *“invasion”* semble aussi relativement fréquent, ce qui est cohérent avec le titre de l’article (“invasive earthworms”). La prépondérance du mot *“earthworm”* sur les autres semble confirmer que les vers de terre sont le sujet principal de cette métaanalyse.

```
library(wordcloud)
library(dplyr)

unnested_tokens1 %>%
  anti_join(stop_words) %>%
  count(word) %>%
  with(wordcloud(word, n, max.words = 50))
```

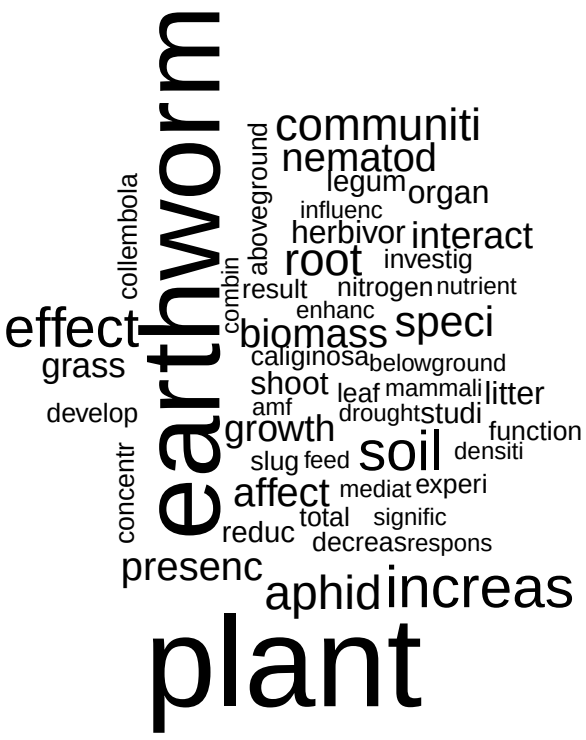



```
cloud1 <- recordPlot()
```

### Wordcloud MA2: “Earthworms affect plant growth and resistance against herbivores: A meta-analysis”

Visuellement, on peut voir sur le nuage de mots que les termes prépondérants dans la MA2 sont : *“plant”,“earthworm”,“aphid”, “soil”,“increase”* et *“effect”*. D’autres mots tel que *“community”, “interaction”, “presence”,“affect”* ou *“biomass”* semblent décrire un processus biologique dynamique. La prépondérance de *“plant”* sur les autres semble confirmer que la biologie végétale en lien avec les vers de terre est le thème principal de cette métaanalyse.

```
unnested_tokens2 %>%
  anti_join(stop_words) %>%
  count(word) %>%
  with(wordcloud(word, n, max.words = 50))
```

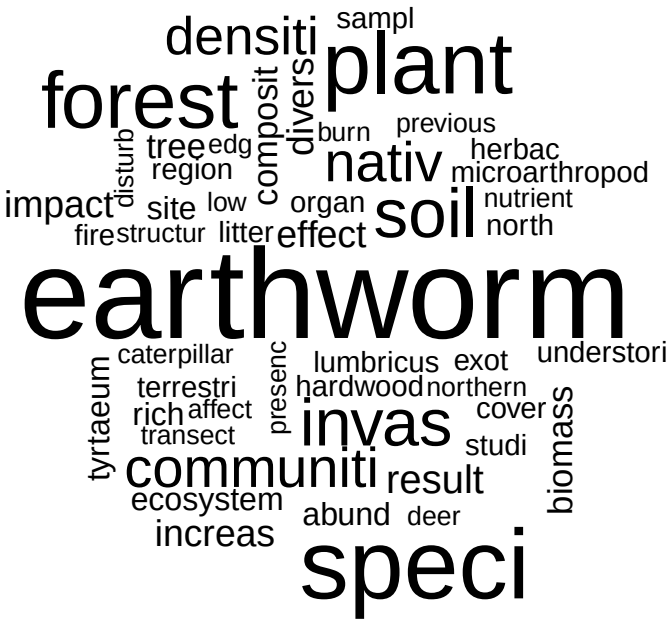



```
cloud2 <- recordPlot()
```

### Wordcloud MA3: "The unseen invaders: introduced earthworms as drivers of change in plant communities in North American forests (a meta-analysis)"

Visuellement, on peut voir sur le nuage de mots que les termes prépondérants dans la MA3 sont : *"earthworm","specie", "plant"* et *"forest"*. La prépondérance des mots *"earthworm"* et *"specie"* sur les autres semble confirmer que la métaanlyse s'intéresse en particulier à la façon dont les différentes **espèces** de vers de terre (natives ou introduite) influent sur l'écosystème global (et notamment l'écosystème **forestier**). D'autres racines telles que *"invasion", "community", "density", "native"* ou *"increase"* semblent décrire un processus écologique dynamique.

```
unnested_tokens3 %>%
  anti_join(stop_words) %>%
  count(word) %>%
  with(wordcloud(word, n, max.words = 50))
```

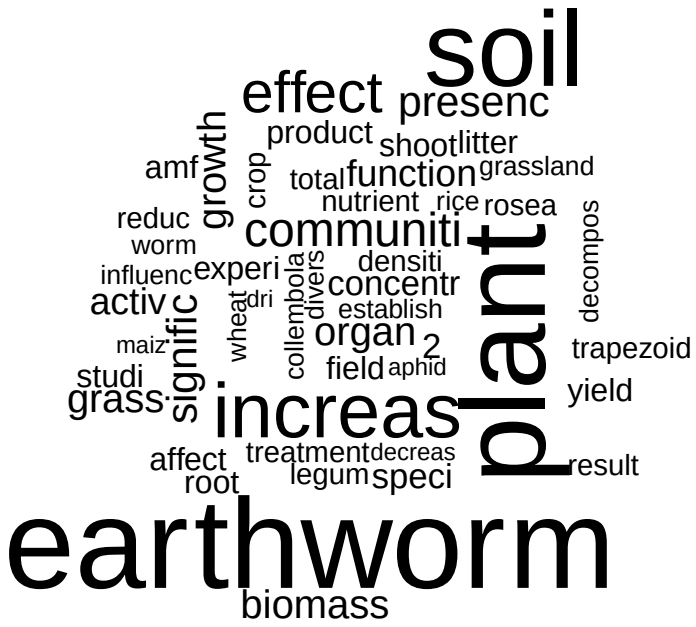



```
cloud3 <- recordPlot()
```

### Wordcloud MA4: "Earthworms increase plant production: a meta-analysis"

Visuellement, on peut voir sur le nuage de mots que les termes prépondérants dans la MA4 sont : *"earthworm","soil", "plant", "effect"* et *"increase"*. Les mots *"presence", "community", "product", "biomass"* et *"growth"* laissent entendre un effet plutôt positif des vers de terre sur le développement végatal.

```
unnested_tokens4 %>%
  anti_join(stop_words) %>%
  count(word) %>%
  with(wordcloud(word, n, max.words = 50))
```



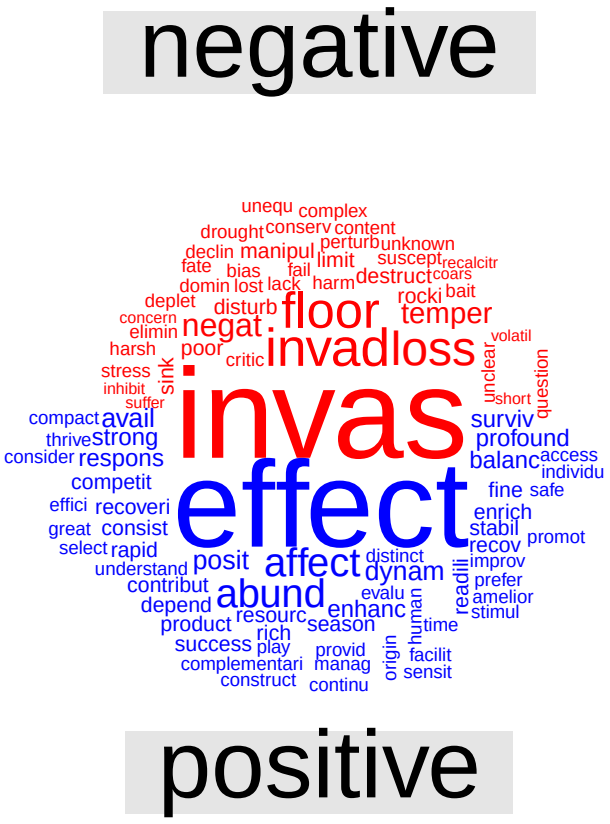

```
cloud4 <- recordPlot()
```

## Wordcloud: coloration en fonction de la valence des mots (bleu: Positif / rouge: Négatif)

### Wordcloud MA1: "Soil chemistry turned upside down: a meta-analysis of invasive earthworm effects on soil chemical properties"

Visuellement, on peut remarquer sur le nuage de mots que les termes positfs prépondérants dans la MA1 sont *"effect", "affect"* et *abundant*, tandis que les termes négatifs prépondérants sont *"invasive"*, *"invade"* et *"loss"*. On remarque tout d'abord que ces résultats sont très différents de ceux trouvés dans la partie précédente pour la même métaanalyse (Wordclouds: Wordcloud MA1) à cause de l'ajout de la variable "sentiment". Grâce à ce nuage de points, on retrouve de manière sythétique les résultats obtenues dans la partie précédante, "Contribution de chaque terme au score final de sentiment", pour la MA1.

```
library(reshape2)

unnested_tokens1 %>%
  anti_join(custom_stop_words, by = c("word")) %>%
  inner_join(bingStem) %>%
  count(word, sentiment, sort = TRUE) %>%
  acast(word ~ sentiment, value.var = "n", fill = 0) %>%
  comparison.cloud(colors = c("red", "blue"),
                   max.words = 100)
```

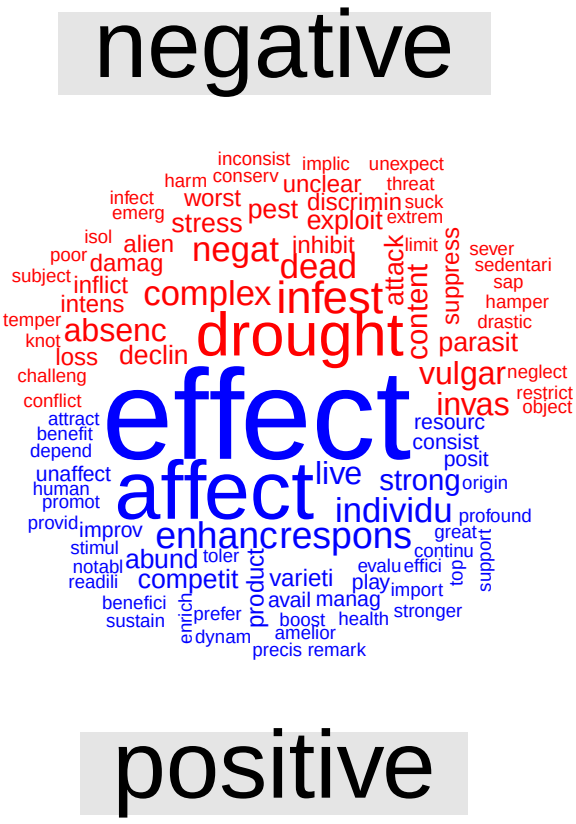



```
cloud_senti1<-recordPlot()
```

### Wordcloud MA2: "Earthworms affect plant growth and resistance against herbivores: A meta-analysis"

Visuellement, on peut remarquer sur le nuage de mots que les termes positfs prépondérants dans la MA1 sont *"effect"*, *"affect"*, *"response"* / *enhance* (approximativement de même taille), tandis que les termes négatifs prépondérants sont *"drought"*, *"infest"* et *"negative"*/*"dead"* ("complex" n'étant pas nécessairement connoté négativement dans un contexte scientifique). On remarque tout d'abord que ces résultats sont très différents de ceux trouvés dans la partie précédente pour la même métaanalyse (Wordclouds: Wordcloud MA2) à cause de l'ajout de la variable "sentiment". Grâce à ce nuage de points, on retrouve de manière sythétique les résultats obtenues dans la partie précédante, "Contribution de chaque terme au score final de sentiment", pour la MA2.

```
unnested_tokens2 %>%
  anti_join(custom_stop_words, by = c("word")) %>%
  inner_join(bingStem) %>%
  count(word, sentiment, sort = TRUE) %>%
  acast(word ~ sentiment, value.var = "n", fill = 0) %>%
  comparison.cloud(colors = c("red", "blue"),
                   max.words = 100)
```




```
cloud_senti2<-recordPlot()
```

### Wordcloud MA3: "The unseen invaders: introduced earthworms as drivers of change in plant communities in North American forests (a meta-analysis)"

Visuellement, on peut remarquer sur le nuage de mots que les termes positfs prépondérants dans la MA1 sont *"effect"*, *"rich"*, et *"abundant"* tandis que les termes négatifs prépondérants sont *"invasive"*, *"disturb"* et *"burn"*. On remarque tout d'abord que ces résultats sont très différents de ceux trouvés dans la partie précédente pour la même métaanalyse (Wordclouds: Wordcloud MA3) à cause de l'ajout de la variable

"sentiment". Grâce à ce nuage de points, on retrouve de manière sythétique les résultats obtenus dans la partie précédante, "Contribution de chaque terme au score final de sentiment", pour la MA3.

```
unnested_tokens3 %>%
  anti_join(custom_stop_words, by = c("word")) %>%
  inner_join(bingStem) %>%
  count(word, sentiment, sort = TRUE) %>%
  acast(word ~ sentiment, value.var = "n", fill = 0) %>%
  comparison.cloud(colors = c("red", "blue"),
                   max.words = 100)
```


```
cloud_senti3<-recordPlot()
```

### Wordcloud MA4: "Earthworms increase plant production: a meta-analysis"

Visuellement, on peut remarquer sur le nuage de mots que les termes positfs prépondérants dans la MA1 sont *"effect"*, *"product"*, et *"affect"* tandis que les termes négatifs prépondérants sont *"infest"*, *"invade"* et *"sever"*. On remarque tout d'abord que ces résultats sont très différents de ceux trouvés dans la partie précédente pour la même métaanalyse (Wordclouds: Wordcloud MA3) à cause de l'ajout de la variable "sentiment". Grâce à ce nuage de points, on retrouve de manière sythétique les résultats obtenus dans la partie précédante, "Contribution de chaque terme au score final de sentiment", pour la MA4.

```
unnested_tokens4 %>%
  anti_join(custom_stop_words, by = c("word")) %>%
  inner_join(bingStem) %>%
  count(word, sentiment, sort = TRUE) %>%
  acast(word ~ sentiment, value.var = "n", fill = 0) %>%
  comparison.cloud(colors = c("red", "blue"),
                   max.words = 100)
```

# negative


# positive

```
cloud_senti4<-recordPlot()
```

```
sentences1 <- tibble(text = abstracts1) %>%
  unnest_tokens(sentence, text, token = "sentences")
sentences2 <- tibble(text = abstracts2) %>%
  unnest_tokens(sentence, text, token = "sentences")
sentences3 <- tibble(text = abstracts3) %>%
  unnest_tokens(sentence, text, token = "sentences")
sentences4 <- tibble(text = abstracts4) %>%
  unnest_tokens(sentence, text, token = "sentences")
```

**Ratios de mots positifs / négatifs pour chaque métaanalyse:**

```
library(dplyr)
library(purrr)
library(kableExtra)

bingnegative <- get_sentiments("bing") %>%
  filter(sentiment == "negative")

bingpositive <- get_sentiments("bing") %>%
  filter(sentiment == "positive")

unnested_tokens1 <- unnested_tokens1 %>%
  mutate(MA = "MA1")
unnested_tokens2 <- unnested_tokens2 %>%
  mutate(MA = "MA2")
```

```
unnested_tokens3 <- unnested_tokens3 %>%
  mutate(MA = "MA3")
unnested_tokens4 <- unnested_tokens4 %>%
  mutate(MA = "MA4")

MA_labels <- c("MA1", "MA2", "MA3", "MA4")
word_counts_MA <- c(length(unnested_tokens1$word),
                    length(unnested_tokens2$word),
                    length(unnested_tokens3$word),
                    length(unnested_tokens4$word))

word_count_tibble <- tibble(
  MA = MA_labels,
  words = word_counts_MA
)

all_ratios <- bind_rows(unnested_tokens1,unnested_tokens2,unnested_tokens3,
                        unnested_tokens4)

joined_tibble <- rbind(inner_join(all_ratios, bingnegative, by = "word"),
                       inner_join(all_ratios, bingpositive, by= "word"))
# Assuming joined_tibble is already defined
sentiment_counts <- joined_tibble %>%
    group_by(MA, sentiment) %>%
    summarise(count = n())
# Transpose the summarized counts
sentiment_counts <- sentiment_counts %>%
    pivot_wider(names_from = sentiment, values_from = count)
```

```
library(dplyr)
library(purrr)
library(kableExtra)

bingnegative <- get_sentiments("bing") %>%
  filter(sentiment == "negative")

bingpositive <- get_sentiments("bing") %>%
  filter(sentiment == "positive")

unnested_tokens1 <- unnested_tokens1 %>%
  mutate(MA = "MA1")
unnested_tokens2 <- unnested_tokens2 %>%
  mutate(MA = "MA2")
unnested_tokens3 <- unnested_tokens3 %>%
  mutate(MA = "MA3")
unnested_tokens4 <- unnested_tokens4 %>%
  mutate(MA = "MA4")

MA_labels <- c("MA1", "MA2", "MA3", "MA4")
word_counts_MA <- c(length(unnested_tokens1$word),
                    length(unnested_tokens2$word),
                    length(unnested_tokens3$word),
                    length(unnested_tokens4$word))
```

```
word_count_tibble <- tibble(
  MA = MA_labels,
  words = word_counts_MA
)

all_ratios <- bind_rows(unnested_tokens1,unnested_tokens2,unnested_tokens3,
                        unnested_tokens4)

joined_tibble <- rbind(inner_join(all_ratios, bingnegative, by = "word"),
                       inner_join(all_ratios, bingpositive, by= "word"))

MA_titles <- c("Soil chemistry turned upside down: a meta-analysis [...]",
               "Earthworms affect plant growth and resistance [...]",
               "The unseen invaders: [...] in north American forests
               (a meta-analysis)",
               "Earthworms increase plant production: a meta-analysis")

positive_count=c(sentiment_counts$positive[1], sentiment_counts$positive[2],
                 sentiment_counts$positive[3], sentiment_counts$positive[4])
negative_count=c(sentiment_counts$negative[1], sentiment_counts$negative[2],
                 sentiment_counts$negative[3], sentiment_counts$negative[4])

summary_tibble <- joined_tibble %>%
  group_by(MA) %>%
  summarize(
    positivewords = positive_count,
    negativewords = negative_count,
    sentimentwords=positive_count + negative_count
  ) %>%
  left_join(word_count_tibble, by = "MA") %>%
  mutate(positive_ratio = signif(positivewords / sentimentwords,digits = 2)) %>%
  mutate(negative_ratio = signif(negativewords / sentimentwords,digits = 2))

summary_tibble <- data.frame(
    MA = unique(summary_tibble$MA),
    positivewords = unique(summary_tibble$positivewords),
    negativewords = unique(summary_tibble$negativewords),
    words = unique(summary_tibble$sentimentwords),
    positive_ratio = unique(summary_tibble$positive_ratio),
    negative_ratio = unique(summary_tibble$negative_ratio)
) %>%head(4)

summary_tibble <- as_tibble(summary_tibble)

summary_tibble %>%
  mutate(MA_title = MA_titles) %>%
  mutate(current_MA = MA) %>%
  select(MA_title, current_MA, words, positivewords,
         negativewords, positive_ratio, negative_ratio)
```

```
## # A tibble: 4 x 7
##   MA_title           current_MA words positivewords negativewords positive_ratio
##   <chr>              <chr>      <int>         <int>         <int>          <dbl>
## 1 "Soil chemistry t~ MA1          231            82           149           0.35
```

| MA | positivewords | negativewords | words | positive_ratio | negative_ratio |
|---|---|---|---|---|---|
| MA1 | 82 | 149 | 231 | 0.35 | 0.65 |
| MA2 | 24 | 91 | 115 | 0.21 | 0.79 |
| MA3 | 30 | 38 | 68 | 0.44 | 0.56 |
| MA4 | 56 | 82 | 138 | 0.41 | 0.59 |

```
## 2 "Earthworms affec~ MA2          115            24            91           0.21
## 3 "The unseen invad~ MA3           68            30            38           0.44
## 4 "Earthworms incre~ MA4          138            56            82           0.41
## # i 1 more variable: negative_ratio <dbl>
```

```
summary_tibble %>%
 kbl() %>%
 kable_material_dark("hover", full_width = F) %>%
 row_spec(0, bold = T, color = "black", background = "#D7261E")
```

Le tablau ci-dessus présente, pour chaque MA, le rapport mot positifs / total sentiment (colonne "positive ratio") et le le rapport mot négatifs / total sentiment (colonne "negative ratio"). On peut donc classer les MA dans cet ordre, la plus positive en dernier: MA2, MA1, MA4, MA3. On peut donc écrire, de manière synthétique, que: * **MA1:** 35% des mots connotés sont positifs, 65% sont négatifs. * **MA2:** 21% des mots connotés sont positifs, 79% sont négatifs. * **MA3:** 44% des mots connotés sont positifs, 56% sont négatifs. * **MA4:** 41% des mots connotés sont positifs, 59% sont négatifs. Ces ratios semblent plutôt surprenants, car des métaanalyses au titre paraissant négatif (comme la MA3), ont finalement le ratio de mots positifs le plus élévé des quatre. A l'inverse, la MA2 semble être la plus négative alors que son titre paraît plutôt optimiste. Une analyse plus approfondie des textes permettrait probablement de résoudre cet apparent paradoxe. Cependant, le ratio de mots négatifs est dans tous les cas $> 50\%$, ce qui semble indiquer que la vision du vers de terre présentée dans ces articles est majoritairement négative.

MA1: "Soil chemistry turned upside down: a meta-analysis of invasive earthworm effects on soil chemical properties" MA2: "Earthworms affect plant growth and resistance against herbivores: A meta-analysis" MA3: "The unseen invaders: introduced earthworms as drivers of change in plant communities in North American forests (a meta-analysis)" MA4: "Earthworms increase plant production: a meta-analysis"

## Approche tf-idf :

**Définitions:** *tf*: term frequency. *idf*: inverse document frequency. Plus élaboré que l'approche en "stop words". Diminue la fréquence des termes rencontrés très souvent et augmente la fréquence des termes rares, en jouant sur la pondération. *tf-idf*: le produit de ces deux indices. Il premet d'identifier les mots qui sont **plus importants pour un document spécifique** parmi un ensemble de documents.

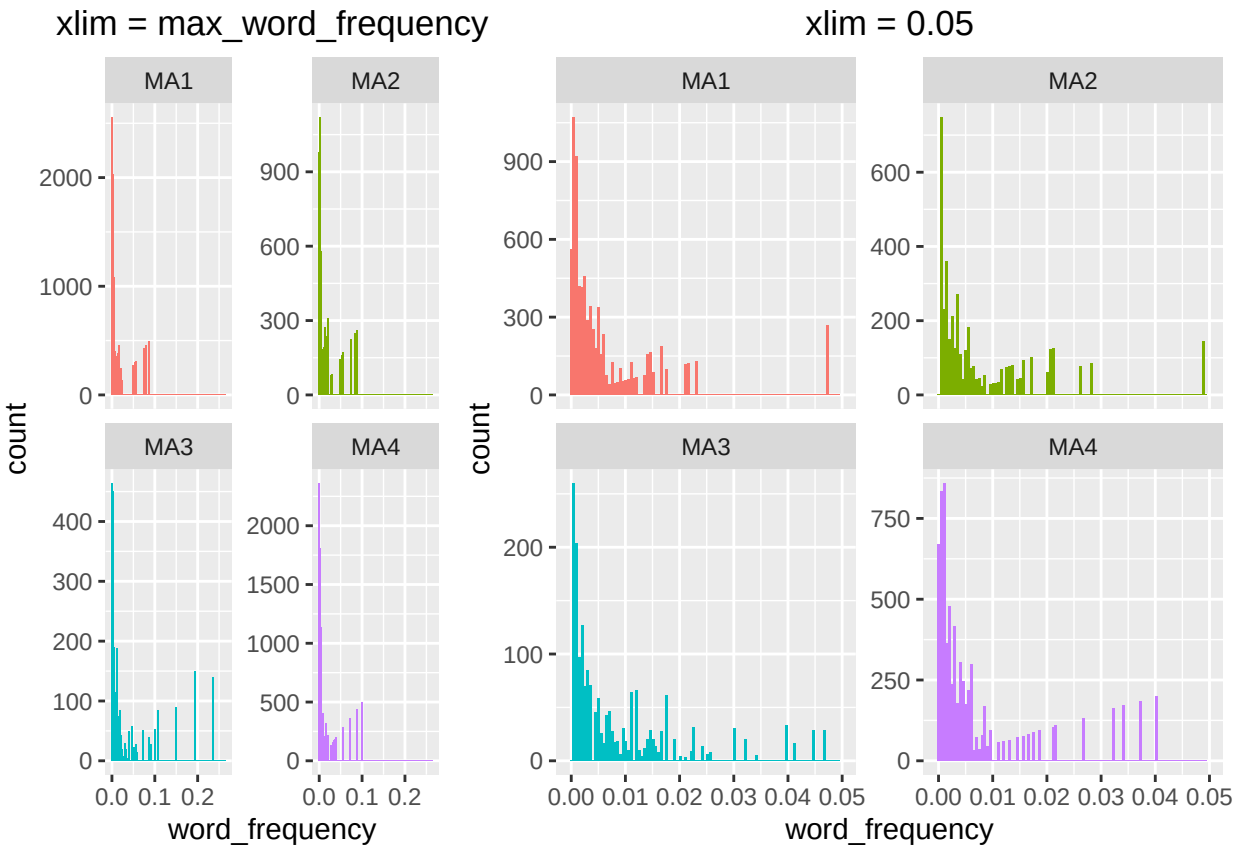
```
unnested_tokens1 <- tibble(abstracts1) %>%
mutate(MA="MA1") %>%
unnest_tokens(word, abstracts1)
unnested_tokens1<-unnested_tokens1 %>% rowwise()%>%
  mutate(word = wordStem(word, language = "en"))

unnested_tokens2 <- tibble(abstracts2) %>%
  mutate(MA="MA2") %>%
  unnest_tokens(word, abstracts2)
unnested_tokens2 <- unnested_tokens2 %>%
  rowwise() %>% mutate(word = wordStem(word, language = "en"))

unnested_tokens3 <- tibble(abstracts3) %>%
  mutate(MA="MA3") %>%
```

```
  unnest_tokens(word, abstracts3)
unnested_tokens3 <- unnested_tokens3 %>% rowwise() %>%
  mutate(word = wordStem(word, language = "en"))

unnested_tokens4 <- tibble(abstracts4) %>%
  mutate(MA="MA4") %>%
  unnest_tokens(word, abstracts4)
unnested_tokens4 <- unnested_tokens4 %>% rowwise() %>%
  mutate(word = wordStem(word, language = "en"))
```

```
word_counts_MA <- c(length(unnested_tokens1$word),
                    length(unnested_tokens2$word),
                    length(unnested_tokens3$word),
                    length(unnested_tokens4$word))

ordre1=unnested_tokens1 %>%
     count(word, sort = TRUE) %>%
     mutate(word = reorder(word, n,decreasing=TRUE))
tokens_ordered_1 <- left_join(unnested_tokens1, ordre1) %>%
  mutate(Total=word_count_tibble$words[1])

ordre2=unnested_tokens2 %>%
     count(word, sort = TRUE) %>%
     mutate(word = reorder(word, n,decreasing=TRUE))
tokens_ordered_2 <- left_join(unnested_tokens2, ordre2) %>%
  mutate(Total=word_count_tibble$words[2])

ordre3=unnested_tokens4 %>%
     count(word, sort = TRUE) %>%
     mutate(word = reorder(word, n,decreasing=TRUE))
tokens_ordered_3 <- inner_join(unnested_tokens3, ordre3) %>%
  mutate(Total=word_count_tibble$words[3])

ordre4=unnested_tokens4 %>%
     count(word, sort = TRUE) %>%
     mutate(word = reorder(word, n,decreasing=TRUE))
tokens_ordered_4 <- left_join(unnested_tokens4, ordre4) %>%
  mutate(Total=word_count_tibble$words[4])

all_freq <- bind_rows(tokens_ordered_1,
                      tokens_ordered_2,
                      tokens_ordered_3,
                      tokens_ordered_4) %>%
  mutate(word_frequency = n/Total)
```

```
library(ggplot2)
library(egg)

word_freq_plot1 <- ggplot(all_freq, aes(x = word_frequency, fill = MA)) +
  geom_histogram(bins=100,show.legend = FALSE) +
  ggtitle("xlim = max_word_frequency") +
  theme(plot.title = element_text(hjust = 0.5)) +
  xlim(NA, max(all_freq$word_frequency)) +
  facet_wrap(~MA, ncol = 2, scales = "free_y")
```

```
word_freq_plot2 <-ggplot(all_freq, aes(x = word_frequency, fill = MA)) +
  geom_histogram(bins=100,show.legend = FALSE) +
  ggtitle("xlim = 0.05") +
  theme(plot.title = element_text(hjust = 0.5)) +
  xlim(NA, 0.05) +
  facet_wrap(~MA, ncol = 2, scales = "free_y")

ggarrange(word_freq_plot1,word_freq_plot2,widths = c(1,2))
```



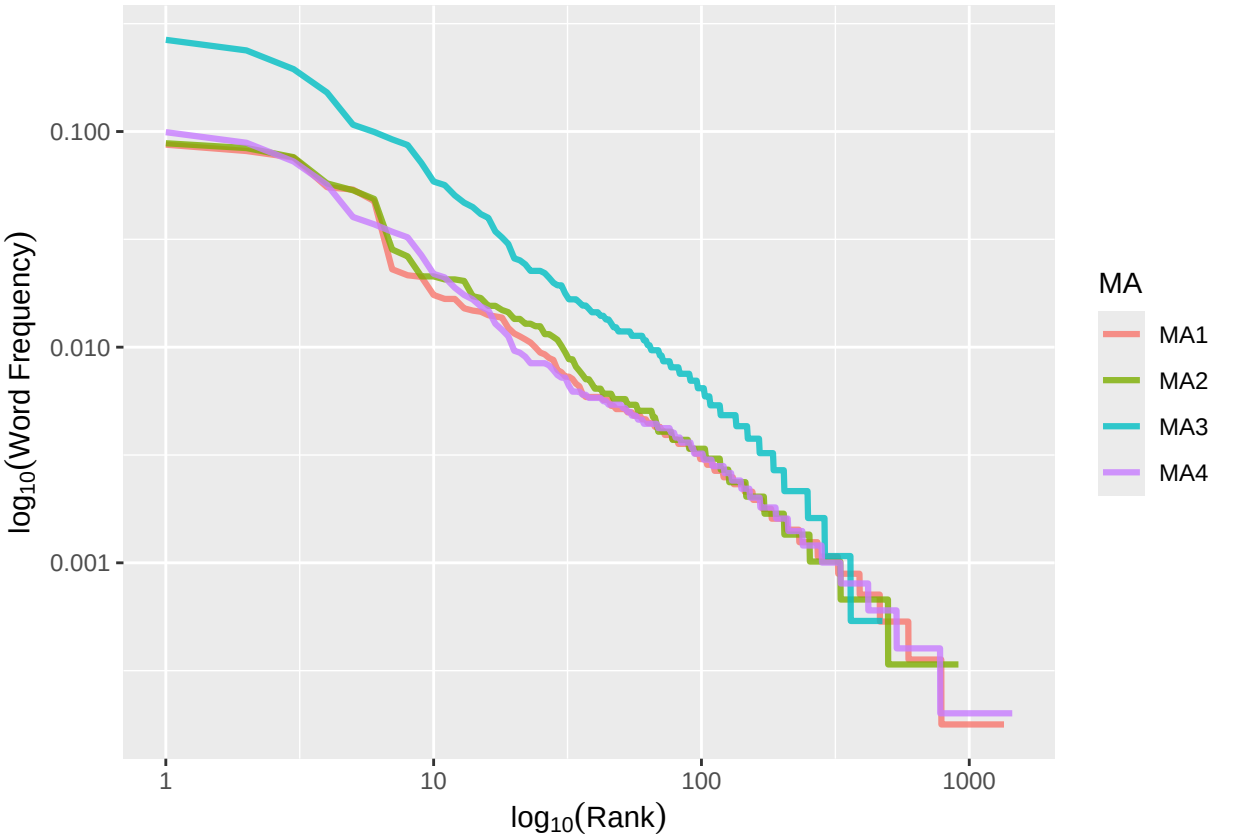

Ces graphiques présentent des distributions proches pour toutes les MA, avec de nombreux mots qui apparaissent rarement et moins de mots qui apparaissent fréquemment (le premier groupe de 4 graphes représentant l'intégralité des données, le deuxième groupe de 4 se focalisant uniquement sur l'intervalle [0;0.05], pour inspecter le massif de pics davantage en détail). En x, on peut voir que les mots apparaissant très souvent (à droite) sont rares, alors que les mots qui apparaissent peu souvent (à gauche) sont fréquents. Ces graphiques présentent des distributions similaires pour toutes les MA, avec de nombreux mots qui apparaissent rarement et moins de mots qui apparaissent fréquemment.

## Loi de Zipf:

La loi de Zipf, découverte empiriquement par Zipf (1949) pour des mots d'un corpus anglais, stipule que si $f$ est la fréquence d'un mot dans le corpus et $r$ le rang, alors : $frequency \propto \frac{1}{rank}$ (la fréquence d'un mot donné est inversement proportionnelle à son rang).

```
library(latex2exp)

freq_and_rank <-  all_freq %>% group_by(MA) %>%
    distinct(word, .keep_all = TRUE) %>%
```

```
    arrange(desc(n)) %>%
    mutate(rank = seq_len(length(word)), word_frequency = n/Total) %>% #
    ungroup()

freq_and_rank %>%
    ggplot(aes(rank, word_frequency, color = MA)) +
    geom_line(linewidth = 1.1, alpha = 0.8, show.legend = TRUE) +
    scale_x_log10() +
    scale_y_log10() +
  xlab(TeX("$\\log_{10}(Rank)$")) +
  ylab(TeX("$\\log_{10}(Word~Frequency)$"))
```



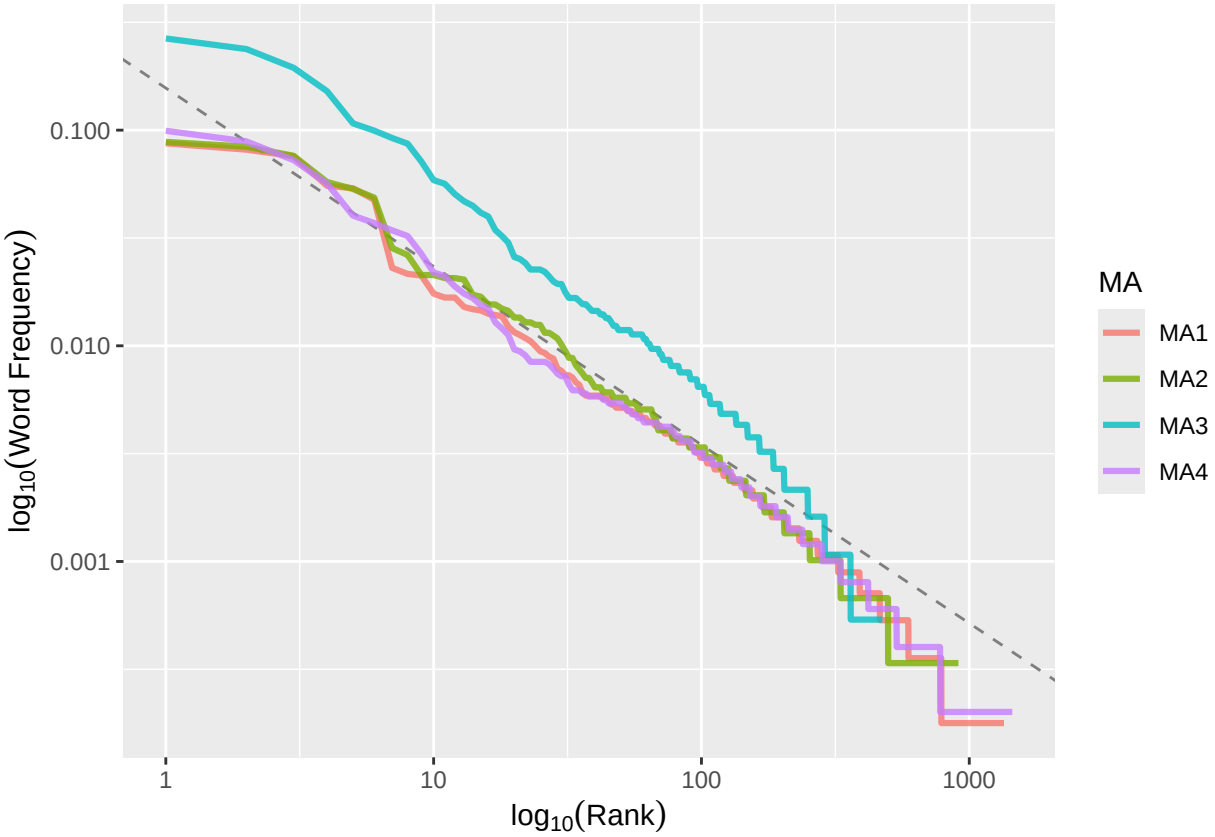

Les courbes de Zipf (log10(f) contre log10(r)) obtenues ci-dessus ont une pente négative. On peut visualiser graphiquement qu'il existe bel et bien une relation inversement proportionnelle entre $f$ et $r$ (échelle logarithmique). On peut cependant remarquer que la décroissance observée n'est pas strictement monotone. On pourrait considérer ces courbes comme des lois de puissance brisées (fonction par morceaux donnée par une séquence de lois de puissance jointes où chaque section a sa propre puissance (indice) et est définie par des "ruptures" limitatives.) et les diviser en trois sections: [1;10]/[10;100]/[100;1000]. Voyons quel est l'exposant de la loi de puissance pour la partie centrale de la plage de rangs.

```
rank_subset <- freq_and_rank %>%
  filter(rank < 100,
         rank > 10)
lm(log10(word_frequency) ~ log10(rank), data = rank_subset)
```

```
##
## Call:
## lm(formula = log10(word_frequency) ~ log10(rank), data = rank_subset)
```

```
##
## Coefficients:
## (Intercept)  log10(rank)
##     -0.7412      -0.8422
```

```
library(latex2exp)

freq_and_rank %>% ggplot(aes(rank, word_frequency, color = MA)) +
  geom_abline(intercept = -0.8048, slope = -0.8272,
              color = "gray50", linetype = 2) +
  geom_line(linewidth = 1.1, alpha = 0.8, show.legend = TRUE) +
  scale_x_log10() +
  scale_y_log10() +
  xlab(TeX("$\\log_{10}(Rank)$")) +
  ylab(TeX("$\\log_{10}(Word~Frequency)$"))
```




On peut remarquer en ragardant la partie supérieure que la MA3 contient davantage de mots “rares” que la valeur prédite par le modèle linéaire, alors que toutes les autres (courbes orange, vertes et blues) en contiennent moins. La MA3 est donc celle qui contient la plus haute fréquence de mots rares. Ce résultat semble logique lorsqu’on considère que c’est aussi elle qui comprends le moins de mots totaux (MA3=1860 < MA2=2959 < MA4=4985 < MA1=5616, *voir table ci-dessous*). Pour ce qui est des mots très fréquents (quart inférieur droit sous 0.001 de fréquence), on peut constater que toutes les métaanalyses en contiennent moins que la valeur prédite. Au centre, les courbes des 4 MA sont relativement proche du modèle linéaire, ce qui n’est pas particulièrement étonnant car il a justement été ajusté pour convenir à l’intervalle correspondant à la partie centrale du graphe.

```
word_count_tibble%>%
  kbl() %>%
```

| MA | words |
|---|---|
| MA1 | 5616 |
| MA2 | 2959 |
| MA3 | 1860 |
| MA4 | 4985 |

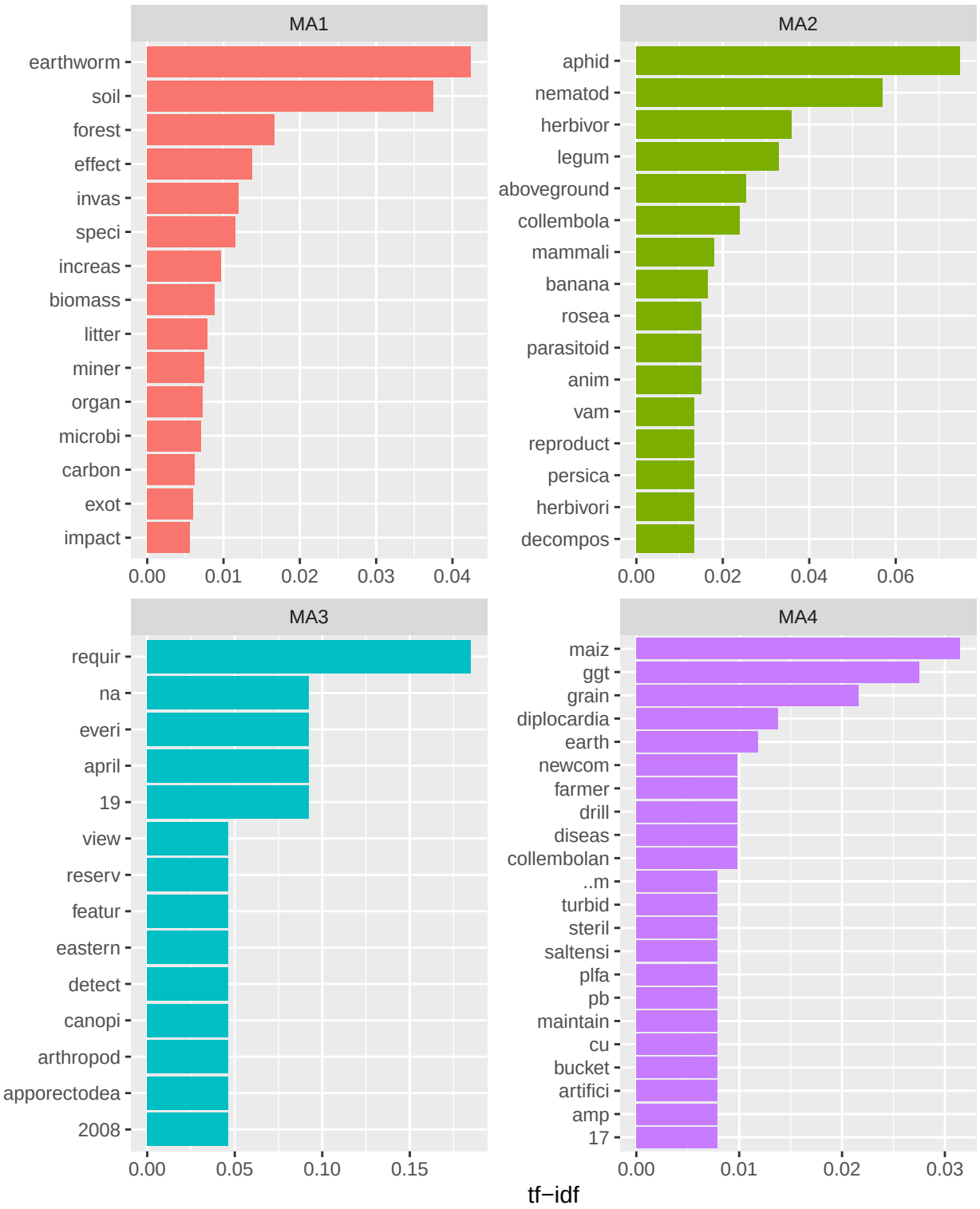
```
  kable_material_dark("hover", full_width = F) %>%
  row_spec(0, bold = T, color = "black", background = "#D7261E")
```

### Fonction bind_tf_idf:

```
MA_tf_idf <- all_freq %>%
  distinct(word, .keep_all = TRUE) %>%
  select(-word_frequency) %>%
  select(-Total) %>%
  bind_tf_idf(word, MA, n) %>%
  arrange(desc(tf_idf))

MA_tf_idf
```

```
## # A tibble: 2,221 x 6
## # Rowwise:
##    MA    word       n     tf   idf tf_idf
##    <chr> <chr>  <int>  <dbl> <dbl>  <dbl>
##  1 MA3   requir     4 0.133   1.39 0.185
##  2 MA3   k          4 0.133   1.39 0.185
##  3 MA3   april      2 0.0667  1.39 0.0924
##  4 MA3   everi      2 0.0667  1.39 0.0924
##  5 MA3   19         2 0.0667  1.39 0.0924
##  6 MA3   open       2 0.0667  1.39 0.0924
##  7 MA3   na         2 0.0667  1.39 0.0924
##  8 MA2   aphid     50 0.0540  1.39 0.0749
##  9 MA1   the      488 0.0494  1.39 0.0685
## 10 MA1   and      456 0.0462  1.39 0.0640
## # i 2,211 more rows
```

```
library(forcats)

MA_tf_idf %>%
  anti_join(stop_words) %>%
  anti_join(custom_stop_words) %>%
  group_by(MA) %>%
  slice_max(tf_idf, n = 15) %>%
  ungroup() %>%
  ggplot(aes(tf_idf, fct_reorder(word, tf_idf), fill = MA)) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~MA, ncol = 2, scales = "free") +
  labs(x = "tf-idf", y = NULL)
```




D'après les graphiques ci-dessus, on peut constater que: - Les mots *"earthworm"*, *"soil"*, *"forest"* et *"effect"* sont plus souvent employés dans la MA1 que dans les autres MA. Le mot *"invasion"* est aussi plus présent dans la MA1 que dans les autres MA. - Les mots *"aphid,*"nematode"*, "herbivore"* et *"legum"* sont plus souvent employés dans la MA2 que dans les autres MA. Le mot *"parasitoid"* est aussi plus présent dans la MA2 que dans les autres MA. - Les mots *"require"*, *"na"*, *"every"* et *"april"* sont plus souvent employés dans la MA3 que dans les autres MA. Les mots *"eastern"*, *"canopy"* et *"arthropod"* sont aussi plus présent dans la

MA3 que dans les autres MA. - Les mots *"maize,*"gtt"*, "grain"* et *"diplocardia"* sont plus souvent employés dans la MA4 que dans les autres MA. Les mots du champs lexical de l'agriculture, comme *"earth"*, *"plfa"*, *"pb"* (probablement employé ici comme le symbole chimique du plomb) et *"cu"* (symbole chimique du cuivre) sont aussi plus présents dans la MA4 que dans les autres MA.

Rappelons les suppositions faites précédemment dans la partie "Analyse de sentiment": - **MA1/MA3:** Métaanalyses globalement défavorables aux vers de terre. Si certains bénéfices, pour la santé des sols notamment, semblent être reconnus, le vers de terre est perçu comme une espèce invasive perturbant les écosystèmes natifs. - **MA2/MA4:** Métaanalyses globalement favorables au vers de terre, mettant en avant les bienfaits physiologiques des vers de terre pour la plante et les avantages de ces interactions pour favoriser la productivité. Cependant, le vers de terre peut malgré tout rester nuisible, car c'est une espèce invasive qui risque de déséquilibrer l'écosystème natif.

En prenant en compte le sujet apparent de chaque métaanalyse, on peut supposer que: - **MA1/MA3:** Métaanalyses globalement défavorables aux vers de terre. Si certains bénéfices, pour la santé des sols notamment, semblent être reconnus, le vers de terre est perçu comme une espèce invasive perturbant les écosystèmes natifs. Les mots "soil", "forest" et "effect" (MA1) et "canopy" (MA3), donnent une idée des environnements où se sont déroulées les études (sols forestiers, forêt). Les mots "exotic" (MA1) et "eastern" (MA3) semblent montrer que les espèces invasives viennent probablement d'Europe ou d'Asie. - **MA2/MA4:** Métaanalyses globalement favorables au vers de terre, mettant en avant les bienfaits physiologiques des vers de terre pour la plante et les avantages de ces interactions pour favoriser la productivité. Cependant, le vers de terre peut malgré tout rester nuisible, car c'est une espèce invasive qui risque de déséquilibrer l'écosystème natif. La MA2 semble plus se focaliser sur les effets observées sur les plantes en tant que telles (pucerons, harbivores) alors que la 4 semble davantage axée sur la pollution des sols, notamment par les métaux lourds (Cu, Pb).

MA1: "Soil chemistry turned upside down: a meta-analysis of invasive earthworm effects on soil chemical properties" MA2: "Earthworms affect plant growth and resistance against herbivores: A meta-analysis" MA3: "The unseen invaders: introduced earthworms as drivers of change in plant communities in North American forests (a meta-analysis)" MA4: "Earthworms increase plant production: a meta-analysis"

### n-grams et corrélations :

```
library(dplyr)
library(tidytext)

all_abstracts_df <- rbind(data.frame(abstract = abstracts1, MA = "MA1"),
      data.frame(abstract = abstracts2, MA = "MA2"),
      data.frame(abstract = abstracts3, MA = "MA3"),
      data.frame(abstract = abstracts4, MA = "MA4"))

tokens_bigrams_all <- tibble(all_abstracts_df) %>%
  unnest_tokens(bigram, abstract, token = "ngrams", n = 2) %>%
  filter(!is.na(bigram))

tokens_bigrams1 <- tibble(abstracts1) %>%
  unnest_tokens(bigram, abstracts1, token = "ngrams", n = 2) %>%
  filter(!is.na(bigram))

tokens_bigrams2 <- tibble(abstracts2) %>%
  unnest_tokens(bigram, abstracts2, token = "ngrams", n = 2) %>%
  filter(!is.na(bigram))

tokens_bigrams3 <- tibble(abstracts3) %>%
  unnest_tokens(bigram, abstracts3, token = "ngrams", n = 2) %>%
```

```
  filter(!is.na(bigram))

tokens_bigrams4 <- tibble(abstracts4) %>%
  unnest_tokens(bigram, abstracts4, token = "ngrams", n = 2) %>%
  filter(!is.na(bigram))

tokens_bigrams1 %>% count(bigram, sort = TRUE)
```

```
## # A tibble: 6,279 x 2
##    bigram                n
##    <chr>             <int>
##  1 in the               81
##  2 of the               73
##  3 effects of           31
##  4 microbial biomass    31
##  5 organic matter       29
##  6 of earthworm         26
##  7 of earthworms        26
##  8 earthworm invasion   25
##  9 forest floor         25
## 10 c and                24
## # i 6,269 more rows
```

```
tokens_bigrams2 %>%
 count(bigram, sort = TRUE)
```

```
## # A tibble: 3,564 x 2
##    bigram           n
##    <chr>        <int>
##  1 in the          35
##  2 of earthworms   34
##  3 effects of      31
##  4 presence of     31
##  5 the presence    30
##  6 of the          28
##  7 on the          18
##  8 plant species   17
##  9 plant growth    16
## 10 affected by     15
## # i 3,554 more rows
```

```
tokens_bigrams3 %>%
 count(bigram, sort = TRUE)
```

```
## # A tibble: 2,448 x 2
##    bigram              n
##    <chr>           <int>
##  1 of the             19
##  2 in the             16
##  3 non native         15
##  4 plant species      15
##  5 native plant       14
##  6 earthworm species  11
##  7 of earthworm       11
##  8 and the            10
```

```
##  9 o tyrtaeum           10
## 10 changes in            9
## # i 2,438 more rows
```

```
tokens_bigrams4 %>%
 count(bigram, sort = TRUE)
```

```
## # A tibble: 6,044 x 2
##    bigram              n
##    <chr>           <int>
##  1 in the             80
##  2 of the             64
##  3 presence of        43
##  4 the presence       41
##  5 of earthworms      34
##  6 on the             28
##  7 the soil           27
##  8 plant community    26
##  9 effects of         25
## 10 in a               23
## # i 6,034 more rows
```

De nombreux bigrammes sont non-informatifs (“in the”, “of the” etc.). On va donc réemployer l’approche en “stop words” des étapes précédentes pour enlever ces mots.

```
# bigrams ok, separated ok,  filtered ok.
bigrams_separated_all <- tokens_bigrams_all %>%
  separate(bigram, c("word1", "word2"), sep = " ")

bigrams_separated1 <- tokens_bigrams1 %>%
  separate(bigram, c("word1", "word2"), sep = " ")

bigrams_filtered_all <- bigrams_separated_all %>%
  filter(!word1 %in% combined_stop_words$word) %>%
  filter(!word2 %in% combined_stop_words$word)

bigrams_filtered1 <- bigrams_separated1 %>%
  filter(!word1 %in% combined_stop_words$word) %>%
  filter(!word2 %in% combined_stop_words$word)

# new bigram counts:
bigram_counts_all <- bigrams_filtered_all %>%
  count(MA, word1, word2, sort = TRUE)

bigram_counts1 <- bigrams_filtered1 %>%
  count(word1, word2, sort = TRUE)

# For bigrams_2
bigrams_separated2 <- tokens_bigrams2 %>%
  separate(bigram, c("word1", "word2"), sep = " ")

bigrams_filtered2 <- bigrams_separated2 %>%
  filter(!word1 %in% combined_stop_words$word) %>%
  filter(!word2 %in% combined_stop_words$word)

bigram_counts2 <- bigrams_filtered2 %>%
```

```
  count(word1, word2, sort = TRUE)

# For bigrams_3
bigrams_separated3 <- tokens_bigrams3 %>%
  separate(bigram, c("word1", "word2"), sep = " ")

bigrams_filtered3 <- bigrams_separated3 %>%
  filter(!word1 %in% combined_stop_words$word) %>%
  filter(!word2 %in% combined_stop_words$word)

bigram_counts3 <- bigrams_filtered3 %>%
  count(word1, word2, sort = TRUE)

# For bigrams_4
bigrams_separated4 <- tokens_bigrams4 %>%
  separate(bigram, c("word1", "word2"), sep = " ")

bigrams_filtered4 <- bigrams_separated4 %>%
  filter(!word1 %in% combined_stop_words$word) %>%
  filter(!word2 %in% combined_stop_words$word)

bigram_counts4 <- bigrams_filtered4 %>%
  count(word1, word2, sort = TRUE)

bigrams_united_all <- bigrams_filtered_all %>%
  unite(bigram, word1, word2, sep = " ")

bigrams_united_all %>% count(MA, bigram, sort = TRUE)
```

```
## # A tibble: 5,506 x 3
##    MA    bigram                n
##    <chr> <chr>             <int>
##  1 MA1   microbial biomass     31
##  2 MA1   organic matter        29
##  3 MA4   plant community       26
##  4 MA1   earthworm invasion    25
##  5 MA1   forest floor          25
##  6 MA1   exotic earthworms     24
##  7 MA1   mineral soil          23
##  8 MA1   earthworm species     18
##  9 MA2   plant species         17
## 10 MA4   plant growth          17
## # i 5,496 more rows
```

```
earthworms_bigrams <- bigrams_united_all %>%
    filter(grepl("earthworms?", bigram, ignore.case = TRUE))
# Count occurrences of MA and word2 combination, sorted in descending order
count_earthworm_bigrams<- earthworms_bigrams %>%
    count(MA, bigram, sort = TRUE) %>%
    group_by(MA) %>%
    filter(n>5) %>%
    mutate(MA = factor(MA, levels = c("MA1", "MA2", "MA3", "MA4"))) %>%
    # Set the order of MA
    arrange(MA, desc(n)) # Sort by MA and then by n in descending order
```

```
count_earthworm_bigrams
```

```
## # A tibble: 22 x 3
## # Groups:   MA [4]
##    MA    bigram                  n
##    <fct> <chr>                <int>
##  1 MA1   earthworm invasion      25
##  2 MA1   exotic earthworms      24
##  3 MA1   earthworm species      18
##  4 MA1   earthworm activity      9
##  5 MA1   earthworm biomass       8
##  6 MA1   earthworm invaded       7
##  7 MA1   european earthworms     7
##  8 MA1   exotic earthworm        7
##  9 MA1   earthworm community     6
## 10 MA2   earthworms increased    7
## # i 12 more rows
```

```
library(dplyr)

# Define the function
count_filtered_bigrams <- function(data, regex_motif,treshold) {
  # Filter the bigrams based on the provided regex motif
  filtered_bigrams <- data %>%
    filter(grepl(regex_motif, bigram, ignore.case = TRUE))

  # Count occurrences of each combination of `MA` and `bigram`,
  # sort in descending order,
  # group by `MA`, and filter counts greater than 5 within each group
  count_bigrams <- filtered_bigrams %>%
    count(MA, bigram, sort = TRUE) %>%
    group_by(MA) %>%
    filter(n > treshold) %>%
    mutate(MA = factor(MA, levels = c("MA1", "MA2", "MA3", "MA4"))) %>%
    arrange(MA, desc(n)) # Sort by MA and then by n in descending order

  return(count_bigrams)
}

count_filtered_bigrams(bigrams_united_all,"earthworms?",5)
```

```
## # A tibble: 22 x 3
## # Groups:   MA [4]
##    MA    bigram                  n
##    <fct> <chr>                <int>
##  1 MA1   earthworm invasion      25
##  2 MA1   exotic earthworms      24
##  3 MA1   earthworm species      18
##  4 MA1   earthworm activity      9
##  5 MA1   earthworm biomass       8
##  6 MA1   earthworm invaded       7
##  7 MA1   european earthworms     7
##  8 MA1   exotic earthworm        7
##  9 MA1   earthworm community     6
```

```
## 10 MA2   earthworms increased     7
## # i 12 more rows
```

```
count_filtered_bigrams(bigrams_united_all,"invasions?",5)
```

```
## # A tibble: 3 x 3
## # Groups:   MA [2]
##   MA    bigram                 n
##   <fct> <chr>              <int>
## 1 MA1   earthworm invasion    25
## 2 MA3   earthworm invasion     7
## 3 MA3   earthworm invasions    6
```

```
count_filtered_bigrams(bigrams_united_all,"invaders?",0)
```

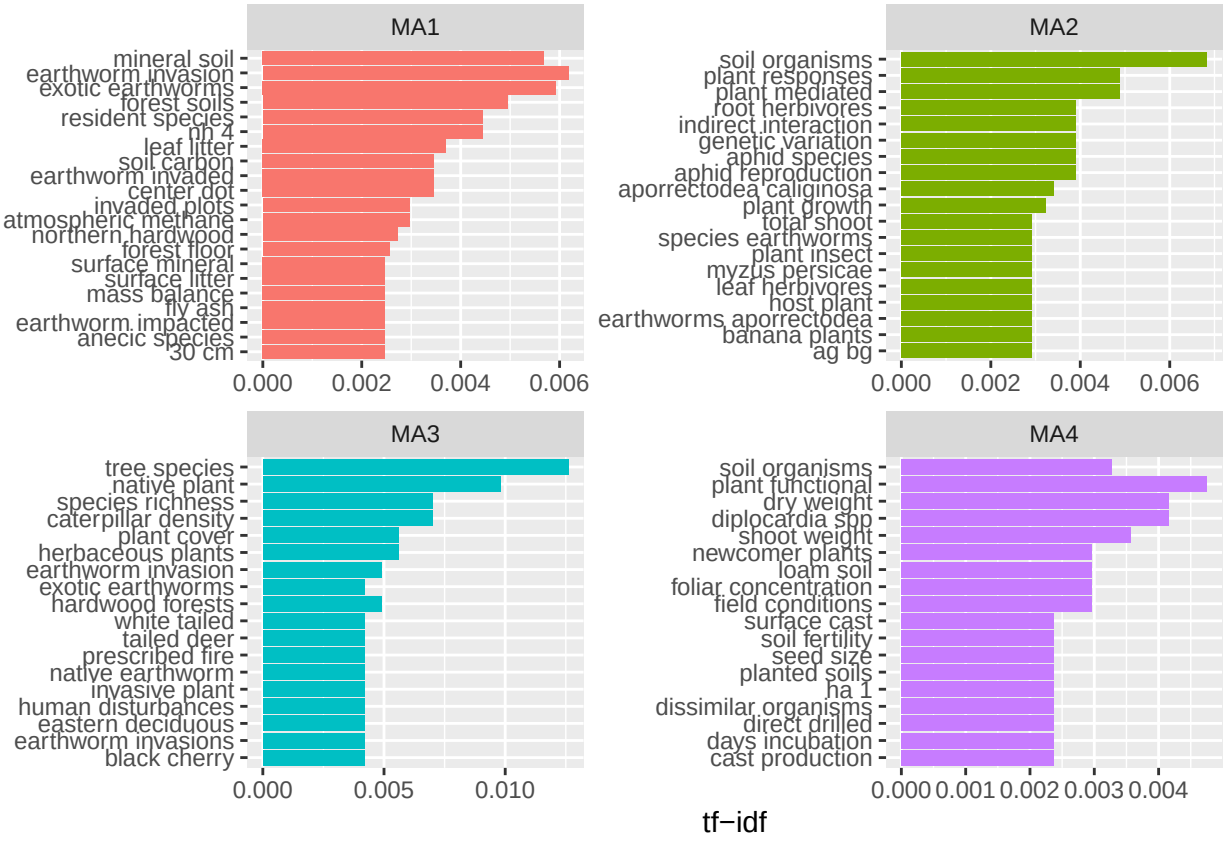
```
## # A tibble: 9 x 3
## # Groups:   MA [2]
##   MA    bigram                n
##   <fct> <chr>             <int>
## 1 MA1   earthworm invaders    1
## 2 MA1   invaders affect       1
## 3 MA4   invader plants        2
## 4 MA4   invader species       2
## 5 MA4   plant invader         2
## 6 MA4   invader biomass       1
## 7 MA4   invader seed          1
## 8 MA4   invader treatments    1
## 9 MA4   plant invaders        1
```

```
count_filtered_bigrams(bigrams_united_all,"europeans?",0)
```

```
## # A tibble: 23 x 3
## # Groups:   MA [4]
##    MA    bigram                 n
##    <fct> <chr>              <int>
##  1 MA1   european earthworms     7
##  2 MA1   european species        5
##  3 MA1   european earthworm      4
##  4 MA1   european settlers       2
##  5 MA1   introduced european     2
##  6 MA1   invasive european       2
##  7 MA1   european lumbricids     1
##  8 MA1   exotic european         1
##  9 MA1   processes european      1
## 10 MA1   time european           1
## # i 13 more rows
```

```
bigram_tf_idf <- bigrams_united_all %>%
  count(MA, bigram) %>%
  bind_tf_idf(bigram, MA, n) %>%
  arrange(desc(tf_idf))

bigram_tf_idf %>%
    group_by(MA) %>%
    slice_max(tf_idf, n = 15) %>%
    ungroup() %>%
    ggplot(aes(tf_idf, fct_reorder(bigram, tf_idf), fill = MA)) +
```

```
    geom_col(show.legend = FALSE) +
    facet_wrap(~MA, ncol = 2, scales = "free") +
    labs(x = "tf-idf", y = NULL)
```




D'après les graphiques ci-dessus, on peut constater que: Les bigrammes *"mineral soil"*, *"earthworm invasion"*, *"exotic earthworms"* et *"forest soils"* sont plus souvent employés dans la MA1 que dans les autres MA. Le bigramme *"earthworm invaded"* est aussi plus présent dans la MA1 que dans les autres MA. Les bigrammes *"soil organisms"*, *"plant responses"*, *"plant mediated"* et *"root herbivore"* sont plus souvent employés dans la MA2 que dans les autres MA. Le bigramme *"aphid species"* est aussi plus présent dans la MA2 que dans les autres MA. Les bigrammes *"tree species"*, *"native plant"*, *"species richness"* et *"caterpillar density"* sont plus souvent employés dans la MA3 que dans les autres MA. Le bigramme *"earthworm invasions"* est aussi plus présent dans la MA3 que dans les autres MA. Les bigrammes *"soil organisms"*, *"plant functionnal"*, *"dry weight"* et *"diplocardia spp"* sont plus souvent employés dans la MA4 que dans les autres MA. Le bigramme *"soil fertility"* est aussi plus présent dans la MA4 que dans les autres MA.

Rappelons les suppositions faites dans la section précédente: - **MA1/MA3:** Métaanalyses globalement défavorables aux vers de terre. Si certains bénéfices, pour la santé des sols notamment, semblent être reconnus, le vers de terre est perçu comme une espèce invasive perturbant les écosystèmes natifs. Les mots "soil", "forest" et "effect" (MA1) et "canopy" (MA3), donnent une idée des environnements où se sont déroulées les études (sols forestiers, forêt). Les mots "exotic" (MA1) et "eastern" (MA3) semblent montrer que les espèces invasives viennent probablement d'Europe ou d'Asie. - **MA2/MA4:** Métaanalyses globalement favorables au vers de terre, mettant en avant les bienfaits physiologiques des vers de terre pour la plante et les avantages de ces interactions pour favoriser la productivité. Cependant, le vers de terre peut malgré tout rester nuisible, car c'est une espèce invasive qui risque de déséquilibrer l'écosystème natif. La MA2 semble plus se focaliser sur les effets observées sur les plantes en tant que telles (pucerons, harbivores) alors que la 4 semble davantage axée sur la pollution des sols, notamment par les métaux lourds (Cu, Pb).

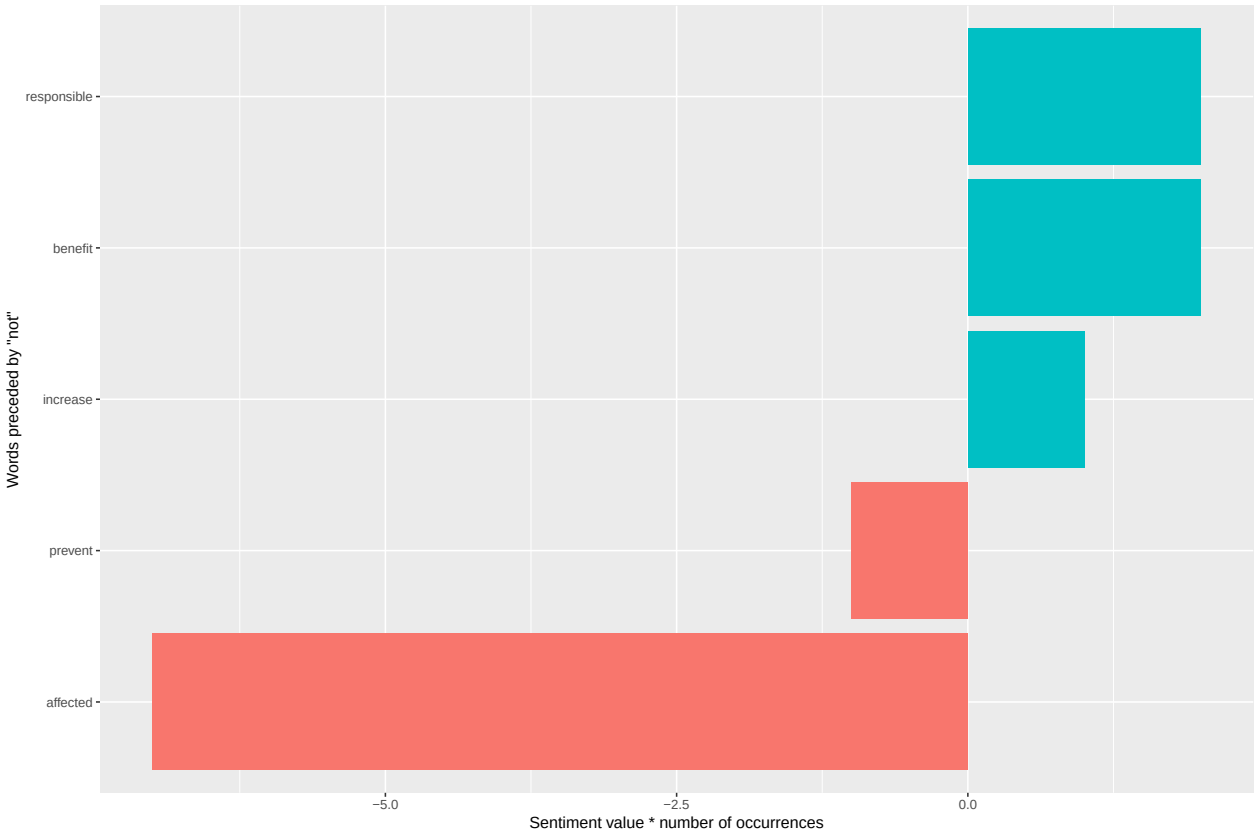
En prenant en compte le sujet apparent de chaque métaanalyse, on peut supposer que: - **MA1/MA3:**

Métaanalyses globalement défavorables aux vers de terre. Si certains bénéfices, pour la santé des sols notamment, semblent être reconnus, le vers de terre est perçu comme une espèce invasive perturbant les écosystèmes natifs. Les mots "soil", "forest" et "effect" (MA1) et "canopy" (MA3), donnent une idée des environnements où se sont déroulées les études (sols forestiers, forêt). Les mots "exotic" (MA1) et "eastern" (MA3) semblent montrer que les espèces invasives viennent probablement d'Europe ou d'Asie. Les bigrammes "forest soils" (MA1) et "tree species" (MA3) confirment que les études s'intéressent aux environnements forestiers. Les bigrammes "earthworm invasion" (MA1) et "earthworm invasions" (MA3) confirment que les **vers de terre asiatiques et européens sont perçus comme des espèces invasives**. - **MA2/MA4:** Métaanalyses globalement favorables au vers de terre, mettant en avant les bienfaits physiologiques des vers de terre pour la plante et les avantages de ces interactions pour favoriser la productivité. Cependant, le vers de terre peut malgré tout rester nuisible, car c'est une espèce invasive qui risque de déséquilibrer l'écosystème natif. La MA2 semble plus se focaliser sur les effets observées sur les plantes en tant que telles (pucerons, harbivores) alors que la 4 semble davantage axée sur la pollution des sols, notamment par les métaux lourds (Cu, Pb). Le bigramme "soil organisms" est le plus représenté dans les deux métaanalyses, reforçant encore le lien thématique (écologie des sols) commun aux MA2 et 4. Par ailleurs, de nombreuses références aux plantes (*"plant responses"*, *"plant mediated"* et *"root herbivore"* pour la MA2 et *"plant functionnal"*, *"dry weight"* et *"diplocardia spp"* pour la MA4) montrent que les deux MA s'intéressent spécifiquement au **rôle des plantes dans l'écologie des sols**.

MA1: "Soil chemistry turned upside down: a meta-analysis of invasive earthworm effects on soil chemical properties" MA2: "Earthworms affect plant growth and resistance against herbivores: A meta-analysis" MA3: "The unseen invaders: introduced earthworms as drivers of change in plant communities in North American forests (a meta-analysis)" MA4: "Earthworms increase plant production: a meta-analysis"

### n-grams et analyse de sentiment:

```
 library(ggplot2)

AFINN <-get_sentiments("afinn")

not_words <- bigrams_separated_all %>%
  filter(word1 == "not") %>%
  inner_join(AFINN, by = c(word2 = "word")) %>%
  count(word2, value, sort = TRUE)

not_words %>%
  mutate(contribution = n * value) %>%
  arrange(desc(abs(contribution))) %>%
  mutate(word2 = reorder(word2, contribution)) %>%
  ggplot(aes(contribution, word2, fill = contribution > 0)) +
  geom_col(show.legend = FALSE) +
  labs(x = "Sentiment value * number of occurrences",
       y = "Words preceded by \"not\"")
```



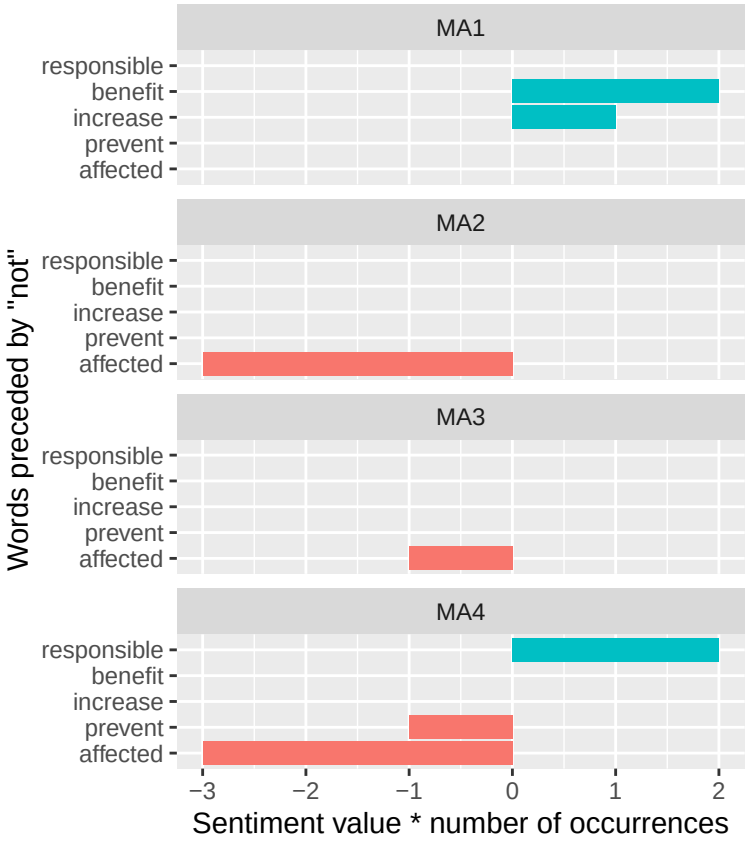

Les bigrammes *"not benefit"* et *"not increase"* sont les plus important facteur d'erreur d'identification, faisant paraître le texte plus **positif** qu'il ne l'est réellement. Le bigramme *"not affected"* est le plus important facteur d'erreur d'identification, faisant paraître le texte plus **négatif** qu'il ne l'est réellement.

```
library(ggplot2)

AFINN <-get_sentiments("afinn")
bigrams_separated1 %>%
  filter(word1 == "not") %>%
  inner_join(AFINN, by = c(word2 = "word")) %>%
  count(word2, value, sort = TRUE)
```

```
## # A tibble: 2 x 3
##   word2    value     n
##   <chr>    <dbl> <int>
## 1 benefit      2     1
## 2 increase     1     1
```

```
not_words_with_factor <- bigrams_separated_all %>%
    filter(word1 == "not") %>%
    inner_join(AFINN, by = c(word2 = "word")) %>%
    count(MA, word2, value, sort = TRUE)

not_words_with_factor %>%
  mutate(contribution = n * value) %>%
  arrange(desc(abs(contribution))) %>%
  mutate(word2 = reorder(word2, contribution)) %>%
  ggplot(aes(contribution, word2, fill = contribution > 0)) +
  geom_col(show.legend = FALSE) +
```

```
  labs(x = "Sentiment value * number of occurrences",
       y = "Words preceded by \"not\"") +
  coord_fixed(ratio = 0.25)+
  facet_wrap(facets = vars(MA), ncol = 1)
```




Lorsque l'on inspecte le jeu de données **pour chacune des MA** avec la fonction R *"facet_grid"*, on remarque que les mots ayant conduit au plus d'erreurs d'identification dans chaque MA sont: - MA1: *"benefit"* et *"increase"* classés à tort en **positif** (comme le mot précédant était "not", "not benefit" et "not increase" ont une **valance négative**), pas de négatif relevé. **La MA1 est moins positive que les analyses précédentes le laissaient penser.** - MA2: Pas de mot positif relevé. *"affected"* classé à tort en **négatif** (comme le mot précédent était "not", "not affected" a une **valence positive**). **La MA2 est plus positive que les analyses précédentes le laissaient penser.** - MA3: Pas de mot positif relevé. *"affected"* classé à tort en **négatif** (comme le mot précédent était "not", "not affected" a une **valence positive**). On remarque cependant que l'effet observé est moindre comparé à la MA2. **La MA3 est plus positive que les analyses précédentes le laissaient penser**. - MA4: *"responsible"* (donc, le bigramme "not responsible"), est classé positif par le dictionnaire AFINN, mais ces seuls deux mots ne suffisent pas à déterminer la valence (positive ou négative), de l'expression. Elles traduisent au contraire une forme de neutralité (quelque-chose qui est **non responsable** d'un effet est à priori neutre). Une analyse plus détaillée du contexte d'utilisation serait requise pour pouvoir décider.

```
AFINN %>% filter(word %in% c("responsible", "prevent"))
```

```
## # A tibble: 2 x 2
##   word        value
##   <chr>       <dbl>
## 1 prevent        -1
## 2 responsible     2
```

Dans AFINN (dictionnaire), *"prevent"* est classé négatif (c'est pourquoi "not prevent" a une valence positive), mais en écologie, "not prevent" semble bel et bien négatif (il n'y a donc pas d'erreur d'identification dans cas précis). Par contre, classé à tort en **négatif**, on retrouve là-aussi *"affected"*, dans ce cas précédé par *"not"*, et donc de valence plutôt **positive**. **La MA4 est plus positive que les analyses précédentes le laissaient penser**.

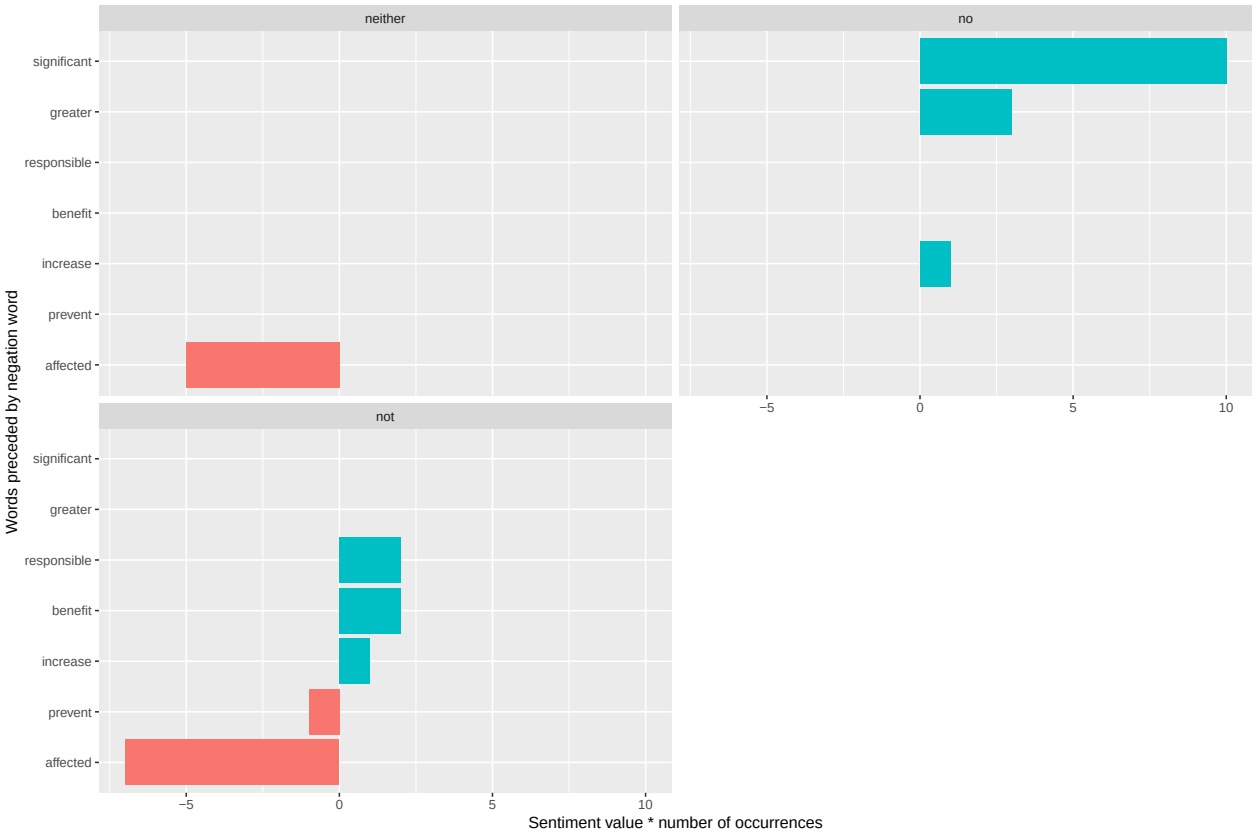
Cependant, d'autres mots négatifs (comme *"never"* etc.) n'ont pas été pris en compte dans l'analyse ci-dessus. La prochaine étape à donc pour but de prendre davantage de mots de négation en considération dans l'analyse de sentiment.

```
negation_words <- c("ain't", "aren't", "can't", "couldn't", "didn't", "doesn't",
                    "don't", "hasn't", "isn't", "mightn't", "mustn't",
                    "neither", "never", "no", "nobody", "nor",
                    "not", "shan't", "shouldn't", "wasn't", "weren't", "won't",
                    "wouldn't","without")
bigrams_separated_all %>%
    filter(word1 %in% negation_words) %>%
    inner_join(AFINN, by = c(word2 = "word"))
```

```
## # A tibble: 28 x 4
##    MA    word1   word2       value
##    <chr> <chr>   <chr>       <dbl>
##  1 MA1   no      significant     1
##  2 MA1   no      significant     1
##  3 MA1   not     increase        1
##  4 MA1   no      greater         3
##  5 MA1   not     benefit         2
##  6 MA1   no      significant     1
##  7 MA1   no      significant     1
##  8 MA1   neither affected       -1
##  9 MA1   no      significant     1
## 10 MA2   not     affected       -1
## # i 18 more rows
```

```
library(dplyr)
library(ggplot2)
library(tidyr)

negation_words_with_factor <- bigrams_separated_all %>%
    filter(word1 %in% negation_words) %>%
    inner_join(AFINN, by = c(word2 = "word")) %>%
    count(MA, word1, word2, value, sort = TRUE)
```

```
negation_words_with_factor %>%
   mutate(contribution = n * value) %>%
   arrange(desc(abs(contribution))) %>%
   mutate(word2 = reorder(word2, contribution)) %>%
   ggplot(aes(contribution, word2, fill = contribution > 0)) +
   geom_col(show.legend = FALSE) +
   labs(x = "Sentiment value * number of occurrences",
        y = "Words preceded by negation word") +
   #coord_fixed(ratio = 0.25)+ : ATTENTION, peut sévèrement interférer
   #avec l'aspect d'un plot si mal utilisé.
   facet_wrap(facets = vars(word1), ncol = 2)
```




Rappelons les suppositions faites dans la section précédente: - **MA1/MA3:** Métaanalyses globalement défavorables aux vers de terre. Si certains bénéfices, pour la santé des sols notamment, semblent être reconnus, le vers de terre est perçu comme une espèce invasive perturbant les écosystèmes natifs. Les mots "soil", "forest" et "effect" (MA1) et "canopy" (MA3), donnent une idée des environnements où se sont déroulées les études (sols forestiers, forêt). Les mots "exotic" (MA1) et "eastern" (MA3) semblent montrer que les espèces invasives viennent probablement d'Europe ou d'Asie. Les bigrammes "forest soils" (MA1) et "tree species" (MA3) confirment que les études s'intéressent aux environnements forestiers. Les bigrammes "earthworm invasion" (MA1) et "earthworm invasions" (MA3) confirment que les vers de terre asiatiques et européens sont perçus comme des espèces invasives. - **MA2/MA4:** Métaanalyses globalement favorables au vers de terre, mettant en avant les bienfaits physiologiques des vers de terre pour la plante et les avantages de ces interactions pour favoriser la productivité. Cependant, le vers de terre peut malgré tout rester nuisible, car c'est une espèce invasive qui risque de déséquilibrer l'écosystème natif. La MA2 semble plus se focaliser sur les effets observées sur les plantes en tant que telles (pucerons, harbivores) alors que la 4 semble davantage axée sur la pollution des sols, notamment par les métaux lourds (Cu, Pb). Le bigramme "soil organisms" est le plus représenté dans les deux métaanalyses, reforçant encore le lien thématique (écologie des sols) commun aux MA2 et 4. Par ailleurs, de nombreuses références aux plantes (*"plant responses"*, *"plant mediated"* et *"root herbivore"* pour la MA2 et *"plant functionnal"*, *"dry weight"* et *"diplocardia spp"* pour la MA4) montrent que les deux MA s'intéressent spécifiquement au rôle des plantes dans l'écologie des sols.

En prenant en compte le sujet apparent de chaque métaanalyse, on peut supposer que: - **MA1/MA3:** Métaanalyses globalement défavorables aux vers de terre. Si certains bénéfices, pour la santé des sols notamment, semblent être reconnus, le vers de terre est perçu comme une espèce invasive perturbant les écosystèmes natifs. Les mots "soil", "forest" et "effect" (MA1) et "canopy" (MA3), donnent une idée des environnements où se sont déroulées les études (sols forestiers, forêt). Les mots "exotic" (MA1) et "eastern" (MA3) semblent montrer que les espèces invasives viennent probablement d'Europe ou d'Asie. Les bigrammes "forest soils" (MA1) et "tree species" (MA3) confirment que les études s'intéressent aux environnements forestiers. Les bigrammes "earthworm invasion" (MA1) et "earthworm invasions" (MA3) confirment que les vers de terre asiatiques et européens sont perçus comme des espèces invasives. La MA1 étant plus négative

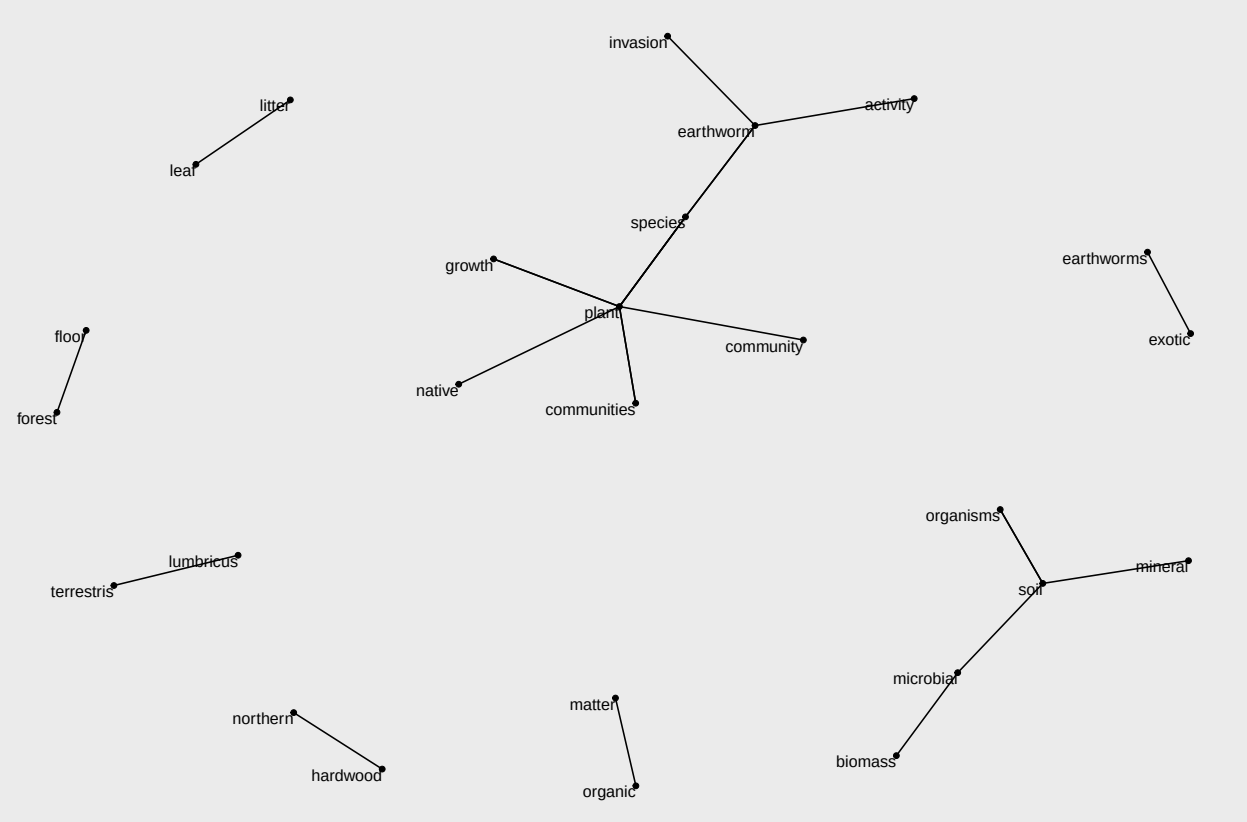
que prévu, sachant que la MA3 (plus positive que prévu) ne semble pas compenser totalement cet effet (cf. “words preceded by not”), on en déduit que ces deux métaanalyses sont un peu plus négatives que présenté précedemment, en raison de l’effet du mot “not” sur le mot suivant. - **MA2/MA4:** Métaanalyses globalement favorables au vers de terre, mettant en avant les bienfaits physiologiques des vers de terre pour la plante et les avantages de ces interactions pour favoriser la productivité. Cependant, le vers de terre peut malgré tout rester nuisible, car c’est une espèce invasive qui risque de déséquilibrer l’écosystème natif. La MA2 semble plus se focaliser sur les effets observées sur les plantes en tant que telles (pucerons, harbivores) alors que la 4 semble davantage axée sur la pollution des sols, notamment par les métaux lourds (Cu, Pb). Le bigramme “soil organisms” est le plus représenté dans les deux métaanalyses, reforçant encore le lien thématique (écologie des sols) commun aux MA2 et 4. Par ailleurs, de nombreuses références aux plantes (*“plant responses”*, *“plant mediated”* et *“root herbivore”* pour la MA2 et *“plant functionnal”*, *“dry weight”* et *“diplocardia spp”* pour la MA4) montrent que les deux MA s’intéressent spécifiquement au rôle des plantes dans l’écologie des sols. La MA2 étant plus positive que prévu, sachant que la MA4 est aussi plus positive que prévu (cf. “words preceded by not”), on en déduit que ces deux métaanalyses sont un peu plus positives que présenté précedemment, en raison de l’effet du mot “not” sur le mot suivant. - **Faux positifs / faux négatifs:** On peut remarquer que le mot le plus important dans les faux négatifs est “not”, et que le plus important dans les faux positfs est “no”. Le terme “neither” a lui-aussi engendré quelques faux négatifs.

MA1: “Soil chemistry turned upside down: a meta-analysis of invasive earthworm effects on soil chemical properties” MA2: “Earthworms affect plant growth and resistance against herbivores: A meta-analysis” MA3: “The unseen invaders: introduced earthworms as drivers of change in plant communities in North American forests (a meta-analysis)” MA4: “Earthworms increase plant production: a meta-analysis”

## Réseaux de bigrammes (ggraph)

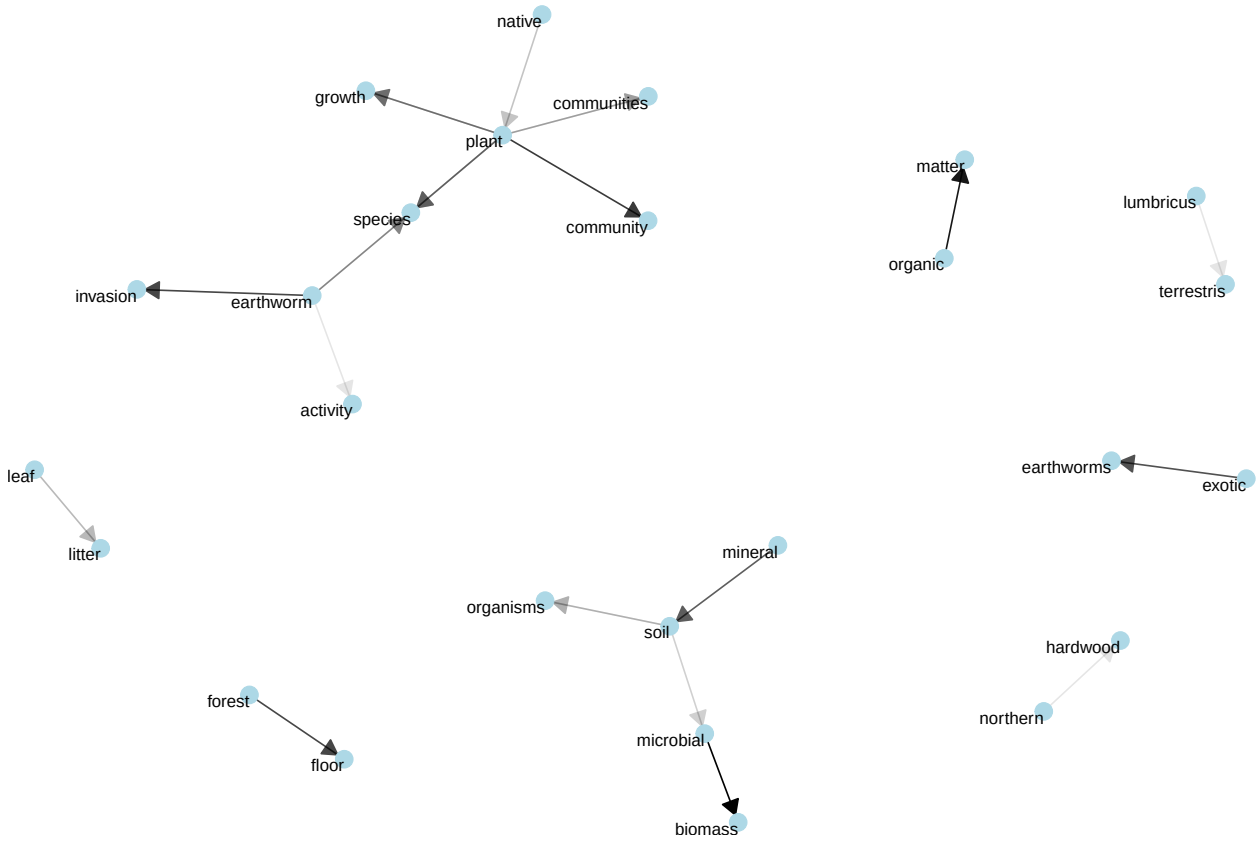
Visualiser les relations entre tous les mots simultanément, sous la forme d’un graphe noeuds / arrêtes. Un graphe est une combinaison de noeuds (mots) connectés, c’est à dire que leurs occurrences sont, dans le texte, proches les unes des autres. C’est une façon d’aller encore plus loin que de simplement regarder le mot précédant (bigrammes).

```
library(igraph)
library(ggraph)

bigram_counts_all_noMA <- bigram_counts_all %>% select(-MA)

bigram_graph <- bigram_counts_all_noMA %>%
  filter(n > 10) %>%
  graph_from_data_frame()

ggraph(bigram_graph, layout = "fr") +
  geom_edge_link() +
  geom_node_point() +
  geom_node_text(aes(label = name), vjust = 1, hjust = 1)
```




Ce premier réseau de bigrammes (toutes MA confondues) nous montre que trois mots sont un centre (un noeud avec au moins trois connections): - **earthworm:** Connecté avec *“species”*, *“invasion”* et *“activity”*. - **soil:** Connecté avec *“mineral”*, *“microbial”* et *“organism”*. - **plant:** Connecté avec *“native”*, *“plant”* et *“community/communities”*. Pour le cluster centré sur “soil”, on remarque que “microbial” et “biomass” sont connectés. De la même façon, les clusters mineurs (2 mots / 1 connexion) indiquent que les mot “lumbricus terrestris”, “leaf litter”, exotic earthworms”, “organic matter”, “forest floor” et “hardwood forest” sont fréquement employés en groupe de mots. Cependant, l’abscence de flèches sur ce graphe nous empêche de savoir avec certitude dans quel ordre ces mots sont utilisés.

```
a <- grid::arrow(type = "closed", length = unit(.15, "inches"))

ggraph(bigram_graph, layout = "fr") +
  geom_edge_link(aes(edge_alpha = n), show.legend = FALSE,
                 arrow = a, end_cap = circle(.07, 'inches')) +
  geom_node_point(color = "lightblue", size = 5) +
  geom_node_text(aes(label = name), vjust = 1, hjust = 1) +
  theme_void()
```



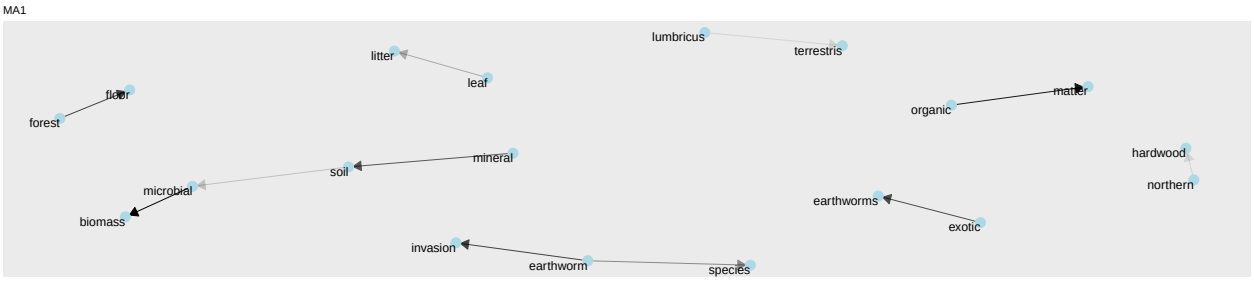

Dans cette nouvelle version du même graphe: - Les liens sont **orientés**, afin de savoir quel mot est le premier de chaque bigramme. - L'opacité des flèches est **proportionnelle à la fréquence d'occurence** du bigramme (haute fréquence/ opacité élevée, basse fréquence/opacité réduite). - Les noeuds sont plus larges et colorés en bleu clair. Les centres retrouvés sont: - **plant:** "plant community", "plant growth", "plant communities", "native plant". - **soil:** "mineral soil", "soil microbial" + "microbial biomass" (forte association), "soil organisms". -**earthworm:** "invasion", "species", "activity". D'autres bigrammes notables sont: "forest floor", "exotic earthworm", "organic matter", "leaf litter", "northern hardwood" et "lumbricus terretris". *Grâce à ce nouveau graphe dont les liens sont orientés, on est maintenant certain de l'ordre dans lequel les mots composant les bigrammes apparaissaient.*

```
library(igraph)
library(ggraph)
library(egg)
library(ggpubr)
library(ggplot2)

bigram_graph1 <- bigram_counts1 %>%
  filter(n > 10) %>%
  graph_from_data_frame()

network1 <- ggraph(bigram_graph1, layout = "fr") +
  geom_edge_link(aes(edge_alpha = n), show.legend = FALSE,
                 arrow = a, end_cap = circle(.07, 'inches')) +
  geom_node_point(color = "lightblue", size = 5) +
  geom_node_text(aes(label = name), vjust = 1, hjust = 1, size = 5) +
  labs(title = "MA1")

bigram_graph2 <- bigram_counts2 %>%
```

```
  filter(n > 10) %>%
  graph_from_data_frame()

network2 <- ggraph(bigram_graph2, layout = "fr") +
  geom_edge_link(aes(edge_alpha = n), show.legend = FALSE,
                 arrow = a, end_cap = circle(.07, 'inches')) +
  geom_node_point(color = "lightblue", size = 5) +
  geom_node_text(aes(label = name), vjust = 1, hjust = 1, size = 5) +
  labs(title = "MA2")

bigram_graph3 <- bigram_counts3 %>%
  filter(n > 10) %>%
  graph_from_data_frame()

network3 <- ggraph(bigram_graph3, layout = "fr") +
  geom_edge_link(aes(edge_alpha = n), show.legend = FALSE,
                 arrow = a, end_cap = circle(.07, 'inches')) +
  geom_node_point(color = "lightblue", size = 5) +
  geom_node_text(aes(label = name), vjust = 1, hjust = 1, size = 5) +
  labs(title = "MA3")

bigram_graph4 <- bigram_counts4 %>%
  filter(n > 10) %>%
  graph_from_data_frame()

network4 <- ggraph(bigram_graph4, layout = "fr") +
  geom_edge_link(aes(edge_alpha = n), show.legend = FALSE,
                 arrow = a, end_cap = circle(.07, 'inches')) +
  geom_node_point(color = "lightblue", size = 5) +
  geom_node_text(aes(label = name), vjust = 1, hjust = 1, size = 5) +
  labs(title = "MA4")
```

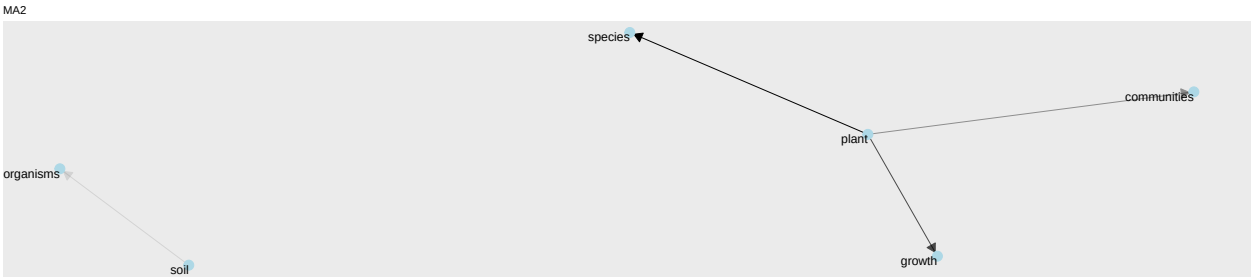
```
library(patchwork)
networks <- network1 + network2 + network3 + network4 + plot_layout(nrow = 4)
```

```
network1
```



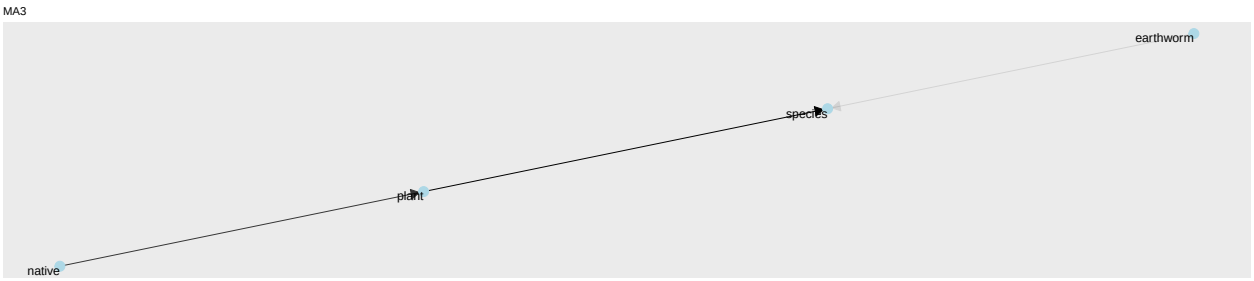

Les centres retrouvées sont: - **earthworm :** “invasion”, “species”. D’autres bigrammes notables sont: “mineral soil” –> “soil microbial” –> “microbial biomass”. Ou encore: “exotic earthworms”, “organic matter”, “forest floor”, “leaf litter”, “northern hardwood”, “lumbricus terrestris”.

```
network2
```



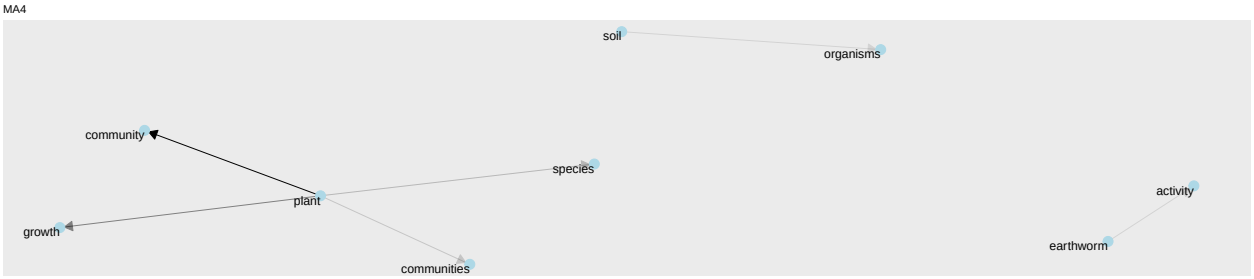

Les centres retrouvées sont: - **plant :** “species”, “growth”, “communities”. “soil organism” est aussi un bigramme fréquent.

```
network3
```




Aucun centre n’est retrouvé pour cette métaanalyse. On peut par contre retrouver une chaîne de mots reliés entre eux: “native plants” –> “plant species” –> “earthworm species”.

```
network4
```




Les centres retrouvées sont: - **plant :** “species”, “growth”, “community/communities”. “soil organism” et “earthworm activity” sont aussi des bigrammes fréquents.

En conclusion, on remarque que les centres retrouvés pour l’ensemble des quatres MA (“earthworm”, “plant”, “soil”), sont souvent retrouvés dans les MA individuellement (1- “earthworm”, 2- “plant”). Le cas de “soil” est remarquable, car il est un centre global, sans pour autant être un centre local dans aucune des quatre MA.