Bibliométrie

Eric Marcon

25 octobre 2018

Résumé

Utilisation de Google Scholar et de Scopus avec R pour analyser les publications d'une structure ou d'un auteur.

Table des matières

1 Google Scholar

Le package *scholar* permet d'accéder à l'API de Google Scholar. L'objectif est d'analyser la production d'un auteur (ou d'une structure) disposant d'un identifiant, donc d'une page, Google Scholar.

Le paramètre de base est l'identifiant de l'auteur :

```
AuthorID <- "4iLBmbUAAAAJ" # Eric Marcon
# AuthorID <- "8IqZyDUAAAAJ" # UMR EcoFoG
```

La vignette du package fournit la majorité du code utile.

```
vignette(topic = "scholar", package = "scholar")
```

1.1 Information sur l'auteur

La fonction `get_profile` retourne une liste avec les informations sur l'auteur.

```
library("scholar")
get_profile(AuthorID)
```

```
## $id
## [1] "4iLBmbUAAAAJ"
##
## $name
## [1] "Eric Marcon"
##
## $affiliation
## [1] "UMR EcoFoG, AgroParisTech"
##
## $total_cites
## [1] 1079
##
## $h_index
## [1] 15
##
## $i10_index
## [1] 18
##
## $fields
## [1] "verified email at ecofog.gf - homepage"
##
## $homepage
## [1] "http://www.ecofog.gf/spip.php?article16"
##
## $coauthors
##  [1] "Puech Florence"
##  [2] "Bruno Hérault"
##  [3] "Gabriel Lang"
##  [4] "Baraloto Christopher"
##  [5] "Sabrina Coste"
##  [6] "Heidy Schimann"
##  [7] "Céline Leroy"
```

```
##  [8] "Jerome Chave"
##  [9] "Lilian Blanc"
## [10] "Sandrine Pavoine"
## [11] "Zhiyi Zhang"
## [12] "Vivien Rossi"
## [13] "Ivan Scotti"
## [14] "Céline Born"
## [15] "François Morneau"
## [16] "Cecile Richard-Hansen"
## [17] "Guitet"
## [18] "Carlo Ricotta"
## [19] "Michael Grabchak"
## [20] "S. T. Buckland"
```

1.2 Liste des publications

La fonction `get_publications` retourne un dataframe contenant toutes les publications. Les colonnes contiennent le titre, la liste des auteurs (séparés par des virgules), le nom du journal, la pagination (sous la forme *Volume (numéro), pages*), le nombre de citations et les années correspondantes (sous la forme de vecteurs), et deux identifiants internes de la publication (`cid` et `pubid`).

```
Publications <- get_publications(AuthorID)
colnames(Publications)
```

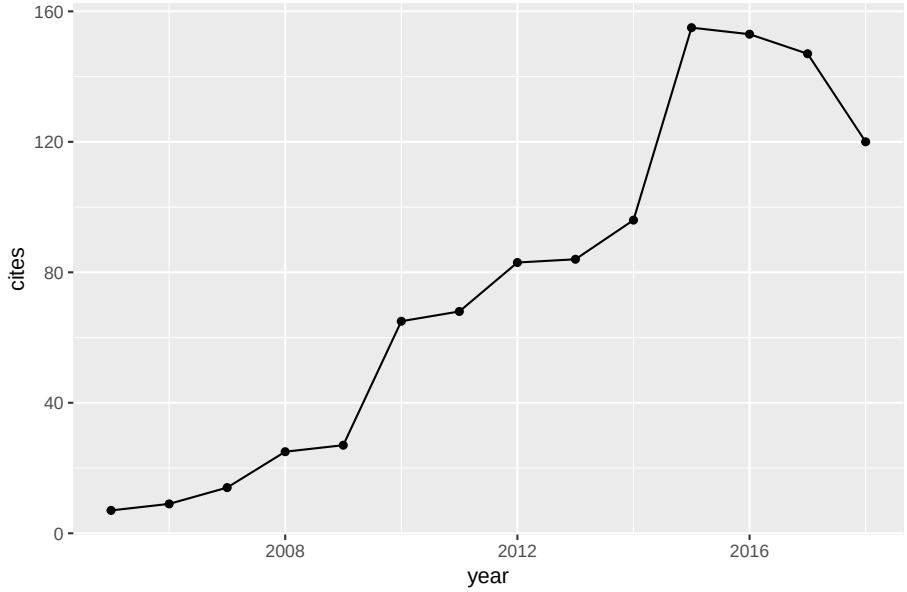
```
## [1] "title"   "author"  "journal" "number"
## [5] "cites"   "year"    "cid"     "pubid"
```

1.3 Citations par année

Evolution du nombre de citations d'un auteur :

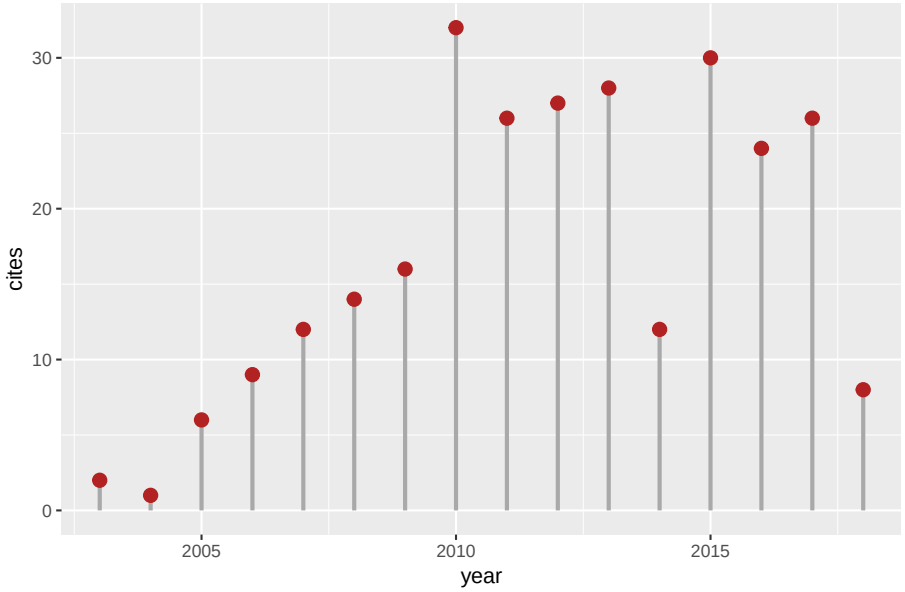
```
library("ggplot2")

get_citation_history(AuthorID) %>%
  ggplot(aes(x = year, y = cites)) +
    geom_line() +
    geom_point() +
    labs(caption= format(Sys.time(), "%Y-%m-%d %H:%M:%S %Z"))
```



Suivi d'un article en particulier (le plus cité : les articles sont classés par ordre décroissant du nombre de citations) :

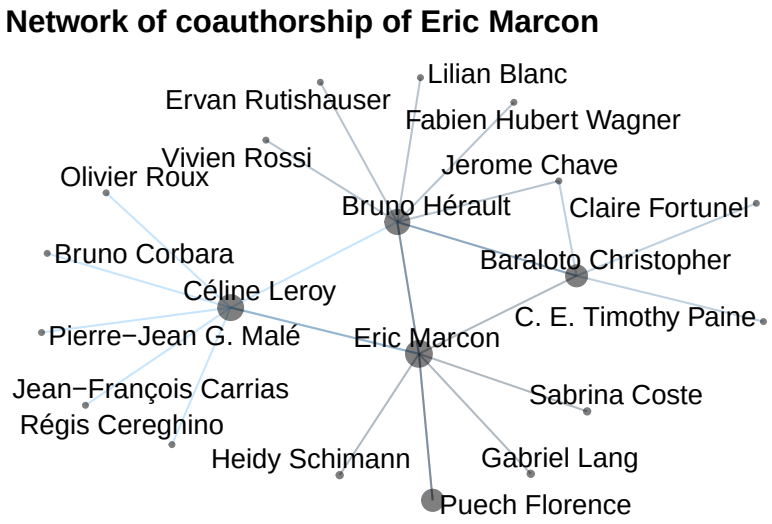
```
NumArticle <- 1
Reference <- with(Publications[NumArticle, ],
  paste(author, " (", year, ") ", journal, ". ", number, sep=""))
get_article_cite_history(AuthorID, Publications$pubid[NumArticle]) %>%
  ggplot(aes(year, cites)) +
    geom_segment(aes(xend = year, yend = 0), size=1, color='darkgrey') +
    geom_point(size=3, color='firebrick') +
    labs(caption = Reference)
```



1.4 Réseau d'auteurs

`get_coauthors` retourne un dataframe contenant les coauteurs déclarés par l'auteur sur sa page et leurs coauteurs. La profondeur `n_deep` du graphe permet d'augmenter le nombre de niveaux de coauteurs mais ne peut pas être mise à 0 pour obtenir seulement les coauteurs directs. Les valeurs par défaut sont 5 coauteurs et une profondeur de 1.

```
get_coauthors(AuthorID, n_coauthors = 7, n_deep=1) %>%
  plot_coauthors
```

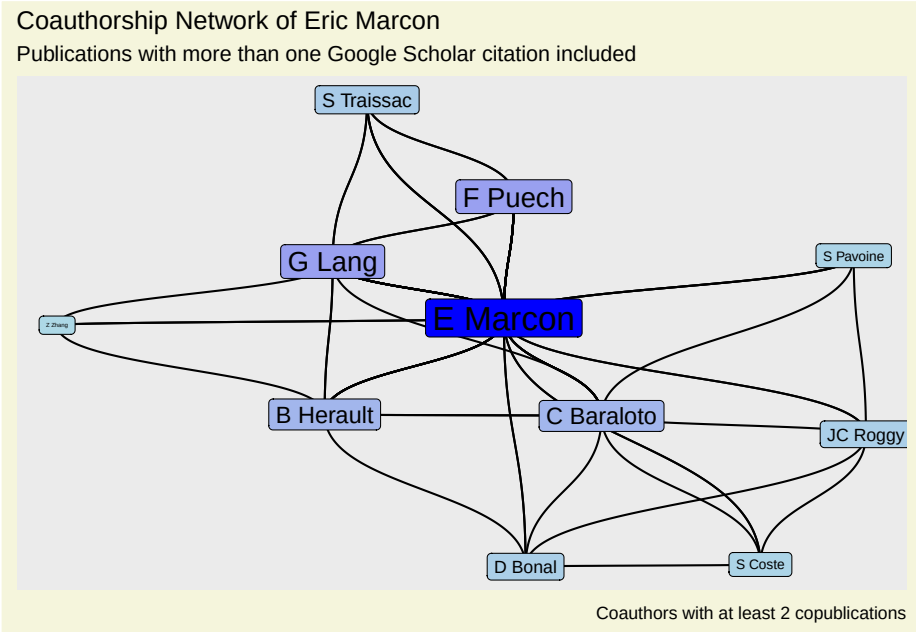


Les coateurs réels, définis par le nombre de publications écrites en commun, est à rechercher dans le tableau des publications.

```
# Paramètres
MinCopublications <- 2
MaxCoauteurs <- 100

library("magrittr")
# Vecteur des coauteurs de publications, sans accents
Publications %>%
  mutate(AuthorsASCII=iconv(author, from="UTF-8", to="ASCII//TRANSLIT")) %$%
  AuthorsASCII ->
  AuthorsASCII
# Auteurs uniques
AuthorsASCII %>%
  paste(collapse=", ") %>%
  str_split(pattern=", ") %>%
  unlist %>%
  unique ->
  UniqueAuthors
# Elimination de ... (= et al.)
UniqueAuthors <- UniqueAuthors[UniqueAuthors != "..."]
# Matrice d'autorat: une ligne par articles, auteurs en colonnes, valeurs logiques
PaperAuthoredBy <- sapply(UniqueAuthors, function(Author) str_detect(AuthorsASCII, Author))
# Filtrage des auteurs
tibble(Author=UniqueAuthors, NbPapers=colSums(PaperAuthoredBy)) %>%
  filter(NbPapers >= MinCopublications) %>%
  arrange(desc(NbPapers)) %>%
  slice(1:MaxCoauteurs) ->
  NbPapersPerAuthor
# Recalcul de la matrice d'autorat réduite
PaperAuthoredBy <- sapply(NbPapersPerAuthor$Author,
                          function(Author) str_detect(AuthorsASCII, Author))
# Matrice d'adjacence
adjacencyMatrix <- t(PaperAuthoredBy) %*% PaperAuthoredBy
# Graphe d'adjacence
# (https://paulvanderlaken.com/2017/10/31/network-visualization-with-igraph-and-ggraph/)
```

```
library("igraph")
g <- graph.adjacency(adjacencyMatrix, mode = "undirected", diag = FALSE)
V(g)$Degree <- degree(g, mode = 'in') # Nombre de liens
V(g)$Name <- NbPapersPerAuthor$Author # Etiquettes des noeuds
# Figure
library("ggraph")
ggraph(g, layout = "auto") +
  geom_edge_diagonal(alpha = 1, label_colour = "blue") +
  geom_node_label(aes(label = Name, size = log(Degree), fill = Degree)) +
  scale_fill_gradient(high = "blue", low = "lightblue") +
  theme(plot.background = element_rect(fill = "beige"),
    panel.border = element_blank(),
    panel.grid = element_blank(),
    legend.position = "none",
    axis.text = element_blank(),
    axis.title = element_blank(),
    axis.ticks = element_blank()) +
  labs(title = paste("Coauthorship Network of", get_profile(AuthorID)$name),
       subtitle = "Publications with more than one Google Scholar citation included",
       caption = paste("Coauthors with at least", MinCopublications, "copublications"))
```



Nombres de publications :

```
knitr::kable(NbPapersPerAuthor, caption="Nombre de documents par auteur",
             longtable = TRUE, booktabs = TRUE) %>%
  kableExtra::kable_styling(bootstrap_options = "striped")
```

TABLE 1: Nombre de documents par auteur

Author	NbPapers

E Marcon	43
F Puech	13
G Lang	9
B Herault	6
C Baraloto	5
S Traissac	3
S Pavoine	3
S Coste	2
D Bonal	2
JC Roggy	2
Z Zhang	2

2 Scopus et Web of Science

Le package *bibliometrix* permet d'exploiter les données des bases de données commerciales majeures.

La vignette du package décrit l'ensemble de ses possibilités.

```
vignette(topic = "bibliometrix-vignette", package = "bibliometrix")
```

2.1 Lecture des données

Voir la première partie de la vignette. Sur le site de Scopus (utlisé en exemple), sélectionner les références utiles et les exporter dans un fichier Bibtex. L'export doit contenir tous les champs, y compris le résumé et les documents cités.

Le fichier est ensuite lu et converti :

```
library(bibliometrix)
# Fichier de données au format bibtex, exporté de Scopus
D <- readFiles("scopus.bib")
# Conversion en dataframe
M <- convert2df(D, dbsource="scopus", format="bibtex")
```

```
##
## Converting your scopus collection into a bibliographic dataframe
##
## Articles extracted   100
```

```
## Articles extracted   200
## Articles extracted   300
## Articles extracted   400
## Articles extracted   500
## Articles extracted   600
## Articles extracted   690
## Done!
##
##
## Generating affiliation field tag AU_UN from C1:  Done!
```

2.2 Analyses basiques

Les analyses de base sont retournées par la fonction `biblioAnalysis`. Le résultat est un objet de type `bibliometrix`. Les méthodes `summary` et `plot` de cet objet sont malheureusement mal écrites et renvoient tous les résultats à l'écran sans possibilité de les afficher séparément. Pour cette raison, le code ci-dessous reproduit le code original, séparément pour chaque figure.

```
k <- 5 # Nombre d'auteurs à afficher
BA <- biblioAnalysis(M)
summary(BA, k)
```

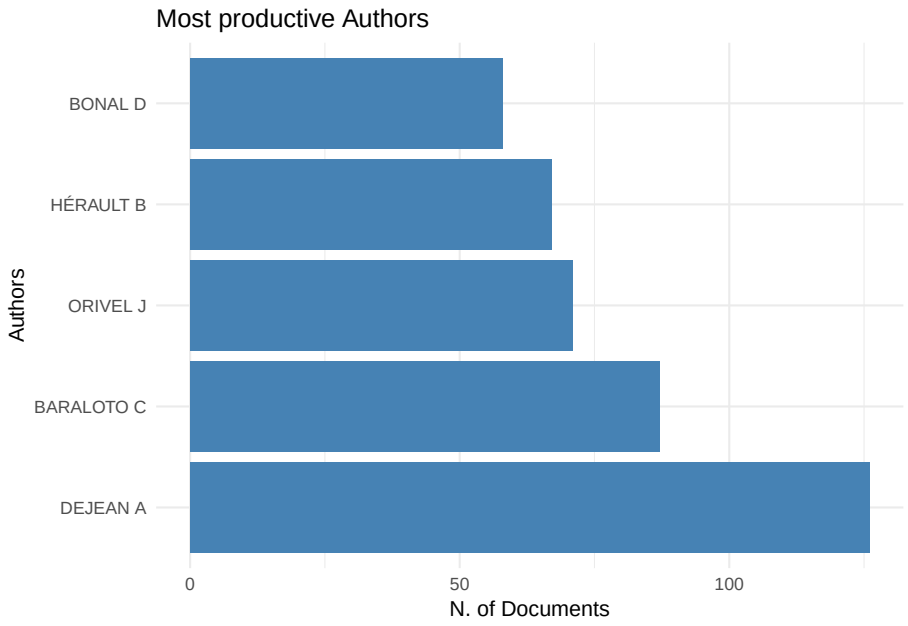
```
##
##
## Main Information about data
##
##  Documents                               690
##  Sources (Journals, Books, etc.)         262
##  Keywords Plus (ID)                      5234
##  Author's Keywords (DE)                  2197
##  Period                                  2001 - 2018
##  Average citations per documents         26.42
##
##  Authors                                 3326
##  Author Appearances                      7054
##  Authors of single-authored documents    2
##  Authors of multi-authored documents     3324
##  Single-authored documents               7
##
##  Documents per Author                    0.207
##  Authors per Document                    4.82
##  Co-Authors per Documents                10.2
##  Collaboration Index                     4.87
##
##  Document types
##  ARTICLE               632
```

```
##  BOOK CHAPTER          1
##  CONFERENCE PAPER      17
##  EDITORIAL             1
##  ERRATUM               2
##  LETTER                3
##  NOTE                  4
##  REVIEW                30
##
##
## Annual Scientific Production
##
##  Year    Articles
##     2001        1
##     2002        4
##     2003       27
##     2004       18
##     2005       16
##     2006       21
##     2007       31
##     2008       26
##     2009       50
##     2010       73
##     2011       63
##     2012       64
##     2013       51
##     2014       49
##     2015       67
##     2016       58
##     2017       46
##     2018       25
##
## Annual Percentage Growth Rate 20.84586
##
##
## Most Productive Authors
##
##    Authors        Articles Authors        Articles Fractionalized
## 1     DEJEAN A         126    DEJEAN A                      23.67
## 2     BARALOTO C        87    BARALOTO C                    14.16
## 3     ORIVEL J          71    HÉRAULT B                     11.91
## 4     HÉRAULT B         67    ORIVEL J                      11.38
## 5     BONAL D           58    LEROY C                        9.29
## 6     LEROY C           56    CORBARA B                      9.09
## 7     CORBARA B         55    BONAL D                        7.57
## 8     CÉRÉGHINO R       45    CÉRÉGHINO R                    7.11
## 9     CHAVE J           38    CLAIR B                        6.84
```

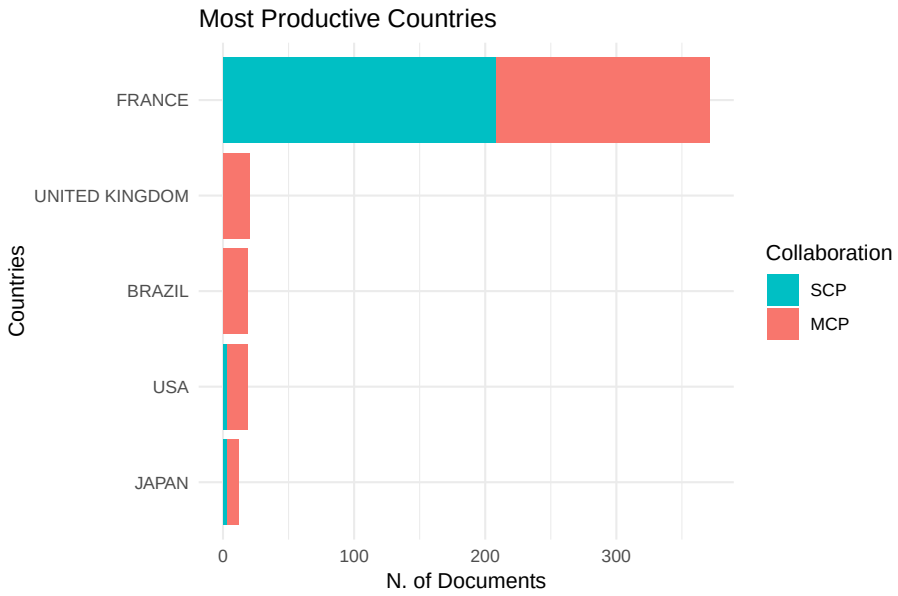
```
## 10    STIEN D           38    SCOTTI I                       6.51
##
##
## Top manuscripts per citations
##
##                            Paper           TC TCperYear
## 1  PHILLIPS OL, 2009, SCIENCE            816      90.7
## 2  LUYSSAERT S, 2007, GLOBAL CHANGE BIOL 502      45.6
## 3  TERSTEEGE H, 2013, SCIENCE            355      71.0
## 4  DÍAZ S, 2016, NATURE                  277     138.5
## 5  MOUILLOT D, 2013, PLOS BIOL           244      48.8
## 6  BRIENEN RJW, 2015, NATURE             219      73.0
## 7  PHILLIPS OL, 2010, NEW PHYTOL         218      27.2
## 8  HARDY OJ, 2006, MOL ECOL              214      17.8
## 9  EVA HD, 2004, GLOBAL CHANGE BIOL      193      13.8
## 10 BASSET Y, 2012, SCIENCE               184      30.7
##
##
## Most Productive Countries (of corresponding authors)
##
##         Country   Articles    Freq SCP MCP MCP_Ratio
## 1  FRANCE              371 0.72745 208 163     0.439
## 2  UNITED KINGDOM       20 0.03922   0  20     1.000
## 3  BRAZIL               19 0.03725   0  19     1.000
## 4  USA                  19 0.03725   3  16     0.842
## 5  JAPAN                12 0.02353   3   9     0.750
## 6  GERMANY              10 0.01961   0  10     1.000
## 7  CANADA                9 0.01765   1   8     0.889
## 8  BELGIUM               7 0.01373   0   7     1.000
## 9  AUSTRALIA             5 0.00980   0   5     1.000
## 10 CAMEROON              4 0.00784   0   4     1.000
##
##
## SCP: Single Country Publications
##
## MCP: Multiple Country Publications
##
##
## Total Citations per Country
##
##      Country      Total Citations Average Article Citations
## 1  FRANCE                   7643                       20.6
## 2  UNITED KINGDOM           1981                       99.0
## 3  BELGIUM                   838                      119.7
## 4  USA                       764                       40.2
## 5  BRAZIL                    399                       21.0
```

```
## 6  NETHERLANDS                358                      179.0
## 7  ITALY                      346                       86.5
## 8  GERMANY                    309                       30.9
## 9  JAPAN                      281                       23.4
## 10 ARGENTINA                  277                      277.0
##
##
## Most Relevant Sources
##
##                    Sources        Articles
## 1  ANNALS OF FOREST SCIENCE             35
## 2  PLOS ONE                             35
## 3  COMPTES RENDUS - BIOLOGIES           17
## 4  BIOTROPICA                           15
## 5  FOREST ECOLOGY AND MANAGEMENT        12
## 6  GLOBAL CHANGE BIOLOGY                12
## 7  JOURNAL OF ECOLOGY                   11
## 8  JOURNAL OF ETHNOPHARMACOLOGY         11
## 9  NATURWISSENSCHAFTEN                  11
## 10 ANNALS OF BOTANY                     10
##
##
## Most Relevant Keywords
##
##    Author Keywords (DE)      Articles Keywords-Plus (ID)     Articles
## 1        FRENCH GUIANA              82          FRENCH GUIANA      174
## 2        TROPICAL FOREST            24          ARTICLE            138
## 3        TROPICAL RAINFOREST        19          ANT                128
## 4        TENSION WOOD               16          NONHUMAN           102
## 5        AMAZONIA                   14          PHYSIOLOGY          97
## 6        FUNCTIONAL DIVERSITY       14          RAINFOREST          92
## 7        FUNCTIONAL TRAITS          13          ECOSYSTEM           89
## 8        AMAZON                     12          BIODIVERSITY        86
## 9        TROPICAL RAIN FOREST       12          TREES               71
## 10       BIODIVERSITY               11          ANTS                69
```

```r
# plot(BA) renvoie tous les graphiques à la suite.
# Le code ci-dessous (copié de plot.bibliometrix) les produit séparément.
x <- BA ; xx<- as.data.frame(x$Authors[1:k])
ggplot(data = xx, aes(x = xx$AU, y = xx$Freq)) + geom_bar(stat = "identity",
       fill = "steelblue") + labs(title = "Most productive Authors",
       x = "Authors") + labs(y = "N. of Documents") + theme_minimal() +
       coord_flip()
```

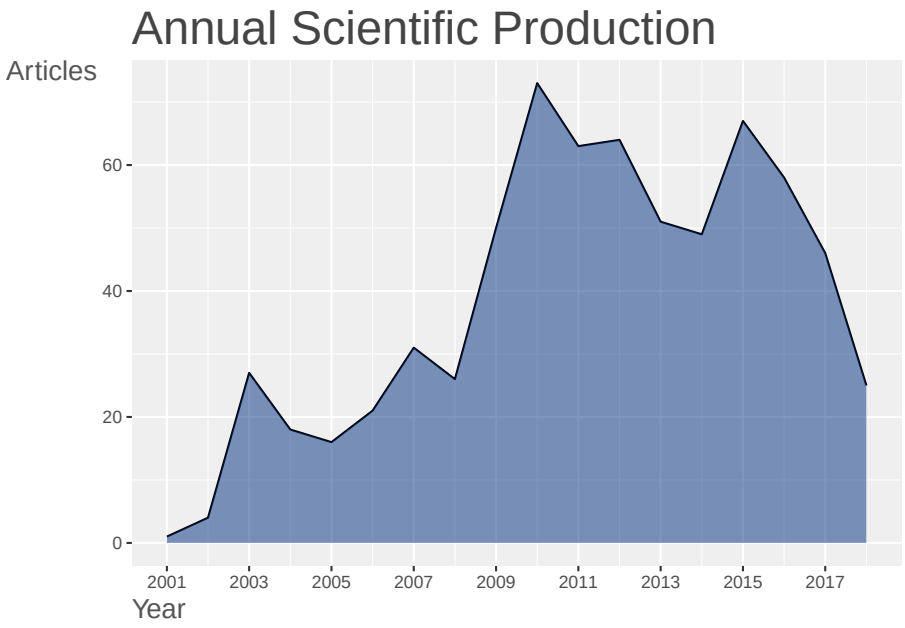



```
if (!is.na(x$CountryCollaboration[1, 1])) {
  xx = x$CountryCollaboration[1:k,]
  xx = xx[order(-(xx$SCP + xx$MCP)),]
  xx1 = cbind(xx[, 1:2], rep("SCP", k))
  names(xx1) = c("Country", "Freq", "Collaboration")
  xx2 = cbind(xx[, c(1, 3)], rep("MCP", k))
  names(xx2) = c("Country", "Freq", "Collaboration")
  xx = rbind(xx2, xx1)
  xx$Country = factor(xx$Country, levels = xx$Country[1:dim(xx2)[1]])
  ggplot(data = xx, aes(x = xx$Country, y = xx$Freq, fill = xx$Collaboration)) +
    geom_bar(stat = "identity") +
    scale_x_discrete(limits = rev(levels(xx$Country))) +
    scale_fill_discrete(name = "Collaboration", breaks = c("SCP", "MCP")) +
    labs( title = "Most Productive Countries",
          x = "Countries",
          y = "N. of Documents",
          caption = "SCP: Single Country Publications, MCP: Multiple Country Publications") +
    theme_minimal() +
    theme(plot.caption = element_text(size = 9, hjust = 0.5, color = "blue", face = "italic")) +
    coord_flip()
}
```



SCP: Single Country Publications, MCP: Multiple Country Publications

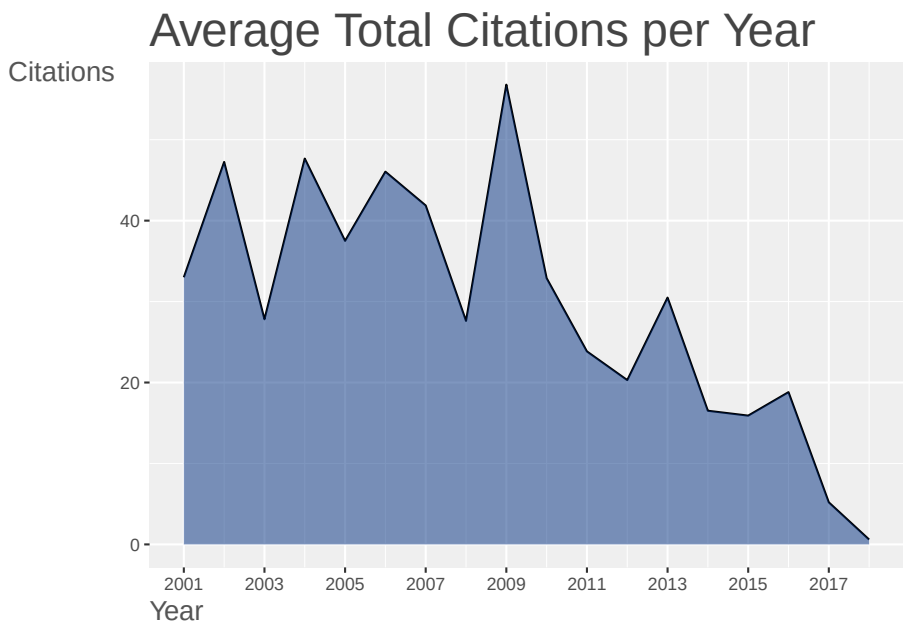
```
Tab = table(x$Years)
YY = setdiff(seq(min(x$Years), max(x$Years)), names(Tab))
Y = data.frame(Year = as.numeric(c(names(Tab), YY)),
               Freq = c(as.numeric(Tab), rep(0, length(YY))))
Y = Y[order(Y$Year),]
names(Y) = c("Year", "Freq")
ggplot(Y, aes(x = Y$Year, y = Y$Freq)) +
  geom_line() +
  geom_area(fill = "#002F80", alpha = 0.5) +
  labs(x = "Year", y = "Articles", title = "Annual Scientific Production") +
  scale_x_continuous(breaks = (Y$Year[seq(1, length(Y$Year), by = 2)])) +
  theme(text = element_text(color = "#444444"),
    panel.background = element_rect(fill = "#EFEFEF"),
    panel.grid.minor = element_line(color = "#FFFFFF"),
    panel.grid.major = element_line(color = "#FFFFFF"),
    plot.title = element_text(size = 24),
    axis.title = element_text(size = 14, color = "#555555"),
    axis.title.y = element_text(vjust = 1, angle = 0),
    axis.title.x = element_text(hjust = 0))
```




```
Table2 = NA
if (!(x$DB %in% c("COCHRANE", "PUBMED"))) {
  Table2 = aggregate(x$TotalCitation, by = list(x$Years), length)
  Table2$xx = aggregate(x$TotalCitation, by = list(x$Years), mean)$x
  Table2$Annual = NA
  d = date()
  d = as.numeric(substring(d, nchar(d) - 3, nchar(d)))
  Table2$Years = d - Table2$Group.1
  Table2$Annual = Table2$xx / Table2$Years
  names(Table2) = c("Year", "N", "MeanTCperArt", "MeanTCperYear", "CitableYears")
  YY = setdiff(seq(min(x$Years), max(x$Years)), Table2$Year)
  if (length(YY > 0)) {
    YY = data.frame(YY, 0, 0, 0, 0)
    names(YY) = c("Year", "N", "MeanTCperArt", "MeanTCperYear", "CitableYears")
    Table2 = rbind(Table2, YY)
    Table2 = Table2[order(Table2$Year),]
    row.names(Table2) = Table2$Year
 }
  ggplot(Table2, aes(x = Table2$Year, y = Table2$MeanTCperYear)) +
    geom_line() +
    geom_area(fill = "#002F80", alpha = 0.5) +
    labs(x = "Year", y = "Citations", title = "Average Article Citations per Year") +
    scale_x_continuous(breaks = (Table2$Year[seq(1, length(Table2$Year), by = 2)])) +
    theme(
      text = element_text(color = "#444444"),
      panel.background = element_rect(fill = "#EFEFEF"),
      panel.grid.minor = element_line(color = "#FFFFFF"),
      panel.grid.major = element_line(color = "#FFFFFF"),
      plot.title = element_text(size = 24),
      axis.title = element_text(size = 14,
      color = "#555555"),
      axis.title.y = element_text(vjust = 1,
      angle = 0),
      axis.title.x = element_text(hjust = 0)
      )

  ggplot(Table2, aes(x = Table2$Year, y = Table2$MeanTCperArt)) +
```

```
    geom_line() +
    geom_area(fill = "#002F80", alpha = 0.5) +
    labs(x = "Year", y = "Citations", title = "Average Total Citations per Year") +
    scale_x_continuous(breaks = (Table2$Year[seq(1, length(Table2$Year), by = 2)])) +
    theme(
      text = element_text(color = "#444444"),
      panel.background = element_rect(fill = "#EFEFEF"),
      panel.grid.minor = element_line(color = "#FFFFFF"),
      panel.grid.major = element_line(color = "#FFFFFF"),
      plot.title = element_text(size = 24),
      axis.title = element_text(size = 14, color = "#555555"),
      axis.title.y = element_text(vjust = 1, angle = 0),
      axis.title.x = element_text(hjust = 0, angle = 0)
    )
}
```



2.3 Documents et auteurs cités

Les documents les plus cités par la base bibliographique sont retournés par la commande `citations`, par article ou par auteur.

```
CAR <- citations(M, field = "article")
CAR$Cited[1:5] %>%
  as_tibble %>%
  rename(Article = CR, Citations=n) %>%
  knitr::kable(caption =
       "Citations les plus fréquentes par les documents de la base de données bibliographique",
       longtable = TRUE, booktabs = TRUE) %>%
  kableExtra::kable_styling(full_width=TRUE, bootstrap_options = "striped")
```

TABLE 2: Citations les plus fréquentes par les documents de la base de données bibliographique

Article	Citations
KRAFT, N.J.B., VALENCIA, R., ACKERLY, D.D., FUNCTIONAL TRAITS AND NICHE-BASED TREE COMMUNITY ASSEMBLY IN AN AMAZONIAN FOREST (2008) SCIENCE, 322, PP. 580-582	14
CHAVE, J., COOMES, D., JANSEN, S., LEWIS, S.L., SWENSON, N.G., ZANNE, A.E., TOWARDS A WORLDWIDE WOOD ECONOMICS SPECTRUM (2009) ECOLOGY LETTERS, 12, PP. 351-366	13
CÉRÉGHINO, R., LEROY, C., DEJEAN, A., CORBARA, B., ANTS MEDIATE THE STRUCTURE OF PHYTOTELM COMMUNITIES IN AN ANT-GARDEN BROMELIAD (2010) ECOLOGY, 91, PP. 1549-1556	11
FINE, P.V.A., MESONES, I., COLEY, P.D., HERBIVORES PROMOTE HABITAT SPECIALIZATION BY TREES IN AMAZONIAN FORESTS (2004) SCIENCE, 305, PP. 663-665	11
KITAJIMA, K., RELATIVE IMPORTANCE OF PHOTOSYNTHETIC TRAITS AND ALLOCATION PATTERNS AS CORRELATES OF SEEDLING SHADE TOLERANCE OF 13 TROPICAL TREES (1994) OECOLOGIA, 98, PP. 419-428	11

Les auteurs les plus cités :

```
CAU <- citations(M, field = "author")
CAU$Cited[1:5] %>%
  as_tibble %>%
  rename(Auteur=CR, Citations=n) %>%
  knitr::kable(
    caption="Auteurs les plus cités par les documents de la base de données bibliographique",
    longtable = TRUE, booktabs = TRUE) %>%
  kableExtra::kable_styling(bootstrap_options = "striped")
```

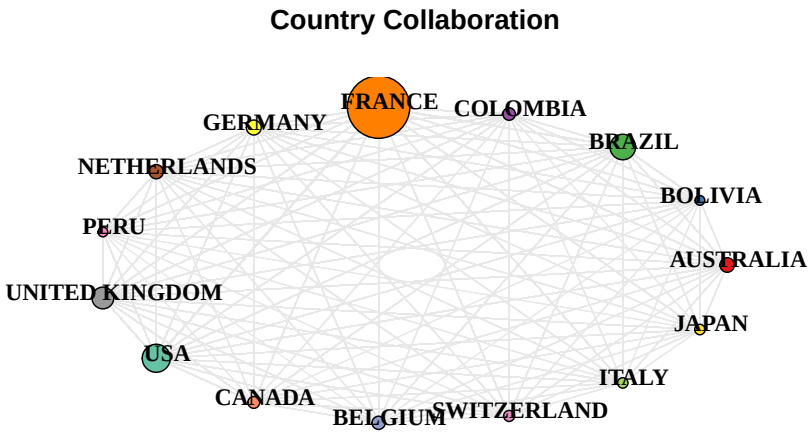
TABLE 3: Auteurs les plus cités par les documents de la base de données bibliographique

Auteur	Citations
DEJEAN, A	725
ORIVEL, J	394
BARALOTO, C	380
BONAL, D	357
PHILLIPS, O.L	335

2.4 Collaborations

Un réseau de collaboration entre les pays des auteurs est retourné par la fonction `biblioNetwork`.

```
NbCountries <- 15
# Create a country collaboration network
mAU_CO <- metaTagExtraction(M, Field = "AU_CO", sep = ";")
NetMatrix <- biblioNetwork(mAU_CO, analysis = "collaboration",
    network = "countries", sep = ";")
# Plot the network
netC <- networkPlot(NetMatrix, n = NbCountries, Title = "Country Collaboration",
    type = "circle", size = TRUE, remove.multiple = FALSE)
```



Le réseau des auteurs est obtenu de la même façon.

```
NbAuthors <- 15
# Réseau d'auteurs
AuthorNet <- biblioNetwork(M, analysis = "collaboration",
    network = "authors", sep = ";")
netA <- networkPlot(AuthorNet, n = NbAuthors, Title = "Author Collaboration",
    type = "circle", size = TRUE, remove.multiple = FALSE)
```

3 Analyse des résumés

Les résumés des publications se trouvent dans la colonne `AB` de la base importée par *bibliometrix*. Ils sont en Anglais.

3.1 Corpus

Le package `tm` permet de constituer un corpus.

```
library("tm")
M$AB %>%
  VectorSource %>%
  VCorpus %>%
  tm_map(PlainTextDocument) %>%
  tm_map(content_transformer(tolower)) ->
  MonCorpus
```

La fonction `tm_map` permet d'appliquer une fonction quelconque à chaque élément du corpus, c'est-à-dire à chaque résumé. Les fonctions standard, n'appartenant pas au package `tm`, doivent être appliquées par l'intermédiaire de la fonction `content_transformer` pour ne pas dégrader la structure du corpus : dans le code précédent, la fonction `tolower` est appliquée à chaque résumé pour le passer en minuscules, alors que la création de corpus est en majuscules.

3.2 Nettoyage du corpus

Des mots sémantiquement identiques ont plusieurs formes. Le traitement le plus rigoureux consiste à les réduire à leur radical mais le résultat n'est pas très lisible. La fonction `stemDocument` permet de le faire : il suffit de l'utiliser à la place de `PlainTextDocument` dans le code ci-dessus. Un bon compromis consiste à supprimer les formes plurielles, par une fonction ad-hoc : ce sera fait plus tard.

Les déterminants, conjonctions, etc. sont les mots les plus fréquents mais n'ont pas d'intérêt pour l'analyse. La fonction `removeWords` permet de retirer une liste de mots. `stopwords` fournit la liste de ces mots dans une langue au choix. `removeNumbers` retire les nombres comme *one*, *two*, etc. et la fonction `removePunctuation` retire la ponctuation.

```
MonCorpus %<>% tm_map(removeWords, stopwords("english")) %>%
  tm_map(removeNumbers) %>%
  tm_map(removePunctuation)
```

Une liste de mots complémentaire est nécessaire pour supprimer des mots inutiles mais fréquents. Elle peut être complétée de façon itérative pour retirer des mots parasites du résultat final.

```
ExtraWords <- c("use", "used", "using", "results",
    "may", "across", "high", "higher", "low", "show",
    "showed", "study", "studies", "studied", "however",
    "can", "our", "based", "including", "within", "total",
    "among", "found", "due", "also", "well", "strong",
    "large", "important", "first", "known")
MonCorpus %<>% tm_map(removeWords, ExtraWords)
```

3.3 Mots du corpus

L'objectif est de transformer le corpus en un vecteur d'abondance des mots utilisés. `TermDocumentMatrix` crée un objet spécifique au package *tm* qui pose des problèmes de traitement. Cet objet est transformé en un vecteur d'abondances.

```
TDM <- TermDocumentMatrix(MonCorpus, control = list(minWordLength = 3))
AbdMots <- sort(rowSums(as.matrix(TDM)), decreasing = TRUE)
```

Le vecteur de mots contient des formes singulières et plurielles. Elles peuvent être regroupées selon un modèle simple : si un mot existe avec et sans *s* ou *es* final, la forme singulière est sans *s* ou *es*. Des pluriels particuliers peuvent être ajoutés selon les besoins.

```
# Adapté de https://github.com/mkfs/misc-text-mining/blob/master/R/wordcloud.R
aggregate_plurals <- function (v) {
    aggr_fn <- function(v, singular, plural) {
        if (! is.na(v[plural])) {
            v[singular] <- v[singular] + v[plural]
            v <- v[-which(names(v) == plural)]
        }
        return(v)
    }
```

```
    for (n in names(v)) {
        n_pl <- paste(n, 's', sep='')
        v <- aggr_fn(v, n, n_pl)
        n_pl <- paste(n, 'es', sep='')
        v <- aggr_fn(v, n, n_pl)
        # cas particuliers
        if (endsWith(n, "y")) {
          n <-  substr(n, 1, nchar(n)-1)
          n_pl <- paste(n, 'ies', sep='')
          }
        if (n == "genus") {
          n_pl <- "genera"
          v <- aggr_fn(v, n, n_pl)
        }
    }
    return(v)
}

AbdMots %<>% aggregate_plurals
```

3.4 Nuage de mots

Le résultat final est un nuage de mots.

```
library("wordcloud")
df <- data.frame(word = names(AbdMots), freq = AbdMots)
wordcloud(df$word, df$freq, max.words = 100, random.order = FALSE,
    rot.per = 0.35, use.r.layout = FALSE, colors = brewer.pal(8,
        "Dark2"))
```