Evolution des thématiques dans les publications de Cybergéo

10 décembre 2015

library(plyr)  
library(stringr)  
library(lubridate)  
library(dplyr)  
library(koRpus)  
library(stringi)  
library(tidyr)  
library(doParallel)  
library(topicmodels)  
library(slam)  
library(tm)  
library(ggplot2)  
library(pdc)

nbCores <- ceiling(detectCores()/2)  
registerDoParallel(cores=nbCores)  
textPath <- "../../Data/cybergeo-articles-1996-2015/texts"  
articlesMetaDataFile <- "cybergeo.csv"

models <- c("LDA")  
nbrReplications <- 1  
nbrFolds <- 10  
k.list <- c(2, 5, 10, 20, seq(25, 36, 2), 50, 100, 200)

k0 <- 20

nbrMotsParThematique <- 20

ts.lag.max <- 8 # Pour le graphique d'entropie  
nbr.dimension <- 3 # Évolution déduite sur une fenêtre de 3 temps  
ts.lag <- 1 # Décalage temporel permis

## Introduction

#-- Lecture des méta-données du corpus  
articles <- read.table(  
 articlesMetaDataFile,   
 sep = ";",   
 quote = "",   
 comment.char = "",   
 header = TRUE  
) %>%   
 tbl\_df() %>%  
 rename(  
 titre = title\_en,   
 auteurs = authors  
 ) %>%  
 select(id, date, langue, auteurs, titre) %>%  
 mutate(  
 langue = tolower(langue),  
 date = as.Date(date)  
 ) %>%  
 arrange(date, auteurs)

[Cybergeo](https://cybergeo.revues.org/) est une revue scientifique européenne de géographie. Depuis sa création en 1996, 1241 documents ont été publiés. L'accès aux textes intégraux de ce large *corpus* sur ce temps long nous permet d'appréhender l'évolution des thématiques abordées par les Géographes : Quelles thématiques sont récurrentes chez les auteurs de Cybergeo ? Quelle est l'évolution de ces thématiques au cours du temps ?

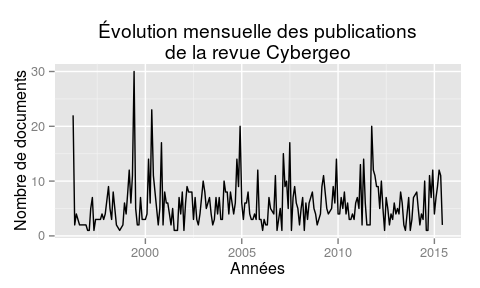
## Méthode

Afin de répondre à ce questionnement, nous proposons tout d'abord d'identifier des thématiques dans le *corpus*, puis d'identifier des typologies d'évolution de ces thématiques au cours du temps.

Afin d'identifier nos thématiques, nous proposons d'utiliser l'Allocation de Dirichlet latente, ou LDA (Blei et al., 2003). La méthode suppose que les mots sont des observations collectées dans les documents du corpus, que chaque document est un mélange d'un faible nombre de thématiques et que chaque mot est attribuable à l'une des thématiques du document. La méthode nécessite au préalable de collecter les mots de chaque document et de les normaliser, puis d'estimer le nombre de thématiques le plus adapté à notre corpus. Chaque mot est ensuite tiré aléatoirement pour constituer des thématiques. Il en résulte une liste des mots les plus probables par thématique et une probabilité de chaque thématique par document.

Une fois les thématiques identifiées, nous réaliserons une classification des séries temporelles d'évolution des thématiques (cf. graphique ci-dessous pour la série temporelle des publications de Cybergeo). Nous utiliserons la méthode dite de *Permutation distribution clustering* (Brandmaier, 2011). Il s'agit de réduire les séries temporelles de présence des thématiques à une évolution relative (augmentation, diminution) en tenant compte d'un possible décalage temporel entre séries temporelles. Il en résulte une classification des séries temporelles selon leur similarité.

ts.articles <- articles %>%  
 select(date) %>%  
 mutate(date = as.Date(cut(date, breaks = "1 month"))) %>%  
 group\_by(date) %>%  
 summarise(counts = n())  
ggplot(  
 ts.articles,   
 aes(x = date, y = counts)  
 ) +   
 geom\_line() +   
 scale\_x\_date()+  
 labs(  
 y = "Nombre de documents",   
 x = "Années",   
 title = "Évolution mensuelle des publications\nde la revue Cybergeo"  
 )



## Présentation du corpus

Ce corpus contient des documents de différents types : des renvois, des bibliographies, des comptes-rendus de lecture, des articles courts et longs. Cette typologie se retrouve dans la distibution des nombres de mots par document (cf. figure ci-après). Ces documents sont écrits dans différentes langues, la langue française étant majoritaire (923 documents, cf. graphique ci-après).

files <- articles %>%  
 filter(langue == "fr") %>%  
 select(id) %>%  
 mutate(  
 name = paste(id, "\_text.txt", sep = ""),  
 path = paste(textPath, name, sep = "/"),  
 size = file.size(path)  
 ) %>%  
 filter(size > 400)

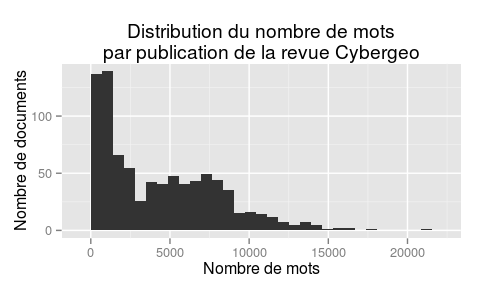
tag.corpus <- foreach(id = files$id) %dopar% {  
 file <- files$path[files$id == id]  
 tagged.text <- treetag(  
 file = file,  
 treetagger = "manual",  
 lang = "fr",  
 encoding = "UTF-8",  
 TT.tknz = FALSE,  
 TT.options = list(path = "~/treetagger", preset = "fr-utf8"),  
 stopwords = tm::stopwords("fr"),  
 stemmer = SnowballC::wordStem  
 )  
 return(list(id = id, tagger = tagged.text))  
}

d0 <- lapply(tag.corpus, function(x) {  
 id <- x$id  
 taggedText(x$tagger) %>%  
 tbl\_df() %>%  
 # Récupération des noms communs et propros, ainsi que des verbes  
 filter(str\_sub(tag,1,3) %in% c("NAM","NOM","VER")) %>%  
 # Si aucun lemme n'est défini (cas des noms propres), alors prendre le mot initial  
 mutate(lemmes = ifelse(lemma == "<unknown>", token, lemma)) %>%  
 # Suppression des nombres et des points  
 mutate(lemmes = gsub("[0-9.]", "", lemmes, perl = TRUE)) %>%  
 # Suppression des termes de moins de 3 caractères  
 filter(nchar(lemmes) >= 3) %>%   
 # Calcul des fréquences de termes  
 group\_by(lemmes) %>%  
 summarise(counts = n()) %>%  
 mutate(DOCID = id)  
}) %>%  
 ldply %>%  
 tbl\_df %>%  
 arrange(lemmes)

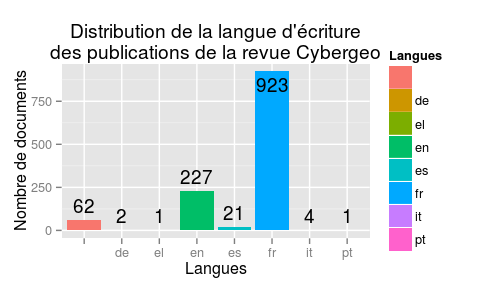
dict0 <- unique(d0$lemmes)  
d1 <- d0 %>%  
 mutate(lemidx = match(lemmes, dict0) - 1)  
l5 <- d1 %>%   
 select(lemidx) %>%  
 group\_by(lemidx) %>%  
 summarise(nbDoc = n()) %>%  
 # Suppression des termes apparaissant dans moins de 5 documents  
 # Car je m'intéresse à l'évolution des thèmes, pas à l'apparition  
 # de thèmes qui ne trouveront pas d'écho ailleurs  
 filter(nbDoc >= 5) %>%  
 select(-nbDoc)  
d2 <- l5 %>%  
 left\_join(d1)  
  
#-- Construction de la matrice document-termes  
vocabulaire <- unique(d2$lemmes)  
d3 <- d2 %>%  
 mutate(termidx = match(lemmes, vocabulaire) - 1)  
doc <- lapply(files$id, function(id){  
 f0 <- d3 %>%  
 filter(DOCID == id) %>%  
 select(-lemmes, -DOCID) %>%  
 arrange(termidx)  
 rbind(f0$termidx, f0$counts)  
})  
dtm <- ldaformat2dtm(doc, vocabulaire)  
  
#-- Calcul du poids TF-IDF  
dim(dtm)  
summary(col\_sums(dtm))  
term\_tfidf <- tapply(dtm$v/row\_sums(dtm)[dtm$i], dtm$j, mean) \* log2(nDocs(dtm)/col\_sums(dtm > 0))  
summary(term\_tfidf)  
dtm <- dtm[, term\_tfidf >= median(term\_tfidf)]  
dtm <- dtm[row\_sums(dtm) > 0,]  
summary(col\_sums(dtm))  
dim(dtm)

mots <- ldply(lapply(tag.corpus, function(x) { x$tagger@TT.res }))  
nbrTokens <- format(nrow(mots), big.mark = " ")  
nbrChars <- ldply(lapply(tag.corpus, function(x) { x$tagger@desc$words }))  
nbrMots <- format(sum(nbrChars$V1), big.mark = " ")  
nbrLemmes <- d0 %>% nrow %>% format(big.mark = " ")  
nbrLemmes0 <- dict0 %>% length %>% format(big.mark = " ")  
nbrLemmes1 <- vocabulaire %>% length %>% format(big.mark = " ")  
nbrLemmes2 <- dim(dtm)[2]

ggplot(nbrChars, aes(x = V1)) +   
 geom\_histogram() +  
 labs(  
 x = "Nombre de mots",  
 y = "Nombre de documents",  
 title = "Distribution du nombre de mots\npar publication de la revue Cybergeo"  
 )



t <- table(articles$langue)  
langues <- data\_frame(  
 lang = factor(names(t)),   
 count = as.numeric(t)  
) %>%  
 arrange(count) %>%  
 mutate(  
 m = max(count),  
 v = ifelse(count < m/2, -0.5, 1.5)  
 )  
ggplot(langues, aes(x = lang, y = count, fill = lang, label = count, vjust = v)) +   
 geom\_bar(stat = "identity") +  
 #coord\_flip() +  
 labs(  
 y = "Nombre de documents",   
 x = "Langues",   
 title = "Distribution de la langue d'écriture\ndes publications de la revue Cybergeo",  
 fill = "Langues"  
 ) +  
 geom\_text()



## Normalisation des textes

Nous traiterons les documents en langue française. Les bibliographies et les renvois, ne constituant pas un texte en soi, ne seront pas traités. Notre corpus se réduit maintenant à 854 documents composés de 3 687 717 mots.

Afin de normaliser ces textes, vu le nombre de documents, nous devons faire appel à un traitement automatique de la langue. Nous utiliserons pour cela un programme pour, tout d'abord, identifier les éléments du discours, puis pour les réduire à leur forme canonique (lemmatisation). Tout d'abord, nous identifions les éléments du discours et leur nature : abréviation, adjectif, adverbe, déterminant, interjection, conjonction, nom commun, nom propre, pronom, préposition, ponctuation, symbole, verbe, temps de conjugaison. Notre corpus se compose ainsi 4 329 637 éléments de discours. Ensuite, nous réduisons ces éléments de discours à leur lemmes, c'est à dire leur forme canonique : les noms au pluriel seront mis au singulier, les verbes conjugués seront mis à l'infinitif, *etc*. Nous retiendrons pour notre analyse que les éléments les plus signifiants, soit les noms communs, les noms propres et les verbes ; chacun possédant au moins trois lettres. À ce stade, notre corpus se compose de 522 123 lemmes, dont 59 394 sont uniques. Comme nous nous intéressons à l'évolution de thématiques au sein de plusieurs articles, nous ne retiendrons que les mots apparaissant au moins dans cinq documents. Notre corpus se réduit ainsi à 10 242 lemmes uniques. Pour finaliser notre processus, nous attribuons un poids à chacun de ces lemmes. Nous utilisons pour cela la méthode de calcul TF-IDF qui permet de cerner les termes les plus significatifs dans un corpus (Sparck Jones, 1972). Nous ne retiendrons que la moitié des lemmes les plus significatifs, soit 5121 lemmes uniques.

## Identification des thématiques

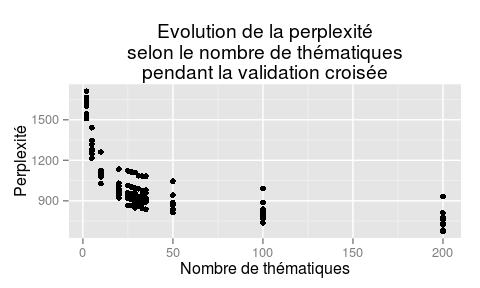
Pour estimer le nombre de thématiques approprié à notre *corpus*, nous allons tester l'adéquation de différentes valeurs. Nous testerons la capacité à rendre compte des thématiques d'un corpus de test constitué par 10% des documents d'après les thématiques identifiées sur le reste du corpus. Nous effectuerons cette validation croisée 10 fois, permettant ainsi de tester l'ensemble du *corpus*. Deux indicateurs - la perplexité et l'entropie - nous permettront de sélectionner le nombre de thématiques adéquat.

Nous constatons dans les graphiques ci-dessous que :

* la perplexité minimale est atteinte avec un nombre de thématiques croissant. Les thématiques traitées dans Cybergeo sont donc très diverses.
* l'entropie maximale est toutefois atteinte aux environs de 20 à 35 thématiques, ce qui signifie que les thématiques identifiées sont assez bien réparties entre les documents.

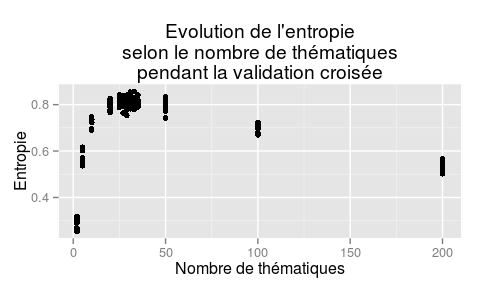
simulation <- function(id) {  
 model <- prog$model[prog$id == id]  
 k <- prog$k[prog$id == id]  
 rep <- prog$rep[prog$id == id]  
 fold <- prog$fold[prog$id == id]  
 FILE <- paste("VEM\_", k, "\_", fold, ".rda", sep = "")  
 training <- LDA(texts[folding != fold,], k = k)  
 testing <- LDA(texts[folding == fold,], model = training, control = list(estimate.beta = FALSE))  
 d <- data.frame(  
 id = id,  
 rep = rep,  
 model = model,  
 k = k,  
 fold = fold,  
 alpha = training@alpha,  
 perplexity = perplexity(testing, texts[folding == fold,], use\_theta = FALSE),  
 entropie = mean(apply(posterior(training)$topics, 1, function(z) - sum(z \* log(z))))  
 )  
 return(d)  
}  
  
texts <- dtm  
folding <- sample(1:nbrFolds, nrow(texts), replace = TRUE)  
prog <- expand.grid(  
 rep = 1:nbrReplications,  
 model = models,  
 k = k.list,   
 fold = 1:nbrFolds  
)  
prog$id <- 1:nrow(prog)  
  
result <- rbind.fill(foreach(id = prog$id) %dopar% simulation(id))

# Perplexity  
ggplot(result, aes(k, perplexity)) +  
 geom\_point() +  
 labs(  
 x = "Nombre de thématiques",  
 y = "Perplexité",  
 title = "Evolution de la perplexité\nselon le nombre de thématiques\npendant la validation croisée"  
 )

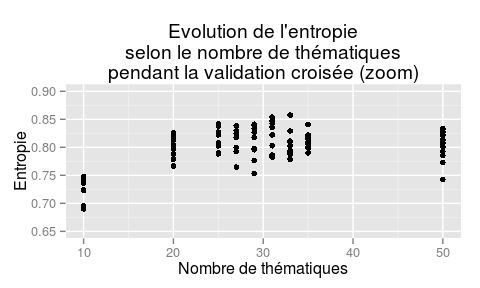


# Alpha  
# The lower α the higher is the percentage of documents which are assigned to one single topic with a high probability  
# Cela implique que les documents consistent un nombre limité de thèmes  
ggplot(result, aes(k, alpha)) +  
 geom\_point() + #+ xlim(0,100)  
 facet\_wrap(~ rep)

# Entropie  
# Higher values indicate that the topic distributions are more evenly spread over the topics.  
g.ent <- ggplot(result, aes(k, entropie)) +  
 geom\_point() +  
 labs(  
 x = "Nombre de thématiques",  
 y = "Entropie"  
 )  
g.ent + labs(title = "Evolution de l'entropie\nselon le nombre de thématiques\npendant la validation croisée")



# Entropie  
# Higher values indicate that the topic distributions are more evenly spread over the topics.  
g.ent + xlim(10, 50) + ylim(0.65, 0.9) + labs(title = "Evolution de l'entropie\nselon le nombre de thématiques\npendant la validation croisée (zoom)")



res.mean <- result %>%  
 group\_by(k) %>%  
 summarise(  
 alpha = mean(alpha),  
 perplexity = mean(perplexity),  
 entropie = mean(entropie)  
 )  
ggplot(res.mean, aes(k, entropie)) +  
 geom\_point()

## Cas avec 20 thématiques

### Analyse des thématiques identifiées

Nous présentons ci-dessous les 20 mots, les plus significatifs, qui ont été alloués à chacun des 20 thèmes ; puis un graphique montrant la probabilité d'allocation d'une thématique par document (un document est représenté par un trait fin vertical).

# Importance de chaque thème (+ CAH)  
# Distribution des probabilités de chaque thème dans les documents (ou liste des 3 premiers titres par thème)  
# Importance de chaque thème dans chaque documents  
# Evolution de chaque théme dans le temps (probabilité)  
# Clustering de séries temporelles  
# Sankey diagram  
# Graphes des thèmes :  
# - un noeud = 1 document représenté par le mot le plus représentatif du thème)  
# - un lien (un thème commun)  
# - Ajouter peu à peu les thèmes majeurs, secondaires, etc.  
  
texts <- dtm  
a <- result %>%  
 filter(k == k0) %>%  
 group\_by(model) %>%  
 summarise(alpha = mean(alpha))  
alpha <- a$alpha  
model <- LDA(texts, k = k0, control = list(alpha = alpha))

# Trois premiers thèmes dans chaque document  
#topics(model, 3)  
# Termes de chaque thème  
thematique <- terms(model, nbrMotsParThematique) %>%  
 as.data.frame() %>%  
 tbl\_df() %>%  
 gather("theme", "mot")  
themes.uniques <- thematique$theme %>% unique  
thematique <- thematique %>%  
 mutate(Thèmes = match(theme, themes.uniques)) %>%  
 select(-theme) %>%  
 group\_by(Thèmes) %>%  
 summarise(Mots = paste(mot, collapse = ", "))  
kable(thematique)

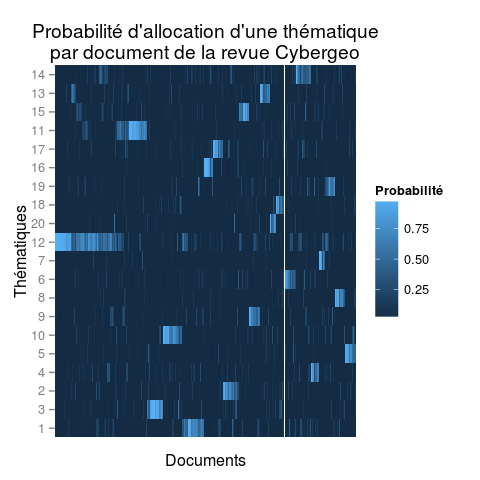
|  |  |
| --- | --- |
| Thèmes | Mots |
| 1 | quartier, ville, logement, ménage, rue, Paris, bâtiment, arrondissement, gentrification, banlieue, centralité, exclusion, parc, réhabilitation, gang, femme, recensement, immeuble, carte, mobilité |
| 2 | déplacement, mobilité, densité, accessibilité, indicateur, simulation, modélisation, scénario, agent, aire, vitesse, simuler, trajet, ville, ménage, indice, étalement, espèce, consommation, agglomération |
| 3 | image, sol, occupation, surface, végétation, carte, résolution, pixel, paysage, feu, forêt, classification, eau, savane, canal, bâti, télédétection, satellite, indice, densité |
| 4 | aménagement, gouvernance, urbanisme, durabilité, document, participation, État, gestion, Loire, barrage, indicateur, consultation, union, citoyen, coopération, ambiance, ouvrage, Lyon, coordination, commission |
| 5 | risque, vulnérabilité, aléa, inondation, ville, eau, catastrophe, gestion, événement, scénario, carte, éruption, prévention, résilience, crue, accident, séisme, écosystème, submersion, tremblement |
| 6 | entreprise, santé, soin, migrant, capital, meubler, client, location, compagnie, ville, aéroport, Internet, usine, revenu, inégalité, télétravail, technologie, province, gestion, investisseur |
| 7 | the, and, The, for, are, that, with, from, géomorphologie, area, this, haver, Paris, not, was, one, their, which, also, between |
| 8 | indice, agent, graphe, sommet, montagne, mangrove, indicateur, simulation, cimetière, cellule, couleur, ségrégation, Humboldt, voisinage, longueur, contraste, surface, canal, nœud, boîte |
| 9 | ville, femme, rue, nuit, paysage, Moscou, plage, jardin, photographie, mer, image, voyage, voyageur, hôtel, monument, Saint-Pétersbourg, art, fête, bâtiment, Russie |
| 10 | eau, exploitation, gestion, agriculteur, agriculture, parcelle, espèce, biodiversité, forêt, exploitant, zonage, parc, commune, trame, sol, bassin, débit, élevage, aménagement, côte |
| 11 | carte, cartographie, revue, http, image, atlas, www, logiciel, auteur, science, format, émission, publication, page, géographe, diaspora, cartographe, discipline, géographie, sémiologie |
| 12 | géographie, géographe, auteur, ouvrage, science, texte, chapitre, livre, discipline, mot, thèse, Géographie, Paris, lecteur, mondialisation, pensée, Reclus, monsieur, page, colloque |
| 13 | île, élève, géographie, enseignement, identité, image, Université, joueur, Japon, enseignant, archipel, magazine, texte, professeur, Paris, Arendt, Maurice, lecteur, enseigner, mot |
| 14 | ville, agglomération, métropole, aire, urbanisation, bureau, aménagement, Paris, entreprise, municipalité, décroissance, densité, urbanisme, recensement, banlieue, étalement, logement, métropolisation, États-Unis, Londres |
| 15 | frontière, Chine, État, États, Brésil, Asie, nation, peuple, Rio, coopération, São, village, Russie, guerre, Etat, identité, république, Québec, colonisation, États-Unis |
| 16 | vote, carte, parti, élection, cluster, score, indice, droit|droite, candidat, liste, Italie, participation, département, CPNT, canton, octobre, électorat, scrutin, circonscription, édition |
| 17 | commune, pôle, département, surface, carte, université, étudiant, recensement, bassin, agglomération, Paris, prévalence, revenu, travailleur, ville, personnel, migration, rang, INSEE, euro |
| 18 | village, saison, pluie, station, vallée, précipitation, vitesse, pêche, carte, paysan, température, pêcheur, eau, fleuve, tronçon, forage, mare, événement, troupeau, remontée |
| 19 | paysage, patrimoine, image, touriste, Afrique, tourisme, vallée, musée, lande, plante, montagne, photo, visiteur, classement, photographie, front, forêt, guide|guides, apartheid, parc |
| 20 | port, mer, vent, station, brise, Tunis, température, côte, trafic, Sfax, surface, indice, bus, littoral, chaleur, surcote, vin, vitesse, ville, dépression |

# Graphique de la répartition de la probabilité de chaque thème dans chaque document  
clustering.order <- function(df) {  
 d <- dist(p, method = "euclidean")  
 h <- hclust(d, method = "ward.D2")  
 return(h)  
}  
  
p <- posterior(model)$topics %>%  
 as.data.frame()  
row.names(p) <- files$id  
doc.clust <- clustering.order(p)  
doc.order <- doc.clust$order  
  
p <- posterior(model)$topics %>%  
 as.data.frame() %>%  
 tbl\_df()  
p$id <- files$id  
p <- p %>%  
 gather("topic", "p", 1:k0) %>%  
 spread(id, p)  
row.names(p) <- p$topic  
topic.clust <- clustering.order(p)  
topic.order <- topic.clust$order  
  
p <- posterior(model)$topics %>%  
 as.data.frame() %>%  
 tbl\_df()  
p$id <- files$id  
p <- p %>%  
 gather("topic", "p", 1:k0)  
p1 <- p  
p$topic <- factor(p$topic, levels = (1:k0)[topic.order])  
p$id <- factor(files$id, levels = files$id[doc.order])

plot(doc.clust)

plot(topic.clust)

ggplot(p, aes(x = id, y = topic, fill = p)) +  
 geom\_tile() +  
 labs(  
 x = "Documents",   
 y = "Thématiques",   
 title = "Probabilité d'allocation d'une thématique\npar document de la revue Cybergeo",  
 fill = "Probabilité"  
 ) +  
 theme(  
 axis.text.x = element\_blank(),   
 axis.ticks.x = element\_blank()  
 )



# Taille des groupes  
table(groupe)  
g1 <- groupe[groupe == 2] %>% names  
articles %>%  
 filter(id %in% g1) %>%  
 select(id, titre)  
  
p$id <- files$id  
p <- p %>%  
 gather("topic", "p", 1:k0)  
  
articles %>% filter(id == 6085)  
p %>% filter(id == 6085) %>% arrange(desc(p)) %>% mutate(p = round(p,digits = 2))  
terms(model, 20)[,32]  
terms(model, 20)[,7]  
terms(model, 20)[,21]  
  
groupe <- cutree(h, k = 43)  
# Taille des groupes  
table(groupe)  
g1 <- groupe[groupe == 2] %>% names  
articles %>%  
 filter(id %in% g1) %>%  
 select(id, titre)

### Analyse des trajectoires des thématiques

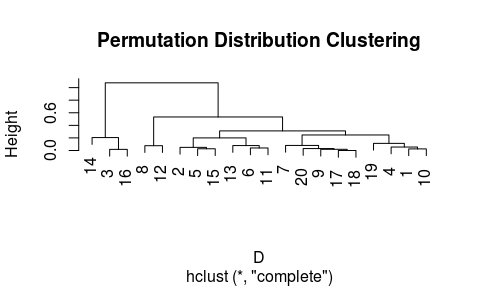
freq <- p1 %>%  
 left\_join(articles) %>%  
 mutate(annee = str\_sub(date, 1, 4)) %>%  
 rename(topic\_id = topic) %>%  
 group\_by(annee, topic\_id) %>%  
 summarise(counts = sum(p)) %>%  
 arrange(annee, topic\_id)

m <- freq %>%  
 mutate(counts = as.numeric(counts)) %>%  
 spread(topic\_id, counts) %>%  
 select(-annee) %>%  
 as.matrix()  
mine <- entropyHeuristic(m, t.max = ts.lag.max)  
#clust <- pdclust(m, m = nbr.dimension, t = ts.lag)  
clust <- pdclust(m)

Nous recherchons maintenant à classer les séries temporelles selon leur similarité. Pour cela, nous étudions les évolutions sur 3 années consécutives, en permettant un retard entre deux séries de 4 années. Un premier graphique montre la classification obtenue ; le deuxième l'évolution temporelle sur 20 ans des thématiques classées selon leurs similarités.

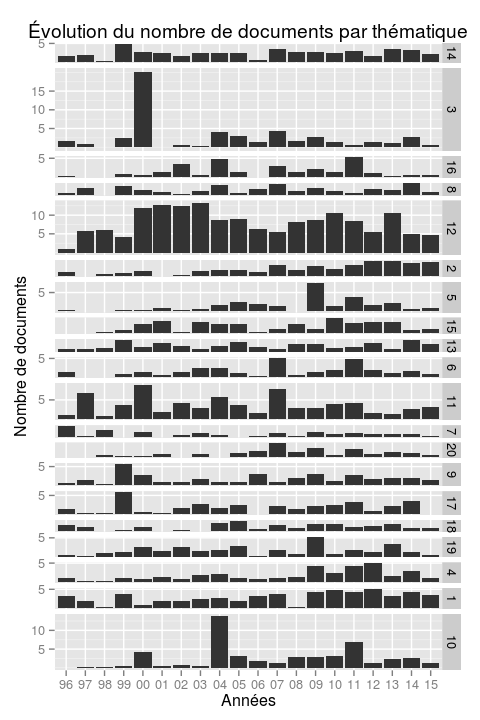
plot(mine)

plot(clust, labels = colnames(m))



plot(clust)

freq$topic\_id <- factor(freq$topic\_id, levels = colnames(m)[clust$order])  
ggplot(freq, aes(annee, counts)) +   
 geom\_bar(stat = "identity") +   
 facet\_grid(topic\_id ~ ., scales = "free\_y", space = "free\_y") +  
 labs(  
 x= "Années",  
 y = "Nombre de documents",  
 title = "Évolution du nombre de documents par thématique"  
 ) +  
 scale\_x\_discrete(labels=c("96", "97", "98", "99", "00", "01", "02", "03", "04", "05", "06", "07", "08", "09", "10", "11", "12", "13", "14", "15")) +  
 scale\_y\_continuous(breaks=c(5,10,15))



k0 <- 35  
a <- result %>%  
 filter(k == k0) %>%  
 group\_by(model) %>%  
 summarise(alpha = mean(alpha))  
alpha <- a$alpha  
model <- LDA(texts, k = k0, control = list(alpha = alpha))

## Cas avec 35 thématiques

### Analyse des thématiques identifiées

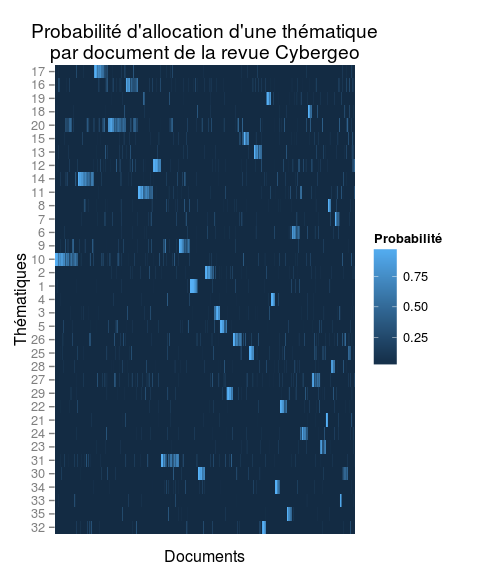
Nous présentons ci-dessous les 20 mots, les plus significatifs, qui ont été alloués à chacun des 35 thèmes ; puis un graphique montrant la probabilité d'allocation d'une thématique par document (un document est représenté par un trait fin vertical).

# Trois premiers thèmes dans chaque document  
#topics(model, 3)  
# Termes de chaque thème  
thematique <- terms(model, nbrMotsParThematique) %>%  
 as.data.frame() %>%  
 tbl\_df() %>%  
 gather("theme", "mot")  
themes.uniques <- thematique$theme %>% unique  
thematique <- thematique %>%  
 mutate(Thèmes = match(theme, themes.uniques)) %>%  
 select(-theme) %>%  
 group\_by(Thèmes) %>%  
 summarise(Mots = paste(mot, collapse = ", "))  
kable(thematique)

|  |  |
| --- | --- |
| Thèmes | Mots |
| 1 | vote, carte, parti, département, élection, droit|droite, candidat, contraste, score, participation, canton, CPNT, électorat, liste, scrutin, circonscription, commune, Front, électeur, ville |
| 2 | sol, occupation, carte, surface, densité, simulation, cellule, modélisation, écosystème, plantation, voisinage, simuler, forêt, matrice, probabilité, scénario, prédire, pixel, Homo, couplage |
| 3 | risque, vulnérabilité, aléa, gestion, catastrophe, inondation, carte, prévention, scénario, eau, accident, résilience, quartier, joueur, ville, séisme, Quito, endommagement, État, événement |
| 4 | côte, mer, vent, littoral, ambiance, magazine, surcote, indice, tempête, dépression, eau, espèce, guide|guides, consultation, submersion, notoriété, événement, bai|baie, Lion, jardin |
| 5 | agent, simulation, modélisation, feu, graphe, déplacement, scénario, simuler, espèce, bâtiment, événement, nœud, cellule, trajectoire, algorithme, voisinage, innovation, paysage, probabilité, module |
| 6 | commune, biodiversité, parc, trame, zonage, espèce, gestion, sol, bureau, indicateur, surface, document, aménagement, aire, occupation, classement, urbanisation, urbanisme, paysage, PLU |
| 7 | aménagement, urbanisme, ville, Loire, eau, île, gestion, Lyon, Montréal, Paris, fleuve, tsunami, Seine, ouvrage, décentralisation, lit, équipe, barrage, urbaniste, inondation |
| 8 | http, www, diaspora, org, Afrique, département, com, touriste, html, htm, maladie, octobre, Internet, revue, page, peste, web, député, and, province |
| 9 | géographie, Université, Paris, carte, voyage, île, géographe, colloque, Vidal, auteur, UMR, suivre|être, texte, planche, livre, science, Blache, récit, Géographie, lire |
| 10 | géographie, élève, géographe, enseignement, texte, science, identité, paysage, image, discipline, enseignant, professeur, pensée, auteur, photographie, mot, enseigner, thème, posture, photo |
| 11 | image, eau, surface, sol, résolution, végétation, pixel, classification, satellite, télédétection, bâti, canal, SPOT, gestion, habitation, réflectance, capteur, carte, estimation, Landsat |
| 12 | mobilité, ménage, logement, déplacement, quartier, ville, trajet, enquêter, accessibilité, recensement, banlieue, migration, domicile, bus, trajectoire, pôle, arrondissement, marche, femme, rue |
| 13 | ville, rue, quartier, nuit, mer, tourisme, plage, hôtel, photo, vendeur, bâtiment, touriste, monument, visiteur, bai|baie, paysage, démolition, eau, port, style |
| 14 | géographie, science, revue, université, géographe, discipline, étudiant, Géographie, mot, géomorphologie, Brésil, Paris, publication, liste, enseignement, institut, langue, Deffontaines, bulletin, auteur |
| 15 | gouvernance, aménagement, commune, durabilité, ZAC, quartier, État, indicateur, Tunis, personnel, participation, logement, coordination, citoyen, document, dépense, union, Paris, discontinuité, auteur |
| 16 | ville, quartier, patrimoine, aire, décroissance, urbanisme, image, métropole, Allemagne, Paris, Nantes, auteur, revalorisation, Bordeaux, marketing, Marseille, urbaine, urbanité, logement, ville-centre |
| 17 | ville, quartier, États-Unis, gentrification, gang, exclusion, logement, front, néolibéralisme, Chicago, Los, Angeles, New, apartheid, Paris, York, auteur, urbanisation, banlieue, carte |
| 18 | bâtiment, vote, meubler, location, logement, Paris, bloc, patrimoine, parc, édition, Wright, occupation, fête, Japonais, PECO, Russie, arrondissement, graphe, Pologne, ménage |
| 19 | eau, station, température, pluie, saison, débit, brise, bassin, mer, fig, vent, étiage, humidité, indice, précipitation, canal, ville, surface, vitesse, chaleur |
| 20 | ouvrage, auteur, chapitre, sommet, lecteur, livre, mangrove, texte, contribution, boîte, inégalité, page, science, centralité, grille, longueur, rue, estuaire, bibliographie, regretter |
| 21 | port, ville, village, surface, prévalence, trafic, santé, pêcheur, pêche, Dakar, rangée, marchandise, carte, tremblement, fleuve, grappe, Bamako, rayonnement, accessibilité, Italie |
| 22 | émission, vitesse, véhicule, consommation, gaz, carte, étudiant, commune, serre, bruit, ville, publicité, déchet, Saint-Pétersbourg, trafic, coefficient, université, gare, Moscou, confection |
| 23 | île, santé, soin, migrant, compagnie, archipel, mare, Maurice, province, transhumance, forage, centralité, noix, aéroport, accueil, hôpital, Air, Natura, tourisme, Tokyo |
| 24 | travailleur, Brésil, Rio, eau, commune, vallée, parcelle, Guyane, frontière, village, carte, bassin, São, exploitation, producteur, canne, étang, État, colonisation, Espagne |
| 25 | entreprise, pôle, innovation, ville, client, cluster, Paris, Internet, télétravail, capital, agglomération, firme, technologie, usine, aéroport, Londres, délocalisation, bassin, salaire, investisseur |
| 26 | frontière, États, État, nation, identité, coopération, union, Chine, Asie, république, Québec, Arendt, Etat, citoyen, réforme, équité, agenda, Allemagne, Ouzbékistan, peuple |
| 27 | ville, densité, agglomération, métropole, indicateur, aire, urbanisation, accessibilité, indice, étalement, recensement, centralité, éruption, corrélation, métropolisation, grille, clinique, commune, municipalité, surface |
| 28 | paysage, indice, montagne, vallée, cimetière, ségrégation, image, jardin, barrage, tronçon, oasis, guide|guides, touriste, versant, palmier, autocorrélation, compagnie, matrice, mot, égalité |
| 29 | exploitation, agriculteur, agriculture, forêt, gestion, village, exploitant, parcelle, sol, paysan, femme, eau, paysage, surface, photographie, savane, contrat, terroir, CTE, commune |
| 30 | géographe, mondialisation, Reclus, reclus, Chine, géographie, peuple, auteur, Asie, civilisation, ouvrage, Méditerranée, guerre, texte, voyage, voyageur, mot, nation, mer, livre |
| 31 | carte, cartographie, image, atlas, logiciel, couleur, géographe, auteur, format, lecteur, SIG, cartographe, sémiologie, dessin, document, graphique, publication, fichier, Bertin, photographie |
| 32 | géographie, montagne, Humboldt, géographe, thèse, Christaller, Gourou, ville, Allemagne, rang, Jean, carte, paysage, Etude, auteur, science, ouvrage, Pierre, Algérie, colonisation |
| 33 | musée, Japon, art, extraction, vallée, restauration, camion, artiste, anse, image, eau, travailleur, réhabilitation, sable, volcan, bloc, North, Manchester, écologie, visiteur |
| 34 | the, and, for, The, are, that, with, from, area, this, haver, octobre, not, was, one, their, Paris, which, Gabon, also |
| 35 | lande, espèce, vignoble, vin, parcelle, végétation, sol, marais, île, mer, forêt, paysage, revenu, Québec, précipitation, plante, ozone, eau, côte, vent |

# Graphique de la répartition de la probabilité de chaque thème dans chaque document  
clustering.order <- function(df) {  
 d <- dist(p, method = "euclidean")  
 h <- hclust(d, method = "ward.D2")  
 return(h)  
}  
  
p <- posterior(model)$topics %>%  
 as.data.frame()  
row.names(p) <- files$id  
doc.clust <- clustering.order(p)  
doc.order <- doc.clust$order  
  
p <- posterior(model)$topics %>%  
 as.data.frame() %>%  
 tbl\_df()  
p$id <- files$id  
p <- p %>%  
 gather("topic", "p", 1:k0) %>%  
 spread(id, p)  
row.names(p) <- p$topic  
topic.clust <- clustering.order(p)  
topic.order <- topic.clust$order  
  
p <- posterior(model)$topics %>%  
 as.data.frame() %>%  
 tbl\_df()  
p$id <- files$id  
p <- p %>%  
 gather("topic", "p", 1:k0)  
p1 <- p  
p$topic <- factor(p$topic, levels = (1:k0)[topic.order])  
p$id <- factor(files$id, levels = files$id[doc.order])

ggplot(p, aes(x = id, y = topic, fill = p)) +  
 geom\_tile() +  
 labs(  
 x = "Documents",   
 y = "Thématiques",   
 title = "Probabilité d'allocation d'une thématique\npar document de la revue Cybergeo",  
 fill = "Probabilité"  
 ) +  
 theme(  
 axis.text.x = element\_blank(),   
 axis.ticks.x = element\_blank()  
 )



### Analyse des trajectoires des thématiques

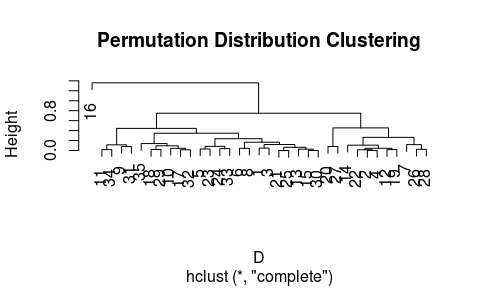
freq <- p1 %>%  
 left\_join(articles) %>%  
 mutate(annee = str\_sub(date, 1, 4)) %>%  
 rename(topic\_id = topic) %>%  
 group\_by(annee, topic\_id) %>%  
 summarise(counts = sum(p)) %>%  
 arrange(annee, topic\_id)

m <- freq %>%  
 mutate(counts = as.numeric(counts)) %>%  
 spread(topic\_id, counts) %>%  
 select(-annee) %>%  
 as.matrix()  
mine <- entropyHeuristic(m, t.max = ts.lag.max)  
#clust <- pdclust(m, m = nbr.dimension, t = ts.lag)  
clust <- pdclust(m)

Nous recherchons maintenant à classer les séries temporelles selon leur similarité. Pour cela, nous étudions les évolutions sur 3 années consécutives, en permettant un retard entre deux séries de 4 années. Un premier graphique montre la classification obtenue ; le deuxième l'évolution temporelle sur 20 ans des thématiques classées selon leurs similarités.

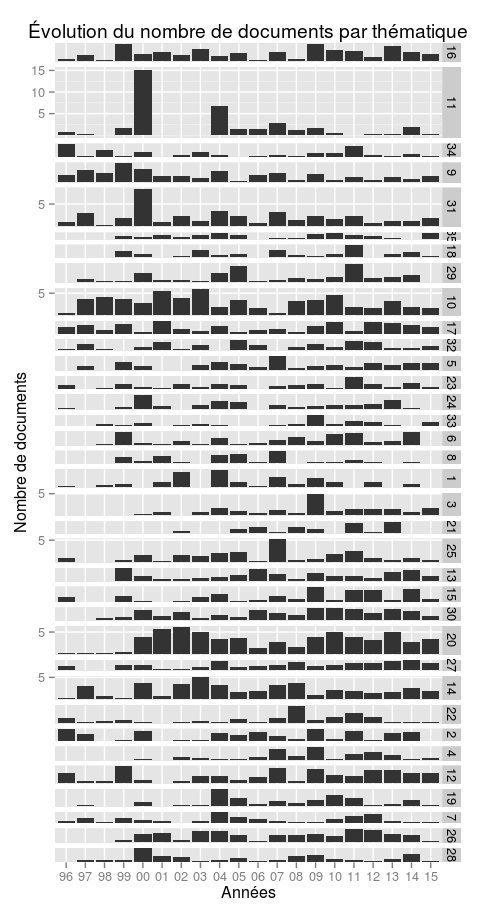
plot(mine)

plot(clust, labels = colnames(m))



plot(clust)

freq$topic\_id <- factor(freq$topic\_id, levels = colnames(m)[clust$order])  
ggplot(freq, aes(annee, counts)) +   
 geom\_bar(stat = "identity") +   
 facet\_grid(topic\_id ~ ., scales = "free\_y", space = "free\_y") +  
 labs(  
 x= "Années",  
 y = "Nombre de documents",  
 title = "Évolution du nombre de documents par thématique"  
 ) +  
 scale\_x\_discrete(labels=c("96", "97", "98", "99", "00", "01", "02", "03", "04", "05", "06", "07", "08", "09", "10", "11", "12", "13", "14", "15")) +  
 scale\_y\_continuous(breaks=c(5,10,15))



## Bibliographie

Blei D.M., Ng A.Y., & Jordan M.I. (2003) Latent dirichlet allocation. *the Journal of machine Learning research*, **3**, 993–1022.

Brandmaier A.M. (2011) *Permutation distribution clustering and structural equation model trees.* Saarbrücken.

Sparck Jones K. (1972) A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, **28**, 11–21.