Theme analysis for camera trap publications

Ada Y. Sánchez-Mercado and José R. Ferrer Paris

2020-04-03

Load required packages

library(bibliometrix) #the library for bibliometrics  
require(topicmodels) #for topic modeling  
library(quanteda) #a library for quantitative text analysis  
require(ggplot2) #visualization  
library(dplyr) #for data munging  
library("RColorBrewer") # user friendly color palettes  
library(tidytext)  
library("ldatuning")

Load data from previous step:

load(file=sprintf("%s/ISI-camera-corpus.rda",Rdata.dir))  
load(file=sprintf("%s/%s.rda",Rdata.dir,data.set.id))

This analysis is based on the data-set created from the search with ID: ISI-20191211.

# Standard bibliometric analysis

This would output several summaries for the dataset, de-activating the output for now.

# Document Term Matrix

Create DTM (Document Term Matrix). Common format for text analysis. A DTM is a matrix in which rows are documents, columns are terms, and cells indicate how often each term occurred in each document.

ISI.camera.dfm <- dfm(ISI.camera.bigram, thesaurus = camera\_thesaurus)  
ISI.camera.dfm

## Document-feature matrix of: 2,415 documents, 179,997 features (99.9% sparse) and 1 docvar.  
## features  
## docs BEHAVIOUR HOME\_RANGE POPULATION DENSITY\_ESTIMATION  
## ISI000498887100006 0 0 0 0  
## ISI000497781700009 0 0 0 0  
## ISI000496589200033 0 0 0 0  
## ISI000496310100006 6 0 0 0  
## ISI000496310100011 1 0 0 0  
## ISI000492419000009 0 0 0 0  
## features  
## docs OCCUPANCY\_MODEL CAPTURE\_RECAPTURE DISTRIBUTION RARITY  
## ISI000498887100006 0 0 0 0  
## ISI000497781700009 0 1 0 0  
## ISI000496589200033 0 0 0 0  
## ISI000496310100006 0 0 0 0  
## ISI000496310100011 1 0 0 0  
## ISI000492419000009 7 0 0 0  
## features  
## docs WIDESPREAD REPROD\_LIFEHISTORY  
## ISI000498887100006 0 0  
## ISI000497781700009 0 0  
## ISI000496589200033 0 0  
## ISI000496310100006 0 0  
## ISI000496310100011 0 0  
## ISI000492419000009 0 0  
## [ reached max\_ndoc ... 2,409 more documents, reached max\_nfeat ... 179,987 more features ]

There are too many features. Lets simplify it by trimming the dfm to include words that have appeared at least 20 times in the corpus.

ISI.camera.dfm <- dfm\_trim(ISI.camera.dfm, min\_termfreq = 20)

Better?

# Feature co-occurrence matrix

ISI.camera.fcm <- fcm(ISI.camera.dfm)

Extract top 50 keywords based on abstracts and create a feature co-occurrence matrix based on the top 50.

topfeatures(ISI.camera.fcm, 50)

## PROTECTED\_AREAS MAMMALS FOREST BEAR   
## 7560 6961 6618 4415   
## PREDATOR\_PREY TIGER CONSERVATION\_PLAN DENSITY\_ESTIMATION   
## 4323 4320 4282 3884   
## CAPTURE\_RECAPTURE HUMAN\_IMPACT FRAGMENTATION OCCUPANCY\_MODEL   
## 3705 3440 3374 3253   
## SPECIES\_STATUS WILDLIFE ANIMAL\_PLANT snow\_leopard   
## 3042 2936 2438 2292   
## HABITAT SEASONALITY wild\_boar LARGE\_MAMMALS   
## 2271 2128 2046 1819   
## SMALL\_MAMMALS BEHAVIOUR HABITAT LOSS CONSERVATION   
## 1765 1696 1673 1648   
## leopard\_cat feral\_cat cloud\_leopard BIOTIC\_INTERACTIONS   
## 1564 1557 1492 1484   
## atlant\_forest red\_fox BIRDS spatial\_tempor   
## 1468 1402 1350 1312   
## PALM\_PLANTATION leopard\_panthera medium\_larg DISTRIBUTION   
## 1285 1255 1247 1245   
## MONITORING panthera\_pardus detect\_speci REPROD\_LIFEHISTORY   
## 1205 1186 1127 1119   
## IND\_IDENTIFICATION RARITY panthera\_tigri panthera\_onca   
## 1101 1083 1053 1032   
## pine\_marten tiger\_panthera DIVERSITY whitetail\_deer   
## 1010 982 981 981   
## line\_transect puma\_concolor   
## 958 946

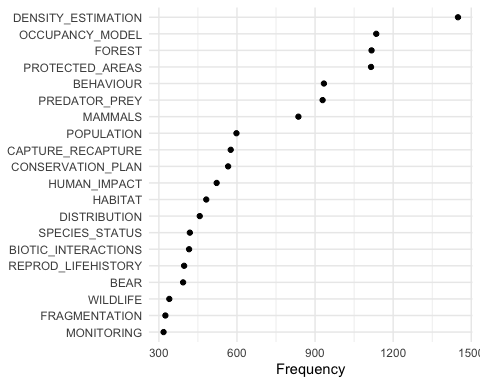
feat <- names(topfeatures(ISI.camera.fcm, 50))  
ISI.camera.fcm <- fcm\_select(ISI.camera.fcm, feat)

We could use this to plot the network, but this step takes time, we are skipping this for now.

size <- log(colSums(dfm\_select(ISI.camera.dfm, feat)))  
  
textplot\_network(ISI.camera.fcm, min\_freq = 0.5, vertex\_size = size/max(size) \* 3)

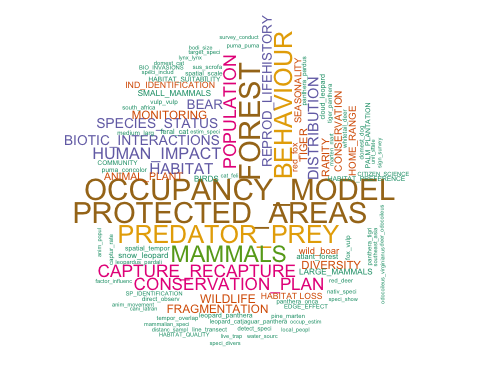
Plot top keywords frequencies instead:

freq <- textstat\_frequency(ISI.camera.dfm, n = 50)  
ISI.camera.dfm %>%  
textstat\_frequency(n = 20) %>%  
ggplot(aes(x = reorder(feature, frequency), y = frequency)) +  
geom\_point() +  
coord\_flip() +  
labs(x = NULL, y = "Frequency") +  
theme\_minimal()



Also plot a word cloud (but note that some words are excluded due to their size).

textplot\_wordcloud(ISI.camera.dfm, max\_words = 100,  
 random.order=FALSE, rot.per=0.35,  
 colors=brewer.pal(8, "Dark2"))



# LDA model

Now,apply Natural Language Processing and Topic Modeling to abstracts to identify the topics published in camera trap research.

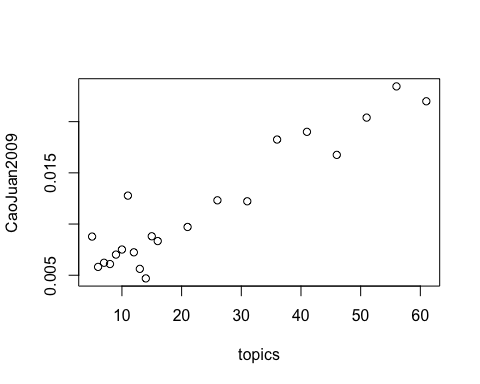
We transform the DFM to DTM

ISI.camera.dtm <- convert(ISI.camera.dfm, to = "topicmodels")

Now we need determine what is the optimal number of topics we should specify in the LDA model

Package ldatuning realizes 4 metrics to select perfect number of topics for LDA model.

if (!exists("result")) {  
 result <- FindTopicsNumber(  
 ISI.camera.dtm,  
 topics = c(5:15,seq(from = 16, to = 61, by = 5)),  
 metrics = c("CaoJuan2009"),  
 method = "Gibbs",  
 control = list(seed = 77),  
 mc.cores = 2L,  
 verbose = TRUE  
 )  
}  
  
plot(CaoJuan2009~topics,result)



We will set the number of terms to 14

Now, we can fit our first LDA model

ISI.camera.lda <- LDA(ISI.camera.dtm, control=list(seed=0), k = 14)

Show top 10 words pertaining to each topic

terms(ISI.camera.lda, 10)

## Topic 1 Topic 2 Topic 3   
## [1,] "DENSITY\_ESTIMATION" "MAMMALS" "PREDATOR\_PREY"   
## [2,] "CAPTURE\_RECAPTURE" "DIVERSITY" "REPROD\_LIFEHISTORY"  
## [3,] "POPULATION" "SMALL\_MAMMALS" "mammalian\_carnivor"  
## [4,] "IND\_IDENTIFICATION" "LARGE\_MAMMALS" "carnivor\_popul"   
## [5,] "HOME\_RANGE" "BIRDS" "south\_africa"   
## [6,] "distanc\_sampl" "COMMUNITY" "larg\_predat"   
## [7,] "individu\_per" "medium\_larg" "predat\_speci"   
## [8,] "captur\_recaptur" "speci\_mammal" "lowland\_tapir"   
## [9,] "estim\_anim" "speci\_divers" "play\_import"   
## [10,] "use\_spatial" "medium\_larges" "predat\_rate"   
## Topic 4 Topic 5 Topic 6   
## [1,] "FOREST" "BIOTIC\_INTERACTIONS" "BEAR"   
## [2,] "FRAGMENTATION" "TIGER" "POPULATION"   
## [3,] "HABITAT LOSS" "ANIMAL\_PLANT" "PROTECTED\_AREAS"   
## [4,] "atlant\_forest" "spatial\_scale" "cloud\_leopard"   
## [5,] "SPECIES\_STATUS" "tree\_speci" "leopard\_panthera"  
## [6,] "EDGE\_EFFECT" "anim\_speci" "panthera\_pardus"   
## [7,] "CONSERVATION" "visit\_rate" "leopard\_cat"   
## [8,] "select\_log" "tiger\_panthera" "SPECIES\_STATUS"   
## [9,] "mammal\_assemblag" "panthera\_tigri" "CONSERVATION"   
## [10,] "bush\_dog" "surviv\_rate" "imperfect\_detect"  
## Topic 7 Topic 8 Topic 9   
## [1,] "WILDLIFE" "BEHAVIOUR" "OCCUPANCY\_MODEL"   
## [2,] "MONITORING" "HUMAN\_IMPACT" "whitetail\_deer"   
## [3,] "target\_speci" "spatial\_tempor" "deer\_odocoileus"   
## [4,] "CITIZEN\_SCIENCE" "puma\_concolor" "occup\_estim"   
## [5,] "detect\_speci" "panthera\_onca" "unit\_state"   
## [6,] "trap\_detect" "jaguar\_panthera" "wild\_pig"   
## [7,] "live\_trap" "HOME\_RANGE" "odocoileus\_virginianus"  
## [8,] "deploy\_trap" "tempor\_overlap" "fox\_squirrel"   
## [9,] "collect\_data" "puma\_puma" "brocket\_deer"   
## [10,] "method\_detect" "HABITAT\_PREFERENCE" "estim\_occup"   
## Topic 10 Topic 11 Topic 12   
## [1,] "red\_fox" "PROTECTED\_AREAS" "CONSERVATION\_PLAN"  
## [2,] "snow\_leopard" "HABITAT" "RARITY"   
## [3,] "fox\_vulp" "CONSERVATION" "SPECIES\_STATUS"   
## [4,] "vulp\_vulp" "line\_transect" "feral\_cat"   
## [5,] "domest\_dog" "mammalian\_speci" "CONSERVATION"   
## [6,] "nativ\_speci" "bodi\_size" "domest\_cat"   
## [7,] "cani\_latran" "local\_peopl" "survey\_conduct"   
## [8,] "marten\_mart" "collar\_peccari" "tiger\_leopard"   
## [9,] "cani\_lupus" "buffer\_zone" "cat\_feli"   
## [10,] "coyot\_cani" "HABITAT\_PREFERENCE" "wild\_dog"   
## Topic 13 Topic 14   
## [1,] "DISTRIBUTION" "SEASONALITY"   
## [2,] "PALM\_PLANTATION" "wild\_boar"   
## [3,] "HABITAT\_SUITABILITY" "captur\_rate"   
## [4,] "SPECIES\_STATUS" "sus\_scrofa"   
## [5,] "direct\_observ" "pine\_marten"   
## [6,] "raccoon\_dog" "red\_deer"   
## [7,] "CLIMATE\_CHANGE" "BIO\_INVASIONS"  
## [8,] "vertebr\_speci" "water\_sourc"   
## [9,] "factor\_affect" "sika\_deer"   
## [10,] "group\_size" "boar\_sus"

Obtain the most likely topics for each document, and show topic allocation for the first documents

tt <- topics(ISI.camera.lda)  
docvars(ISI.camera.dfm, 'topic') <- tt[match(row.names(ISI.camera.dfm),names(tt))]  
  
  
head(topics(ISI.camera.lda), 5)

## ISI000498887100006 ISI000497781700009 ISI000496589200033 ISI000496310100006   
## 14 4 11 8   
## ISI000496310100011   
## 12

lists the document to (primary) topic assignments:

prd.topic <- topics(ISI.camera.lda)  
table(prd.topic)

## prd.topic  
## 1 2 3 4 5 6 7 8 9 10 11 12 13 14   
## 307 207 154 144 99 137 208 241 162 155 148 118 99 122

The *tidytext* package provides this method for extracting the per-topic-per-word probabilities, called β(“beta”), from the model.

ap\_topics <- tidy(ISI.camera.lda, matrix = "beta")  
ap\_topics

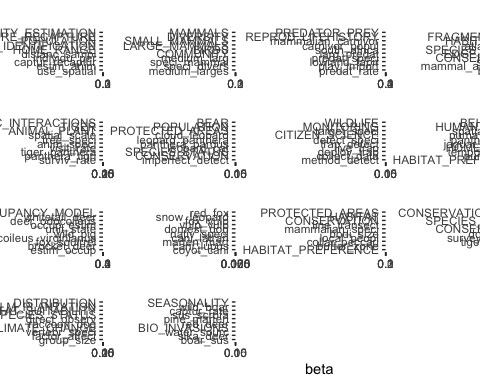
## # A tibble: 8,106 x 3  
## topic term beta  
## <int> <chr> <dbl>  
## 1 1 BEHAVIOUR 1.08e-42  
## 2 2 BEHAVIOUR 4.47e-35  
## 3 3 BEHAVIOUR 9.70e-31  
## 4 4 BEHAVIOUR 2.09e-39  
## 5 5 BEHAVIOUR 2.41e-42  
## 6 6 BEHAVIOUR 5.30e-31  
## 7 7 BEHAVIOUR 1.51e-45  
## 8 8 BEHAVIOUR 2.90e- 1  
## 9 9 BEHAVIOUR 2.81e-36  
## 10 10 BEHAVIOUR 5.99e-32  
## # … with 8,096 more rows

Now we can build the tidy data frame for the keywords. For this one, we need to use unnest() from tidyr, because they are in a list-column.

ap\_top\_terms <- ap\_topics %>%  
group\_by(topic) %>%  
top\_n(10, beta) %>%  
ungroup() %>%  
arrange(topic, -beta)

Visualization

ap\_top\_terms %>%  
mutate(term = reorder\_within(term, beta, topic)) %>%  
ggplot(aes(term, beta, fill = factor(topic))) +  
geom\_col(show.legend = FALSE) +  
facet\_wrap(~ topic, scales = "free") +  
coord\_flip() +  
scale\_x\_reordered()

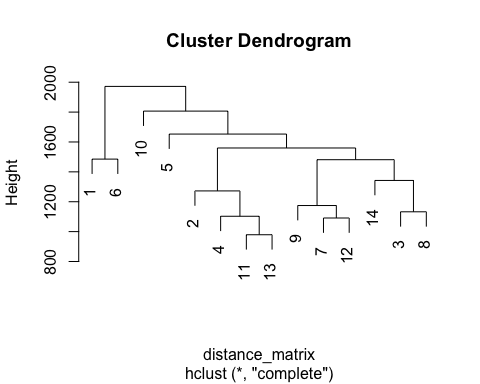


as.data.frame(ap\_top\_terms)

## topic term beta  
## 1 1 DENSITY\_ESTIMATION 0.337097896  
## 2 1 CAPTURE\_RECAPTURE 0.134001648  
## 3 1 POPULATION 0.078627921  
## 4 1 IND\_IDENTIFICATION 0.044434574  
## 5 1 HOME\_RANGE 0.039677099  
## 6 1 distanc\_sampl 0.011166804  
## 7 1 individu\_per 0.010468877  
## 8 1 captur\_recaptur 0.010003595  
## 9 1 estim\_anim 0.009198425  
## 10 1 use\_spatial 0.009073028  
## 11 2 MAMMALS 0.287385694  
## 12 2 DIVERSITY 0.102148414  
## 13 2 SMALL\_MAMMALS 0.057752149  
## 14 2 LARGE\_MAMMALS 0.056377098  
## 15 2 BIRDS 0.054314521  
## 16 2 COMMUNITY 0.029907363  
## 17 2 medium\_larg 0.028188549  
## 18 2 speci\_mammal 0.014781800  
## 19 2 speci\_divers 0.012739026  
## 20 2 medium\_larges 0.012719223  
## 21 3 PREDATOR\_PREY 0.466517434  
## 22 3 REPROD\_LIFEHISTORY 0.199362133  
## 23 3 mammalian\_carnivor 0.022597669  
## 24 3 carnivor\_popul 0.021091208  
## 25 3 south\_africa 0.020775064  
## 26 3 larg\_predat 0.015063037  
## 27 3 predat\_speci 0.014960656  
## 28 3 lowland\_tapir 0.014418101  
## 29 3 play\_import 0.014060805  
## 30 3 predat\_rate 0.014060805  
## 31 4 FOREST 0.463055208  
## 32 4 FRAGMENTATION 0.134729582  
## 33 4 HABITAT LOSS 0.075863118  
## 34 4 atlant\_forest 0.056793701  
## 35 4 SPECIES\_STATUS 0.034730915  
## 36 4 EDGE\_EFFECT 0.027360469  
## 37 4 CONSERVATION 0.015683939  
## 38 4 select\_log 0.012022010  
## 39 4 mammal\_assemblag 0.011604745  
## 40 4 bush\_dog 0.009949261  
## 41 5 BIOTIC\_INTERACTIONS 0.246900116  
## 42 5 TIGER 0.170930849  
## 43 5 ANIMAL\_PLANT 0.159654161  
## 44 5 spatial\_scale 0.051041688  
## 45 5 tree\_speci 0.026707945  
## 46 5 anim\_speci 0.025520924  
## 47 5 visit\_rate 0.025520529  
## 48 5 tiger\_panthera 0.021994179  
## 49 5 panthera\_tigri 0.021993161  
## 50 5 surviv\_rate 0.017211787  
## 51 6 BEAR 0.182902686  
## 52 6 POPULATION 0.120906190  
## 53 6 PROTECTED\_AREAS 0.062070418  
## 54 6 cloud\_leopard 0.056313550  
## 55 6 leopard\_panthera 0.043687435  
## 56 6 panthera\_pardus 0.040489893  
## 57 6 leopard\_cat 0.035370494  
## 58 6 SPECIES\_STATUS 0.022268931  
## 59 6 CONSERVATION 0.020965908  
## 60 6 imperfect\_detect 0.020470415  
## 61 7 WILDLIFE 0.153404737  
## 62 7 MONITORING 0.143478560  
## 63 7 target\_speci 0.026620236  
## 64 7 CITIZEN\_SCIENCE 0.025717855  
## 65 7 detect\_speci 0.021147430  
## 66 7 trap\_detect 0.018878508  
## 67 7 live\_trap 0.018171679  
## 68 7 deploy\_trap 0.015763640  
## 69 7 collect\_data 0.015394677  
## 70 7 method\_detect 0.015340475  
## 71 8 BEHAVIOUR 0.289862467  
## 72 8 HUMAN\_IMPACT 0.161984698  
## 73 8 spatial\_tempor 0.031619875  
## 74 8 puma\_concolor 0.026227364  
## 75 8 panthera\_onca 0.023896575  
## 76 8 jaguar\_panthera 0.022344076  
## 77 8 HOME\_RANGE 0.020622360  
## 78 8 tempor\_overlap 0.019862096  
## 79 8 puma\_puma 0.018310370  
## 80 8 HABITAT\_PREFERENCE 0.012782816  
## 81 9 OCCUPANCY\_MODEL 0.408405439  
## 82 9 whitetail\_deer 0.028066629  
## 83 9 deer\_odocoileus 0.017970560  
## 84 9 occup\_estim 0.016549419  
## 85 9 unit\_state 0.016306468  
## 86 9 wild\_pig 0.016164823  
## 87 9 odocoileus\_virginianus 0.015206828  
## 88 9 fox\_squirrel 0.013673486  
## 89 9 brocket\_deer 0.012953829  
## 90 9 estim\_occup 0.012409840  
## 91 10 red\_fox 0.101266176  
## 92 10 snow\_leopard 0.082901844  
## 93 10 fox\_vulp 0.047222569  
## 94 10 vulp\_vulp 0.046173179  
## 95 10 domest\_dog 0.038825728  
## 96 10 nativ\_speci 0.031994193  
## 97 10 cani\_latran 0.025710065  
## 98 10 marten\_mart 0.025494305  
## 99 10 cani\_lupus 0.022561894  
## 100 10 coyot\_cani 0.022037199  
## 101 11 PROTECTED\_AREAS 0.343556607  
## 102 11 HABITAT 0.174015558  
## 103 11 CONSERVATION 0.030697229  
## 104 11 line\_transect 0.025277336  
## 105 11 mammalian\_speci 0.016971506  
## 106 11 bodi\_size 0.015060543  
## 107 11 local\_peopl 0.014187753  
## 108 11 collar\_peccari 0.012638668  
## 109 11 buffer\_zone 0.012485263  
## 110 11 HABITAT\_PREFERENCE 0.011246274  
## 111 12 CONSERVATION\_PLAN 0.264813129  
## 112 12 RARITY 0.127252035  
## 113 12 SPECIES\_STATUS 0.093170234  
## 114 12 feral\_cat 0.057547731  
## 115 12 CONSERVATION 0.027057285  
## 116 12 domest\_cat 0.026668461  
## 117 12 survey\_conduct 0.021047974  
## 118 12 tiger\_leopard 0.019182577  
## 119 12 cat\_feli 0.018217005  
## 120 12 wild\_dog 0.017778974  
## 121 13 DISTRIBUTION 0.237841259  
## 122 13 PALM\_PLANTATION 0.054646603  
## 123 13 HABITAT\_SUITABILITY 0.049962608  
## 124 13 SPECIES\_STATUS 0.042101946  
## 125 13 direct\_observ 0.035078132  
## 126 13 raccoon\_dog 0.020817753  
## 127 13 CLIMATE\_CHANGE 0.019005814  
## 128 13 vertebr\_speci 0.018735978  
## 129 13 factor\_affect 0.018735978  
## 130 13 group\_size 0.018215532  
## 131 14 SEASONALITY 0.126928127  
## 132 14 wild\_boar 0.124661554  
## 133 14 captur\_rate 0.040798327  
## 134 14 sus\_scrofa 0.039665040  
## 135 14 pine\_marten 0.037965110  
## 136 14 red\_deer 0.035698536  
## 137 14 BIO\_INVASIONS 0.031165388  
## 138 14 water\_sourc 0.026632241  
## 139 14 sika\_deer 0.025498954  
## 140 14 boar\_sus 0.024365667

Dendogram to evaluate how similar are the topics

#str(ISI.camera.lda)  
my.model <- ISI.camera.lda@beta  
distance\_matrix <- dist(my.model, method="euclidean")  
plot(hclust(distance\_matrix), cex = 1)



Document-topic probabilities: Besides estimating each topic as a mixture of words, LDA also models each document as a mixture of topics.

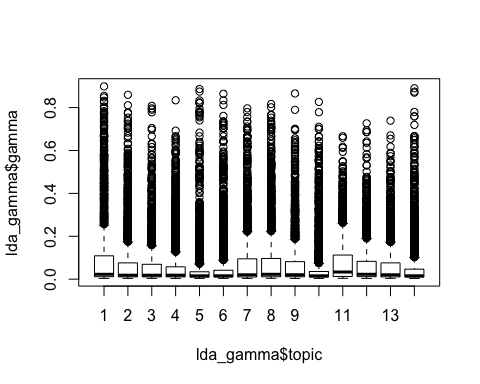
We can examine the per-document-per-topic probabilities, called γ (“gamma”), with the matrix = “gamma” argument to tidy()

lda\_gamma <- tidy(ISI.camera.lda, matrix = "gamma")  
lda\_gamma

## # A tibble: 32,214 x 3  
## document topic gamma  
## <chr> <int> <dbl>  
## 1 ISI000498887100006 1 0.00918  
## 2 ISI000497781700009 1 0.0492   
## 3 ISI000496589200033 1 0.0417   
## 4 ISI000496310100006 1 0.00673  
## 5 ISI000496310100011 1 0.0156   
## 6 ISI000492419000009 1 0.00468  
## 7 ISI000491627700009 1 0.0128   
## 8 ISI000498146900001 1 0.432   
## 9 ISI000497464700001 1 0.00747  
## 10 ISI000497712500001 1 0.0119   
## # … with 32,204 more rows

How are the probabilities distributed? Let’s visualize them

boxplot(lda\_gamma$gamma ~ lda\_gamma$topic)

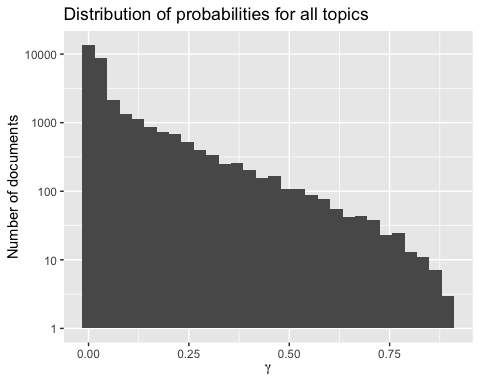


Hmm, few documents are classified in a given topic with high γ…seems like documents are asigned randomly to a given topic

Another way to see this is:

ggplot(lda\_gamma, aes(gamma)) +  
geom\_histogram() +  
scale\_y\_log10() +  
labs(title = "Distribution of probabilities for all topics",  
y = "Number of documents", x = expression(gamma))

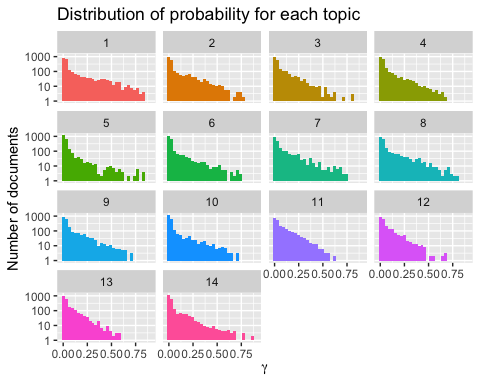
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



There are many values near zero, which means there are many documents that do not belong in each topic. Also, there are few values near γ= 1 these are the documents that do belong in those topics. This distribution shows that documents are being not well discriminated as belonging to a topic or not. We can also look at how the probabilities are distributed within each topic

ggplot(lda\_gamma, aes(gamma, fill = as.factor(topic))) +  
geom\_histogram(show.legend = FALSE) +  
facet\_wrap(~ topic, ncol = 4) +  
scale\_y\_log10() +  
labs(title = "Distribution of probability for each topic",  
y = "Number of documents", x = expression(gamma))

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



We can use this information to decide how many topics for our topic modeling procedure. When we tried options higher than 6, the distributions for γ started to look very flat toward γ= 1; documents were not getting sorted into topics very well.

Convert to a data frame

prd.topic <- as.data.frame(as.table(prd.topic))  
colnames(prd.topic)<- c("UT", "lda\_topic")  
str(prd.topic)

## 'data.frame': 2301 obs. of 2 variables:  
## $ UT : Factor w/ 2301 levels "ISI000498887100006",..: 1 2 3 4 5 6 7 8 9 10 ...  
## $ lda\_topic: int 14 4 11 8 12 9 3 1 4 8 ...

Now get back to the complete collection of literature review and combine it

ISI.topic.df <- merge(ISI.search.df, prd.topic, by = "UT", all.x = T)  
M3 <- subset(ISI.topic.df,search.group %in% "cameratrap")  
M3$topic <- ifelse(is.na(M3$lda\_topic),99,M3$lda\_topic)  
  
summary(M3$lda\_topic)

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 1.000 3.000 7.000 6.739 10.000 14.000 168

dim(M3)

## [1] 2469 45

# Topic temporal trends

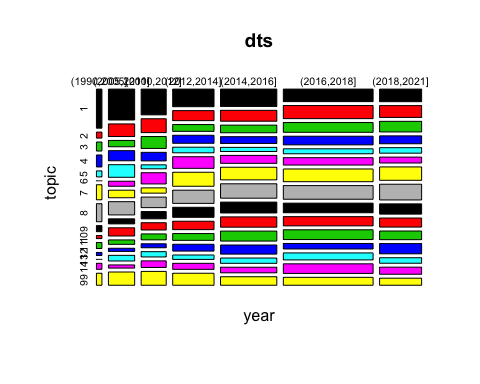
We can do this by: 1\* Calculating the total number of articles that have been published on a topic over a particular period. This will provide information on total research effort within a corpus. 2\* by investigating changes in topic popularity over that period. This allow us evaluate which topics are hot (i.e., show positive growth) versus cold (negative growth) within a given research community.

Lets see the first approach

(yrange <- range(M3$PY,na.rm=T))

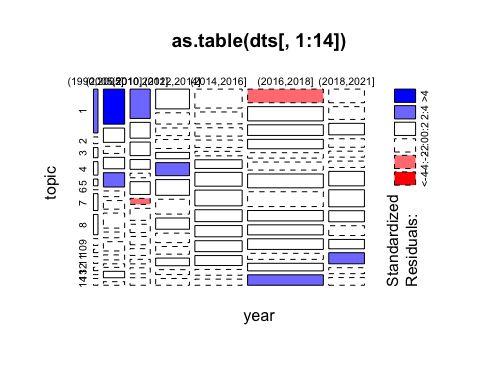
## [1] 1994 2019

M3$period <- cut(M3$PY,c(1990,2005,2010,2012,2014,2016,2018,2021))  
  
dts <- with(M3,tapply(UT,list(year=period,topic=topic), length))  
dts[is.na(dts)] <- 0  
 mosaicplot(dts,col=1:15)



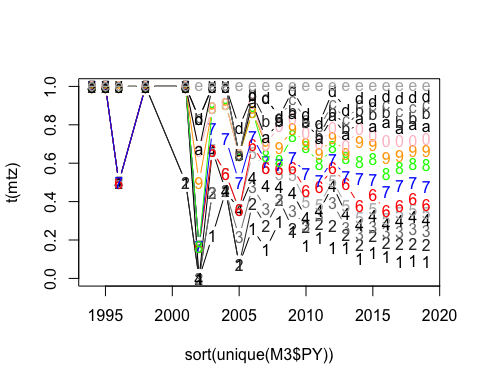
We can check the standardized residuals of the contingency table:

mosaicplot(as.table(dts[,1:14]), shade = TRUE)



Or relative frequencies (This graphs is difficult to interpret)

dts <- with(M3,tapply(UT,list(year=PY,topic=topic), length))  
dts[is.na(dts)] <- 0  
  
 mtz <- apply(t(apply(dts,1,function(x) x/sum(x))),1,cumsum)  
  
matplot(sort(unique(M3$PY)),t(mtz),type="b",lty=1,cex=1,col=c("black","grey20", "grey50", "black", "grey70", "red","blue", "green", "orange", "pink"))



#legend(2000, 90,sprintf("topic %s",1:14), pch =c (3,15,17,18,19, 3,15,17,18,19),col=c("black","grey20", "grey50", "black", "grey70", "red","blue", "green", "orange", "pink"), cex=1 )