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

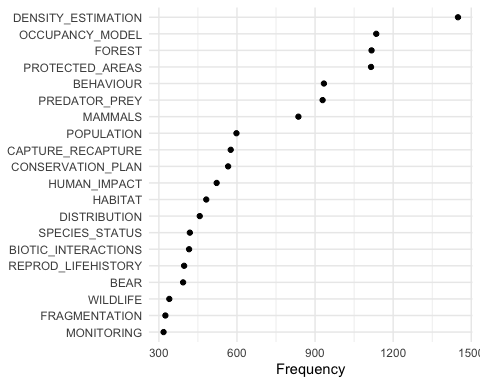
feat <- names(topfeatures(ISI.camera.fcm, 50))  
ISI.camera.fcm <- fcm\_select(ISI.camera.fcm, feat)

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)

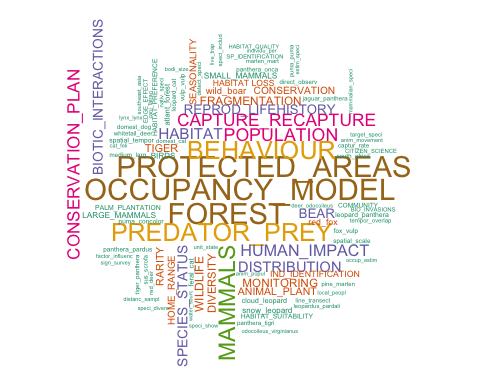
Plot top keywords frequencies instead:

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")

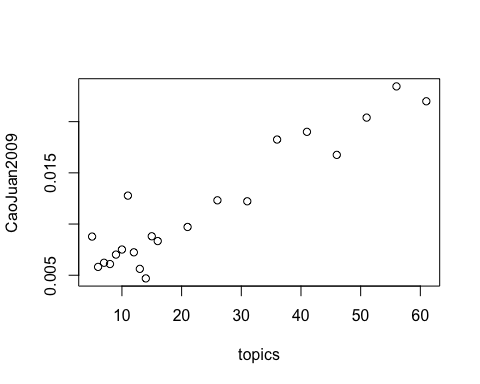
Now we need determine what is the optimal number of topics we should specify in the LDA model

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  
 )  
}

## fit models... done.  
## calculate metrics:  
## CaoJuan2009... done.

plot(CaoJuan2009~topics,result)



We will set the number of terms to 10

Now, we can fit our first LDA model

ISI.camera.lda <- LDA(ISI.camera.dtm, control=list(seed=0), k = 10)

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,] "POPULATION" "DIVERSITY" "REPROD\_LIFEHISTORY"  
## [3,] "CAPTURE\_RECAPTURE" "SMALL\_MAMMALS" "pine\_marten"   
## [4,] "IND\_IDENTIFICATION" "LARGE\_MAMMALS" "panthera\_pardus"   
## [5,] "HOME\_RANGE" "BIRDS" "leopard\_panthera"   
## [6,] "CONSERVATION\_PLAN" "DISTRIBUTION" "mammalian\_carnivor"  
## [7,] "RARITY" "PROTECTED\_AREAS" "sika\_deer"   
## [8,] "SPECIES\_STATUS" "COMMUNITY" "carnivor\_popul"   
## [9,] "distanc\_sampl" "medium\_larg" "south\_africa"   
## [10,] "individu\_per" "captur\_rate" "wild\_dog"   
## Topic 4 Topic 5 Topic 6   
## [1,] "FOREST" "BIOTIC\_INTERACTIONS" "BEAR"   
## [2,] "FRAGMENTATION" "TIGER" "PROTECTED\_AREAS"   
## [3,] "HABITAT" "ANIMAL\_PLANT" "CONSERVATION\_PLAN"   
## [4,] "HABITAT LOSS" "wild\_boar" "SPECIES\_STATUS"   
## [5,] "atlant\_forest" "PROTECTED\_AREAS" "SEASONALITY"   
## [6,] "PALM\_PLANTATION" "panthera\_tigri" "cloud\_leopard"   
## [7,] "SPECIES\_STATUS" "tiger\_panthera" "CONSERVATION"   
## [8,] "DISTRIBUTION" "sus\_scrofa" "WILDLIFE"   
## [9,] "EDGE\_EFFECT" "red\_deer" "HABITAT\_SUITABILITY"  
## [10,] "HABITAT\_QUALITY" "tree\_speci" "leopard\_cat"   
## Topic 7 Topic 8 Topic 9   
## [1,] "MONITORING" "BEHAVIOUR" "OCCUPANCY\_MODEL"   
## [2,] "WILDLIFE" "HUMAN\_IMPACT" "whitetail\_deer"   
## [3,] "feral\_cat" "HABITAT" "unit\_state"   
## [4,] "RARITY" "spatial\_tempor" "deer\_odocoileus"   
## [5,] "line\_transect" "puma\_concolor" "water\_sourc"   
## [6,] "target\_speci" "panthera\_onca" "occup\_estim"   
## [7,] "direct\_observ" "HOME\_RANGE" "odocoileus\_virginianus"  
## [8,] "CITIZEN\_SCIENCE" "jaguar\_panthera" "wild\_pig"   
## [9,] "CONSERVATION\_PLAN" "tempor\_overlap" "survey\_site"   
## [10,] "live\_trap" "puma\_puma" "fox\_squirrel"   
## Topic 10   
## [1,] "PROTECTED\_AREAS"   
## [2,] "red\_fox"   
## [3,] "snow\_leopard"   
## [4,] "fox\_vulp"   
## [5,] "vulp\_vulp"   
## [6,] "CONSERVATION\_PLAN"  
## [7,] "domest\_dog"   
## [8,] "DISTRIBUTION"   
## [9,] "domest\_cat"   
## [10,] "cani\_latran"