Topic Modelling

NLP4StatRef

10/4/2021

## Topic modelling: tests with the Latent Semantic Analysis (LSA) algorithm.

### 1. Initialization of the R environment.

The first step is to load the required libraries. The code chunk below automatically installs these libraries if they are missing. Then we set the working folder to the one containing the R Markdown document and the input datasets. The commented-out code:

*current\_working\_dir <- dirname(rstudioapi::getActiveDocumentContext()$path)*

works only from within RStudio when running the document chunk-by-chunk. If this is not the case (e.g. when knitting the document), the user has to set the working directory manually.

rm(list=ls()) ## clear objects from memory  
  
## install libraries if missing  
list.of.packages <- c('tm','ggplot2','topicmodels','tidytext','dplyr')  
new.packages <- list.of.packages[!(list.of.packages %in% installed.packages()[,"Package"])]  
if(length(new.packages)) install.packages(new.packages)  
  
library(tm)  
library(ggplot2)  
library(topicmodels)  
library(tidytext)  
library(dplyr)  
  
#current\_working\_dir <- dirname(rstudioapi::getActiveDocumentContext()$path)  
#print(current\_working\_dir)  
#setwd(current\_working\_dir)  
  
## ADJUST THIS   
setwd('D://Kimon//Documents//Quantos-new//NLP4StatRef//Deliverable D2.2')

### 2. Data input.

We read two of the files extracted from the database, with the glossary articles definitions in *ESTAT\_dat\_concepts\_2021\_04\_08.csv* and their titles in *ESTAT\_dat\_link\_info\_2021\_04\_08.csv*. The common key is *id*. **At a later stage, the reading of the files will be directly from the KD**.

We then drop articles with missing titles and/or definitions and also de-duplicate the records of the resulting file based on these two fields.

dat1 <- read.csv2('ESTAT\_dat\_concepts\_2021\_04\_08.csv')  
dat2 <- read.csv2('ESTAT\_dat\_link\_info\_2021\_04\_08.csv')  
dat <- merge(dat1,dat2,by=c('id'),all=FALSE)  
dat <- dat[,c('title','definition')]  
  
dels <- which(is.na(dat$title))  
if(length(dels)>0) dat <- dat[-dels,]  
  
dels <- which(is.na(dat$definition))  
if(length(dels)>0) dat <- dat[-dels,]  
  
dels <-which(duplicated(dat$title))  
if(length(dels)>0) dat <- dat[-dels,]  
  
dels <- which(duplicated(dat$definition))  
if(length(dels)>0) dat <- dat[-dels,]  
  
rm(dat1,dat2)

### 3. Data cleaning.

In the next step we do some data cleaning:

* Replace multiple spaces with single ones in definitions.
* Discard spaces at the start of definitions and titles.
* Replace space-comma-space by comma-space in definitions.

dat$definition <- gsub(' +',' ',dat$definition) ## discard multiple spaces  
dat$definition <- gsub('^ +','',dat$definition) ## discard spaces at start  
dat$definition <- gsub(' \\, ','\\, ',dat$definition) ## space-comma-space -> comma-space  
  
dat$title <- gsub('^ +','',dat$title) ## discard spaces at start

### 4. Creating tm objects.

Next we create a corpus *texts* from the articles. This has initially 1285 text entries. We apply the standard pre-processing steps to the texts:

* Remove punctuation and numbers.
* Convert all to lower case.
* Strip whitespace and apply an English stemmer.

In the end we obtain 331 terms.

We then create a document-to-term matrix *dmat*, keeping words with minimum length 5, each one in at least 2% of documents and in at most 30% of the documents. We remove documents without terms and convert the matrix to a 1278 x 331 dataframe for inspection.

Note that in the construction of the document-to-term matrix, we do not request any weights, such as tf-idf. This is a requirement of the LDA algorith.

texts <- Corpus(VectorSource(dat$definition))  
ndocs <- nrow(dat)  
cat('ndocs = ',ndocs,'\n')

> ndocs = 1285

## apply several pre-processing steps (see package tm)  
texts <- tm\_map(texts, removePunctuation)   
texts <- tm\_map(texts, removeNumbers)   
texts <- tm\_map(texts, tolower)  
  
texts <- tm\_map(texts, removeWords, stopwords(kind='SMART'))   
texts <- tm\_map(texts, stripWhitespace)   
texts <- tm\_map(texts, stemDocument, language='english')  
  
## create document-to-term matrix (tf-idf)  
## min word length: 5, each term in at least 2% of documents   
## and at most in 40% of documents  
dtm <- DocumentTermMatrix(texts,  
 control=list(weighting=weightTf,   
 wordLengths=c(5, Inf),bounds =   
 list(global = c(0.02\*ndocs,  
 0.3\*ndocs))))  
  
dels <- which(apply(dtm,1,sum)==0) #remove all texts without terms   
if(length(dels)>0) {  
 dtm <- dtm[-dels, ]   
 dat <- dat[-dels,]  
}  
  
nTerms(dtm)

> [1] 331

Terms(dtm)

> [1] "appli" "charg" "collect" "context" "count"   
> [6] "countri" "defin" "establish" "european" "limit"   
> [11] "locat" "regul" "resid" "statist" "union"   
> [16] "characterist" "common" "develop" "framework" "member"   
> [21] "particip" "product" "purpos" "refer" "report"   
> [26] "state" "suppli" "classifi" "employ" "entiti"   
> [31] "environ" "geograph" "person" "place" "abbrevi"   
> [36] "cooper" "europ" "inform" "norway" "process"   
> [41] "econom" "growth" "nation" "peopl" "region"   
> [46] "social" "world" "agreement" "benefit" "integr"   
> [51] "manag" "network" "organis" "origin" "respons"   
> [56] "trade" "januari" "order" "treati" "group"   
> [61] "account" "amount" "compris" "cover" "electr"   
> [66] "except" "facil" "financ" "financi" "health"   
> [71] "household" "indic" "month" "payment" "purchas"   
> [76] "reason" "relat" "similar" "total" "water"   
> [81] "andor" "balanc" "deduct" "interest" "singl"   
> [86] "specif" "intern" "polici" "territori" "extern"   
> [91] "carri" "chang" "consid" "consist" "control"   
> [96] "custom" "direct" "enterpris" "expenditur" "export"   
> [101] "final" "grant" "import" "legal" "occur"   
> [106] "physic" "principl" "procedur" "record" "servic"   
> [111] "signific" "system" "transact" "transfer" "period"   
> [116] "construct" "materi" "perman" "agricultur" "build"   
> [121] "communic" "exclud" "industri" "public" "technic"   
> [126] "transport" "categori" "class" "correspond" "distinguish"   
> [131] "field" "observ" "surfac" "communiti" "energi"   
> [136] "protect" "definit" "differ" "number" "privat"   
> [141] "averag" "divid" "calcul" "effect" "instanc"   
> [146] "posit" "popul" "variabl" "harvest" "natur"   
> [151] "describ" "repres" "resourc" "small" "special"   
> [156] "adopt" "central" "consum" "continu" "council"   
> [161] "function" "implement" "market" "object" "point"   
> [166] "price" "produc" "provid" "scheme" "standard"   
> [171] "support" "worker" "general" "plant" "regular"   
> [176] "activ" "belong" "classif" "consumpt" "distinct"   
> [181] "engag" "famili" "identifi" "institut" "intend"   
> [186] "research" "separ" "surplus" "annual" "exclus"   
> [191] "higher" "human" "maintain" "condit" "environment"   
> [196] "improv" "subsidi" "independ" "individu" "temporari"   
> [201] "assess" "commiss" "concern" "decis" "present"   
> [206] "aggreg" "applic" "basic" "compar" "comparison"   
> [211] "conduct" "eurostat" "input" "labour" "legisl"   
> [216] "level" "method" "requir" "status" "structur"   
> [221] "survey" "detail" "estim" "found" "measur"   
> [226] "obtain" "result" "volum" "diseas" "death"   
> [231] "deriv" "organ" "train" "compil" "distribut"   
> [236] "qualiti" "corpor" "stock" "actual" "express"   
> [241] "factor" "weight" "access" "allow" "remov"   
> [246] "typic" "complet" "convent" "equip" "author"   
> [251] "increas" "innov" "knowledg" "major" "properti"   
> [256] "depend" "exampl" "local" "equal" "practic"   
> [261] "sourc" "combin" "primari" "reduc" "secondari"   
> [266] "analysi" "contribut" "ratio" "exist" "design"   
> [271] "domest" "initi" "normal" "asset" "capit"   
> [276] "compos" "comput" "gross" "instrument" "perform"   
> [281] "current" "sector" "equival" "determin" "economi"   
> [286] "incom" "involv" "liabil" "nonfinanci" "secur"   
> [291] "compon" "reflect" "output" "vehicl" "pension"   
> [296] "concept" "project" "section" "publish" "dispos"   
> [301] "provis" "creat" "generat" "power" "princip"   
> [306] "addit" "employe" "invest" "receiv" "figur"   
> [311] "monetari" "previous" "share" "criteria" "technolog"   
> [316] "minus" "agenc" "offici" "adjust" "administr"   
> [321] "short" "intermedi" "percentag" "govern" "offic"   
> [326] "formal" "relev" "index" "exchang" "unemploy"   
> [331] "format"

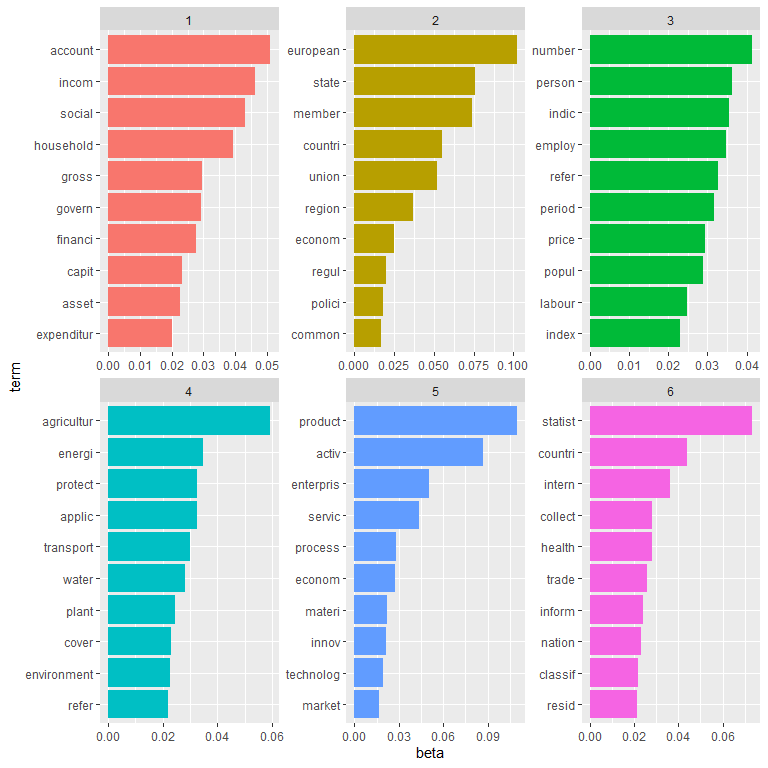
## convert to dataframe for inspection  
dtm.dat <- as.data.frame(as.matrix(dtm))  
rownames(dtm.dat)<- dat$title  
  
print(inspect(dtm))

> <<DocumentTermMatrix (documents: 1278, terms: 331)>>  
> Non-/sparse entries: 20782/402236  
> Sparsity : 95%  
> Maximal term length: 12  
> Weighting : term frequency (tf)  
> Sample :  
> Terms  
> Docs activ countri econom european member product refer servic state statist  
> 1267 0 0 2 0 0 0 0 0 1 0  
> 159 2 0 1 0 0 3 0 0 0 0  
> 192 6 6 2 8 1 2 0 1 0 6  
> 195 0 1 0 3 1 0 2 0 1 2  
> 272 0 1 3 3 8 1 0 0 8 1  
> 305 0 1 0 1 6 0 8 0 5 2  
> 306 0 7 0 0 3 0 9 0 3 2  
> 599 14 0 9 3 0 13 1 3 0 1  
> 665 0 0 0 1 0 12 0 5 0 4  
> 777 3 0 6 3 2 9 0 2 2 0  
> Terms  
> Docs activ countri econom european member product refer servic state statist  
> 1267 0 0 2 0 0 0 0 0 1 0  
> 159 2 0 1 0 0 3 0 0 0 0  
> 192 6 6 2 8 1 2 0 1 0 6  
> 195 0 1 0 3 1 0 2 0 1 2  
> 272 0 1 3 3 8 1 0 0 8 1  
> 305 0 1 0 1 6 0 8 0 5 2  
> 306 0 7 0 0 3 0 9 0 3 2  
> 599 14 0 9 3 0 13 1 3 0 1  
> 665 0 0 0 1 0 12 0 5 0 4  
> 777 3 0 6 3 2 9 0 2 2 0

### 5. Application of the LDA algorithm.

We apply the LDA algorithm with k=6 topics.

lda\_model <- LDA(dtm, k = 6, control = list(seed = 1234))  
topics <- tidy(lda\_model, matrix = "beta")  
  
top\_terms <- topics %>%  
 group\_by(topic) %>%  
 slice\_max(beta, n = 10) %>%   
 ungroup() %>%  
 arrange(topic, -beta)  
  
top\_terms %>%  
 mutate(term = reorder\_within(term, beta, topic)) %>%  
 ggplot(aes(beta, term, fill = factor(topic))) +  
 geom\_col(show.legend = FALSE) +  
 facet\_wrap(~ topic, scales = "free") +  
 scale\_y\_reordered()



The results with the top 10 terms by topic can be interpreted as follows:

* Topic 1 is about demography and regions.
* Topic 2 is about EU countries and economies.
* Topic 3 is about persons and employment.
* Topic 4 is about household incomes and social aspects.
* Topic 5 is about enterprises and business activities.
* Topic 6 is about primary production, energy and resources.