# topic-modelling

Elisa Bankl

2023-01-13

```
library(quanteda)
```

```
## Warning: Paket 'quanteda' wurde unter R Version 4.2.2 erstellt
```

```
## Package version: 3.2.4
## Unicode version: 13.0
## ICU version: 69.1
```

```
## Parallel computing: 4 of 4 threads used.
```

```
## See https://quanteda.io for tutorials and examples.
```

```
library(quanteda.textmodels)
```

```
## Warning: Paket 'quanteda.textmodels' wurde unter R Version 4.2.2 erstellt
```

```
library(text2vec)
```

```
## Warning: Paket 'text2vec' wurde unter R Version 4.2.2 erstellt
```

```
library(LDAvis)
```

```
## Warning: Paket 'LDAvis' wurde unter R Version 4.2.2 erstellt
```

```
library(LSAfun)
```

```
## Warning: Paket 'LSAfun' wurde unter R Version 4.2.2 erstellt
```

```
## Lade nötiges Paket: lsa
```

```
## Warning: Paket 'lsa' wurde unter R Version 4.2.2 erstellt
```

```
## Lade nötiges Paket: SnowballC
```

```
## Lade nötiges Paket: rgl
```

```
## Warning: Paket 'rgl' wurde unter R Version 4.2.2 erstellt
```

```
##
## Attache Paket: 'LSAfun'
```

```
## Die folgenden Objekte sind maskiert von 'package:text2vec':
##
##     coherence, normalize
```

Read in the dataset

```
textdata <- base::readRDS(url("https://slcladal.github.io/data/sotu_paragraphs.rda", "rb"))
head(textdata)
```

```
##   doc_id speech_doc_id                speech_type         president       date
## 1      1             1 State of the Union Address George Washington 1790-01-08
```

```
## 2      2             1 State of the Union Address George Washington 1790-01-08
## 3      3             1 State of the Union Address George Washington 1790-01-08
## 4      4             1 State of the Union Address George Washington 1790-01-08
## 5      5             1 State of the Union Address George Washington 1790-01-08
## 6      6             1 State of the Union Address George Washington 1790-01-08
##
## 1
## 2 I embrace with great satisfaction the opportunity which now presents itself\nof congratulating you
## 3                                                    In resuming your consultations for the genera
## 4
## 5
## 6
```

```
textdata[2,6]
```

```
## [1] "I embrace with great satisfaction the opportunity which now presents itself\nof congratulating y
```

In this notebook, the packages text2vec are used to create a Document Feature Matrix / a sparse Document Term Matrix from the dataset.

## text2vec

```
prep_fun = tolower

stem_tokenizer =function(x){
  lapply(word_tokenizer(x),SnowballC::wordStem)
}

it = itoken(textdata$text,
                  preprocessor = prep_fun,
                  tokenizer = stem_tokenizer,
                  ids = textdata$doc_id,
                  progressbar = FALSE)
vocab = create_vocabulary(it,stopwords=quanteda::stopwords())
vocab = prune_vocabulary(vocab, doc_proportion_max = 0.1, term_count_min = 5)



vectorizer = vocab_vectorizer(vocab)
text2vec_DTM = create_dtm(it, vectorizer)
```

```
## as(<dgTMatrix>, "dgCMatrix") is deprecated since Matrix 1.5-0; do as(., "CsparseMatrix") instead
```

### text2vec lsa

```
text2vec_lsa = text2vec::LSA$new(n_topics = 40)
```

```
doc_embeddings =  fit_transform(text2vec_DTM, text2vec_lsa)
```

```
## INFO  [15:09:47.485] soft_als: iter 001, frobenious norm change 1250.630 loss NA
## INFO  [15:09:47.817] soft_als: iter 002, frobenious norm change 1.007 loss NA
## INFO  [15:09:48.319] soft_als: iter 003, frobenious norm change 0.092 loss NA
## INFO  [15:09:48.548] soft_als: iter 004, frobenious norm change 0.029 loss NA
## INFO  [15:09:48.770] soft_als: iter 005, frobenious norm change 0.013 loss NA
## INFO  [15:09:48.998] soft_als: iter 006, frobenious norm change 0.007 loss NA
```

```
## INFO  [15:09:49.209] soft_als: iter 007, frobenious norm change 0.004 loss NA
## INFO  [15:09:49.425] soft_als: iter 008, frobenious norm change 0.002 loss NA
## INFO  [15:09:49.639] soft_als: iter 009, frobenious norm change 0.002 loss NA
## INFO  [15:09:49.859] soft_als: iter 010, frobenious norm change 0.001 loss NA
## INFO  [15:09:50.079] soft_als: iter 011, frobenious norm change 0.001 loss NA
## INFO  [15:09:50.082] soft_impute: converged with tol 0.001000 after 11 iter
```

### plot the nearest neighbors of a word based on the LSA space

```
LSAfun::plot_neighbors("freedom",10,tvectors=t(text2vec_lsa$components))
```

```
##                 x         y         z
## freedom 0.4976968 0.6562894 0.4840652
## liberti 0.4418735 0.7159738 0.4804579
## frown   0.7285862 0.5513995 0.3213180
## speech  0.5733606 0.7169188 0.2373297
## deepest 0.5713171 0.4356582 0.6303988
## love    0.4374453 0.7373814 0.4222131
## god     0.3950292 0.4368652 0.7862949
## prayer  0.7454032 0.3800393 0.5105585
## sacr    0.3076312 0.8389164 0.3703607
## bright  0.7487533 0.4276524 0.4560960
```

### text2vec lda

```
text2vec_lda = text2vec::LDA$new(n_topics = 40, doc_topic_prior = 0.1, topic_word_prior = 0.01)
```

```
doc_embeddings =  fit_transform(text2vec_DTM, text2vec_lda)
```

```
## INFO  [15:10:07.961] early stopping at 230 iteration
## INFO  [15:10:13.078] early stopping at 60 iteration
```

### text2vec Visualization

```
text2vec_lda$plot()
```

```
## Lade nötigen Namensraum: servr
```

## quanteda

```
quanteda_corpus = quanteda::corpus(textdata,docid_field="doc_id",text_field="text")

quanteda_DFM <- quanteda::dfm(quanteda::tokens(quanteda_corpus,remove_punct=TRUE,remove_symbols=TRUE))

quanteda_DFM <- quanteda::dfm_select(quanteda_DFM,pattern=quanteda::stopwords("en"),selection="remove")

quanteda_DFM
```

```
## Document-feature matrix of: 8,833 documents, 20,182 features (99.77% sparse) and 4 docvars.
##     features
## docs fellow-citizens senate house representatives embrace great satisfaction
##    1               1      1     1               1       0     0            0
##    2               0      0     0               0       1     1            1
```

```
##    3              0      0     0              0       0     0           0
##    4              0      0     0              0       0     0           0
##    5              0      0     0              0       0     0           0
##    6              0      0     0              0       0     0           0
##     features
## docs opportunity now presents
##    1           0   0        0
##    2           1   1        1
##    3           0   0        0
##    4           0   0        0
##    5           0   0        0
##    6           0   0        0
## [ reached max_ndoc ... 8,827 more documents, reached max_nfeat ... 20,172 more features ]
```

### quanteda lsa

```
quanteda_lsa = quanteda.textmodels::textmodel_lsa(quanteda_DFM,40)
```

```
LSAfun::neighbors("ruler",10,quanteda_lsa$features)
```

```
##      ruler     bright      lover       love     unborn    prayers     lovers
##  1.0000000  0.8194659  0.8151539  0.8031053  0.8019067  0.7870610  0.7868993
##   universe canvassing      hosts
##  0.7828270  0.7816520  0.7787708
```

```
LSAfun::plot_neighbors("island",10,tvectors=quanteda_lsa$features)
```

```
##                      x         y          z
## island       0.6868020 0.4880632 -0.5079702
## 19,931       0.7288128 0.5179632 -0.4153315
## 22,187       0.7288128 0.5179632 -0.4153315
## cuba         0.5039780 0.4197821 -0.7342709
## pacification 0.7432854 0.2997635 -0.4670306
## spain        0.3016197 0.4309325 -0.8223211
## cuban        0.3348761 0.4687611 -0.7554761
## puerto       0.4140639 0.8436646 -0.2569414
## rico         0.3004133 0.8544707 -0.3832542
## strife       0.4801723 0.1390565 -0.8121494
```

#topicModel

```
sel_idx <- slam::row_sums(quanteda_DFM) > 0
quanteda_DFM <- quanteda_DFM[sel_idx, ]
textdata <- textdata[sel_idx, ]
```

```
topicModel_lda <- topicmodels::LDA(quanteda_DFM, 40, method="Gibbs", control=list(iter=400)) #only 400
```

```
# the code is from here: https://gist.github.com/trinker/477d7ae65ff6ca73cace
topicmodels2LDAvis <- function(x, ...){
    post <- topicmodels::posterior(x)
    if (ncol(post[["topics"]]) < 3) stop("The model must contain > 2 topics")
    mat <- x@wordassignments
    LDAvis::createJSON(
        phi = post[["terms"]],
        theta = post[["topics"]],
        vocab = colnames(post[["terms"]]),
```

```
        doc.length = slam::row_sums(mat, na.rm = TRUE),
        term.frequency = slam::col_sums(mat, na.rm = TRUE)
    )
}
```

```
LDAvis::serVis(topicmodels2LDAvis(topicModel_lda))
```

```
## Lade nötigen Namensraum: servr
```