Untitled

TA: Leslie Huang

Course: Text as Data

Date: 4/5/2018

Recitation 10: Unsupervised Learning I, continued

```
# Set up workspace
rm(list = ls())

setwd("/Users/Lingyi/TAD/lab/Text-as-Data-Lab-Spr2018/W10_04_05_18/")

library(quanteda)
```

```
## quanteda version 1.0.0
## Using 3 of 4 threads for parallel computing
##
## Attaching package: 'quanteda'
## The following object is masked from 'package:utils':
##
##     View
```

```
library(quanteda.corpora)
library(lsa)
```

```
## Loading required package: SnowballC
```

1 More LSA (Questions from last time)

1.1 From last time:

```
# Load the SOTU data again
data("data_corpus_sotu")

SOTU_dfm <- dfm(data_corpus_sotu[145:223,],
                stem = T,
                remove_punct = T,
                remove = stopwords("english")
)

SOTU_mat <- convert(SOTU_dfm, to = "matrix")
```

```
# 1.2 LSA without local or global weighting -- SVD of term-document matrix

SOTU_lsa_auto <- lsa(t(SOTU_mat))
SOTU_lsa_auto_mat <- t(as.textmatrix(SOTU_lsa_auto))

# Inspect how specific terms in a specific speech have been transformed
SOTU_dfm@Dimnames$docs[9]
```

```
## [1] "Roosevelt-1942"
```

```
topfeatures(SOTU_dfm[9,])
```

```
##    war  peopl nation   must  fight   unit  world  shall   year    one
##     31     19     19     18     18     17     16     15     14     14
```

```
# LSA transform
sort(SOTU_lsa_auto_mat[9,], decreasing=T)[1:10]
```

```
##      war   nation     must    world    peopl      can     forc     peac
## 31.25254 18.46004 15.93679 13.71454 13.35374 12.44615 11.83062 11.39943
## american       us
## 10.90135 10.77034
```

1.3 LSA with local weighting

```
# local weight is log TF
# global weight is IDF
# Other options listed in the documentation:
#?gw_idf

# Transform the document-term matrix
SOTU_dfm_weighted <- lw_logtf(SOTU_mat) * gw_idf(SOTU_mat)

# Run LSA (auto number of dimensions)
SOTU_lsa_weighted <- lsa(t(SOTU_dfm_weighted))
SOTU_lsa_weighted_mat <- t(as.textmatrix(SOTU_lsa_weighted))


# Inspect the values
SOTU_dfm@Dimnames$docs[9]
```

```
## [1] "Roosevelt-1942"
```

```
topfeatures(SOTU_dfm[9,])
```

```
##    war  peopl nation   must  fight   unit  world  shall   year    one
##     31     19     19     18     18     17     16     15     14     14
```

```
sort(SOTU_lsa_weighted_mat[9,], decreasing=T)[1:10]
```

```
##      war   nation    peopl    fight     year     must american    world
## 15.50230 13.94278 13.77519 12.30838 12.00423 11.95189 11.92328 11.70011
##    enemi       us
## 11.61531 11.33156
```

```
# You can also compare how values for certain words change when you run LSA with different numbers of d
```

2 LSA on n-grams

```
SOTU_bigrams_tokens <- tokens(data_corpus_sotu[145:223,], ngrams = 2, remove_punct = T)
SOTU_bigrams_dfm <- dfm(SOTU_bigrams_tokens) # obviously we cannot remove stopwords etc
SOTU_bigrams_mat <- convert(SOTU_bigrams_dfm, to = "matrix")

SOTU_bigrams_auto <- lsa(t(SOTU_bigrams_mat))
SOTU_bigrams_auto_mat <- t(as.textmatrix(SOTU_bigrams_auto))

# Inspect the values
SOTU_dfm@Dimnames$docs[9]
```

```
## [1] "Roosevelt-1942"
```

```
topfeatures(SOTU_dfm[9,])
```

```
##    war  peopl nation   must  fight   unit  world  shall   year    one
##     31     19     19     18     18     17     16     15     14     14
```

```
# The bigrams won't match up with unigrams, but...

# Most common: plagued with stopwords!
sort(SOTU_bigrams_auto_mat[9,], decreasing=T)[1:10]
```

```
##    of_the    in_the   and_the    to_the   we_have    we_are the_world
## 36.537131 19.555641 12.350223 10.925018 10.659282  9.597845  9.367608
##   for_the    on_the    of_our
##  8.329147  8.152554  7.920150
```

```
# Least common
sort(SOTU_bigrams_auto_mat[9,], decreasing=F)[1:10]
```

```
##          50_years             a_new        the_number
##         -0.6307995        -0.5929454        -0.5634193
##           in_iraq        must_trust       and_empower
##         -0.5311001        -0.4887292        -0.4887292
##       be_achieved federal_government      empower_them
##         -0.4379422        -0.3936211        -0.3801227
##         the_growth
##         -0.3742823
```