Using heatmaps to determine the number of topics in a document to set to zero

Salfo Bikienga

April 30, 2018

(Scroll down to see the figures)

Description of the exploratory works. Here, I assume a ten topics decomposition of the blogs. The goal of this plot is assess if, for clustering purpose, I should consider the ten topics in each blog, or if I should set some topics to zero for some blogs. For instance, for each blog topics distribution, I may determine a threshold and set to zero all proportion less than the threshold.

```r
library(topicmodels)
library(tm)
library(tidyverse)
options(digits = 3)
load('data/constructed_data/blogs_lda_10.RData')
K = 10
theta_matrix <- posterior(blogs_lda_10)$topics # Extract the theta matrix
theta_matrix <- round(as.data.frame(theta_matrix), digits = 3)
names(theta_matrix) <- paste0("Topic.", 1:K) # Name the columns
```

Because the number of observations is high, I will consider a random sample for exploratory purpose.

```r
# Take a random sample of theta
set.seed(314)
n_s <- sample(x = 1:nrow(theta_matrix),
              size = 50)
sub_theta <- theta_matrix[n_s, ]
head(sub_theta)
```

```
##                                  Topic.1 Topic.2 Topic.3 Topic.4 Topic.5
## b_24_218-August 28, 2013.txt       0.107   0.029   0.049   0.335   0.057
## b_25_1989-September 13, 2009.txt   0.130   0.216   0.071   0.062   0.035
## b_26_3547-Oct 24.txt               0.022   0.017   0.024   0.076   0.030
## b_25_1503-September 1, 2010.txt    0.023   0.198   0.132   0.026   0.094
## b_25_1274-April 3, 2011.txt        0.289   0.325   0.019   0.010   0.100
## b_25_2320-January 6, 2009.txt      0.010   0.225   0.217   0.005   0.017
##                                  Topic.6 Topic.7 Topic.8 Topic.9 Topic.10
## b_24_218-August 28, 2013.txt       0.032   0.177   0.163   0.044    0.007
## b_25_1989-September 13, 2009.txt   0.029   0.102   0.313   0.011    0.031
## b_26_3547-Oct 24.txt               0.026   0.017   0.032   0.401    0.355
## b_25_1503-September 1, 2010.txt    0.083   0.288   0.062   0.034    0.060
## b_25_1274-April 3, 2011.txt        0.029   0.164   0.041   0.018    0.005
## b_25_2320-January 6, 2009.txt      0.223   0.283   0.010   0.004    0.007
```

A helper function to set some proportions to zero

```r
ranK_row <- function(x, n_top){
  threshold = sort(x, decreasing = TRUE)[n_top]
  return(threshold)
}
```

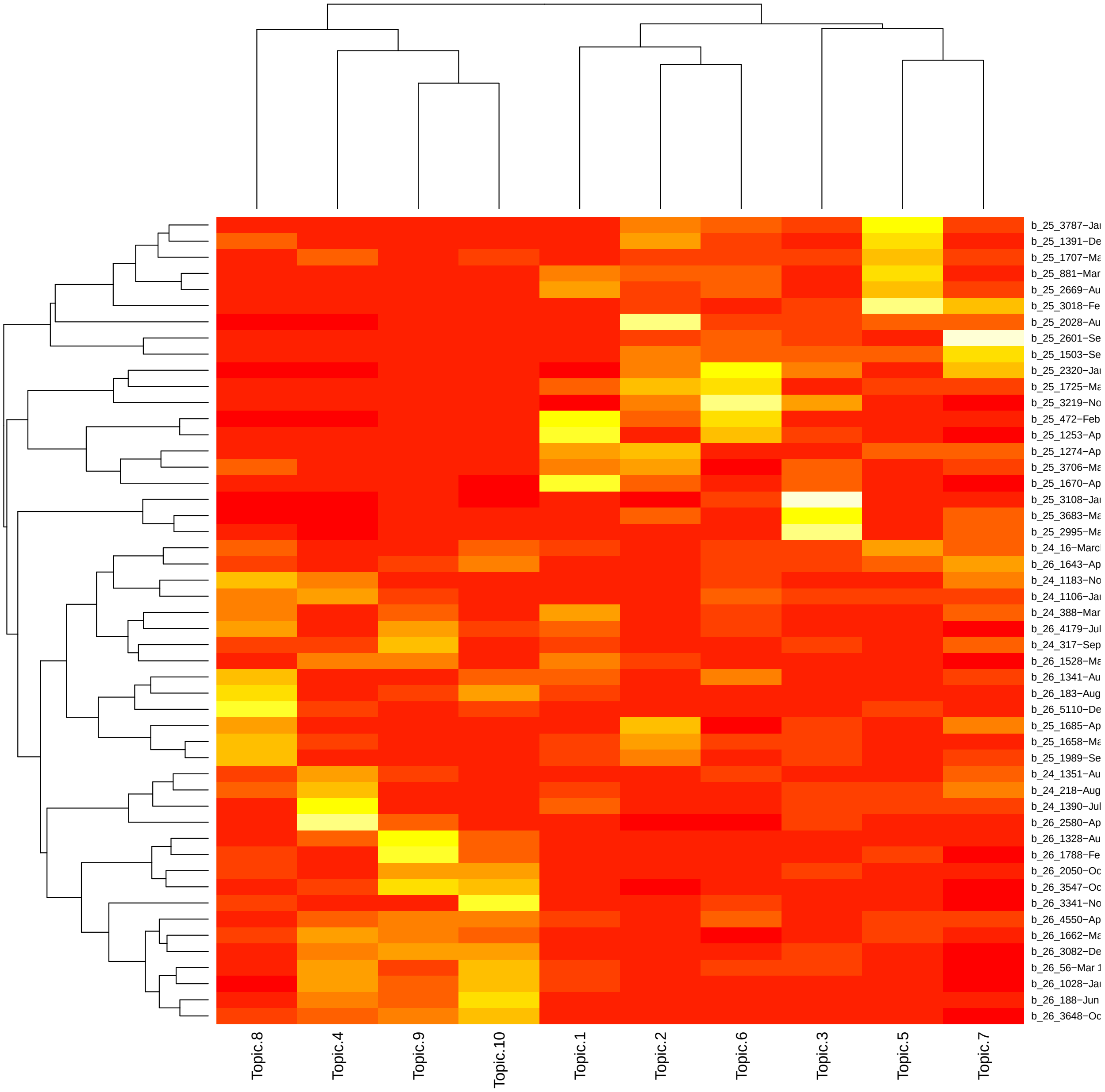
```
n_top_by_doc <- function(theta, n_top = 3){
  thresholds <- apply(theta, MARGIN = 1, FUN = ranK_row, n_top = n_top)
  thresholds <- data.frame(thresholds = thresholds)
  new_theta <- theta*(theta >= thresholds$thresholds)
}

sub_theta2 <- n_top_by_doc(theta = sub_theta, n_top = 3)
head(sub_theta2)
```

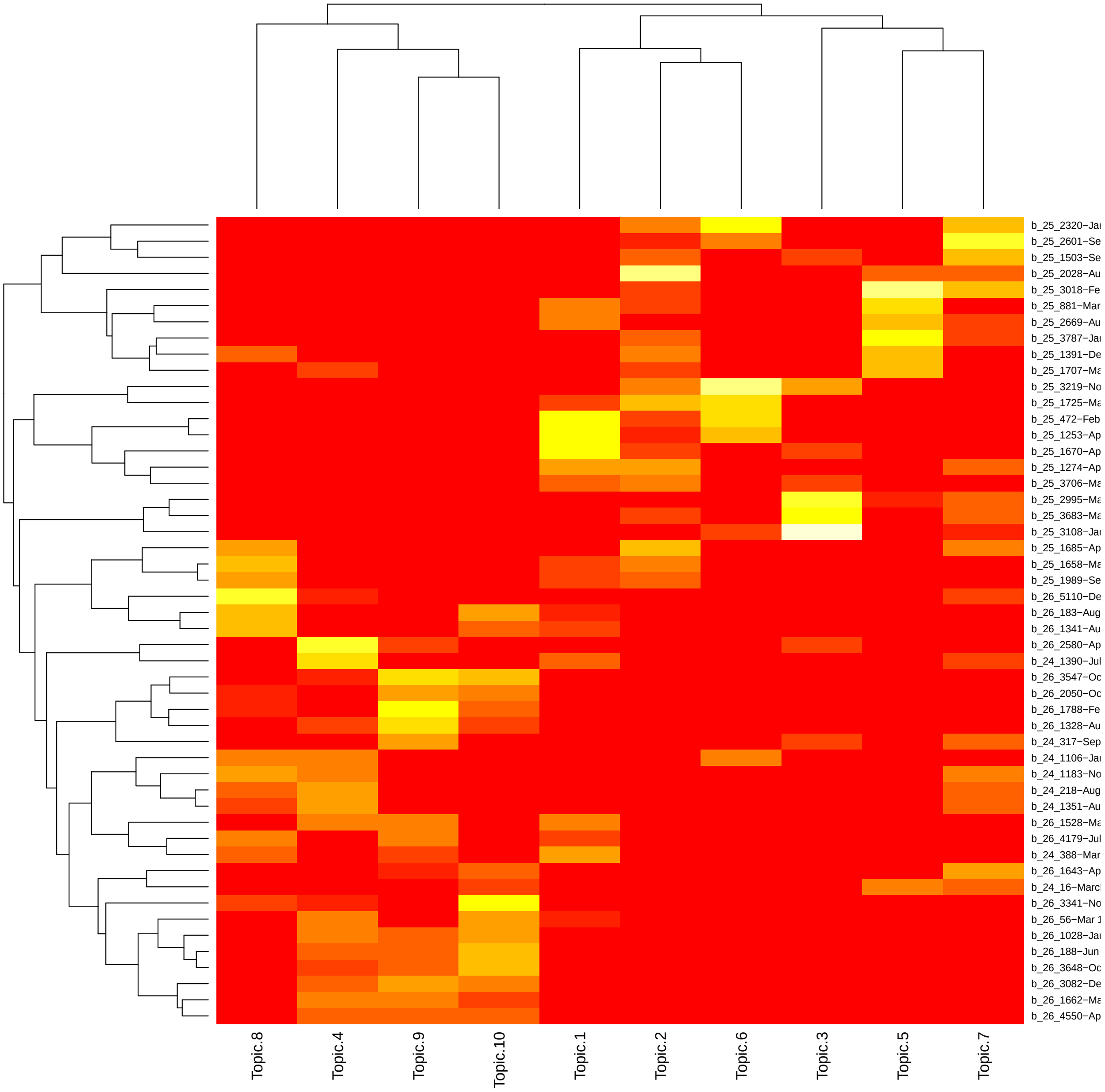
```
##                                  Topic.1 Topic.2 Topic.3 Topic.4 Topic.5
## b_24_218-August 28, 2013.txt       0.000   0.000   0.000   0.335       0
## b_25_1989-September 13, 2009.txt   0.130   0.216   0.000   0.000       0
## b_26_3547-Oct 24.txt               0.000   0.000   0.000   0.076       0
## b_25_1503-September 1, 2010.txt    0.000   0.198   0.132   0.000       0
## b_25_1274-April 3, 2011.txt        0.289   0.325   0.000   0.000       0
## b_25_2320-January 6, 2009.txt      0.000   0.225   0.000   0.000       0
##                                  Topic.6 Topic.7 Topic.8 Topic.9 Topic.10
## b_24_218-August 28, 2013.txt       0.000   0.177   0.163   0.000    0.000
## b_25_1989-September 13, 2009.txt   0.000   0.000   0.313   0.000    0.000
## b_26_3547-Oct 24.txt               0.000   0.000   0.000   0.401    0.355
## b_25_1503-September 1, 2010.txt    0.000   0.288   0.000   0.000    0.000
## b_25_1274-April 3, 2011.txt        0.000   0.164   0.000   0.000    0.000
## b_25_2320-January 6, 2009.txt      0.223   0.283   0.000   0.000    0.000
```

Plot two heatmaps. One, with the estimated topics distributions, and another with some topic values set to zero.

```
sub_scale <- as.matrix(scale(sub_theta))
heatmap(sub_scale, Colv=F, scale='none')
```



```
sub_scale2 <- as.matrix(scale(sub_theta2))
heatmap(sub_scale2, Colv=F, scale='none')
```



From the two plots, we note that by setting some topics values to zero, we get a more separated cluster as shown in the second plot, which has more isolated lighter areas than the first plot. This suggest that for clustering purpose, setting some topics proportions to zero may improve the separation of the documents.