Reading Obama Search on NY Times via Lexis Nexis

Fred Boehm

March 17, 2016

I downloaded the 500 most recent articles that appeared when searching NY Times on Lexis Nexis for the term "Obama".

I downloaded the resulting txt file directly from Lexis Nexis and saved it in the "data" directory.

We now read that file into R.

```
tx <- scan(file = "~/bret-research/twitter/aejmc-manuscript-2016/data/The_New_York_Times2016-03-17_16-0
```

We now partition the object `tx` into separate files.

```
library(wordtools) #install my R package 'wordtools'
tx_list <- split_tx(tx = tx, patt = "Copyright 20")
```

We examined the first 3 articles. It looks like the 8th line is always "LENGTH: _____ words". Let's verify this:

```
library(stringr)
sapply(FUN = function(x)which(str_detect(string = x, pattern = "LENGTH")), X = tx_list)
```

```
##   [1]  8  8  8 10  8  9  8  8  8  9  8  9  8  8  8  8  9  8  8  9 12  8  8
##  [24]  9  8  9  9  8  8  8  8  8  9  8  8  8  9  9  8  9  8  8  8  8  8  8
##  [47]  9  8  8  9  9 11  8  9  8  8  8  9  8  9  8  9  8  8  9  8  8  8  8
##  [70] 11  8 10  8  9 10  8  8 10  8  9  8  8  8  8  9  7  9  7  9  8  7  9
##  [93]  8  9  8  8  8  8  8 10  8  8  8  9  7  8  9  8 10  8  8  8  9  9  9
## [116]  9  8  9  8  8  9  9  9  8  8  9  8  9  9  9  8  8  9  9  9  9  9 10
## [139]  8  9  9  8 11  8  8  8  9  8  9  8  8  9  9  8  8  9  8  9 10  8  9
## [162]  8  9  9  8  9  9  9  9  9  9  9  8  8  9  9 19  8  9  8  9  8  9  9
## [185]  8  9  9  9 11  8  8  9  8  9  9  9  8 10 10  9  9  8  8  9  9  9  9
## [208]  9  8  8  9  9  8  9 10  8  8  8  9  8 12  9  8  9  9  8  9  9  9  9
## [231]  9  9  9  9  8  9  8  9  9  9  9  9  9  9  9 10  8 12  8  9  9  9  8
## [254]  8  8  8  8  9  8  9 10  8  9  8  8  9  9  9 12  8  8  8  8  8  9  9
## [277]  9  9  9  9 10  8  9  9  9 10  8  9  9 12  8  9  8  9  9  9  9 10  8
## [300]  8  8  9  9  9  8  9  9  9  9  9  8  9  9  9  8  8  8  9  9  9  8  9
## [323]  9  9  8  9  8  9  9  9  9  9 10  8  9  9  9  8  9  9  9  9  8  9 12
## [346]  8  9  9  9  8  9  8  9  9  8  9  9  8  9  8  9  9  9  9  9  8  9  8
## [369]  9  9 11  8  8  9  9  9 11  8  8  9  9  8 10  8  9  8  8  9  9  9 11
## [392]  8  8  8  9  9  8  9  9  8  9  8  9  8  9  9 10  8  9  9 10  8  9  9
## [415]  9  9  9  9  9  8  9  9  9  9  9 10  8  8  9  9  9 10  8  9  9  9  8
## [438]  9  9  9  8  9  9  8 10  8  8  8  8  9  8  9  8  8  9  9  9  9  9  9
## [461] 10  8  8  8  8  8  8  8  9  8  8  8 10  8  9  8  9  8  9  9 10 10  8
## [484]  9  8 11  8  8  9  8  8 10  8  9  8  8  8  8  9  8
```

```
# We were wrong!!
tx_list2<- sapply(FUN = function(x)x[-(1:which(str_detect(string = x, pattern = "LENGTH")))], X = tx_lis
```

We now need to remove text lines at the end of each article.

```
tx_list3 <- sapply(FUN = function(x)x[1:(-1 + which(str_detect(string = x, pattern = "URL")))], X = tx_
```

We saw that the code above gives an error, since some articles don't have a URL line.

Let's remove the blog posts then see if remaining articles have URL field.

```
myfun <- function(x, pattern = "Web Blog"){
  collapsed <- paste(x, collapse = " ")
  !stringr::str_detect(collapsed, pattern = pattern)
}


indswb <- sapply(FUN = myfun, X = tx_list2)
indsurl <- sapply(FUN = myfun, pattern = "URL", X = tx_list2)
cbind(indswb, indsurl)
```

```
##        indswb indsurl
##   [1,]   TRUE   FALSE
##   [2,]   TRUE   FALSE
##   [3,]   TRUE   FALSE
##   [4,]   TRUE   FALSE
##   [5,]   TRUE   FALSE
##   [6,]   TRUE   FALSE
##   [7,]   TRUE   FALSE
##   [8,]   TRUE   FALSE
##   [9,]   TRUE   FALSE
##  [10,]   TRUE   FALSE
##  [11,]  FALSE    TRUE
##  [12,]   TRUE   FALSE
##  [13,]   TRUE   FALSE
##  [14,]   TRUE   FALSE
##  [15,]   TRUE   FALSE
##  [16,]  FALSE    TRUE
##  [17,]   TRUE   FALSE
##  [18,]   TRUE   FALSE
##  [19,]   TRUE   FALSE
##  [20,]   TRUE   FALSE
##  [21,]   TRUE   FALSE
##  [22,]   TRUE   FALSE
##  [23,]  FALSE    TRUE
##  [24,]   TRUE   FALSE
##  [25,]  FALSE    TRUE
##  [26,]  FALSE    TRUE
##  [27,]   TRUE   FALSE
##  [28,]   TRUE   FALSE
##  [29,]   TRUE   FALSE
##  [30,]   TRUE   FALSE
##  [31,]   TRUE   FALSE
##  [32,]  FALSE    TRUE
##  [33,]   TRUE   FALSE
##  [34,]   TRUE   FALSE
##  [35,]   TRUE   FALSE
##  [36,]  FALSE    TRUE
```

```
##  [37,]   TRUE   FALSE
##  [38,]   TRUE   FALSE
##  [39,]  FALSE    TRUE
##  [40,]   TRUE   FALSE
##  [41,]   TRUE   FALSE
##  [42,]   TRUE   FALSE
##  [43,]   TRUE   FALSE
##  [44,]   TRUE   FALSE
##  [45,]   TRUE   FALSE
##  [46,]  FALSE    TRUE
##  [47,]   TRUE   FALSE
##  [48,]   TRUE   FALSE
##  [49,]  FALSE    TRUE
##  [50,]  FALSE    TRUE
##  [51,]   TRUE   FALSE
##  [52,]   TRUE   FALSE
##  [53,]   TRUE   FALSE
##  [54,]   TRUE   FALSE
##  [55,]   TRUE   FALSE
##  [56,]   TRUE   FALSE
##  [57,]   TRUE   FALSE
##  [58,]   TRUE   FALSE
##  [59,]  FALSE    TRUE
##  [60,]   TRUE   FALSE
##  [61,]  FALSE    TRUE
##  [62,]   TRUE   FALSE
##  [63,]   TRUE   FALSE
##  [64,]  FALSE    TRUE
##  [65,]   TRUE   FALSE
##  [66,]   TRUE   FALSE
##  [67,]   TRUE   FALSE
##  [68,]   TRUE   FALSE
##  [69,]   TRUE   FALSE
##  [70,]   TRUE   FALSE
##  [71,]   TRUE   FALSE
##  [72,]   TRUE   FALSE
##  [73,]  FALSE    TRUE
##  [74,]   TRUE   FALSE
##  [75,]   TRUE   FALSE
##  [76,]   TRUE   FALSE
##  [77,]   TRUE   FALSE
##  [78,]   TRUE   FALSE
##  [79,]  FALSE    TRUE
##  [80,]   TRUE   FALSE
##  [81,]   TRUE   FALSE
##  [82,]   TRUE   FALSE
##  [83,]   TRUE   FALSE
##  [84,]   TRUE   FALSE
##  [85,]   TRUE   FALSE
##  [86,]  FALSE    TRUE
##  [87,]   TRUE   FALSE
##  [88,]  FALSE    TRUE
##  [89,]   TRUE   FALSE
##  [90,]   TRUE   FALSE
```

```
##  [91,]  FALSE    TRUE
##  [92,]   TRUE   FALSE
##  [93,]   TRUE   FALSE
##  [94,]   TRUE   FALSE
##  [95,]   TRUE   FALSE
##  [96,]   TRUE   FALSE
##  [97,]   TRUE   FALSE
##  [98,]   TRUE   FALSE
##  [99,]  FALSE    TRUE
## [100,]   TRUE   FALSE
## [101,]   TRUE   FALSE
## [102,]   TRUE   FALSE
## [103,]  FALSE    TRUE
## [104,]   TRUE   FALSE
## [105,]   TRUE   FALSE
## [106,]   TRUE   FALSE
## [107,]   TRUE   FALSE
## [108,]   TRUE   FALSE
## [109,]   TRUE   FALSE
## [110,]   TRUE   FALSE
## [111,]   TRUE   FALSE
## [112,]  FALSE    TRUE
## [113,]  FALSE    TRUE
## [114,]  FALSE    TRUE
## [115,]  FALSE    TRUE
## [116,]   TRUE   FALSE
## [117,]   TRUE   FALSE
## [118,]   TRUE   FALSE
## [119,]   TRUE   FALSE
## [120,]   TRUE   FALSE
## [121,]   TRUE   FALSE
## [122,]   TRUE   FALSE
## [123,]   TRUE   FALSE
## [124,]   TRUE   FALSE
## [125,]   TRUE   FALSE
## [126,]   TRUE   FALSE
## [127,]  FALSE    TRUE
## [128,]  FALSE    TRUE
## [129,]  FALSE    TRUE
## [130,]   TRUE   FALSE
## [131,]   TRUE   FALSE
## [132,]  FALSE    TRUE
## [133,]  FALSE    TRUE
## [134,]  FALSE    TRUE
## [135,]  FALSE    TRUE
## [136,]  FALSE    TRUE
## [137,]  FALSE    TRUE
## [138,]   TRUE   FALSE
## [139,]  FALSE    TRUE
## [140,]  FALSE    TRUE
## [141,]   TRUE   FALSE
## [142,]  FALSE    TRUE
## [143,]   TRUE   FALSE
## [144,]   TRUE   FALSE
```

```
## [145,]   TRUE   FALSE
## [146,]  FALSE    TRUE
## [147,]   TRUE   FALSE
## [148,]  FALSE    TRUE
## [149,]   TRUE   FALSE
## [150,]   TRUE   FALSE
## [151,]  FALSE    TRUE
## [152,]  FALSE    TRUE
## [153,]   TRUE   FALSE
## [154,]   TRUE   FALSE
## [155,]  FALSE    TRUE
## [156,]   TRUE   FALSE
## [157,]  FALSE    TRUE
## [158,]  FALSE    TRUE
## [159,]   TRUE   FALSE
## [160,]   TRUE   FALSE
## [161,]   TRUE   FALSE
## [162,]  FALSE    TRUE
## [163,]  FALSE    TRUE
## [164,]   TRUE   FALSE
## [165,]  FALSE    TRUE
## [166,]  FALSE    TRUE
## [167,]  FALSE    TRUE
## [168,]  FALSE    TRUE
## [169,]  FALSE    TRUE
## [170,]  FALSE    TRUE
## [171,]  FALSE    TRUE
## [172,]   TRUE   FALSE
## [173,]   TRUE   FALSE
## [174,]  FALSE    TRUE
## [175,]   TRUE   FALSE
## [176,]   TRUE   FALSE
## [177,]   TRUE   FALSE
## [178,]  FALSE    TRUE
## [179,]   TRUE   FALSE
## [180,]  FALSE    TRUE
## [181,]   TRUE   FALSE
## [182,]  FALSE    TRUE
## [183,]  FALSE    TRUE
## [184,]   TRUE   FALSE
## [185,]  FALSE    TRUE
## [186,]  FALSE    TRUE
## [187,]  FALSE    TRUE
## [188,]  FALSE    TRUE
## [189,]   TRUE   FALSE
## [190,]   TRUE   FALSE
## [191,]  FALSE    TRUE
## [192,]   TRUE   FALSE
## [193,]  FALSE    TRUE
## [194,]  FALSE    TRUE
## [195,]  FALSE    TRUE
## [196,]   TRUE   FALSE
## [197,]  FALSE    TRUE
## [198,]  FALSE    TRUE
```

```
## [199,]   TRUE   FALSE
## [200,]   TRUE   FALSE
## [201,]   TRUE   FALSE
## [202,]   TRUE   FALSE
## [203,]  FALSE    TRUE
## [204,]  FALSE    TRUE
## [205,]  FALSE    TRUE
## [206,]  FALSE    TRUE
## [207,]  FALSE    TRUE
## [208,]   TRUE   FALSE
## [209,]   TRUE   FALSE
## [210,]  FALSE    TRUE
## [211,]  FALSE    TRUE
## [212,]   TRUE   FALSE
## [213,]  FALSE    TRUE
## [214,]  FALSE    TRUE
## [215,]   TRUE   FALSE
## [216,]   TRUE   FALSE
## [217,]   TRUE   FALSE
## [218,]  FALSE    TRUE
## [219,]   TRUE   FALSE
## [220,]  FALSE    TRUE
## [221,]   TRUE   FALSE
## [222,]   TRUE   FALSE
## [223,]  FALSE    TRUE
## [224,]  FALSE    TRUE
## [225,]   TRUE   FALSE
## [226,]  FALSE    TRUE
## [227,]  FALSE    TRUE
## [228,]  FALSE    TRUE
## [229,]  FALSE    TRUE
## [230,]  FALSE    TRUE
## [231,]  FALSE    TRUE
## [232,]  FALSE    TRUE
## [233,]  FALSE    TRUE
## [234,]   TRUE   FALSE
## [235,]  FALSE    TRUE
## [236,]   TRUE   FALSE
## [237,]  FALSE    TRUE
## [238,]  FALSE    TRUE
## [239,]  FALSE    TRUE
## [240,]  FALSE    TRUE
## [241,]  FALSE    TRUE
## [242,]  FALSE    TRUE
## [243,]  FALSE    TRUE
## [244,]  FALSE    TRUE
## [245,]  FALSE    TRUE
## [246,]   TRUE   FALSE
## [247,]   TRUE   FALSE
## [248,]   TRUE   FALSE
## [249,]  FALSE    TRUE
## [250,]  FALSE    TRUE
## [251,]  FALSE    TRUE
## [252,]   TRUE   FALSE
```

```
## [253,]   TRUE   FALSE
## [254,]   TRUE   FALSE
## [255,]   TRUE   FALSE
## [256,]   TRUE   FALSE
## [257,]  FALSE    TRUE
## [258,]   TRUE   FALSE
## [259,]  FALSE    TRUE
## [260,]  FALSE    TRUE
## [261,]   TRUE   FALSE
## [262,]  FALSE    TRUE
## [263,]   TRUE   FALSE
## [264,]   TRUE   FALSE
## [265,]  FALSE    TRUE
## [266,]  FALSE    TRUE
## [267,]  FALSE    TRUE
## [268,]  FALSE    TRUE
## [269,]   TRUE   FALSE
## [270,]   TRUE   FALSE
## [271,]   TRUE   FALSE
## [272,]   TRUE   FALSE
## [273,]   TRUE   FALSE
## [274,]  FALSE    TRUE
## [275,]  FALSE    TRUE
## [276,]  FALSE    TRUE
## [277,]  FALSE    TRUE
## [278,]  FALSE    TRUE
## [279,]  FALSE    TRUE
## [280,]  FALSE    TRUE
## [281,]   TRUE   FALSE
## [282,]  FALSE    TRUE
## [283,]   TRUE   FALSE
## [284,]   TRUE   FALSE
## [285,]   TRUE   FALSE
## [286,]   TRUE   FALSE
## [287,]  FALSE    TRUE
## [288,]  FALSE    TRUE
## [289,]  FALSE    TRUE
## [290,]   TRUE   FALSE
## [291,]  FALSE    TRUE
## [292,]   TRUE   FALSE
## [293,]  FALSE    TRUE
## [294,]  FALSE    TRUE
## [295,]  FALSE    TRUE
## [296,]  FALSE    TRUE
## [297,]  FALSE    TRUE
## [298,]   TRUE   FALSE
## [299,]   TRUE   FALSE
## [300,]   TRUE   FALSE
## [301,]  FALSE    TRUE
## [302,]  FALSE    TRUE
## [303,]   TRUE   FALSE
## [304,]   TRUE   FALSE
## [305,]  FALSE    TRUE
## [306,]  FALSE    TRUE
```

```
## [307,]  FALSE    TRUE
## [308,]  FALSE    TRUE
## [309,]  FALSE    TRUE
## [310,]   TRUE   FALSE
## [311,]  FALSE    TRUE
## [312,]  FALSE    TRUE
## [313,]  FALSE    TRUE
## [314,]   TRUE   FALSE
## [315,]   TRUE   FALSE
## [316,]   TRUE   FALSE
## [317,]  FALSE    TRUE
## [318,]  FALSE    TRUE
## [319,]  FALSE    TRUE
## [320,]   TRUE   FALSE
## [321,]  FALSE    TRUE
## [322,]  FALSE    TRUE
## [323,]   TRUE   FALSE
## [324,]   TRUE   FALSE
## [325,]  FALSE    TRUE
## [326,]   TRUE   FALSE
## [327,]  FALSE    TRUE
## [328,]  FALSE    TRUE
## [329,]  FALSE    TRUE
## [330,]  FALSE    TRUE
## [331,]   TRUE   FALSE
## [332,]   TRUE   FALSE
## [333,]   TRUE   FALSE
## [334,]  FALSE    TRUE
## [335,]  FALSE    TRUE
## [336,]  FALSE    TRUE
## [337,]   TRUE   FALSE
## [338,]  FALSE    TRUE
## [339,]  FALSE    TRUE
## [340,]  FALSE    TRUE
## [341,]  FALSE    TRUE
## [342,]   TRUE   FALSE
## [343,]  FALSE    TRUE
## [344,]  FALSE    TRUE
## [345,]   TRUE   FALSE
## [346,]  FALSE    TRUE
## [347,]  FALSE    TRUE
## [348,]  FALSE    TRUE
## [349,]   TRUE   FALSE
## [350,]  FALSE    TRUE
## [351,]   TRUE   FALSE
## [352,]  FALSE    TRUE
## [353,]   TRUE   FALSE
## [354,]   TRUE   FALSE
## [355,]  FALSE    TRUE
## [356,]  FALSE    TRUE
## [357,]   TRUE   FALSE
## [358,]  FALSE    TRUE
## [359,]   TRUE   FALSE
## [360,]  FALSE    TRUE
```

```
## [361,]  FALSE    TRUE
## [362,]  FALSE    TRUE
## [363,]  FALSE    TRUE
## [364,]  FALSE    TRUE
## [365,]   TRUE   FALSE
## [366,]  FALSE    TRUE
## [367,]   TRUE   FALSE
## [368,]  FALSE    TRUE
## [369,]  FALSE    TRUE
## [370,]  FALSE    TRUE
## [371,]   TRUE   FALSE
## [372,]   TRUE   FALSE
## [373,]  FALSE    TRUE
## [374,]  FALSE    TRUE
## [375,]  FALSE    TRUE
## [376,]  FALSE    TRUE
## [377,]   TRUE   FALSE
## [378,]  FALSE    TRUE
## [379,]  FALSE    TRUE
## [380,]  FALSE    TRUE
## [381,]   TRUE   FALSE
## [382,]  FALSE    TRUE
## [383,]   TRUE   FALSE
## [384,]  FALSE    TRUE
## [385,]   TRUE   FALSE
## [386,]   TRUE   FALSE
## [387,]  FALSE    TRUE
## [388,]  FALSE    TRUE
## [389,]  FALSE    TRUE
## [390,]  FALSE    TRUE
## [391,]   TRUE   FALSE
## [392,]   TRUE   FALSE
## [393,]   TRUE   FALSE
## [394,]  FALSE    TRUE
## [395,]  FALSE    TRUE
## [396,]   TRUE   FALSE
## [397,]  FALSE    TRUE
## [398,]  FALSE    TRUE
## [399,]   TRUE   FALSE
## [400,]   TRUE   FALSE
## [401,]   TRUE   FALSE
## [402,]  FALSE    TRUE
## [403,]   TRUE   FALSE
## [404,]  FALSE    TRUE
## [405,]  FALSE    TRUE
## [406,]  FALSE    TRUE
## [407,]   TRUE   FALSE
## [408,]  FALSE    TRUE
## [409,]  FALSE    TRUE
## [410,]  FALSE    TRUE
## [411,]   TRUE   FALSE
## [412,]  FALSE    TRUE
## [413,]  FALSE    TRUE
## [414,]  FALSE    TRUE
```

```
## [415,]  FALSE    TRUE
## [416,]  FALSE    TRUE
## [417,]  FALSE    TRUE
## [418,]   TRUE   FALSE
## [419,]   TRUE   FALSE
## [420,]  FALSE    TRUE
## [421,]  FALSE    TRUE
## [422,]  FALSE    TRUE
## [423,]  FALSE    TRUE
## [424,]  FALSE    TRUE
## [425,]  FALSE    TRUE
## [426,]   TRUE   FALSE
## [427,]   TRUE   FALSE
## [428,]  FALSE    TRUE
## [429,]  FALSE    TRUE
## [430,]  FALSE    TRUE
## [431,]  FALSE    TRUE
## [432,]   TRUE   FALSE
## [433,]  FALSE    TRUE
## [434,]  FALSE    TRUE
## [435,]  FALSE    TRUE
## [436,]   TRUE   FALSE
## [437,]  FALSE    TRUE
## [438,]  FALSE    TRUE
## [439,]  FALSE    TRUE
## [440,]   TRUE   FALSE
## [441,]  FALSE    TRUE
## [442,]  FALSE    TRUE
## [443,]   TRUE   FALSE
## [444,]  FALSE    TRUE
## [445,]   TRUE   FALSE
## [446,]   TRUE   FALSE
## [447,]   TRUE   FALSE
## [448,]   TRUE   FALSE
## [449,]  FALSE    TRUE
## [450,]   TRUE   FALSE
## [451,]  FALSE    TRUE
## [452,]   TRUE   FALSE
## [453,]  FALSE    TRUE
## [454,]  FALSE    TRUE
## [455,]  FALSE    TRUE
## [456,]  FALSE    TRUE
## [457,]  FALSE    TRUE
## [458,]  FALSE    TRUE
## [459,]  FALSE    TRUE
## [460,]  FALSE    TRUE
## [461,]   TRUE   FALSE
## [462,]   TRUE   FALSE
## [463,]   TRUE   FALSE
## [464,]   TRUE   FALSE
## [465,]   TRUE   FALSE
## [466,]   TRUE   FALSE
## [467,]   TRUE   FALSE
## [468,]   TRUE   FALSE
```

```
## [469,]   TRUE   FALSE
## [470,]   TRUE   FALSE
## [471,]   TRUE   FALSE
## [472,]   TRUE   FALSE
## [473,]   TRUE   FALSE
## [474,]   TRUE   FALSE
## [475,]   TRUE   FALSE
## [476,]   TRUE   FALSE
## [477,]   TRUE   FALSE
## [478,]   TRUE   FALSE
## [479,]   TRUE   FALSE
## [480,]   TRUE   FALSE
## [481,]   TRUE   FALSE
## [482,]   TRUE   FALSE
## [483,]   TRUE   FALSE
## [484,]   TRUE   FALSE
## [485,]  FALSE    TRUE
## [486,]   TRUE   FALSE
## [487,]   TRUE   FALSE
## [488,]   TRUE   FALSE
## [489,]   TRUE   FALSE
## [490,]   TRUE   FALSE
## [491,]   TRUE   FALSE
## [492,]   TRUE   FALSE
## [493,]   TRUE   FALSE
## [494,]   TRUE   FALSE
## [495,]   TRUE   FALSE
## [496,]   TRUE   FALSE
## [497,]   TRUE   FALSE
## [498,]   TRUE   FALSE
## [499,]   TRUE   FALSE
## [500,]   TRUE   FALSE
```

It looks like only the blog posts have no URL field. Let's remove the blog posts and then pipe it to remove the lines that contain URL and any later lines.

```
library(magrittr)
good_art <- tx_list2[indswb] %>%
  sapply(FUN = function(x){x[-(which(str_detect(x, "URL")):length(x))]})
```

Now we need to separate strings into individual words & remove punctuation.

```
library(tm)
```

```
## Loading required package: NLP
```

```
stopwords <- tm::stopwords("SMART")
good2 <- sapply(FUN = function(x)paste(x, collapse = " "), X = good_art)  %>%
  stringr::str_split( pattern = " ") %>%
  sapply(FUN = function(x) gsub("'", "", x)) %>% # remove apostrophes
  sapply(FUN = function(x) gsub("[[:punct:]]", " ", x)) %>%  # replace punctuation with space
  sapply(FUN = function(x) gsub("[[:cntrl:]]", " ", x)) %>% # replace control characters with space
```

```
  sapply(FUN = function(x) gsub("^[[:space:]]+", "", x)) %>% # remove whitespace at beginning of docume
  sapply(FUN = function(x) gsub("[[:space:]]+$", "", x)) %>% # remove whitesp
  sapply(FUN = function(x)tolower(x)) %>%
  sapply(FUN = function(x)x[!(x == "")]) %>% # remove elements that are ""
  sapply(FUN = function(x)x[!(x %in% stopwords)]) # remove stopwords
```

```
# from http://cpsievert.github.io/LDAvis/reviews/reviews.html
# compute the table of terms:
n_min <- 3
term_table <- table(unlist(good2)) %>%
  sort(decreasing = TRUE)
term_table <- term_table[term_table >= n_min]
vocab <- names(term_table)
get_terms <- function(x) {
  index <- match(x, vocab)
  index <- index[!is.na(index)]
  rbind(as.integer(index - 1), as.integer(rep(1, length(index))))
}
documents <- lapply(good2, get_terms)
```

```
# Compute some statistics related to the data set:
D <- length(documents)  # number of documents (2,000)
W <- length(vocab)  # number of terms in the vocab (14,568)
doc.length <- sapply(documents, function(x) sum(x[2, ]))  # number of tokens per document [312, 288, 17
N <- sum(doc.length)  # total number of tokens in the data (546,827)
term.frequency <- as.integer(term_table)  # frequencies of terms in the corpus [8939, 5544, 2411, 2410,
```

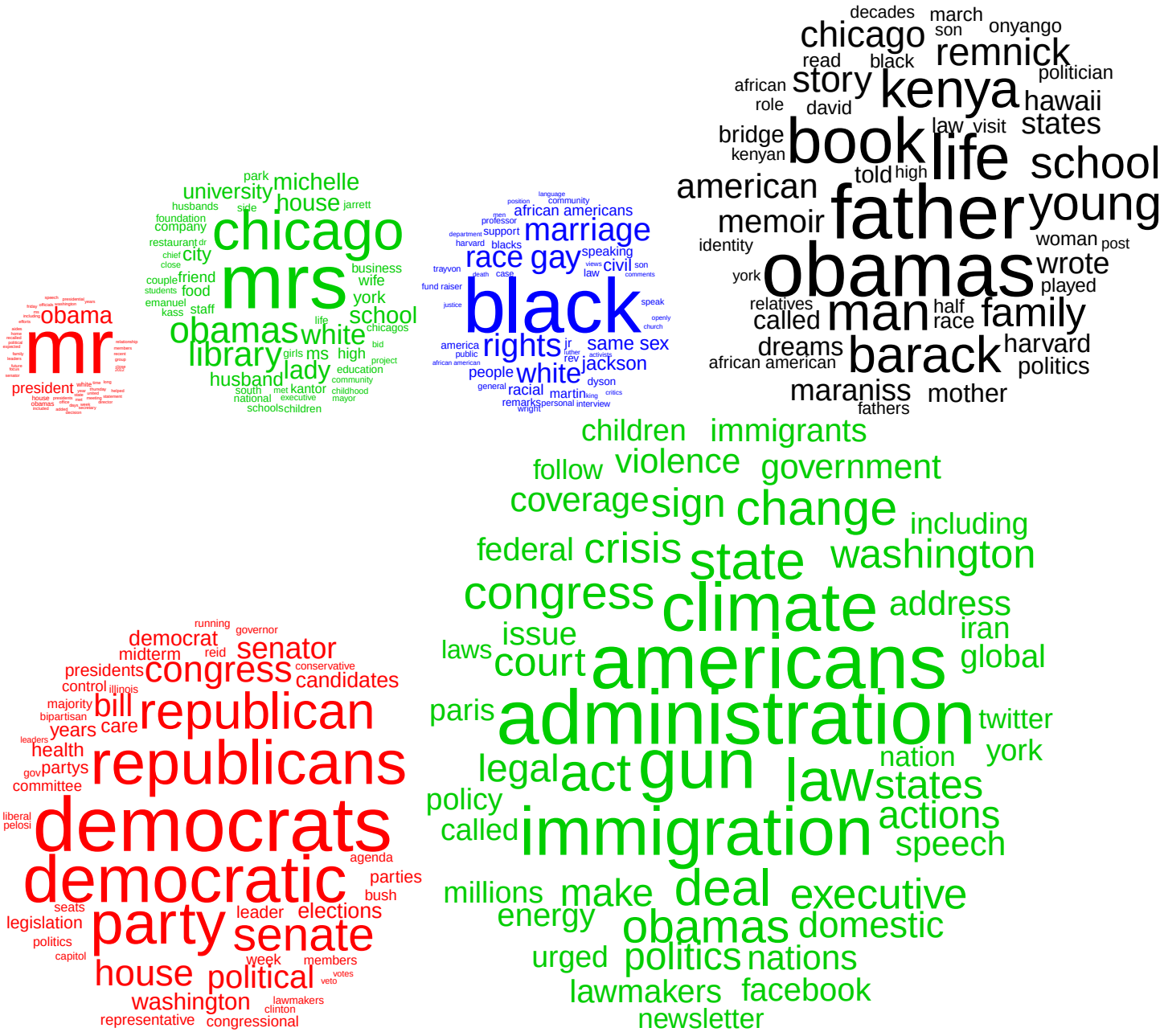
```
# MCMC and model tuning parameters:
K <- 20
G <- 5000
alpha <- 0.02
eta <- 0.02

# Fit the model:
library(lda)
set.seed(357)
t1 <- Sys.time()
fit <- lda.collapsed.gibbs.sampler(documents = documents, K = K, vocab = vocab,
                                   num.iterations = G, alpha = alpha,
                                   eta = eta, initial = NULL, burnin = 0,
                                   compute.log.likelihood = TRUE)
t2 <- Sys.time()
t2 - t1  # about 12 minutes on laptop
```

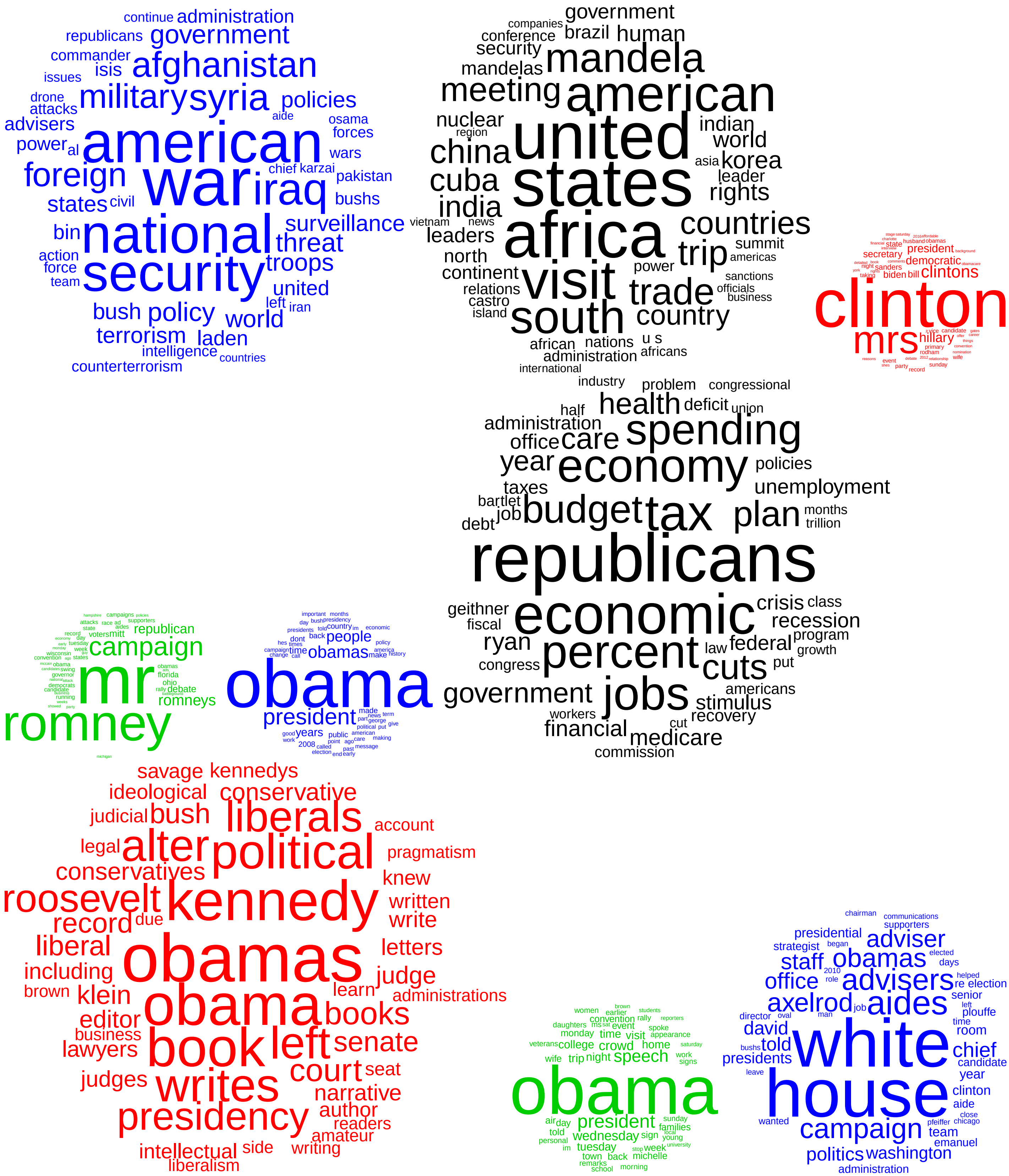
```
## Time difference of 9.28813 mins
```

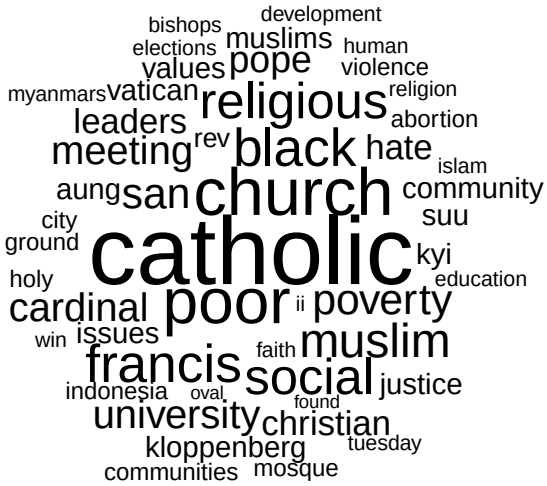
```
library(wordcloud)
```

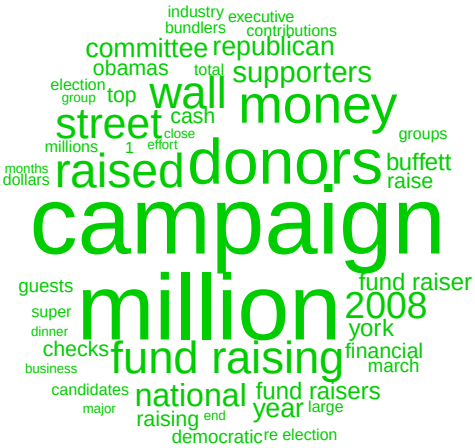
```
## Loading required package: RColorBrewer
```

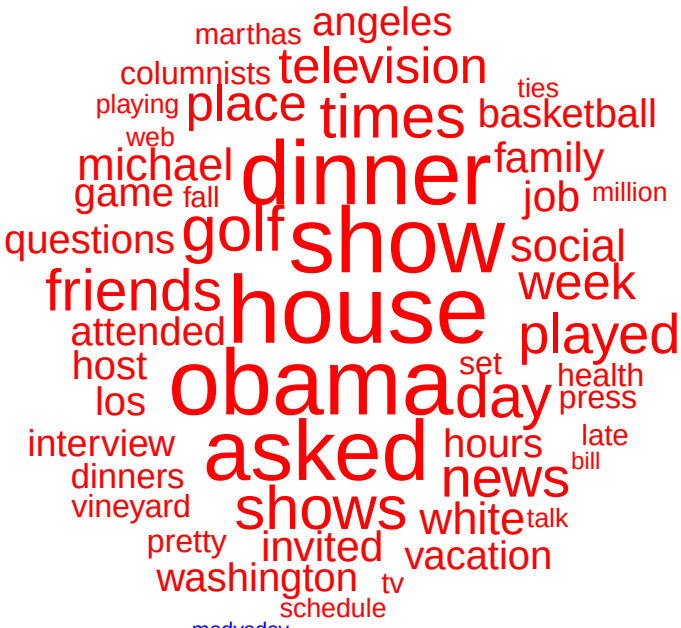
```
for (i in 1:K){
  cloud.data<-sort(fit$topics[i,], decreasing=TRUE)[1:50]
  wordcloud(names(cloud.data), freq=cloud.data, scale=c(3,.10), min.freq=1, rot.per=0, random.order=FALS
}
```



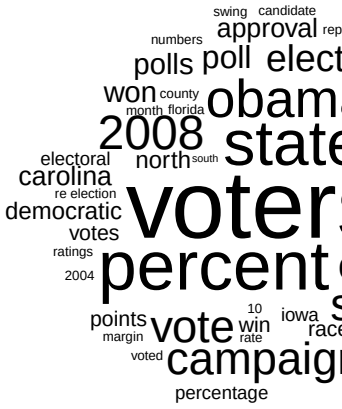

continue administration republicans government commander isis afghanistan issues drone attacks military syria policies aide advisers osama forces power al american wars foreign chief karzai war iraq pakistan states civil bushs bin national surveillance threat action force team security troops united left iran bush policy world terrorism laden intelligence countries counterterrorism
companies government conference brazil human security mandelas mandela meeting american nuclear region united indian world china asia korea cuba states leader rights india vietnam news africa countries leaders summit north trip americas continent power sanctions relations visit trade officials castro business island south country african nations u s administration africans international
state president secretary democratic sanders clintons biden bill clinton mrs hillary event party sunday wife record
industry problem congressional half health deficit union administration spending office care year economy policies taxes unemployment bartlet job budget tax plan months debt trillion republicans geithner economic crisis class fiscal recession ryan program law federal growth congress percent cuts put government jobs americans workers stimulus cut recovery financial medicare commission
campaign republican mitt voters mr obama florida ohio debate romneys romney
people country obamas time obama president years made public
savage kennedys ideological conservative judicial bush liberals account legal alter political pragmatism conservatives knew roosevelt kennedy written record due write liberal letters obamas including judge brown klein learn administrations editor obama books business lawyers book left senate judges court seat writes narrative author presidency readers amateur intellectual side writing liberalism
women brown students earlier convention rally daughters event monday time visit veterans college crowd home wife trip night speech obama air day president told wednesday sign tuesday week town back michelle remarks school morning
chairman communications supporters presidential adviser strategist staff obamas office advisers re election axelrod aides senior director david white room told chief presidents candidate house year wanted clinton aide campaign team emanuel politics washington administration


development
bishops
muslims
elections
human
values
pope
violence
myanmars
vatican
religious
religion
leaders
abortion
meeting
rev
black
hate
islam
aung
san
church
community
city
suu
ground
catholic
kyi
holy
education
cardinal
poor
ii
poverty
win
issues
faith
muslim
francis
social
justice
indonesia
oval
found
university
christian
kloppenberg
tuesday
communities
mosque


industry
executive
bundlers
contributions
committee
republican
obamas
total
supporters
election
group
top
wall
money
street
cash
close
groups
millions
1
effort
months
dollars
raised
donors
buffett
raise
campaign
guests
fund raiser
million
2008
super
dinner
york
checks
fund raising
financial
business
march
candidates
national
fund raisers
major
raising
end
year
large
democratic
re election


marthas
angeles
columnists
television
ties
playing
place
times
basketball
web
michael
dinner
family
game
fall
job
million
questions
golf
show
social
friends
house
week
attended
played
host
obama
set
health
los
day
press
interview
asked
hours
late
bill
dinners
news
vineyard
shows
white
talk
pretty
invited
vacation
washington
tv
schedule


medvedev
britain
nuclear
policy
arms
meeting
security
deal
officials
jewish
leaders
united
foreign
issues
relationship
netanyahu
world
senior
israeli
israel
east
agreement
palestinians
putin
state
secretary
middle
visit
mr
russia
arab
minister
iran
egypt
russian
american
prime
israelis
treaty
moscow
kerry
states
peace
defense
leader
administration
palestinian
talks
ukraine
missile


swing
candidate
numbers
approval
rep
polls
poll
elect
won
county
month
florida
obama
2008
state
north
south
electoral
carolina
re election
democratic
votes
ratings
voter
2004
percent
points
10
vote
win
iowa
margin
rate
voted
campaign
percentage