Preprocessing Text in R

Loren Collingwood

Once you have harvested text, the cleaning and preprocessing stage is next before you can proceed to any serious statistical analysis.

Step 1

Install then load the dplyr and quanteda packages. The latter is the main tool we will work with.

```
    options(scipen = 999, digits = 4)
    #install.packages("dplyr")
    library(dplyr)
```

```
##
## Attaching package: 'dplyr'
## The following objects are masked from 'package:stats':
##
##     filter, lag
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
    #install.packages("quanteda")
    library(quanteda)
```

```
## Package version: 2.1.1
## Parallel computing: 2 of 8 threads used.
## See https://quanteda.io for tutorials and examples.
##
## Attaching package: 'quanteda'
## The following object is masked from 'package:utils':
##
##     View
```

Step 2

Load the Muslim Ban NYTimes article database

```
# Set Directory #
setwd("~/Dropbox/collingwood_research/posc_fall_20/POSC-207/lecture/")

load("nyt_muslim_ban_2017.RData")
```

```
# Subset to relevant columns #
nyts <- dplyr::select(nyt_muslim_ban_2017, response.docs.pub_date,
                      response.docs.headline.main,
                      response.docs.web_url,
                      response.docs.byline.original,
                      response.docs.section_name,
                      response.docs.word_count,
                      text)
```

Step 3

Generate a corpus to hold all your text

```
# Convert Text into a corpus #
nyt_corpus <- corpus(nyts$text)

# Look at the texts in the corpus; in this case text document 3

texts(nyt_corpus)[3]
```

```
##
## "Advertisement Supported by By Nicholas Kristof Whenever an extremist in the Muslim world does someth
```

```
#####################################################
# Create new docvar onto the data.frame summary()  #
#####################################################

docvars(nyt_corpus, "headline") <- nyts$response.docs.headline.main
docvars(nyt_corpus, "date") <- as.Date(nyts$response.docs.pub_date)
docvars(nyt_corpus, "author") <- nyts$response.docs.byline.original

# Look at Number of: Types, Tokens, and Sentences #
# Look at top 6 documents                         #
head(summary(nyt_corpus))
```

```
##    Text Types Tokens Sentences
## 1 text1   220    435        19
## 2 text2    25     46         1
## 3 text3   475    984        50
## 4 text4   622   1799        75
## 5 text5   489   1055        43
## 6 text6   647   1665        61
##                                                                    headline
## 1 China Bans ‘Muhammad’ and ‘Jihad’ as Baby Names in Heavily Muslim Region
## 2                                        Muslim World Denounces Travel Ban
## 3                                  An Apology to Muslims for President Trump
## 4                                        Traveling to America While Muslim
## 5                                     Quebec’s Anti-Muslim Ban on the Veil
## 6         Fears That Trump’s Visa Ban Betrays Friends and Bolsters Enemies
##         date                author
## 1 2017-04-25 By Javier C. Hernández
## 2 2017-01-30 By AINARA TIEFENTHÄLER
## 3 2017-02-02    By Nicholas Kristof
## 4 2017-07-21      By Ismail Einashe
```

```
## 5 2017-11-08    By Martin Patriquin
## 6 2017-01-28        By Declan Walsh
```

```
# Subset corpus to keep only articles by Adam Liptak #
adam_corp <- subset(nyt_corpus, nyt_corpus$author=="By Adam Liptak")
summary(adam_corp)
```

```
## Corpus consisting of 9 documents, showing 9 documents:
##
##    Text Types Tokens Sentences
##  text18   396   1007        43
##  text25   296    704        30
##  text29   338    757        29
##  text42   308    717        28
##  text53   537   1525        73
##  text54   525   1354        52
##  text61   467   1159        50
##  text74   436   1071        41
##  text78   619   1732        72
##                                                                  headline
##                       The Supreme Court’s Options in the Travel Ban Case
##                     Trump Asks Supreme Court to Dismiss Travel Ban Cases
##              Trump Administration Asks Supreme Court to Revive Travel Ban
##              Supreme Court Dismisses Appeal of Case on Expired Travel Ban
##               Appeals Court Panel Appears Skeptical of Trump’s Travel Ban
##               Appeals Court Will Not Reinstate Trump’s Revised Travel Ban
##    3 Judges Weigh Trump’s Revised Travel Ban, but Keep Their Poker Faces
##                     Supreme Court Allows Trump Travel Ban to Take Effect
##  Court Refuses to Reinstate Travel Ban, Dealing Trump Another Legal Loss
##        date         author
##  2017-06-02 By Adam Liptak
##  2017-10-05 By Adam Liptak
##  2017-06-02 By Adam Liptak
##  2017-10-11 By Adam Liptak
##  2017-02-08 By Adam Liptak
##  2017-05-25 By Adam Liptak
##  2017-05-15 By Adam Liptak
##  2017-12-04 By Adam Liptak
##  2017-02-09 By Adam Liptak
```

```
# Tokenize words then sentences #

#tokenize_word(nyt_corpus)
tokenize_sentence(nyt_corpus[1])
```

```
## $text1
##  [1] "Advertisement Supported by By Javier C."
##  [2] "Hernández BEIJING - The Chinese government, further tightening its grip on Muslims in western C
##  [3] "Officials described the ban, introduced this month, as part of an effort to "curb religious fer
##  [4] "The government considers Xinjiang a hotbed of Islamic extremism, violence and separatist thoug
##  [5] "But many Uighurs say the government’s strict limits on worship and speech are responsible for t
##  [6] "The list of names, a copy of which was provided to The New York Times by Uighur activists, is t
##  [7] "It bans more than two dozen names, including "Mujahid" and "Medina.""
##  [8] "Security officials in Urumqi and other cities in Xinjiang confirmed the ban."
##  [9] "Some said in interviews that if residents did not comply, they risked forfeiting critical benef
## [10] "Rights advocates said the ban showed the lengths to which the government would go to limit the
```

```
## [11] ""China's policies are increasingly hostile," said Dilxat Raxit, a spokesman for the World Uyghu
## [12] ""Uighur people have to be cautious if they want to give their children names they are happy wi
## [13] "Sophie Richardson, the China director of Human Rights Watch, said that choosing baby names shou
## [14] ""This is the latest absurd restriction that the Chinese government has imposed on people in Xi
## [15] "To combat what officials describe as extremism in Xinjiang, the Chinese government has put in p
## [16] "Earlier this month, for example, security officials imposed bans on long beards and veils in pu
## [17] "This year, officials held large rallies of paramilitary and police forces as a show of force i
## [18] "The region has struggled with clashes between residents and security officials and occasional a
## [19] "Advertisement"
```

Step 4

Create a document term (or frequency) matrix. This is a very important step and requires some art in the cleaning process. In general, you should aim to thin the matrix as much as possible without losing any useful/useable information.

```
###################################
# Document Frequency/Term Matrix #
###################################

# Convert text to lower, remove stopwords, stem words, and remove punctution #
# Rows are the 'Documents', columns are 'features'
nyt_dfm <- dfm(nyt_corpus,
               tolower=T,
               remove = stopwords("english"),
               stem = T,
               verbose = T,
               remove_punct = T
               )
```

```
## Creating a dfm from a corpus input...
##  ...lowercasing
##  ...found 90 documents, 8,733 features
##  ...removed 155 features
##  ...stemming types (English)
##  ...complete, elapsed time: 0.305 seconds.
## Finished constructing a 90 x 5,871 sparse dfm.
```

```
# Look at the feature frequency counts, etc.
head ( featfreq(nyt_dfm) )
```

```
##  advertis   support    javier         c hernández      beij
##       175       153         1         4         1         1
```

Step 5

Now you want to trim the matrix even more (real good).

```
#############################################
# Trim the document -- 20, 5 are arbitrary #
#############################################
```

```
# Word has to show up at least 20 times and at least in 5 unique documents #
smalldfm <- dfm_trim(nyt_dfm,minCount=20,minDoc=5)
smalldfm <- dfm_trim(nyt_dfm, sparsity = 0.8)

# Print it out to see what it looks like
smalldfm
```

```
## Document-feature matrix of: 90 documents, 349 features (65.5% sparse) and 3 docvars.
##        features
## docs    advertis support govern muslim prohibit name like offici ban month
##   text1        2       1      7      2        1    7    1      6   6     2
##   text2        1       0      0      1        0    0    0      0   1     0
##   text3        2       1      0     13        0    0    0      0   4     0
##   text4        2       1      0     16        1    0    5      5  11     4
##   text5        2       1      6     12        1    0    2      1   8     1
##   text6        2       1      1     21        1    0    2      1   6     0
## [ reached max_ndoc ... 84 more documents, reached max_nfeat ... 339 more features ]
```

```
# Look at the top features real good #
topfeatures(smalldfm, n = 50)
```

```
##      said       ban        mr    muslim     trump     state     court     order
##       700       680       679       640       587       571       498       481
##   countri      unit    presid    travel       new    refuge    nation     peopl
##       454       440       370       350       235       234       230       228
##    execut    immigr     secur administr    govern      judg  advertis      issu
##       200       190       183       183       181       176       175       171
##      rule       one    offici   support      also     group      like      case
##       167       164       162       153       150       142       141       138
##  american    appeal     legal       law    religi      mani       say      year
##       137       137       133       132       131       127       127       125
##     decis     islam    suprem     women    justic     polit     feder      call
##       122       121       121       115       109       109       109       106
##     first     right
##       106       105
```

```
# Further drop out words that don't tell us much #
# I typically do this iteratively using 'common. sense.' #

smalldfm2 <- dfm_select(smalldfm,
                        pattern = c("said", "mr", "advertis", "also", "like",
                                    "say", "call", "now", "can", "ms", "make",
                                    "just", "may", "go", "ask", "use", "way",
                                    "still", "wrote", "week", "want", "sinc",
                                    "parti", "ad", "version", "argu", "made",
                                    "alreadi", "whether", "might", "came", "see",
                                    "need", "set", "tri", "accord", "show",
                                    "yet", "well", "anoth", "must", "mean", "put",
                                    "turn", "refer", "previous", "becom", "decid",
                                    "rather", "expect", "good", "part", "effort",
                                    "consid", "often", "today", "mani"),
                        selection= "remove")

topfeatures(smalldfm2, n = 50)
```

```
##       ban    muslim     trump     state     court     order   countri      unit
##       680       640       587       571       498       481       454       440
##    presid    travel       new    refuge    nation     peopl    execut    immigr
##       370       350       235       234       230       228       200       190
##     secur administr    govern      judg      issu      rule       one    offici
##       183       183       181       176       171       167       164       162
##   support     group      case  american    appeal     legal       law    religi
##       153       142       138       137       137       133       132       131
##      year     decis     islam    suprem     women    justic     polit     feder
##       125       122       121       121       115       109       109       109
##     first     right    includ    polici      time       day      visa  restrict
##       106       105       103       103       100        99        97        94
##     enter     block
##        94        91
```

Step 6

Take a look at feature co-occurence. This is sort of akin to descriptive statistics.

```
################################
# Feature Co-Occurrence Matrix #
################################

nyt_co <- fcm(smalldfm2,
              ordered=T)

# Convert this to  data frame #
nyt_dat <- convert(nyt_co, to = "data.frame")

muslim_vec <- nyt_dat[3,]
muslim_vec <- muslim_vec[-1]

# top 50 co-occurring words with muslim
rev(sort(muslim_vec))[1:50]
```

```
##   muslim state trump  ban countri court unit order women presid travel  law
## 3   5046  4308  4000 3962    3326  3197 3078  2919  2625   2332   2159 1851
##   peopl  new islam refuge nation group right  one year religi immigr rule
## 3  1831 1738  1586   1530   1514  1487  1472 1435 1426   1389   1292 1207
##   execut administr issu american offici secur polit judg case religion even
## 3   1172      1115 1080     1075   1036  1025   994  970  964      963  955
##   time legal suprem includ citizen polici constitut terrorist world public last
## 3  887   885    878    831     816    797       796       781   753    746  731
##   visa come attack day
## 3  729  715    705 702
```

Step 7

Understanding term frequency, inverse document frequencies. Basically, the higher the number the more distinct that word is to that particular document. The Tf-idf will become very important later (although under the hood).

```
### Tf-idf weighting ###
# The higher the number the more unique that word is to that document #

tf.idf <- dfm_tfidf(smalldfm2,
                    scheme_tf = 'prop')

tf.idf
```

```
## Document-feature matrix of: 90 documents, 292 features (66.2% sparse) and 3 docvars.
##        features
## docs       support    govern    muslim prohibit    name    offici ban    month
##   text1 0.00032219 0.0173093 0.0007562 0.006579 0.04917 0.0182029   0 0.006268
##   text2 0          0         0.0070329 0        0       0           0 0
##   text3 0.00018496 0         0.0028218 0        0       0           0 0
##   text4 0.00008277 0         0.0015542 0.001690 0       0.0038970   0 0.003221
##   text5 0.00017835 0.0082131 0.0025117 0.003642 0       0.0016794   0 0.001735
##   text6 0.00007783 0.0005973 0.0019181 0.001589 0       0.0007328   0 0
##        features
## docs      religi     home
##   text1 0.003034 0.006796
##   text2 0        0
##   text3 0        0.003901
##   text4 0        0.001746
##   text5 0.006718 0
##   text6 0.001466 0.001642
## [ reached max_ndoc ... 84 more documents, reached max_nfeat ... 282 more features ]
```

```
# Term Frequency: Frequency of times word shows up in document divided by
# total number of words in the document

docfreq(smalldfm2)
```

```
##      support       govern       muslim     prohibit         name       offici
##           84           53           83           22           20           47
##          ban        month       religi         home        minor        group
##           90           46           47           21           20           47
##        islam       extrem      violenc        limit         list       provid
##           40           18           18           22           26           21
##          new         time          two       includ        secur         citi
##           66           57           47           52           49           26
##        resid         risk       critic         care        right        civil
##           21           18           36           19           44           26
##      liberti        fight       terror       polici        world        peopl
##           19           24           29           37           41           72
##         give     director        human       latest     restrict        impos
##           22           28           22           18           31           23
##        place       recent         year      earlier         long       public
##           43           32           52           27           29           34
##         larg         forc          act       travel         take     american
##           23           20           24           54           52           47
##           us       presid        trump      america         iraq     campaign
##           33           69           70           32           34           41
##       around          bar         unit        state        women      univers
##           25           38           68           73           20           36
##          one          far        anyon     challeng      problem        obama
```

```
##          68          19          19          41          18          18
##         vet        issu      effect        took      target       seven
##          22          61          41          24          29          42
##      countri      person    terrorist      attack          11       order
##          79          29          38          39          23          64
##        help       legal     certain       yemen       other      differ
##          22          42          20          31          28          28
##      threat     similar      nation       advis        hear    religion
##          28          19          69          21          26          44
##       found      refuge      intern        fear      immigr        full
##          32          53          27          20          60          18
##        come       admit        rais       studi     somalia      import
##          49          20          18          24          32          26
##       refus      direct       thing        work       stand      happen
##          19          26          21          38          23          21
##     announc      leader         day       faith     airport     protest
##          27          25          51          19          27          26
##    christian        plan        last       visit    administr     visitor
##          27          24          48          20          53          18
##         six     predomin      report      number     concern     general
##          35          38          34          31          35          35
##       organ       among     citizen muslim-major        feel         get
##          23          24          36          22          19          21
##      suprem       court        rule    temporari       block       lower
##          30          53          41          23          44          18
##        live     question         end      follow      enforc      origin
##          24          27          24          27          22          26
##      execut       enter       allow       relat        move        back
##          56          46          43          26          31          26
##      famili        war        later       entri      center        visa
##          36          22          24          32          22          30
##        even      return       march        know        case        face
##          45          20          24          18          43          34
##        keep     individu        view        2015        told       major
##          19          24          25          18          22          38
##       north        lead     protect        head        stop       chang
##          23          24          38          20          19          24
##        open        busi        meet        hold       appli       first
##          26          23          22          21          26          53
##      second      reject       point       posit     without       remov
##          22          18          29          21          27          18
##       cover        much       debat         law      receiv      reason
##          21          25          20          38          18          20
##        pass       least     express        valu       elect    institut
##          21          21          22          20          25          26
##      though      remain      justic        sign      measur        file
##          25          23          32          32          22          21
##      violat        iran        news      across     complet      friday
##          25          37          26          21          24          18
##       think      former       sever      action        seek      promis
##          25          23          36          33          22          20
##      affect       syria       sudan       libya          90     potenti
##          30          41          23          30          19          18
##    indefinit    discrimin     continu        leav       union        kill
```

```
##           20           41           26           22           20           19
##        polit         base      foreign     argument      suggest      process
##           45           30           27           28           20           24
##      program        whose       screen         hous        three         seem
##           22           23           18           36           30           22
##       member    communiti        littl         sept        decis        white
##           33           19           24           18           40           36
##        offic       depart     district    statement         look       monday
##           36           34           21           35           28           23
##       repres        close       author        clear     interest       defend
##           27           19           25           24           20           19
##         note       appear        feder         judg   washington       appeal
##           21           26           39           30           28           33
##       perman         find    constitut      attempt        revis       matter
##           18           24           35           19           19           18
##        power      twitter      instead        addit    establish       inform
##           30           18           23           18           19           20
##       review        reach     attorney        brief       lawyer        claim
##           26           19           21           20           26           21
##   presidenti     homeland       requir     thursday
##           20           21           21           19
```

```
doc_df <- convert(smalldfm2, "data.frame")

# Let's just look at the first textual document #
doc1 <- doc_df[1,]

# Drop the first doc_id entity
doc1 <- doc1[-1]

# Calculate tf #
tf <- doc_df$govern[1]/sum(doc1)

# Calculate idf #
n <- nrow(doc_df)
s <- table( ifelse(doc_df$govern >= 1, 1, 0) )[2]
idf <- log10(n/s)

# Calculate tf-idf #
tf*idf
```

```
##       1
## 0.01731
```