Creating DFM for Mission and Programs

Just using Quanteda at this point

```
library(quanteda)
```

```
## Package version: 1.3.14
```

```
## Parallel computing: 2 of 4 threads used.
```

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

```
##
## Attaching package: 'quanteda'
```

```
## The following object is masked from 'package:utils':
##
##     View
```

Mission

First doing mission with more explanation, before re-running code on programs

```
mission <- read.csv("~/Dropbox (ASU)/USC Mission Paper/Data and Analysis/Sample Framework/DATA/MISSION.csv")
```

Let's see how many unique missions we have in the data

```
nrow(mission)
```

```
## [1] 3446
```

```
length(unique(mission$EIN))
```

```
## [1] 2135
```

```
length(unique(mission$EIN, mission$TAXYR))
```

```
## [1] 2135
```

```
length(unique(mission$F9_03_PZ_MISSION))
```

```
## [1] 2296
```

A bit of overlap, I'm going to remove repeated missions since they wont add any additional information. We may need to look closer at the ones that filed multiple forms in the same year

```
mission <- mission[!duplicated(mission[c('EIN', 'F9_03_PZ_MISSION')]),]
mission <- mission[!duplicated(mission[c('EIN', 'TAXYR')]),]
nrow(mission)
```

```
## [1] 2334
```

Removing extra varaibles to create a smaller dataset, can be expanded in the future.

```
mission.lim <- mission[, c("EIN","F9_03_PZ_MISSION")]
```

In addition, need to ensure all varaibles are characters in order to change to corpus

```
mission.lim <- data.frame(lapply(mission.lim, as.character), stringsAsFactors=FALSE)
```

Converting data to a corpus using 'corpus' command from quanteda, text_field indicates which column holds the text data we want to analyze. Also creating a label for each listing in order ot ensure the data is labeled thorugh to the end of the analysis.

```
mission.corp <- corpus(mission.lim,
                       text_field = "F9_03_PZ_MISSION")
docid <- paste(mission$EIN,
               mission$TAXYR,  sep = " ")
docnames(mission.corp) <- docid
summary(mission.corp, 5)
```

```
## Corpus consisting of 2334 documents, showing 5 documents:
##
##           Text Types Tokens Sentences      EIN
##  10716217 2016    35     50         1 10716217
##  10842551 2015     6      6         1 10842551
##  20792368 2011     4      4         1 20792368
##  30555726 2015     4      4         1 30555726
##  43611860 2016    26     33         1 43611860
##
## Source: /Users/ericholm/Dropbox (ASU)/* on x86_64 by ericholm
## Created: Fri Feb  8 13:26:21 2019
## Notes:
```

We can also add some inforamtion about the data as part of this step, particularly if we plan to publish the corpus online

```
metadoc(mission.corp, "docsource") <- "IRS EZ forms"
metadoc(mission.corp, "notes") <- "caveat emptor"
metadoc(mission.corp, "citation") <- "Lecy et. al,"
```

We can lok at the corpus to see how it's structured

```
mission.corp
```

```
## Corpus consisting of 2,334 documents and 1 docvar.
```

```
mission.corp[1]
```

```
##

                                                                                                             107
16217 2016
## "THE CORPORATION'S SPECIFIC PURPOSE IS TO SUPPORTS AFFORDABLE HOUSING, COMMUNITY DEVELOPMENT AND ECONOMIC DEVE
LOPMENT OF THE CITY AND COUNTY OF SAN FRANCISCO'S ECONOMICALLY DISADVANTAGED INDIVIDUALS AND COMMUNITIES, BY LEND
ING TO, INVESTING IN, AND DIRECTLY ACQUIRING SUCH AFFORDABLE HOUSING AND RELATED COMMUNITY DEVELOPMENT REAL ESTAT
E ASSETS."
```

```
summary(mission.corp)[1:10,]
```

```
##             Text Types Tokens Sentences      EIN
## 1  10716217 2016    35     50         1 10716217
## 2  10842551 2015     6      6         1 10842551
## 3  20792368 2011     4      4         1 20792368
## 4  30555726 2015     4      4         1 30555726
## 5  43611860 2016    26     33         1 43611860
## 6  43771703 2017    51     85         3 43771703
## 7  50549622 2016     6      6         1 50549622
## 8  50581787 2016     6      6         1 50581787
## 9  50618564 2015    30     34         1 50618564
## 10 61582376 2016    16     19         1 61582376
```

Preprocessing steps from last week, making lower case, tokenizing into words and removing stop words. Adding in padding between words where stopwords are to prevent finding artificial Ngrams

```
mission.corp2 <- tolower(mission.corp)
mission.corp3 <- tokens(mission.corp2, remove_punct = TRUE)
mission.corp4 <- tokens_remove(tokens(mission.corp3), stopwords("english"), padding  = TRUE)
```

Now looking at Ngrams. Looking for combinations of 2 and 3 words. I've exported the lists that were produces for us all to look over to decide what we want to capture into a dictionary. This code can be updated once we have a larger list.

```
myNgram2 <- tokens(mission.corp4) %>%
  tokens_ngrams(n = 2) %>%
  dfm()
myNgram3 <- tokens(mission.corp4) %>%
  tokens_ngrams(n = 3) %>%
  dfm()
topfeatures(myNgram2)
```

```
##             501_c               c_3       high_school     mental_health
##                53                53                52                32
##     united_states       raise_funds      young_people provide_financial
##                30                26                24                23
##      jesus_christ   local_community
##                23                22
```

```
topfeatures(myNgram3)
```

```
##                       501_c_3                  section_501_c
##                            53                             19
##          internal_revenue_code                see_schedule_o
##                            15                             11
##                 c_3_non-profit           high_school_students
##                            10                              8
##   provide_financial_assistance      provide_financial_support
##                             8                              7
##            low_income_families science_technology_engineering
##                             7                              7
```

```
my_dict <- dictionary(list(five01_c_3= c("501 c 3","section_501_c") ,
                           united_states = "united states"))
mission.corp5 <- tokens_compound(mission.corp4, pattern = my_dict)
```

Now removing teh extra white space created earlier

```
mission.corp6 <- sapply(mission.corp5, paste, collapse=" ")
```

converting to a document frequency matrix as a final step, and removing stems.

```
mission.dfm <- dfm(mission.corp6,
                   stem = T)
```

This is from quanteda's website, so this should be what we need for the next stemps.

"Once constructed, a quanteda "dfm"" can be easily passed to other text-analysis packages for additional analysis of topic models or scaling, such as:

topic models (including converters for direct use with the topicmodels, LDA, and stm packages)

document scaling (using quanteda's own functions for the "wordfish" and "Wordscores" models, direct use with the ca package for correspondence analysis, or scaling with the austin package)

document classification methods, using (for example) Naive Bayes, k-nearest neighbour, or Support Vector Machines

more sophisticated machine learning through a variety of other packages that take matrix or matrix-like inputs.

graphical analysis, including word clouds and strip plots for selected themes or words.""

```
mission.dfm
```

```
## Document-feature matrix of: 2,334 documents, 4,828 features (99.7% sparse).
```

```
topfeatures(mission.dfm, 20)
```

```
##    provid      educ communiti     organ   support    promot   mission
##       922       692       611       528       496       340       311
##   program    servic      help  children   develop    famili      nbsp
##       295       273       262       260       258       257       249
##      need    purpos     youth    school    assist     activ
##       242       241       236       217       201       192
```

And exporting for our future uses/training

```
mission.dfm.df <- convert(mission.dfm, to = "data.frame")
mission.corpus.df <- as.data.frame(mission.corp6, row.names=docid)
```

Now doing the same with Program

but with less explanation

```
programs <- read.csv("~/Dropbox (ASU)/USC Mission Paper/Data and Analysis/Sample Framework/DATA/PROGRAMS.csv")
nrow(programs)
```

```
## [1] 4346
```

```
length(unique(programs$DESCRIPTION))
```

```
## [1] 3121
```

```
programs <- programs[!duplicated(programs[c('EIN', 'DESCRIPTION')]),]
programs <- programs[!duplicated(programs[c('EIN', 'TAXYR')]),]
programs.lim <- programs[, c("EIN","DESCRIPTION")]
programs.lim <- data.frame(lapply(programs.lim, as.character), stringsAsFactors=FALSE)
```

```
programs.corp <- corpus(programs.lim,
                        text_field = "DESCRIPTION")
docid <- paste(programs$EIN,
               programs$TAXYR,  sep = " ")
docnames(programs.corp) <- docid
summary(programs.corp, 5)
```

```
## Corpus consisting of 2536 documents, showing 5 documents:
##
##           Text Types Tokens Sentences      EIN
##  10716217 2016   112    217         7 10716217
##  10842551 2015    20     28         3 10842551
##  20792368 2011     4      4         1 20792368
##  30555726 2016     8      8         1 30555726
##  43611860 2016    17     17         1 43611860
##
## Source: /Users/ericholm/Dropbox (ASU)/* on x86_64 by ericholm
## Created: Fri Feb  8 13:26:24 2019
## Notes:
```

```
programs.corp
```

```
## Corpus consisting of 2,536 documents and 1 docvar.
```

```
programs.corp[1]
```

```
##
```

```
                   10716217 2016
## "SFHAF CLOSED A LOAN TO BRIDGE HOUSING CORPORATION FOR A VACANT SITE AT 4840 MISSION STREET IN SAN FRANCISCO.
THE LOAN WAS CLOSED ON JUNE 7, 2017 AND TOTALED $9.0M. THE LOAN PROCEEDS WILL BE USED FOR 175 NEW AFFORDABLE HOUS
ING UNITS AND A GROUND FLOOR COMMUNITY HEALTH CLINIC. WITHOUT THE FUND, BRIDGE'S PURCHASE AGREEMENT ON THIS STRAT
EGIC SITE WOULD HAVE EXPIRED.SFHAF CLOSED A LOAN TO MISSION ECONOMIC DEVELOPMENT AGENCY (MEDA) FOR ACQUISITION AN
D REHABILITATION OF A 6-UNIT PROPERTY AT 1411 FLORIDA STREET IN SAN FRANCISCO. THIS LOAN WAS CLOSED ON MAY 24, 20
17 AND TOTALED $3.5M. THE LOAN PROCEEDS WILL BE USED TO REHABILITATE AND PRESERVE 6 AFFORDABLE HOUSING UNITS, AND
BUILD AN ADDITIONAL ACCESSORY DWELLING UNIT (ADU). THE ADU IS A FIRST FOR THE CITY'S SMALL SITES PROGRAM, ALLOWIN
G LONGTIME ELDERLY TENANTS TO AGE IN PLACE IN A NEW GROUND FLOOR APARTMENT.SFHAF CLOSED ITS FIRST ROUND OF CAPITA
L IN APRIL, $37 MILLION IN TOTAL LED BY INVESTMENTS OF $20 MILLION FROM CITI COMMUNITY CAPITAL, $10 MILLION FROM
THE CITY OF SAN FRANCISCO, AND $6.5 MILLION PHILANTHROPIC CAPITAL FROM DIGNITY HEALTH, THE SAN FRANCISCO FOUNDATI
ON, AND THE HEWLETT FOUNDATION."
```

```
summary(programs.corp)[1:10,]
```

```
##              Text Types Tokens Sentences      EIN
## 1  10716217 2016   112    217         7 10716217
## 2  10842551 2015    20     28         3 10842551
## 3  20792368 2011     4      4         1 20792368
## 4  30555726 2016     8      8         1 30555726
## 5  43611860 2016    17     17         1 43611860
## 6  43771703 2017    23     28         2 43771703
## 7  50549622 2016    11     12         1 50549622
## 8  50581787 2016     2      2         1 50581787
## 9  50618564 2015    21     28         1 50618564
## 10 61582376 2016    15     18         1 61582376
```

```
programs.corp2 <- tolower(programs.corp)
programs.corp3 <- tokens(programs.corp2, remove_punct = TRUE)
programs.corp4 <- tokens_remove(tokens(programs.corp3), stopwords("english"), padding  = TRUE)
```

```
myNgram2 <- tokens(programs.corp4) %>%
  tokens_ngrams(n = 2) %>%
  dfm()
myNgram3 <- tokens(programs.corp4) %>%
  tokens_ngrams(n = 3) %>%
  dfm()
topfeatures(myNgram2)
```

```
##          high_school        united_states                501_c
##                   68                   31                   30
##                  c_3          raise_funds financial_assistance
##                   30                   26                   24
##         young_people    provide_financial           low_income
##                   22                   19                   18
##      school_students
##                   18
```

```
topfeatures(myNgram3)
```

```
##                     501_c_3         high_school_students
##                          30                           15
##               section_501_c provide_financial_assistance
##                          14                           13
##    carmel_international_film        internal_revenue_code
##                          10                           10
##   provided_financial_support  international_film_festival
##                           9                            9
##          low_income_families             c_3_organization
##                           8                            7
```

```
my_dict <- dictionary(list(five01_c_3= c("501 c 3","section_501_c") ,
                           united_states = "united states"))
programs.corp5 <- tokens_compound(programs.corp4, pattern = my_dict)
programs.corp6 <- sapply(programs.corp5, paste, collapse=" ")
```

```
programs.dfm <- dfm(programs.corp6,
                    stem = T)
programs.dfm
```

```
## Document-feature matrix of: 2,536 documents, 6,411 features (99.7% sparse).
```

```
topfeatures(programs.dfm, 20)
```

```
##    provid   program communiti      educ     organ   support    servic
##      1001       663       582       582       559       439       377
##    school  children    famili     train   student      fund      help
##       367       319       307       271       269       256       254
##   develop     youth      need    activ    assist    promot
##       237       236       233       230       227       210
```

```
programs.dfm.df <- convert(programs.dfm, to = "data.frame")
programs.corpus.df <- as.data.frame(programs.corp6)
```