Homework 2

Beau Harrison

Libraries

```
library(NLP) # Required for tm
library(tm) # Corpus
library(data.table) # rbindlist
library(quanteda) # tokenize
```

```
## quanteda version 0.9.9.17
```

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

```
## The following objects are masked from 'package:tm':
##
##     as.DocumentTermMatrix, stopwords
```

```
## The following object is masked from 'package:NLP':
##
##     ngrams
```

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

```
## The following object is masked from 'package:base':
##
##     sample
```

```
library(plyr) # join
```

function

```
parseCorpus <- function(files) {
  docs <- Corpus(files)
  if(length(files) > 1) {
    docFrame <- lapply(docs, function(doc) data.frame(doc$content))

  } else {

  }

}
```

Parse into R structures

```
# Get files into R
clintonDocs <- Corpus(DirSource('CampaignSpeeches',pattern='clinton'))
trumpDocs <- Corpus(DirSource('CampaignSpeeches',pattern='trump'))
```

```
# Parse file content into data.frames
clintonFrames <- lapply(clintonDocs, function(corpus) data.frame(corpus$content))
trumpFrames <- lapply(trumpDocs, function(corpus) data.frame(corpus$content))
# Join data.frames
clintonFrame <- rbindlist(clintonFrames)
trumpFrame <- rbindlist(trumpFrames)
```

Clean up

Remove speaker's name and applause

```
# Remove names
clintonFrame <- lapply(clintonFrame, function(x) gsub('CLINTON: ', '', x))
trumpFrame <- lapply(trumpFrame, function(x) gsub('TRUMP: ', '', x))
# Remove notes
clintonFrame <- lapply(clintonFrame, function(x) gsub('\\([A-Z ]+\\)', '', x))
trumpFrame <- lapply(trumpFrame, function(x) gsub('\\([A-Z ]+\\)', '', x))
```

Tokenize

```
# Seperate words and remove punctuation
clintonUnigramTokens <- tokenize(paste(clintonFrame, collapse=''), removePunct=TRUE, removeNumbers=TRUE
trumpUnigramTokens <- tokenize(paste(trumpFrame, collapse=''), removePunct=TRUE, removeNumbers=TRUE, re
clintonBigramTokens <- tokenize(paste(clintonFrame, collapse=''), removePunct=TRUE, removeNumbers=TRUE,
trumpBigramTokens <- tokenize(paste(trumpFrame, collapse=''), removePunct=TRUE, removeNumbers=TRUE, rem
# Without stopwords
clintonUnigramTokensNoStopwords <- removeFeatures(tokenize(paste(clintonFrame, collapse=''), removePunc
trumpUnigramTokensNoStopwords <- removeFeatures(tokenize(paste(trumpFrame, collapse=''), removePunct=TR
clintonBigramTokensNoStopwords <- removeFeatures(tokenize(paste(clintonFrame, collapse=''), removePunct
trumpBigramTokensNoStopwords <- removeFeatures(tokenize(paste(trumpFrame, collapse=''), removePunct=TRU
# Put lower case versions in data.table
clintonUnigrams <- data.table(token=tolower(unlist(clintonUnigramTokens)))
trumpUnigrams <- data.table(token=tolower(unlist(trumpUnigramTokens)))
clintonBigrams <- data.table(token=tolower(unlist(clintonBigramTokens)))
trumpBigrams <- data.table(token=tolower(unlist(trumpBigramTokens)))
# Without stopwords
clintonUnigramsNoStopwords <- data.table(token=tolower(unlist(clintonUnigramTokensNoStopwords)))
trumpUnigramsNoStopwords <- data.table(token=tolower(unlist(trumpUnigramTokensNoStopwords)))
clintonBigramsNoStopwords <- data.table(token=tolower(unlist(clintonBigramTokensNoStopwords)))
trumpBigramsNoStopwords <- data.table(token=tolower(unlist(trumpBigramTokensNoStopwords)))
# Count instances
clintonUnigramCount <- clintonUnigrams[, .N, by=token][order(N, decreasing=TRUE)]
trumpUnigramCount <- trumpUnigrams[, .N, by=token][order(N, decreasing=TRUE)]
clintonBigramCount <- clintonBigrams[, .N, by=token][order(N, decreasing=TRUE)]
trumpBigramCount <- trumpBigrams[, .N, by=token][order(N, decreasing=TRUE)]
# Without stopwords
clintonUnigramCountNoStopwords <- clintonUnigramsNoStopwords[, .N, by=token][order(N, decreasing=TRUE)]
trumpUnigramCountNoStopwords <- trumpUnigramsNoStopwords[, .N, by=token][order(N, decreasing=TRUE)]
clintonBigramCountNoStopwords <- clintonBigramsNoStopwords[, .N, by=token][order(N, decreasing=TRUE)]
trumpBigramCountNoStopwords <- trumpBigramsNoStopwords[, .N, by=token][order(N, decreasing=TRUE)]
# Add canidate names
clintonUnigramCount$canidate <- 'Clinton'
```

```
trumpUnigramCount$canidate <- 'Trump'
clintonBigramCount$canidate <- 'Clinton'
trumpBigramCount$canidate <- 'Trump'
# Without stopwords
clintonUnigramCountNoStopwords$canidate <- 'Clinton'
trumpUnigramCountNoStopwords$canidate <- 'Trump'
clintonBigramCountNoStopwords$canidate <- 'Clinton'
trumpBigramCountNoStopwords$canidate <- 'Trump'
# Combine data.tables
unigramCount <- rbind(clintonUnigramCount, trumpUnigramCount)[order(N, decreasing=TRUE)]
bigramCount <- rbind(clintonBigramCount, trumpBigramCount)[order(N, decreasing=TRUE)]
# Without stopwords
unigramCountNoStopwords <- rbind(clintonUnigramCountNoStopwords, trumpUnigramCountNoStopwords)[order(N,
bigramCountNoStopwords <- rbind(clintonBigramCountNoStopwords, trumpBigramCountNoStopwords)[order(N, de
```

Chi^2 calculation

```
chiSquared <- function(input) {
  DT <- data.table(join(input[canidate == "Clinton"][, list(token, clintonCount = as.numeric(N))], inpu
  DT[is.na(clintonCount)]$clintonCount <- 0
  DT[is.na(trumpCount)]$trumpCount <- 0
  DT[, `:=`(totalCount, clintonCount + trumpCount)]
  DT <- DT[order(totalCount, decreasing=TRUE)][totalCount > 5]
  DT[, `:=`(totalClinton, sum(clintonCount))]
  DT[, `:=`(totalTrump, sum(trumpCount))]
  DT[, `:=`(chi2, (totalClinton + totalTrump) * (trumpCount * (totalClinton - clintonCount) - clintonCou

  return(DT[order(chi2, decreasing=TRUE)])
}
```

Question 1 Results

```
unigramChi <- chiSquared(unigramCount)
```

```
## Joining by: token
```

```
unigramChi[1:10]
```

```
##        token clintonCount trumpCount totalCount totalClinton totalTrump
##  1:    going           40        233        273        18868      22108
##  2:     very           18        127        145        18868      22108
##  3:    women           70         10         80        18868      22108
##  4: families           46          2         48        18868      22108
##  5:    let's           48          3         51        18868      22108
##  6:  economy           46          4         50        18868      22108
##  7:  hillary            9         73         82        18868      22108
##  8:      i'm           31        119        150        18868      22108
##  9:  clinton            5         60         65        18868      22108
## 10:    don't           24        103        127        18868      22108
##          chi2
##  1: 109.03235
##  2:  66.25431
```

```
##  3:  55.44286
##  4:  47.94725
##  5:  47.49681
##  6:  42.55219
##  7:  40.67807
##  8:  39.03410
##  9:  38.54886
## 10:  37.79530
```

```
bigramChi <- chiSquared(bigramCount)
```

```
## Joining by: token
```

```
bigramChi[1:10]
```

```
##               token clintonCount trumpCount totalCount totalClinton
##  1:        going to           31        208        239         5944
##  2: hillary clinton            2         47         49         5944
##  3:      each other           18          0         18         5944
##  4:    young people           19          1         20         5944
##  5:       i believe           19          1         20         5944
##  6:         look at            4         42         46         5944
##  7:       we should           20          2         22         5944
##  8:          i mean            1         29         30         5944
##  9:           do it            2         32         34         5944
## 10:         the top           14          0         14         5944
##     totalTrump     chi2
##  1:       7927 88.66913
##  2:       7927 30.18287
##  3:       7927 24.03624
##  4:       7927 22.24129
##  5:       7927 22.24129
##  6:       7927 21.98726
##  7:       7927 20.78045
##  8:       7927 19.17317
##  9:       7927 19.02230
## 10:       7927 18.68946
```

```
# Without stopwords
unigramChiNoStopwords <- chiSquared(unigramCountNoStopwords)
```

```
## Joining by: token
```

```
unigramChiNoStopwords[1:10]
```

```
##        token clintonCount trumpCount totalCount totalClinton totalTrump
##  1:    going           40        233        273         8415      10238
##  2:    women           70         10         80         8415      10238
##  3: families           46          2         48         8415      10238
##  4:    let's           48          3         51         8415      10238
##  5:  economy           46          4         50         8415      10238
##  6: together           65         17         82         8415      10238
##  7:  hillary            9         73         82         8415      10238
##  8:  clinton            5         60         65         8415      10238
##  9:      i'm           31        119        150         8415      10238
## 10:    don't           24        103        127         8415      10238
##          chi2
```

```
##  1: 103.82284
##  2:  58.29646
##  3:  49.99723
##  4:  49.59696
##  5:  44.51041
##  6:  38.80267
##  7:  38.76376
##  8:  36.88843
##  9:  36.49784
## 10:  35.49142
```

```
bigramChiNoStopwords <- chiSquared(bigramCountNoStopwords)
```

```
## Joining by: token
```

```
bigramChiNoStopwords[1:10]
```

```
##               token clintonCount trumpCount totalCount totalClinton
##  1:        going to           31        208        239         5944
##  2: hillary clinton            2         47         49         5944
##  3:      each other           18          0         18         5944
##  4:    young people           19          1         20         5944
##  5:       i believe           19          1         20         5944
##  6:         look at            4         42         46         5944
##  7:       we should           20          2         22         5944
##  8:          i mean            1         29         30         5944
##  9:           do it            2         32         34         5944
## 10:         the top           14          0         14         5944
##     totalTrump     chi2
##  1:       7927 88.66913
##  2:       7927 30.18287
##  3:       7927 24.03624
##  4:       7927 22.24129
##  5:       7927 22.24129
##  6:       7927 21.98726
##  7:       7927 20.78045
##  8:       7927 19.17317
##  9:       7927 19.02230
## 10:       7927 18.68946
```

Parse into R structures

```
# Get files into R
clintonOrlando <- Corpus(DirSource('CampaignSpeeches',pattern='clinton-orlando'))
trumpOrlando <- Corpus(DirSource('CampaignSpeeches',pattern='trump-orlando'))
# Parse file content into data.frames
clintonOrlandoFrames <- clintonOrlando$content
trumpOrlandoFrames <- trumpOrlando$content
```

Clean up

Remove speaker's name and applause

```
# Remove names
clintonOrlandoFrames <- lapply(clintonOrlandoFrames, function(x) gsub('CLINTON: ', '', x))
trumpOrlandoFrames <- lapply(trumpOrlandoFrames, function(x) gsub('TRUMP: ', '', x))
# Remove notes
clintonOrlandoFrames <- lapply(clintonOrlandoFrames, function(x) gsub('\\([A-Z ]+\\)', '', x))
trumpOrlandoFrames <- lapply(trumpOrlandoFrames, function(x) gsub('\\([A-Z ]+\\)', '', x))
```

Tokenize

```
# Seperate words and remove punctuation
clintonOrlandoUnigramTokens <- tokenize(paste(clintonOrlandoFrames, collapse=''), removePunct=TRUE, reme
trumpOrlandoUnigramTokens <- tokenize(paste(trumpOrlandoFrames, collapse=''), removePunct=TRUE, removeNu
clintonOrlandoBigramTokens <- tokenize(paste(clintonOrlandoFrames, collapse=''), removePunct=TRUE, remov
trumpOrlandoBigramTokens <- tokenize(paste(trumpOrlandoFrames, collapse=''), removePunct=TRUE, removeNum
# Without stopwords
clintonOrlandoUnigramTokensNoStopwords <- removeFeatures(tokenize(paste(clintonOrlandoFrames, collapse=
trumpOrlandoUnigramTokensNoStopwords <- removeFeatures(tokenize(paste(trumpOrlandoFrames, collapse=''),
clintonOrlandoBigramTokensNoStopwords <- removeFeatures(tokenize(paste(clintonOrlandoFrames, collapse='
trumpOrlandoBigramTokensNoStopwords <- removeFeatures(tokenize(paste(trumpOrlandoFrames, collapse=''), r
# Put lower case versions in data.table
clintonOrlandoUnigrams <- data.table(token=tolower(unlist(clintonOrlandoUnigramTokens)))
trumpOrlandoUnigrams <- data.table(token=tolower(unlist(trumpOrlandoUnigramTokens)))
clintonOrlandoBigrams <- data.table(token=tolower(unlist(clintonOrlandoBigramTokens)))
trumpOrlandoBigrams <- data.table(token=tolower(unlist(trumpOrlandoBigramTokens)))
# Without stopwords
clintonOrlandoUnigramsNoStopwords <- data.table(token=tolower(unlist(clintonOrlandoUnigramTokensNoStopwo
trumpOrlandoUnigramsNoStopwords <- data.table(token=tolower(unlist(trumpOrlandoUnigramTokensNoStopwords)
clintonOrlandoBigramsNoStopwords <- data.table(token=tolower(unlist(clintonOrlandoBigramTokensNoStopwor
trumpOrlandoBigramsNoStopwords <- data.table(token=tolower(unlist(trumpOrlandoBigramTokensNoStopwords)))
# Count instances
clintonOrlandoUnigramCount <- clintonOrlandoUnigrams[, .N, by=token][order(N, decreasing=TRUE)]
trumpOrlandoUnigramCount <- trumpOrlandoUnigrams[, .N, by=token][order(N, decreasing=TRUE)]
clintonOrlandoBigramCount <- clintonOrlandoBigrams[, .N, by=token][order(N, decreasing=TRUE)]
trumpOrlandoBigramCount <- trumpOrlandoBigrams[, .N, by=token][order(N, decreasing=TRUE)]
# Without stopwords
clintonOrlandoUnigramCountNoStopwords <- clintonOrlandoUnigramsNoStopwords[, .N, by=token][order(N, deci
trumpOrlandoUnigramCountNoStopwords <- trumpOrlandoUnigramsNoStopwords[, .N, by=token][order(N, decreasi
clintonOrlandoBigramCountNoStopwords <- clintonOrlandoBigramsNoStopwords[, .N, by=token][order(N, decrea
trumpOrlandoBigramCountNoStopwords <- trumpOrlandoBigramsNoStopwords[, .N, by=token][order(N, decreasing
# Add canidate names
clintonOrlandoUnigramCount$canidate <- 'Clinton'
trumpOrlandoUnigramCount$canidate <- 'Trump'
clintonOrlandoBigramCount$canidate <- 'Clinton'
trumpOrlandoBigramCount$canidate <- 'Trump'
# Without stopwords
clintonOrlandoUnigramCountNoStopwords$canidate <- 'Clinton'
trumpOrlandoUnigramCountNoStopwords$canidate <- 'Trump'
clintonOrlandoBigramCountNoStopwords$canidate <- 'Clinton'
trumpOrlandoBigramCountNoStopwords$canidate <- 'Trump'
# Combine data.tables
unigramOrlandoCount <- rbind(clintonOrlandoUnigramCount, trumpOrlandoUnigramCount)[order(N, decreasing=
bigramOrlandoCount <- rbind(clintonOrlandoBigramCount, trumpOrlandoBigramCount)[order(N, decreasing=TRU
```

```
# Without stopwords
unigramOrlandoCountNoStopwords <- rbind(clintonOrlandoUnigramCountNoStopwords, trumpOrlandoUnigramCount
bigramOrlandoCountNoStopwords <- rbind(clintonOrlandoBigramCountNoStopwords, trumpOrlandoBigramCountNoS
```

Question 2 Results

```
unigramOrlandoChi <- chiSquared(unigramOrlandoCount)
```

```
## Joining by: token
```

```
unigramOrlandoChi[1:10]
```

```
##           token clintonCount trumpCount totalCount totalClinton totalTrump
##  1: immigration            0         21         21         2105       2968
##  2:       don't            1         21         22         2105       2968
##  3:         and          164        159        323         2105       2968
##  4:        that           52         36         88         2105       2968
##  5:          as           28         14         42         2105       2968
##  6:         she            2         21         23         2105       2968
##  7:       those           11          2         13         2105       2968
##  8:    together           11          2         13         2105       2968
##  9:     clinton            0         14         14         2105       2968
## 10:          up            7          0          7         2105       2968
##          chi2
##  1: 14.955778
##  2: 12.425759
##  3: 12.236735
##  4: 11.422463
##  5: 11.054157
##  6: 10.238206
##  7:  9.982817
##  8:  9.982817
##  9:  9.956723
## 10:  9.883471
```

```
bigramOrlandoChi <- chiSquared(bigramOrlandoCount)
```

```
## Joining by: token
```

```
bigramOrlandoChi[1:10]
```

```
##                token clintonCount trumpCount totalCount totalClinton
##  1: first responders            6          0          6          326
##  2:            a lot            6          0          6          326
##  3:          as well            6          0          6          326
##  4:       each other            6          0          6          326
##  5:          that we           10          3         13          326
##  6:           we are            7          1          8          326
##  7:           i have            7          1          8          326
##  8:  hillary clinton            0         12         12          326
##  9:         we don't            0         12         12          326
## 10:          will be            0         11         11          326
##     totalTrump      chi2
##  1:        547 10.137156
```

```
##  2:        547 10.137156
##  3:        547 10.137156
##  4:        547 10.137156
##  5:        547  8.835823
##  6:        547  8.681275
##  7:        547  8.681275
##  8:        547  7.251413
##  9:        547  7.251413
## 10:        547  6.639417
```

```
# Without stopwords
unigramOrlandoChiNoStopwords <- chiSquared(unigramOrlandoCountNoStopwords)
```

```
## Joining by: token
```

```
unigramOrlandoChiNoStopwords[1:10]
```

```
##           token clintonCount trumpCount totalCount totalClinton totalTrump
##  1:    together           11          2         13          588       1018
##  2:        well            9          1         10          588       1018
##  3: immigration            0         21         21          588       1018
##  4:       first            8          1          9          588       1018
##  5:       learn            6          0          6          588       1018
##  6:  responders            6          0          6          588       1018
##  7:         lot            6          0          6          588       1018
##  8:       don’t            1         21         22          588       1018
##  9:        back            7          1          8          588       1018
## 10:     clinton            0         14         14          588       1018
##          chi2
##  1: 13.012781
##  2: 12.358180
##  3: 12.290375
##  4: 10.657538
##  5: 10.426709
##  6: 10.426709
##  7: 10.426709
##  8:  9.883310
##  9:  8.971055
## 10:  8.157556
```

```
bigramOrlandoChiNoStopwords <- chiSquared(bigramOrlandoCountNoStopwords)
```

```
## Joining by: token
```

```
bigramOrlandoChiNoStopwords[1:10]
```

```
##                token clintonCount trumpCount totalCount totalClinton
##  1: first responders            6          0          6          326
##  2:            a lot            6          0          6          326
##  3:          as well            6          0          6          326
##  4:       each other            6          0          6          326
##  5:          that we           10          3         13          326
##  6:           we are            7          1          8          326
##  7:           i have            7          1          8          326
##  8:  hillary clinton            0         12         12          326
##  9:         we don’t            0         12         12          326
## 10:          will be            0         11         11          326
```

```
##     totalTrump      chi2
##  1:        547 10.137156
##  2:        547 10.137156
##  3:        547 10.137156
##  4:        547 10.137156
##  5:        547  8.835823
##  6:        547  8.681275
##  7:        547  8.681275
##  8:        547  7.251413
##  9:        547  7.251413
## 10:        547  6.639417
```