Text Similarity and Difference

Lingyi Zhang

2/27/2018

```
# Clear Global Environment and Set working directory
rm(list = ls())
#getwd()
setwd("/Users/Lingyi/TAD/HW1")
set.seed(8888)

#import needed libraries
library(quanteda)
library(quanteda.corpora)
library(tidyverse)
library(stringr)
library(tm)
```

1. First we'll use the data from the State Of the Union speeches available in quanteda.corpora. Let's first look at the States of the Union given by Franklin D. Roosevelt in 1936 and 1945.

```
#loading sotu data
data("data_corpus_sotu", package = "quanteda.corpora")

#subset all speeches by Franklin D. Roosevelt
Rooseveltcorpus <- corpus_subset(data_corpus_sotu, FirstName == "Franklin D.")
#summary(Rooseveltcorpus)

#Find the speeches of 1936 and 1945
Rooseveltcorpus_1936 <- Rooseveltcorpus[3]
Rooseveltcorpus_1945 <- Rooseveltcorpus[ndoc(Rooseveltcorpus)]
```

(a) Calculate the TTR of each of these speeches.

```
#tokenize
Roosevelttokens_1936 <- tokens(Rooseveltcorpus_1936, remove_punct = TRUE)
Roosevelttokens_1945 <- tokens(Rooseveltcorpus_1945, remove_punct = TRUE)

#calculate the TTR
Roosevelttokens_1936_TTR <- textstat_lexdiv(dfm(
    Roosevelttokens_1936, tolower = FALSE,
                          remove_punct = TRUE), measure = "TTR")

#calculate the TTR
Roosevelttokens_1945_TTR <- textstat_lexdiv(dfm(
    Roosevelttokens_1945, tolower = FALSE,
```

```
                          remove_punct = TRUE), measure = "TTR")

cat(paste0(Rooseveltokens_1936_TTR), "\n", paste0(Rooseveltokens_1945_TTR))
```

```
## Roosevelt-1936 0.315913865546218
##  Roosevelt-1945 0.251593137254902
```

(b) Create a document feature matrix of the two speeches, with no pre-processing other than to remove the punctuation–be sure to check the options on "dfm" in R as appropriate. Calculate the cosine distance between the two documents with quanteda.

```
#calculate the cosine distance.
#no pre-processing other than to remove the punctuation
Roosevelt2Dfm <- dfm(c(Rooseveltcorpus_1936, Rooseveltcorpus_1945),
                     stem = FALSE,
                     tolower = FALSE,
                     remove_punct = TRUE)

RooseveltSimil <- textstat_simil(Roosevelt2Dfm,
                                 margin = "documents",
                                 method = "cosine")

as.matrix(RooseveltSimil)
```

```
##                Roosevelt-1936 Roosevelt-1945
## Roosevelt-1936      1.0000000      0.9628502
## Roosevelt-1945      0.9628502      1.0000000
```

2. Consider different pre-processing choices could be made,calculate the following:

- (i) a theoretical argument for how it should affect the TTR of each document and the similarity of the two documents, and
- (ii) re-do (1a) with the pre-processing option indicated, and
- (iii) redo (1b) with the pre-processing option indicated.

(a) Stemming the words?

(i)

It will decrease the TTR of each document because this method will combine words like "like" and "likes" as the same type. It's hard to determine the change of similarity, as fewer types are identified in the dfm, but the occurrence of a certain type might increase.

(ii)

```
Rooseveltokens_1936_TTR <- textstat_lexdiv(dfm(Rooseveltokens_1936,
                    stem = TRUE,
                    tolower = FALSE,
                    remove_punct = TRUE), measure = "TTR")

Rooseveltokens_1945_TTR <- textstat_lexdiv(dfm(Rooseveltokens_1945,
                    stem = TRUE,
                    tolower = FALSE,
                    remove_punct = TRUE), measure = "TTR")

cat(" ", paste0(Rooseveltokens_1936_TTR), "\n", paste0(Rooseveltokens_1945_TTR))
```

```
##   Roosevelt-1936 0.272321428571429
##  Roosevelt-1945 0.206004901960784
```

(iii)

```
Roosevelt2Dfm <- dfm(c(Rooseveltcorpus_1936, Rooseveltcorpus_1945),
                     stem = TRUE,
                     tolower = FALSE,
                     remove_punct = TRUE)

RooseveltSimil <- textstat_simil(Roosevelt2Dfm,
                                 margin = "documents",
                                 method = "cosine")
as.matrix(RooseveltSimil)
```

```
##                Roosevelt-1936 Roosevelt-1945
## Roosevelt-1936       1.000000       0.961592
## Roosevelt-1945       0.961592       1.000000
```

(b) Removing stop words?

(i)

It will increase the TTR of each document because this method will remove the stopwords from both types and tokens. The similarity will decrease as less common words for types and tokens are left.

(ii)

```
Rooseveltokens_1936_TTR <- textstat_lexdiv(dfm(Rooseveltokens_1936,
                    stem = FALSE,
                    tolower = FALSE,
                    remove_punct = TRUE,
                    remove = stopwords("english")), measure = "TTR")

Rooseveltokens_1945_TTR <- textstat_lexdiv(dfm(Rooseveltokens_1945,
                    stem = FALSE,
                    tolower = FALSE,
                    remove_punct = TRUE,
                    remove = stopwords("english")), measure = "TTR")

cat(" ", paste0(Rooseveltokens_1936_TTR), "\n", paste0(Rooseveltokens_1945_TTR))
```

```
##   Roosevelt-1936 0.576362655153805
##  Roosevelt-1945 0.449786830885836
```

(iii)

```
Roosevelt2Dfm <- dfm(c(Rooseveltcorpus_1936, Rooseveltcorpus_1945),
                     stem = FALSE,
                     tolower = FALSE,
                     remove_punct = TRUE,
                     remove = stopwords("english"))

RooseveltSimil <- textstat_simil(Roosevelt2Dfm,
                                 margin = "documents",
                                 method = "cosine")
as.matrix(RooseveltSimil)
```

```
##                Roosevelt-1936 Roosevelt-1945
## Roosevelt-1936      1.0000000      0.4864789
## Roosevelt-1945      0.4864789      1.0000000
```

(c) Converting all words to lowercase?

(i)

It will decrease the TTR of each document because this method will combine words like “like” and “Like” as the same type. It’s hard to determine the change of similarity, as fewer types are identified in the dfm, but the occurrence of a certain type might increase.

(ii)

```
Roosevelttokens_1936_TTR <- textstat_lexdiv(dfm(Roosevelttokens_1936,
                    stem = FALSE,
                    tolower = TRUE,
                    remove_punct = TRUE), measure = "TTR")

Roosevelttokens_1945_TTR <- textstat_lexdiv(dfm(Roosevelttokens_1945,
                    stem = FALSE,
                    tolower = TRUE,
                    remove_punct = TRUE), measure = "TTR")

cat(" ", paste0(Roosevelttokens_1936_TTR), "\n", paste0(Roosevelttokens_1945_TTR))
```

```
##   Roosevelt-1936 0.301207983193277
##  Roosevelt-1945 0.238848039215686
```

(iii)

```
Roosevelt2Dfm <- dfm(c(Rooseveltcorpus_1936, Rooseveltcorpus_1945),
                     stem = FALSE,
                     tolower = TRUE,
                     remove_punct = TRUE)

RooseveltSimil <- textstat_simil(Roosevelt2Dfm,
                                 margin = "documents",
                                 method = "cosine")
as.matrix(RooseveltSimil)
```

```
##                Roosevelt-1936 Roosevelt-1945
## Roosevelt-1936      1.0000000      0.9627462
## Roosevelt-1945      0.9627462      1.0000000
```

(d) Does tf-idf weighting make sense here? Explain why or why not.

```
## Roosevelt-1936 type: 1203; lengths: 3808
## Roosevelt-1945 type: 2053; lengths: 8160
```

Yes, it makes sense. As these two speeches have a big difference in token sizes. Roosevelt-1936:3808, Roosevelt-1945:8160. Using tf-idf weighting is better than the absolute counts, as it will help unique words stand out.

> "when a word is common in a given document, but rare in the corpus as whole... So presence of that word is indicative of difference, and it is weighted up. – Arthur Spirling PPT"

3. Calculate the MLTD of each of the two speeches by FDR, with the TTR limit set at .72. Rather than covering the entire speech, you can find the Mean Lengths starting from 25 different places in each speech, as long is there is no overlap between the snippets.

```
#splitting up strings by word
words_1936 <- str_split(Rooseveltcorpus_1936, boundary("word"))[[1]]
words_1945 <- str_split(Rooseveltcorpus_1945, boundary("word"))[[1]]

#This function calculate the number of words (NoW) before hitting TTR=0.72
#record the number in a list and start calculate again
NoW_Calculate <- function(words) {
    if (length(words) == 0) {
        return("Empty input!")
    } else {
        wordlist <- character()
        word_count_list <- vector()
        word_count <- 0
        wordType <- 0
        wordToken <- 0

        for (i in words) {

            if (i %in% wordlist) {
                wordToken <- wordToken + 1
                word_count <- word_count + 1
            } else {
                wordType <- wordType + 1
                wordToken <- wordToken + 1
                wordlist <- c(wordlist, i)
                word_count <- word_count + 1
            }

            ttr = wordType / wordToken

            if (ttr <= 0.72) {
                #print("hitting 0.72")
                word_count_list <- append(word_count_list, word_count-1)
                word_count <- 1
                wordType <- 1
                wordToken <- 1
                wordlist <- character()
                wordlist <- c(wordlist, i)
            }
        }
        word_count_list
    }

}
```

```
#This function takes in the tokenized text
#divide the text into 25 parts
```

```r
#calculate the average MTLD of these 25 parts
MTLD_selfdefined <- function(text) {

    listForNoW <- vector()
    chunk_length <- length(text) %/% 25

    for (num in 1:24) {
        start <- 1 + (num-1) * chunk_length
        end <- num * chunk_length
        words <- text[start:end]
        listForNoW <- c(listForNoW, NoW_Calculate(words))
    }

    start <- 1 + 24 * chunk_length
    end <- length(text)
    listForNoW <- c(listForNoW, NoW_Calculate(words))
    mean(listForNoW)
}

cat(paste0(" ", "MTLD of Roosevelt_1936: ",
           MTLD_selfdefined(Rooseveltokens_1936[[1]])), "\n",
    paste0("MTLD of Roosevelt_1945: ",
           MTLD_selfdefined(Rooseveltokens_1945[[1]])))
```

```
##  MTLD of Roosevelt_1936: 51.9803921568627
##  MTLD of Roosevelt_1945: 74.9891304347826
```

4. Take the following two sentences:

> "Mr. and Mrs. Dursley, of number four, Privet Drive, were proud to say that they were normal, thank you very much."
>
> "The Dursleys had everything they wanted, but they also had a secret, and their greatest fear was that somebody would discover it."

(a) Calculate the Euclidean distance between these sentences by hand—that is.

In the pre-processing, I set `stem=TRUE` because the "Dursley" and "Dursleys" refer to the same thing. I set `tolower=TRUE` which makes no difference in this example.

```
sentence1 <- paste0("Mr. and Mrs. Dursley, of number four, Privet Drive, ",
             "were proud to say that they were normal, thank you very much.")
sentence2 <- paste0("The Dursleys had everything they wanted, but they also had a ",
             "secret, and their greatest fear was that somebody would discover it.")

two_sentence <- c(sentence1, sentence2)

two_sentence_dfm <- dfm(two_sentence,
                     stem = TRUE,
                     tolower = TRUE,
                     remove_punct = TRUE,
                     remove = stopwords("english"))

as.matrix(two_sentence_dfm)
```

```
##        features
## docs    mr mrs dursley number four privet drive proud say normal thank
##   text1  1   1       1      1    1      1     1     1   1      1     1
##   text2  0   0       1      0    0      0     0     0   0      0     0
##        features
## docs    much everyth want also secret greatest fear somebodi discov
##   text1    1       0    0    0      0        0    0        0      0
##   text2    0       1    1    1      1        1    1        1      1
```

```
sent1 <- c(1,1,1,1,1,1,1,1,1,1,1,1,0,0,0,0,0,0,0,0)
sent2 <- c(0,0,1,0,0,0,0,0,0,0,0,0,1,1,1,1,1,1,1,1)

print(paste0("Euclidean distance: ", sqrt(sum((sent1-sent2)^2))))
```

```
## [1] "Euclidean distance: 4.35889894354067"
```

(b) Calculate the Manhattan distance between these sentences.

```
print(paste0("Manhattan distance: ", sum(abs(sent1-sent2))))
```

```
## [1] "Manhattan distance: 19"
```

(c) Calculate the cosine similarity between these sentences.

```
print(paste0("Cosine similarity: ",
             sum(sent1 * sent2) / (sqrt(sum((sent1)^2)) * sqrt(sum((sent2)^2)))))
```

```
## [1] "Cosine similarity: 0.0962250448649376"
```

5. One of the earliest and most famous applications of statistical textual analysis was to determine the authorship of texts. You now get to do the same! You've been given 10 machine- readable texts written by Charles Dickens and Jane Austen, as well as a mysterious 11th document. Your task is to determine which of these authors wrote it.

(a) Before getting started, consider the outliers discussed in Figure 3 of Peng & Hengartner (2002). You need to make sure that the texts you analyze are actually what you want to analyze. Summarize your decisions about how to address this.

Peng & Hengartner (2002) talked about they found out some outliers when running PCA on function words they chose. They found out it was because some blocks of words were footnotes and commentary at the ends. Similarly, in our cases, some words in the very front and end might not be written by the authors. So I will delete them as Peng and Hengartner did.

Through manually check, the piece of mystery_text has APPENDIX in the end, and some pieces have the table of content in the front. I manually deleted them.

(b) Now choose the features of the texts to analyze.

I will use the function words as features I will analyze. Because function words have little contextual meaning. They vary obviously across authors while having little variance among certain author's different pieces.

```
function_word <- c("a", "all", "also", "an", "and", "any", "are", "as", "at",
               "be", "been", "but", "by", "can", "do", "down", "even", "every",
               "for", "from", "had", "has", "have", "her", "his", "if", "in",
               "into", "is", "it", "its", "may", "more", "must", "my", "no",
               "not", "now", "of", "on", "one", "only", "or", "our", "should",
               "so", "some", "such", "than", "that", "the", "their", "then", "there",
               "things", "this", "to", "up", "upon", "was", "were", "what", "when",
               "which", "who", "will", "with", "would", "your")
```

(c) Prepare and divide up the texts in the way that P&H do.

They divided the texts into blocks in size of 1700 words each. This size was considered neither too large or too small to balance two conflicting goals: decreasing the dependence between the counts of function words and ensuring the style in each block being consistent.

```
#all files we will use
filenames <- c("austen_emma.txt", "austen_northanger.txt", "austen_persuasion.txt",
               "austen_pride.txt", "austen_sense.txt", "dickens_bleak.txt",
               "dickens_copperfield.txt", "dickens_great.txt", "dickens_oliver.txt",
               "dickens_tale.txt", "mystery_text.txt")

#import all the 11 pieces of documents
content <- character()

for (i in 1:11) {
```

```
    #readChar allocates space for the number of bytes you specify
    file <- readChar(filenames[i], file.info(filenames[i])$size)
    content <- c(content, file)
}

#name of the 11 pieces of documents
textName <- character()

for (i in 1:11) {
    Name <- str_sub(filenames[i], 1, -5)
    textName <- c(textName, Name)
}

#store name and content into data.frame
corpus_df <- data.frame(textName, content, stringsAsFactors = FALSE)

# tokenize all documents
for (i in 1:11) {
  corpus_df$token[[i]] <- tokens(
      char_tolower(corpus_df$content[i]), remove_punct = TRUE)

  #store the length of each document
  corpus_df$doclength[i] <- length(corpus_df$token[[i]][[1]])
}

#create 3 containers for divided blocks of austen, dickens, and the unknown one
block_austen <- list()
block_dickens <- list()
block_unknown <- list()

#store corresponding blocks into right containers
for (i in 1:11) {

    #austen
    if (substr(corpus_df$textName[i], start=1, stop=3) == "aus") {
        for (j in 1:(corpus_df$doclength[i] %/% 1700)) {
            block <- paste(
                corpus_df$token[[i]][[1]][(1700*(j-1)+1):(1700*j)],
                collapse = " ")
            block_austen <- c(block_austen, block)
        }

    #dickens
    } else if (substr(corpus_df$textName[i], start=1, stop=3) == "dic") {
        for (j in 1:(corpus_df$doclength[i] %/% 1700)) {
            block <- paste(
                corpus_df$token[[i]][[1]][(1700*(j-1)+1):(1700*j)],
                collapse = " ")
            block_dickens <- c(block_dickens, block)
        }

    #unknown
    } else {
```

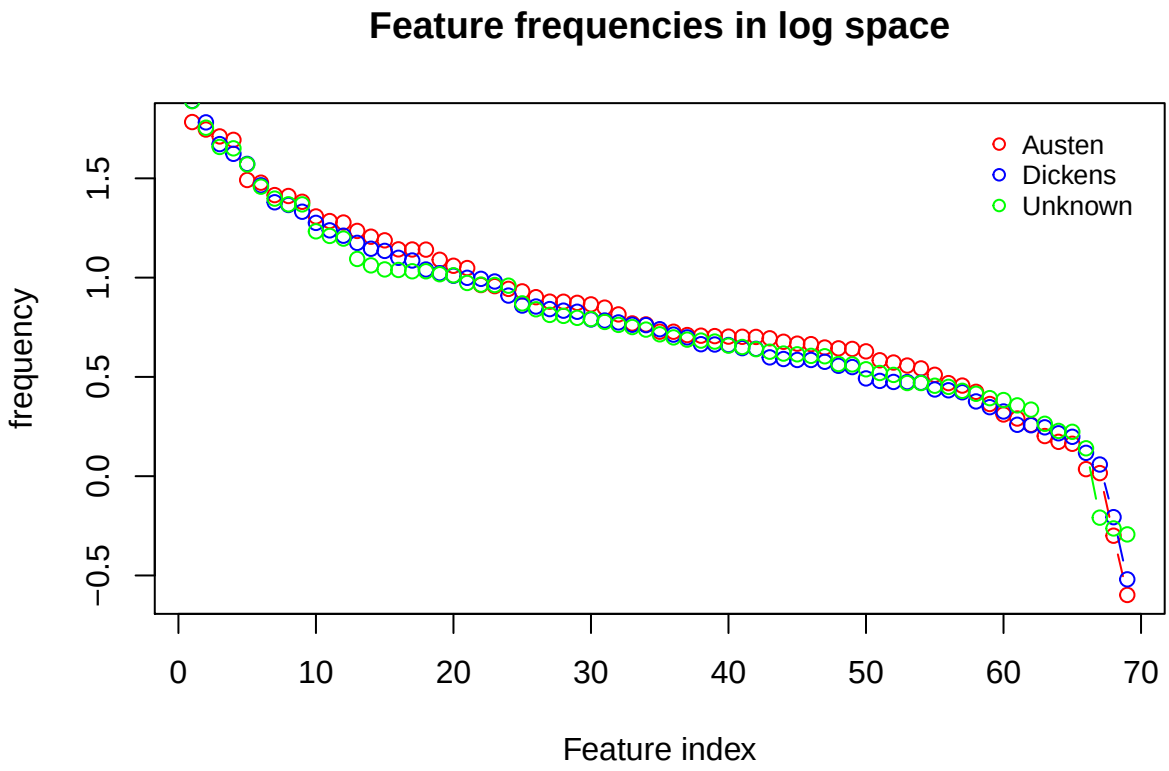
```
        for (j in 1:(corpus_df$doclength[i] %/% 1700)) {
            block <- paste(
                corpus_df$token[[i]][[1]][(1700*(j-1)+1):(1700*j)],
                collapse = " ")
            block_unknown <- c(block_unknown, block)
        }
    }
}

#convert the block list I got into DFM
getDFM <- function (blocks, auther) {

    #turn list of characters back into corpus
    corp <- Corpus(VectorSource(blocks))
    corp <- corpus(corp)
    docvars(corp, "auther") <- auther

    #tokenize
    block_token <- tokens(corp)

    #keep the function words I need
    left <- tokens_select(block_token, function_word, selection = "keep")
    function_dfm <- dfm(left)
    function_dfm
}

#evaluate using function defined above
block_austen_dfm <- getDFM(block_austen, "austen")
block_dickens_dfm <- getDFM(block_dickens, "dickens")
block_unknown_dfm <- getDFM(block_unknown, "unknown")

#show how many documents in each dfm
cat(paste0(" Austen: ",
           as.integer(summary(block_austen_dfm)[["Length"]])/69, " documents\n"),
    paste0("Dickens: ",
           as.integer(summary(block_dickens_dfm)[["Length"]])/69, " documents\n"),
    paste0("Unknown: ",
           as.integer(summary(block_unknown_dfm)[["Length"]])/69, " documents\n"))
```

```
##  Austen: 329 documents
##  Dickens: 698 documents
##  Unknown: 55 documents
```

(d) Using the DFM, calculate the average occurrence of the features that you chose in part (b) for each author and the mystery text. Generate a graph that compares the average frequency of the terms you chose in the mystery text and for each author.

```
plot(log10(topfeatures(block_austen_dfm/329, n = 69)), type = "b", col="red",
     xlab="Feature index", ylab="frequency")
points(log10(topfeatures(block_dickens_dfm/698, n = 69)), type = "b", col="blue")
points(log10(topfeatures(block_unknown_dfm/55, n = 69)), type = "b", col='green')
```

```
legend("topright", legend=c("Austen", "Dickens", "Unknown"),
       col=c("red", "blue", "green"), cex=0.8, inset=.02, bty="n", pch = 21:21)
title("Feature frequencies in log space")
```

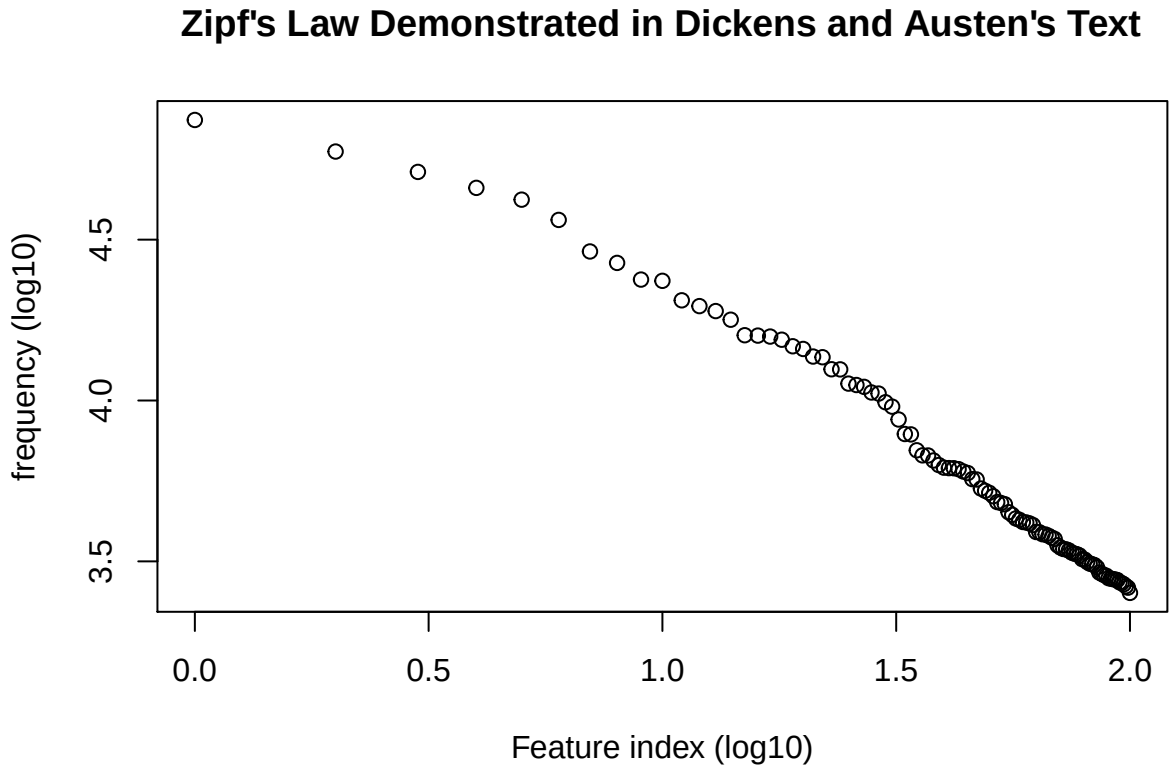


(e) which author do you believe wrote the mystery text?

In the log space, we can see that the average term frequency points of mystery text are mostly closer to the "Dickens" than the "Austen". Based on this finding, I think Dickens wrote this mystery text.

6. Using the 10 labeled Dickens and Austen texts, make a graph demonstrating Zipf's law.

```
corp <- Corpus(VectorSource(content[1:10])) #choose the first 10 texts
corp <- corpus(corp)
token10 <- tokens(corp, remove_numbers = FALSE, remove_punct = TRUE)
dfm10 <- dfm(token10, tolower = TRUE, stem = TRUE)

plot(log10(1:100), log10(topfeatures(dfm10, n = 100)), type = "p",
     xlab="Feature index (log10)", ylab="frequency (log10)")
title("Zipf's Law Demonstrated in Dickens and Austen's Text")
```



I removed punctuation, as the use of punctuation is inevitable in every document, deleting them will not lose a lot of information.I also lowered all words, stemmed them, because in this way words that have the same meaning will be combined together.

7. Find the value of b that best fit the 10 labeled Dickens and Austen texts to Heap's law, fixing k = 44.

```
corp <- Corpus(VectorSource(content[1:10]))
corp <- corpus(corp)
token10 <- tokens(corp, remove_numbers = FALSE, remove_punct = TRUE)
dfm10 <- dfm(token10, tolower = TRUE, stem = TRUE)

#calculate values of T and M
T_token <- token10 %>% lengths() %>% unlist() %>% sum()
M_type <- dfm10 %>% ntype() %>% unlist() %>% sum()

k <- 44

#find the optimal value of b
#as M = k * T^b
#so b = log10(M/k) / log10(T)
b <- log10(M_type / k) / log10(T_token)
print(paste0("Optimal b is: ", b))
```

```
## [1] "Optimal b is: 0.503706379600561"
```

I removed punctuation, as the use of punctuation is inevitable in every document, deleting them will not lose a lot of information.I also lowered all words, stemmed them, because in this way words that have the same meaning will be combined together.

8. Both Dickens' Tale of Two Cities and Austen's Pride and Prejudice examine the role of class in British society, but in very different ways. Choose a few Key Words in Context and discuss the different context in which those words are used by each author.

Dickens' Tale of Two Cities showed the contrast between very high and low classes. Austen's Pride and Prejudice more focus on the second class to fourth class.

'death, marriage': Pride: Related to inheriting the wealth. Tale: used in more broad context.

"work": Pride: not "jobs", as they didn't work. Instead, their income comes from the rent of land. Tale: have more meanings in this term.

"poor": Pride: mainly appear as poor someone. Tale: have more meanings in this term.

```
dickens_tale <- corpus_df %>% select(content) %>% filter(
    textName == "dickens_tale")
austen_pride <- corpus_df %>% select(content) %>% filter(
    textName == "austen_pride")

corp_dickens_tale <- corpus(Corpus(VectorSource(dickens_tale)))
corp_austen_pride <- corpus(Corpus(VectorSource(austen_pride)))
```

```
#Here I only showed the code
#please find the evaluated results in the appendix
kwic(corp_dickens_tale, "death", 2)
kwic(corp_austen_pride, "death", 2)

kwic(corp_dickens_tale, "marriage", 2)
kwic(corp_austen_pride, "marriage", 2)

kwic(corp_dickens_tale, "work", 2)
kwic(corp_austen_pride, "work", 2)

kwic(corp_dickens_tale, "poor", 2)
kwic(corp_austen_pride, "poor", 2)
```

9. Consider the bootstrapping of the texts we used to calculate the standard errors of the Flesch reading scores of Irish budget speeches in Recitation 4.

(a) Obtain the UK Labour Party's manifestos from quanteda.corpora. Generate estimates of the FRE scores of these manifestos over time, using sentence-level bootstraps instead of the speech-level bootstraps.

```
Document_Lab <- corpus_subset(data_corpus_ukmanifestos, Party == "Lab")

sentences <- tokens(Document_Lab, what = "sentence")

#store sentences in a data.frame
df_sentences <- data.frame(year = numeric(),
                           sentence = character(),
                           stringsAsFactors=FALSE)

count <- 0

#store each sentence with its corresponding year
for (i in 1:length(sentences)) {
    count <- count + length(sentences[[i]])

    df_temp <- data.frame(year = numeric(),
                          sentence = character(),
                          stringsAsFactors=FALSE)

    for (j in 1:length(sentences[[i]])) {
        df_temp[j,1] <- Document_Lab[["Year"]]$Year[i]
        df_temp[j,2] <- sentences[[i]][j]
    }
    df_sentences <- rbind(df_sentences, df_temp)
}


#some are numbers or headings, filter them out
df <- filter(df_sentences, grepl(
        "^\\d", df_sentences$sentence)==FALSE & (grepl(     #NOT start with numbers
        "\\.$", df_sentences$sentence)==TRUE | grepl(       #(end with .
        "\\?$", df_sentences$sentence)==TRUE | grepl(       #or end with ?
        "\\!$", df_sentences$sentence)==TRUE) & ntoken(     #or end with !)
        df_sentences$sentence)>2)                           #length longer than 2

iters <- 50

#create a data.frame storing FRE
year_FRE <- data.frame(matrix(ncol = 16, nrow = iters))
colnames(year_FRE) <- names(table(Document_Lab[["Year"]]$Year))


# run the bootstrap
```

```
for(i in 1:iters) {

  sentences_grouped <- group_by(df, year)

  # take a sample of 100 sentences per level (year)
  bootstrap_sample <- sample_n(sentences_grouped, 100, replace = TRUE)

  readability_results <- textstat_readability(
      bootstrap_sample$sentence, measure = "Flesch")

  #store results
  readability_grouped <- group_by(
      readability_results, bootstrap_sample$year)

  readability_means <- summarize(readability_grouped, mean(Flesch))

  year_FRE[i, ] <- t(readability_means[, 2])

}
```

```
# Define the standard error function
std <- function(x) sd(x)/sqrt(length(x))

year_ses <- apply(year_FRE, 2, std)
year_means <- apply(year_FRE, 2, mean)

# store for plotting results
coefs <- year_means
ses <- year_ses

#95% confidence intervals of the mean
min <- min(coefs - 2*ses)
max <- max(coefs + 2*ses)
var.names <- colnames(year_FRE)
adjust <- 0

plot(var.names, coefs, type = "p", pch = 19, cex = .8,
     ylim=c(min,max),xlim = c(1940,2010), main = "", axes = F,
     xlab="Year", ylab="FRE")

rect(1940,min,1945,max, col = c("grey97"), border="grey90", lty = 2)
rect(1945,min,1950,max, col = c("grey95"), border="grey90", lty = 2)
rect(1950,min,1955,max, col = c("grey97"), border="grey90", lty = 2)
rect(1955,min,1960,max, col = c("grey95"), border="grey90", lty = 2)
rect(1960,min,1965,max, col = c("grey97"), border="grey90", lty = 2)
rect(1965,min,1970,max, col = c("grey95"), border="grey90", lty = 2)
rect(1970,min,1975,max, col = c("grey97"), border="grey90", lty = 2)
rect(1975,min,1980,max, col = c("grey95"), border="grey90", lty = 2)
rect(1980,min,1985,max, col = c("grey97"), border="grey90", lty = 2)
rect(1985,min,1990,max, col = c("grey95"), border="grey90", lty = 2)
rect(1990,min,1995,max, col = c("grey97"), border="grey90", lty = 2)
rect(1995,min,2000,max, col = c("grey95"), border="grey90", lty = 2)
rect(2000,min,2005,max, col = c("grey97"), border="grey90", lty = 2)
rect(2005,min,2010,max, col = c("grey95"), border="grey90", lty = 2)
```

```
axis(1, at=seq(1940,2010,5),
    labels=c(1940,1945,1950,1955,1960,1965,1970,1975,1980,1985,1990,1995,2000,2005,2010),
    tick=T, cex.axis=.75, mgp=c(2,.7,0))
axis(2, at=seq(min,max,(max-min)/10),
    labels = c(round(min+0*((max-min)/10),0),
               round(min+1*((max-min)/10),0),
               round(min+2*((max-min)/10),0),
               round(min+3*((max-min)/10),0),
               round(min+4*((max-min)/10),0),
               round(min+5*((max-min)/10),0),
               round(min+6*((max-min)/10),0),
               round(min+7*((max-min)/10),0),
               round(min+8*((max-min)/10),0),
               round(min+9*((max-min)/10),0),
               round(max,0)),tick = T,cex.axis = .75, mgp = c(2,.7,0))

segments(as.numeric(var.names)+2*adjust, coefs-qnorm(.975)*ses,
         as.numeric(var.names)+2*adjust, coefs+qnorm(.975)*ses, lwd=1)
segments(as.numeric(var.names)+2*adjust-.3, coefs-qnorm(.95)*ses,
         as.numeric(var.names)+2*adjust+.3, coefs-qnorm(.95)*ses, lwd=.9)
segments(as.numeric(var.names)+2*adjust-.3, coefs+qnorm(.95)*ses,
         as.numeric(var.names)+2*adjust+.3, coefs+qnorm(.95)*ses, lwd=.9)
points(var.names, coefs, pch=21, cex=.8, bg="white")
title("FRE Score by Year of UK Labour Party (bootstrapped)")
```

(b) The means of the bootstrapped results and the means observed in the data.

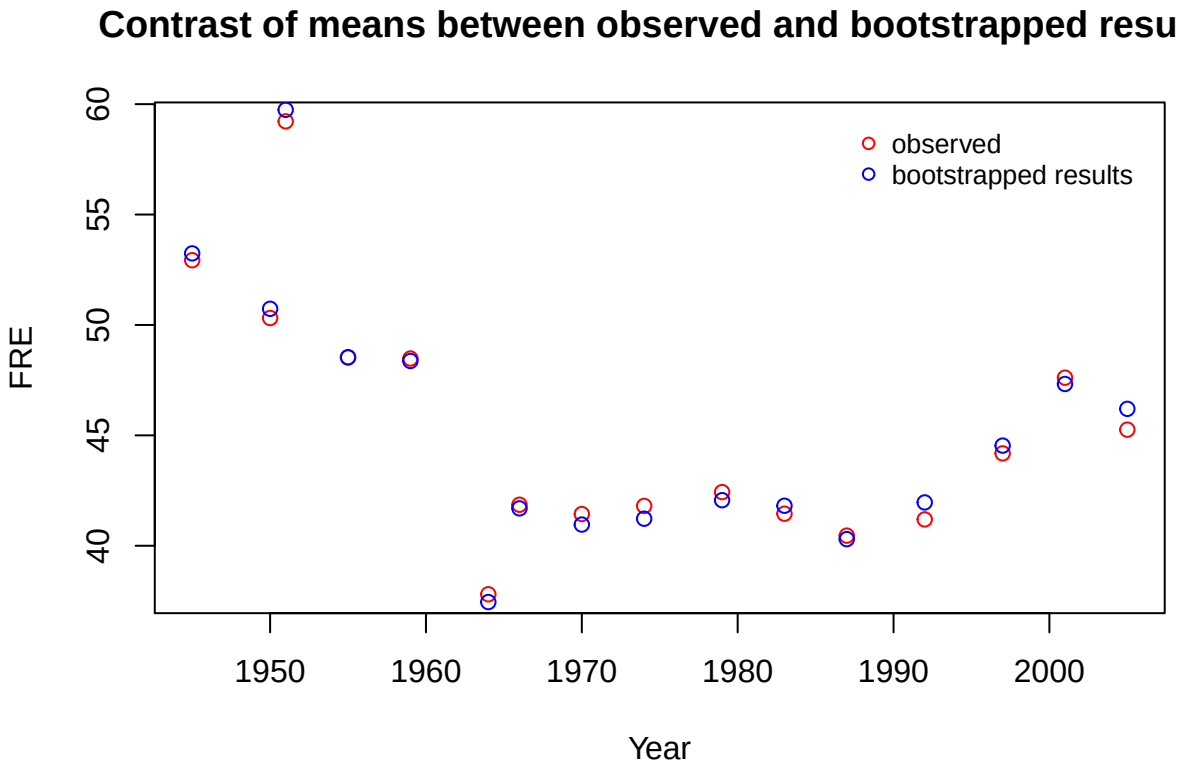
```
#observed means
summarize(group_by(textstat_readability(df$sentence, measure = "Flesch"), df$year),
          mean(Flesch))
```

```
## # A tibble: 16 x 2
##    `df$year` `mean(Flesch)`
##        <dbl>          <dbl>
##  1      1945           52.9
##  2      1950           50.3
##  3      1951           59.2
##  4      1955           48.5
##  5      1959           48.5
##  6      1964           37.8
##  7      1966           41.9
##  8      1970           41.4
##  9      1974           41.8
## 10      1979           42.4
## 11      1983           41.5
## 12      1987           40.5
## 13      1992           41.2
## 14      1997           44.2
## 15      2001           47.6
## 16      2005           45.3
```

```
#bootstrapped means
coefs
```

```
##     1945     1950     1951     1955     1959     1964     1966     1970
## 53.24097 50.73132 59.74270 48.52824 48.36847 37.45020 41.69351 40.95926
##     1974     1979     1983     1987     1992     1997     2001     2005
## 41.22441 42.06353 41.81270 40.29928 41.96207 44.53537 47.32743 46.20064
```

```
#plot the difference
plot(summarize(group_by(
    textstat_readability(df$sentence, measure = "Flesch"), df$year),
          mean(Flesch)),
     col="red", xlab="Year", ylab="FRE")
points(var.names, coefs, col="blue")

legend("topright", legend=c("observed", "bootstrapped results"),
       col=c("red", "blue"), cex=0.8, inset=.02, bty="n", pch = 21:21)
title("Contrast of means between observed and bootstrapped results")
```



They are all concluded based on samples. But bootstrapped means are supposed to be more close to the real mean.

(c) For the empirical values of each text, calculate the FRE score and the Dale-Chall score.

```
readability_means_f <- df$sentence %>% textstat_readability(
    "Flesch") %>% group_by(df$year) %>% summarize(mean(Flesch))
readability_means_d <- df$sentence %>% textstat_readability(
    "Dale.Chall.old") %>% group_by(df$year) %>% summarize(mean(Dale.Chall.old))

readability_matrix <- cbind(readability_means_f[['mean(Flesch)']],
                            readability_means_d[['mean(Dale.Chall.old)']])

cor(readability_matrix)
```

```
##            [,1]       [,2]
## [1,]  1.0000000 -0.6075837
## [2,] -0.6075837  1.0000000
```

Appendix

```
kwic(corp_dickens_tale, "death", 2)
```

```
##
##     [text1, 539]         put to | death | , already
##     [text1, 630]        Farmer, | Death | , had
##    [text1, 3771]        even of | Death | itself,
##   [text1, 10780]         her to | death | ? Look
##   [text1, 13608]      bread and | death | . Is
##   [text1, 21676]     putting to | death | was a
##   [text1, 21697]     Tellson's. | Death | is Nature's
##   [text1, 21717]         put to | Death | ; the
##   [text1, 21728]         put to | Death | ; the
##   [text1, 21739]         put to | Death | ; the
##   [text1, 21751]         put to | Death | ; the
##   [text1, 21771]         put to | Death | ; the
##   [text1, 21782]         put to | Death | ; the
##   [text1, 21801]         put to | Death | . Not
##   [text1, 32877]    escape from | death | . It
##   [text1, 39715]          me to | death | ,"
##   [text1, 47817]       ran into | death | according to
##   [text1, 48417]  captivity and | Death | in the
##   [text1, 50435] his benefactor | Death | , had
##   [text1, 51296]          me to | death | I hope
##   [text1, 51309]         Not to | death | ,"
##   [text1, 51325]           , to | death | ."
##   [text1, 51352]       brink of | death | , you
##   [text1, 51918]       life and | death | over the
##   [text1, 52657]             "" | Death | has done
##   [text1, 57401]         to the | death | . Not
##   [text1, 65633]          of my | death | , I
##   [text1, 73443]   condemned to | death | he will
##   [text1, 73470]         by the | death | of his
##   [text1, 94817]   with instant | death | if any
##   [text1, 96214]       means of | death | ;"
##  [text1, 106558]      danger of | death | , to
##  [text1, 107723]           , or | Death | . A
##  [text1, 109053]         all to | death | who return
##  [text1, 111748]         by the | death | they had
##  [text1, 112436]     body after | death | ."
##  [text1, 112872]           , or | Death | , Monseigneur's
##  [text1, 114385]       Life and | Death | in the
##  [text1, 119339]           , or | Death | , declared
##  [text1, 119345]     victory or | death | against the
##  [text1, 120412]           , or | death | ;-
##  [text1, 120657]      shadow of | death | --
##  [text1, 121790]           , or | Death | ! The
##  [text1, 121830]   had squeezed | Death | in with
##  [text1, 122990]           , or | Death | ! Who
##  [text1, 123998]        pain of | Death | . It
##  [text1, 125275]        sign of | Death | --
##  [text1, 126072]         put to | death | on vague
```

```
##  [text1, 126395]           , or | Death | , that
##  [text1, 127693]       still as | Death | ."
##  [text1, 129094]         be the | death | of me
##  [text1, 129575]         be the | death | of me
##  [text1, 133718] having feigned | death | and come
##  [text1, 133792]      ducked to | death | , and
##  [text1, 137954]         put to | death | , and
##  [text1, 138111]           to a | death | which had
##  [text1, 138151]       life and | death | of the
##  [text1, 143522]     shadows of | death | , as
##  [text1, 147130]          , and | Death | within four-and-twenty
##  [text1, 154105]          of my | death | !"
##  [text1, 155211]       add your | death | to the
##  [text1, 156046]         to the | death | ."
##  [text1, 156998]          by my | death | ; but
##  [text1, 159694]         of his | death | . She
##  [text1, 165674]         put to | death | , that
```

```
kwic(corp_austen_pride, "death", 3)
```

```
##
##   [text1, 29607]     before my father's | death | , Mr.
##   [text1, 30023]     Since her father's | death | , her home
##   [text1, 39594]       estate after the | death | of your honoured
##   [text1, 52890]        there since the | death | of Darcy's father
##   [text1, 56728] till her grandfather's | death | made her mistress
##   [text1, 82959]           would be the | death | of half the
##   [text1, 84483]              it as the | death | warrant of all
##  [text1, 108608]              mind. The | death | of your daughter
```

```
kwic(corp_dickens_tale, "marriage", 2)
```

```
##
##   [text1, 58576]        on your | marriage | morning.
##   [text1, 61390]      myself in | marriage | to your
##   [text1, 81264]          if my | marriage | were so
##   [text1, 81485]   through your | marriage | , than
##   [text1, 82918]         to the | marriage | but Mr
##   [text1, 82938]          . The | marriage | was to
##   [text1, 86785] his daughter's | marriage | had taken
##  [text1, 104447]         On the | marriage | morning,
##  [text1, 146081]         in her | marriage | . How
##  [text1, 153592]       of their | marriage | . He
```

```
kwic(corp_austen_pride, "marriage", 2)
```

```
##
##    [text1, 7419]    Happiness in | marriage | is entirely
##   [text1, 18501]  their supposed | marriage | , and
##   [text1, 23295]           of in | marriage | . This
##   [text1, 32339]        of their | marriage | was extremely
##   [text1, 36501]         which a | marriage | of true
##   [text1, 40577]        offer of | marriage | may ever
##   [text1, 41169]          in the | marriage | state.
##   [text1, 41559]        offer of | marriage | . Is
##   [text1, 41581]        offer of | marriage | you have
```

```
##   [text1, 42242]        offer of | marriage | in this
##   [text1, 45898]      matrimony, | marriage | had always
##   [text1, 46959]    entering the | marriage | state.
##   [text1, 47035]       offers of | marriage | within three
##   [text1, 48088]    approved his | marriage | , that
##   [text1, 49902]  is Charlotte's | marriage | . It
##   [text1, 51614]        point of | marriage | , and
##   [text1, 52852]      before her | marriage | , she
##   [text1, 53789]           . His | marriage | was now
##   [text1, 57509]          by his | marriage | ; his
##   [text1, 68097]  most imprudent | marriage | , but
##   [text1, 71222]        offer of | marriage | from Mr
##   [text1, 72367]        of their | marriage | . He
##   [text1, 72640]          to the | marriage | were not
##   [text1, 73044]   prevented the | marriage | , had
##   [text1, 79277]     felicity in | marriage | . My
##   [text1, 86617]        in their | marriage | put an
##   [text1, 86877]    unsuitable a | marriage | , nor
##   [text1, 99685]          as the | marriage | between Mr
##  [text1, 100074]      upon their | marriage | ; he
##  [text1, 101976]    intention of | marriage | , she
##  [text1, 103032]      must their | marriage | be private
##  [text1, 103273]      other than | marriage | ?"
##  [text1, 104788]  announce their | marriage | . Mrs
##  [text1, 106243]          him in | marriage | , because
##  [text1, 113188]      settled by | marriage | articles on
##  [text1, 113669]           . The | marriage | of a
##  [text1, 113994]       which her | marriage | would scarcely
##  [text1, 114082]       since her | marriage | would so
##  [text1, 114195]     Had Lydia's | marriage | been concluded
##  [text1, 114530]      such happy | marriage | could now
##  [text1, 114710]          as his | marriage | was fixed
##  [text1, 115252]          on her | marriage | by her
##  [text1, 115769]         and his | marriage | been exactly
##  [text1, 118752]      expedite a | marriage | , which
##  [text1, 118905]    benefited by | marriage | . But
##  [text1, 118930]      fortune by | marriage | in some
##  [text1, 120944]         for his | marriage | with Miss
##  [text1, 131080]        offer of | marriage | ?"
##  [text1, 131384]        in their | marriage | , to
##  [text1, 131525]    planning the | marriage | . Its
##  [text1, 132227]      make their | marriage | at all
##  [text1, 132736]           by my | marriage | with Mr
##  [text1, 133246]        that the | marriage | of her
##  [text1, 133348]   prevent their | marriage | , it
##  [text1, 133438]            of a | marriage | with _one_
##  [text1, 134471]         of this | marriage | to her
##  [text1, 134566]          into a | marriage | which has
##  [text1, 134609]      before the | marriage | took place
##  [text1, 138885]      in Lydia's | marriage | . All
##  [text1, 139956]      an unequal | marriage | . You
##  [text1, 142371] his approaching | marriage | , were
##  [text1, 143337]        from the | marriage | of her
##  [text1, 143402]          on her | marriage | , explained
```

```
##  [text1, 143772]       which her | marriage | had given
##  [text1, 143866]      by Darcy's | marriage | ; but
##  [text1, 144065]          on the | marriage | of her
```

```
kwic(corp_dickens_tale, "work", 2)
```

```
##
##     [text1, 653]    though they | work | unceasingly,
##     [text1, 656]   unceasingly, | work | silently,
##    [text1, 4374]   like Smith's | work | , so
##    [text1, 7081]         out of | work | . Mr
##   [text1, 15977]         do any | work | requiring nicety
##   [text1, 15986]           Yet, | work | of that
##   [text1, 16103]        hard at | work | , I
##   [text1, 16313]      of silent | work | had passed
##   [text1, 16401]    stopped his | work | ; looked
##   [text1, 16726]         in his | work | . He
##   [text1, 16851]         of his | work | , and
##   [text1, 16944]         to his | work | , and
##   [text1, 17008]       from his | work | ."
##   [text1, 17183]         had no | work | to hold
##   [text1, 17346]        bent to | work | again,
##   [text1, 17919]    resumed his | work | ."
##   [text1, 18056]       over his | work | . It
##   [text1, 18108]    stooping to | work | again,
##   [text1, 18398]        fell to | work | at his
##   [text1, 18446]       down his | work | , put
##   [text1, 33538]      a night's | work | to do
##   [text1, 36654]         Get to | work | , get
##   [text1, 36658]         get to | work | ."
##   [text1, 42587]       his last | work | , three
##   [text1, 45020]     setting to | work | in earnest
##   [text1, 53330]             -- | work | ."
##   [text1, 54753]     already at | work | on the
##   [text1, 55378]         to his | work | besides mere
##   [text1, 58824] old unfinished | work | , were
##   [text1, 64229]         at her | work | , alone
##   [text1, 68136]        I shall | work | you for
##   [text1, 71751]         was at | work | ."
##   [text1, 72053]     leaving my | work | on the
##   [text1, 72405]       again at | work | upon the
##   [text1, 73130]          to my | work | . There
##   [text1, 73392]       when the | work | of the
##   [text1, 74173]          " All | work | is stopped
##   [text1, 75376]          " You | work | hard,
##   [text1, 78713]        off her | work | as her
##   [text1, 79806]       from her | work | and her
##   [text1, 80700]       with her | work | in her
##   [text1, 80755] the mechanical | work | was a
##   [text1, 81151]      enough to | work | and read
##   [text1, 81163]      her usual | work | , nor
##   [text1, 84838]         at his | work | before,
##   [text1, 84895]       over his | work | again.
##   [text1, 84924]       did that | work | ; and
##   [text1, 84966]         at the | work | in his
```

```
##   [text1, 85056]         in his | work | ."
##   [text1, 85839]         and to | work | . On
##   [text1, 85909]       with her | work | , several
##   [text1, 86247]         on his | work | , and
##   [text1, 86948]       his late | work | , the
##   [text1, 88818] - Blacksmith's | work | , Blacksmith's
##   [text1, 88821] , Blacksmith's | work | . We
##   [text1, 88850]           , to | work | at a
##   [text1, 91522]       when her | work | would slowly
##   [text1, 94291]             !" | Work | , comrades
##   [text1, 94296]           all, | work | ! Work
##   [text1, 94298]          work! | Work | , Jacques
##   [text1, 94335]             -- | work | !"
##   [text1, 94465]         , hard | work | at neighbouring
##   [text1, 96840]   been without | work | before,
##   [text1, 96845]       had this | work | always ready
##   [text1, 98596]      day's bad | work | , for
##  [text1, 100146]       cease to | work | ?"
##  [text1, 110785]       going to | work | . In
##  [text1, 114535]          in to | work | at the
##  [text1, 114568]     such awful | work | ! The
##  [text1, 115495]       spell of | work | was feeble
##  [text1, 117035]         in her | work | for the
##  [text1, 121296]           . My | work | is my
##  [text1, 121427]         was at | work | , and
##  [text1, 121519]         in its | work | ."
##  [text1, 135951] little helpful | work | for her
##  [text1, 137273]         him at | work | ?"
##  [text1, 139593]       with its | work | ."
##  [text1, 142958]     obliged to | work | for him
##  [text1, 151142]        with my | work | ? Time
##  [text1, 151187]         get to | work | . Give
##  [text1, 151192]          me my | work | ."
##  [text1, 151240]          me my | work | ! What
##  [text1, 151316]       have his | work | presently.
##  [text1, 152861]      bench and | work | were hidden
##  [text1, 165563]       in their | Work | , count
```

```
kwic(corp_austen_pride, "work", 2)
```

```
##
##   [text1, 13018]      been the | work | of many
##   [text1, 15423]     their own | work | ; _my_
##   [text1, 19185]           , at | work | in the
##   [text1, 21144]        day to | work | on Jane
##   [text1, 38827] gathering her | work | together,
##   [text1, 57814]          . To | work | in this
##   [text1, 61868]     either at | work | in the
##   [text1, 62163]      at their | work | , and
##   [text1, 66519]  seriously to | work | to find
##   [text1, 73139]  scarcely the | work | of a
##   [text1, 76922]      been the | work | of her
##   [text1, 78692]       all the | work | of the
##   [text1, 88113]       of this | work | to give
##   [text1, 92989]     could not | work | such a
```

```
##   [text1, 93409]        be the | work | of her
##  [text1, 123575]   intently at | work | , striving
##  [text1, 123678]        to her | work | , with
##  [text1, 127892]      again at | work | to get
##  [text1, 139556]    admire her | work | said in
##  [text1, 140051]       not the | work | of a
```

```
kwic(corp_dickens_tale, "poor", 2)
```

```
##
##    [text1, 7763]           of my | poor | father,
##    [text1, 9496]          if the | poor | lady had
##    [text1, 9518]     sparing the | poor | child the
##   [text1, 10119]          to the | poor | wronged gentleman
##   [text1, 10562]        from the | poor | young lady
##   [text1, 13705]       that such | poor | cattle always
##   [text1, 16248]          into a | poor | weak stain
##   [text1, 16652]         all his | poor | tatters of
##   [text1, 19223]      while your | poor | heart pined
##   [text1, 19394]           of my | poor | mother hid
##   [text1, 22887]        and this | poor | boy had
##   [text1, 22967]          ever a | poor | devil of
##   [text1, 23250]  disturbed that | poor | woman at
##   [text1, 42598]       some very | poor | instrument,
##   [text1, 44533]     was growing | poor | . Hence
##   [text1, 44577] Farmer-General, | poor | in family
##   [text1, 47173]         for the | poor | little plaything
##   [text1, 47880]      Patches of | poor | rye where
##   [text1, 47890]      patches of | poor | peas and
##   [text1, 48210]         its one | poor | street,
##   [text1, 48215]        with its | poor | brewery,
##   [text1, 48218]        brewery, | poor | tannery,
##   [text1, 48221]        tannery, | poor | tavern,
##   [text1, 48224]         tavern, | poor | stable-yard for
##   [text1, 48231]    post-horses, | poor | fountain,
##   [text1, 48236]       all usual | poor | appointments.
##   [text1, 48242]         had its | poor | people too
##   [text1, 48250]     people were | poor | , and
##   [text1, 48303]       made them | poor | , were
##   [text1, 49440]           was a | poor | figure in
##   [text1, 49654]         heap of | poor | grass.
##   [text1, 49672]        heaps of | poor | grass?
##   [text1, 49865]         heap of | poor | grass.
##   [text1, 54171]        heaps of | poor | grass were
##   [text1, 54520]          to the | poor | live stock
##   [text1, 60429]         ill and | poor | ; you
##   [text1, 63395]          on the | poor | father's account
##   [text1, 64729]        drunken, | poor | creature of
##   [text1, 70147]           " the | poor | woman protested
##   [text1, 78783]         are so | poor | ."
##   [text1, 79008]           ! the | poor | Gaspard!
##   [text1, 79114]    touching the | poor | fellow?
##   [text1, 79335]         fate of | poor | Gaspard.
##   [text1, 80023]             ah, | poor | Gaspard!
##   [text1, 82054]          or the | poor | mother's shock
```

```
##   [text1, 82661]          but my | poor | history pervaded
##   [text1, 84343]        with the | poor | shoemaker's white
##   [text1, 89034]         of this | poor | man's mind
##   [text1, 90999]        Charles, | poor | Mr.
##   [text1, 92300]              ." | Poor | Carton!
##   [text1, 95523]             ' a | poor | physician.
##   [text1, 98794]        and then | poor | lights began
##   [text1, 99101]        hold his | poor | ignorant soul
##   [text1, 99106]         and his | poor | reduced body
##   [text1, 99200]  shrivelled and | poor | as the
##  [text1, 100320]         to very | poor | account.
##  [text1, 100728]       taken its | poor | supper,
##  [text1, 107502]         was the | poor | prisoner's cry
##  [text1, 109189]      cluster of | poor | cottages,
##  [text1, 115389]          of the | poor | wife!
##  [text1, 117072]          is our | poor | prisoner's darling
##  [text1, 117218]           to my | poor | husband.
##  [text1, 117736]       with many | poor | souls.
##  [text1, 118891]        rich and | poor | , bad
##  [text1, 120909]            , my | poor | child,
##  [text1, 122544]            , my | poor | dear?
##  [text1, 125934]        laid his | poor | head on
##  [text1, 135797]       return to | poor | Darnay,
##  [text1, 138685]         all his | poor | blindnesses and
##  [text1, 142419]            , my | poor | fellow,
##  [text1, 143062]         made so | poor | , that
##  [text1, 143174]           time, | poor | fellow,
##  [text1, 146175]       make what | poor | amends I
##  [text1, 147732]    discharge my | poor | mother's trust
##  [text1, 151221]       torture a | poor | forlorn wretch
##  [text1, 156872]            am a | poor | little seamstress
##  [text1, 156928]          with a | poor | little weak
##  [text1, 156992]           to us | poor | , will
##  [text1, 157015]          Such a | poor | weak little
##  [text1, 157196]            , my | poor | sister;
##  [text1, 160959]          " them | poor | things well
##  [text1, 161047]          : them | poor | things well
##  [text1, 161120]            O my | poor | darlings!
##  [text1, 161292]        for them | poor | creeturs now
##  [text1, 162857]           " you | poor | wretch!
##  [text1, 165650]     naturally a | poor | little thing
##  [text1, 166009]          to the | poor | , and
```

```
kwic(corp_austen_pride, "poor", 2)
```

```
##
##     [text1, 842]      for my | poor | nerves.
##    [text1, 5877]          -- | poor | Eliza!
##    [text1, 6854]      be but | poor | consolation to
##   [text1, 23393]       to my | poor | girls,
##   [text1, 29818] relieve the | poor | . Family
##   [text1, 30616]  thought of | poor | Miss Bingley
##   [text1, 42161]      for my | poor | nerves.
##   [text1, 52194]       felt. | Poor | Jane!
##   [text1, 56784] was equally | poor | ?"
```

```
##   [text1, 56939]      a very | poor | opinion of
##   [text1, 62282]      or too | poor | , she
##   [text1, 67321]  others are | poor | . I
##   [text1, 73484]    , always | poor | from the
##   [text1, 79376]     sprang. | Poor | Charlotte!
##   [text1, 80402]      by one | poor | regiment of
##   [text1, 82184]    this for | poor | Jane!
##   [text1, 82396]       . And | poor | Mr.
##   [text1, 82523]          "" | Poor | Wickham!
##   [text1, 84707]     lovers? | Poor | little Lizzy
##   [text1, 85130] luckily too | poor | to be
##   [text1, 90293]      to the | poor | ."
##   [text1, 90521]      to our | poor | friend.
##   [text1, 94024]     done by | poor | Wickham.
##   [text1, 96405]   among the | poor | . With
##   [text1, 99280]  relates to | poor | Lydia.
##   [text1, 99422]       . Our | poor | mother is
##   [text1, 99537]     from my | poor | mother.
##   [text1, 99692]     and our | poor | Lydia would
##  [text1, 100101]        . My | poor | mother is
##  [text1, 100146]   affected. | Poor | Kitty has
##  [text1, 104893]       ; but | poor | dear Lydia
##  [text1, 104976]         am. | Poor | dear child
##  [text1, 106872]        . My | poor | father!
##  [text1, 109243] and without | poor | Lydia?
##  [text1, 112799]    freedom. | Poor | Lydia's situation
##  [text1, 120810]    suppose? | Poor | Reynolds,
##  [text1, 122449]   that this | poor | man cannot
##  [text1, 135342]       to my | poor | sister.
```