Gutenberg Data via Neural Net Documentation

Sebastian Knigge

5 8 2019

Contents

1 Setup

Following packages were used in this script:

```
# loading packages
library(keras)
library(gutenbergr)
library(dplyr)
library(tensorflow)
library(tidyr)
library(stringr)
library(tidytext)
library(udpipe)
library(sampling)
```

2 Example 1 (6 books)

2.1 Get data - Sampling

This is the essential step for setting up the neural net. These functions include the sampling procedure of the *gutenbergr* library.

```
sampling_books <- function(seed=1234, n=20){
  # sample n books from the whole library
  set.seed(seed)
  gutenberg_works() %>%
    # select works with title
    dplyr::filter(!is.na(title)) %>%
    # set the sample sitze
    sample_n(n) %>%
    # set a special download link
    gutenberg_download(
      mirror = "http://mirrors.xmission.com/gutenberg/")
```

```r
}

set_up_books <- function(n_books=4, seed=1992){
  # initial book sample
  books <- sampling_books(n=n_books, seed=seed)
  by_chapter <- books %>%
    group_by(gutenberg_id) %>%
    # split in chapters
    mutate(chapter = cumsum(str_detect(text, regex("^chapter ", ignore_case = TRUE)))) %>%
    ungroup() %>%
    # exclude books without chapters
    dplyr::filter(chapter > 0)
  return(by_chapter)
}

shorten_titles <- function(titles){
  # shorten very long book titles by setting
  # a subset of characters of the first line
  # of the title
  sub_inds <- titles %>%
    regexpr(pattern="\\n|\\r")-1
  sub_inds[sub_inds<0] <- nchar(titles)[sub_inds<0]
  titles %>%
    substr(1,sub_inds)
}

get_titles <- function(x, n_books){
  # get the sampled gutenberg_ids
  unique_ids <- x %>%
    select(gutenberg_id) %>%
    unique() %>% unlist()
  # get the titles
  titles <- gutenberg_works() %>%
    dplyr::filter(gutenberg_id %in% unique_ids) %>%
    select(gutenberg_id, title, author) %>%
    mutate(title=shorten_titles(title))
  # get the number of gutenberg ids
  len <- nrow(titles)
  if(n_books!=len) warning(paste("--- ",n_books-len,
                                 " books have 0 chapters --- "))
  # the output as a list
  ret <- list(
    titles=titles,
    len=len
  )
  return(ret)
}

append_by_chapter <- function(x=by_chapter, n_books, seed_index=1){
  # append the books matrix until
  # we get the desired number of books n_books
  titles <- get_titles(x, n_books)
  n <- titles$len
```

```r
  while (n<n_books) {
    book2add <- sampling_books(n=1, seed=seed_index)
    by_chapter_add <- book2add %>%
      group_by(gutenberg_id) %>%
      # split in chapters
      mutate(chapter = cumsum(str_detect(text, regex("^chapter ", ignore_case = TRUE)))) %>%
      ungroup() %>%
      # exclude books without chapters
      dplyr::filter(chapter > 2)
    titles2add <- get_titles(by_chapter_add, 1)
    # adding the book to by_chapter if there are chapters in the
    # book plus it is not in the data already
    if (titles2add$len==1) if(!titles2add$titles$gutenberg_id%in%titles$titles$gutenberg_id) {
      x <- bind_rows(x, by_chapter_add)
    }
    n<-get_titles(x, n)$len
    seed_index <- seed_index+1
  }
  return(x)
}

exclude_stop_words <- function(x){
  # unite chapter and document title
  by_chapter_word <- x %>%
    unite(document, gutenberg_id, chapter) %>%
    # split into words
    unnest_tokens(word, text)
  # import tibble stop words
  data(stop_words)
  # find document-word counts
  word_counts <- by_chapter_word %>%
    # exclude stop words
    anti_join(stop_words) %>%
    # count each word by chapter
    count(document, word, sort = TRUE) %>%
    ungroup()
  return(word_counts)
}

convert_to_dtm <- function(x, minfq = 2){
  # get into a format lda can handle
  chapters_dtm <- x %>%
    select(doc_id=document, term=word, freq=n) %>%
    document_term_matrix() %>%
    # reduce by low frequencies
    dtm_remove_lowfreq(minfreq = minfq)
  return(chapters_dtm)
}


convert_to_dtm_2 <- function(x, n=n, minfq = 2, top=10000){
  # get into a format lda can handle
  chapters_dtm <- x %>%
```

```
    select(doc_id=document, term=word, freq=n) %>%
    document_term_matrix() %>%
    # reduce by low frequencies
    dtm_remove_tfidf(top=top)
  return(chapters_dtm)
}

# convert x matrix into a form such that it can be used for tensorflow
adjust_tensor_format <- function(x){
  x_chapters <- apply(x, 1, function(x) as.matrix(x)) %>% t()
  topics <- x %>% rownames() %>% as_tibble() %>%
    separate(value, c("gutenberg_id", "chapter"), sep = "_", convert = TRUE) %>%
    select(gutenberg_id) %>%
    # split joint name of book and chapter
    as.matrix %>% as.factor() %>% as.integer()
  # one hot encoding for the chapters (y)
  topics_categorical <- topics %>% -1 %>%
    to_categorical()
  ret <- list(
    x=x_chapters,
    y=topics_categorical,
    topics=topics
  )
  return(ret)
}
```

Now we can use all these functions to get to the initial corpus sample. the corpus of this example is equivalent to the corpus in Example 1 of the LDA Gutenberg Documentation.

```
n_books <- 6
by_chapter <- set_up_books(n_books=n_books, seed=222)
get_titles(by_chapter, n_books)
```

```
## Warning in get_titles(by_chapter, n_books): --- 1 books have 0 chapters ---
```

```
## $titles
## # A tibble: 5 x 3
##   gutenberg_id title                                      author
##          <int> <chr>                                      <chr>
## 1           11 Alice's Adventures in Wonderland           Carroll, Lewis
## 2         3096 Beatrice                                   Haggard, H. Rider~
## 3        25603 Detailed Minutiae of Soldier life in the~ McCarthy, Carlton
## 4        47402 Along Alaska's Great River                 Schwatka, Frederi~
## 5        49675 Hawkins Electrical Guide v. 5 (of 10)      Hawkins, N. (Nehe~
##
## $len
## [1] 5
```

The function set_up_books() (defined above) returns a warning that one book seems to consist of only one chapter. In order to get a copus consisting out of 6 books, the function append_by_chaper() is used, which fills up the corpus to the desired number of books.

```
appended_by_chapter <- append_by_chapter(x=by_chapter, n_books = n_books)
word_counts <- exclude_stop_words(appended_by_chapter)
```

```
## Joining, by = "word"
```

In table 1 the sampled titles for the book sample with the seed 222 are displayed. It appears through the function append_by_chapter() one book was added, called “My Novel” — Volume 04“.

```
titles <- get_titles(appended_by_chapter, n_books)
titles$titles %>% stargazer(summary=FALSE, font.size = "footnotesize",
                            header=FALSE, title="Book-titles", rownames=FALSE,
                            label="titles:6books")
```

Table 1: Book-titles

gutenberg_id	title	author
11	Alice’s Adventures in Wonderland	Carroll, Lewis
3096	Beatrice	Haggard, H. Rider (Henry Rider)
7705	"My Novel" — Volume 04	Lytton, Edward Bulwer Lytton, Baron
25603	Detailed Minutiae of Soldier life in the Army of Northern Virginia, 1861-1865	McCarthy, Carlton
47402	Along Alaska’s Great River	Schwatka, Frederick
49675	Hawkins Electrical Guide v. 5 (of 10)	Hawkins, N. (Nehemiah)

2.2 Reduction of the dimensionality

As mentioned in the documentation for the LDA using Gutenberg data, also here there are different possible embedding methods. The function *convert_to_dtm* takes the parameter *minfq*, which is used to reduce the “bag of words” (i.e. dimensionality). *minfq* is the minimum frequency for the bag of words dictionary. I will refer to this as “embedding”. Let us set it to 2 in this case, meaning that we include a word only if the frequency is 2 or more.

```
chapters_dtm <- convert_to_dtm(word_counts, minfq=2)
adjusted_format <- adjust_tensor_format((chapters_dtm))
( M2 <- ncol(chapters_dtm) )
```

```
## [1] 10685
```

Let us compare it to the case if we include all words.

```
chapters_dtm_all <- convert_to_dtm(word_counts, minfq=0)
adjusted_format_all <- adjust_tensor_format((chapters_dtm_all))
( M <- ncol(chapters_dtm_all) )
```

```
## [1] 17961
```

We also want to compare this to a reduction of the word dictionary by *tf-idf*. We reduce to 50% of the original size of the bag-of-words.

```
chapters_dtm_tfidf <- convert_to_dtm_2(word_counts, top=(M*0.5))
adjusted_format_tfidf <- adjust_tensor_format((chapters_dtm_tfidf))
ncol(chapters_dtm_tfidf)
```

```
## [1] 8980
```

2.3 Splitting and Fitting

The following function splits the sample in an manner, such that each cluster is eually to its size represented in the test data and the validation data.

```
sample_cluster_wise <- function(data, test_ratio=0.1, val_ratio=0.2, seed=1234){
  X <- data$x; y <- data$y
```

```
  cluster=data$topics
  set.seed(seed)
  {
  # setting the absolute number of observations for the sample of each cluster
  n_test <- (table(cluster)*test_ratio) %>% floor() %>%
    # use at least one observation of each cluster
    sapply(., function(x) max(x,1))
  n_val <- (table(cluster)*val_ratio) %>% floor() %>%
    # use at least one observation of each cluster
    sapply(., function(x) max(x,1))
  # function to get the correct sample indices for validation and test sample
  samp_ind <- function(i, n_list) which(cluster==i) %>% sample(n_list[i])
  test_indices <- unique(cluster) %>% sort() %>%
    sapply(function(i) samp_ind(i, n_test)) %>%
    unlist()
  val_indices <- unique(cluster) %>% sort() %>%
    sapply(function(i) samp_ind(i, n_val)) %>%
    unlist()
  }
  ret <- list(partial_x_train=X[-c(val_indices,test_indices),],
              partial_y_train=y[-c(val_indices,test_indices),],
              x_val=X[val_indices,], y_val = y[val_indices,],

              x_test=X[test_indices,], y_test = y[test_indices,])
  return(ret)
}
```

The following two functions are setting up the model in training, validation and testing sets and evaluate the goodness of fit.

```
# The whole model is set up and trained within this function
set_up_n_fit <- function(split, books_n=n_books){
  # starting with 64 neurons and scaling it down to 46 in the
  # mid layer turned out to be a well predicting model
  model <- keras_model_sequential() %>%
    layer_dense(units=64, activation="relu", input_shape=ncol(split$partial_x_train)) %>%
    layer_dense(units=46, activation="relu") %>%
    # we want to classify for as many categories as books
    layer_dense(units=books_n, activation="softmax")

  model %>% compile(
    optimizer="rmsprop",
    loss="categorical_crossentropy",
    metrics=c("accuracy"))

  history <- model %>% fit(
    split$partial_x_train,
    split$partial_y_train,
    # from experience the model tends to
    # overfit for more than 5 epochs
    epochs=5,
    batch_size=512,
    validation_data=list(split$x_val,split$y_val)
  )
```

```r
  return(
    list(history=history,
         model=model))
}


# making a prediction on the test data and calculating the
# mispecification rate; we also want to save the true categories and the predicted ones
evaluate_model <- function(model_fit, y=split$y_test, x=split$x_test) {
  prediction <- model_fit %>% predict(x)
  pred <- apply(prediction, 1, which.max)
  true_value <- apply(y, 1, which.max)
  mispecified <- sum(!pred==true_value)/length(pred)
  ret <- list(mispecified=mispecified,
              # the function also dicloses the true and the predicted
              # values for exact evaluation, if needed
              pred=pred, true_value=true_value)
  return(ret)
}
```

Now we can start applying the model. We will measure the misclassification ratio, the fitting time and store the results for evaluation.

```r
tim1 <- Sys.time()
n <- 59
results <- rep(NA,n)
for(i in 1:n){
  split <- sample_cluster_wise(adjusted_format, seed=i*2)
  results[i] <- set_up_n_fit(split) %>% .$model %>%
    evaluate_model() %>% .$mispecified
}
tim2 <- Sys.time()
```

We will present some statistics of the results of the evaluation. We consider the misclassification rate als primary measure to evaluate the fit. The results of the misclassification rate over 59 splits and fits of the model are:

```r
results %>% summary
```

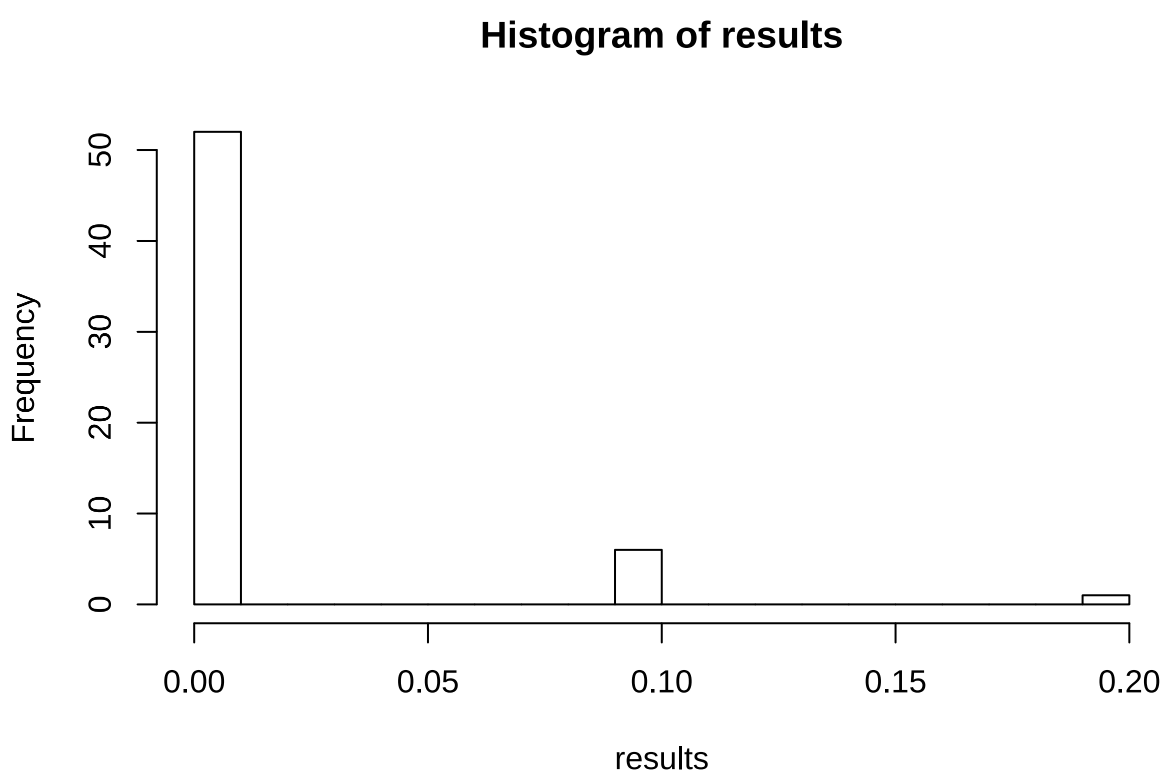
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00000 0.00000 0.00000 0.01356 0.00000 0.20000
```

```r
results %>% var
```

```
## [1] 0.001537113
```

A histogram might be helpful in visualizing the results.

```r
hist(results, breaks = 20)
```



The fitting time might be interesting as well.

```
# mfreq=2
(u1 <- tim2-tim1)
```

```
## Time difference of 4.353287 mins
```

We now apply this very same procedure using the full bag of words:

```
tim1_all <- Sys.time()
n <- 59
results_all <- rep(NA,n)
for(i in 1:n){
  split <- sample_cluster_wise(adjusted_format_all, seed=i*2)
  results_all[i] <- set_up_n_fit(split) %>% .$model %>%
    evaluate_model() %>% .$mispecified
}
tim2_all <- Sys.time()
```

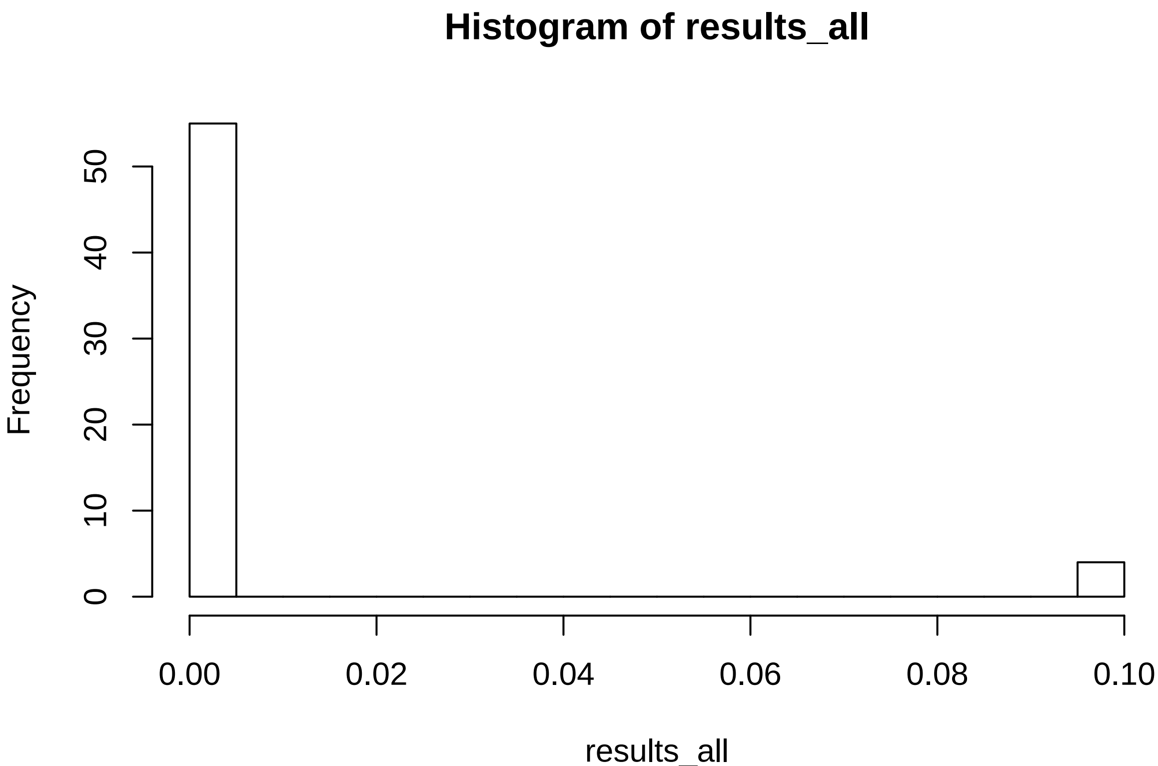
```
results_all %>% summary
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00000 0.00000 0.00000 0.00678 0.00000 0.10000
```

```
results_all %>% var
```

```
## [1] 0.0006428989
```

```
hist(results_all, breaks = 20)
```




```
# mfreq=2
(u2 <- tim2_all-(tim1_all))
```

```
## Time difference of 11.24102 mins
```

In this step we use the embeding via *tf-idf* and fit the model. The *tf-idf* embedding equals a reduction by 50%.

```
tim1_tfidf <- Sys.time()
n <- 59
results_tfidf <- rep(NA,n)
for(i in 1:n){
  split <- sample_cluster_wise(adjusted_format_tfidf, seed=i*2)
  results_tfidf[i] <- set_up_n_fit(split) %>% .$model %>%
    evaluate_model() %>% .$mispecified
}
tim2_tfidf <- Sys.time()
```

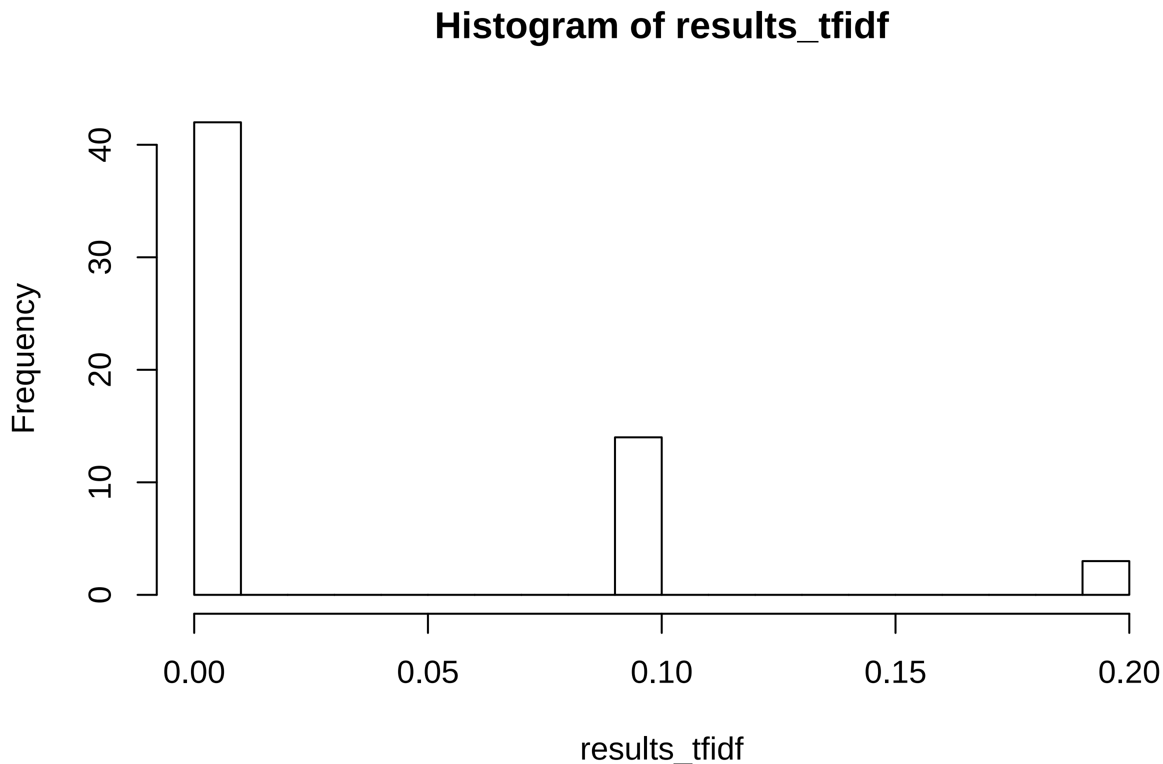
```
results_tfidf %>% summary
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  0.0000  0.0339  0.1000  0.2000
```

```
results_tfidf %>% var
```

```
## [1] 0.003313852
```

```
hist(results_tfidf, breaks = 20)
```




```
# mfreq=2
(u3 <- tim2_tfidf-(tim1_tfidf))
```

```
## Time difference of 17.41653 mins
```

Table 2 summarises the results of all 3 different embeding mthods.

```
performance_matrix <- data.frame(freq2.embedding=c(results%>%mean, u1),
           all.embedding=c(results_all%>%mean, u2),
           tfidf=c(results_tfidf%>%mean, u3))
rownames(performance_matrix) <- c("misc. rate", "time")
performance_matrix %>%  stargazer(summary=FALSE, header=F,
                                  title="Performance of embeddings",
                                  label="performancematrix")
```

Table 2: Performance of embeddings

	freq2.embedding	all.embedding	tfidf
misc. rate	0.014	0.007	0.034
time	4.353	11.241	17.417

The time for the embedding using more than frequency 2 is very short. Whereas the embedding via *tf-idf* with the same dimensionality takes very long. It certainly makes sense that the mislcassification rate is better for embedding with the full vocabulary than for the other two embeddings.

3 Example 2 (6 Books)

Also the second example is based on the same sample of Example 2 of the LDA Gutenberg Documentation. This is $seed = 101$. We will use the same calculations as done for Example 1. This second example shall halp providing more valid results.

```
n_books_sec <- 6
by_chapter_sec <- set_up_books(n_books=n_books_sec, seed=101)
appended_by_chapter_sec <- append_by_chapter(x=by_chapter_sec, n_books = n_books_sec)
word_counts_sec <- exclude_stop_words(appended_by_chapter_sec)
```

```
## Joining, by = "word"
```

```
titles_sec <- get_titles(appended_by_chapter_sec, n_books)
gbids_sec <- titles_sec$titles$gutenberg_id
categories_sec <- gutenberg_works() %>%
  filter(gutenberg_id %in% gbids_sec) %>%
  select(gutenberg_id, gutenberg_bookshelf)
```

Table 3 displays the sampled books for this example.

```
titles <- get_titles(appended_by_chapter_sec, n_books)
titles$titles %>% stargazer(summary=FALSE, font.size = "footnotesize",
                            header=FALSE, title="Book-titles for Example 2 (6 Books)", rownames=FALSE,
                            label="titles:6books_sec")
```

Table 3: Book-titles for Example 2 (6 Books)

gutenberg_id	title	author
2029	Lahoma	Ellis, J. Breckenridge (John Breckenridge)
12035	Progressive Morality: An Essay in Ethics	Fowler, Thomas
14625	Military Instructors Manual	
18633	My Lady of Doubt	Parrish, Randall
34037	Under the Southern Cross	Ballou, Maturin Murray
36697	Revisiting the Earth	Hill, James Langdon

```
chapters_dtm_sec <- convert_to_dtm(word_counts_sec, minfq=2)
adjusted_format_sec <- adjust_tensor_format((chapters_dtm_sec))
( M2 <- ncol(chapters_dtm_sec) )
```

```
## [1] 12704
```

```
chapters_dtm_all_sec <- convert_to_dtm(word_counts_sec, minfq=0)
adjusted_format_all_sec <- adjust_tensor_format((chapters_dtm_all_sec))
( M <- ncol(chapters_dtm_all_sec) )
```

```
## [1] 20634
```

```
chapters_dtm_tfidf_sec <- convert_to_dtm_2(word_counts_sec, top=(M*0.5))
adjusted_format_tfidf_sec <- adjust_tensor_format((chapters_dtm_tfidf_sec))
ncol(chapters_dtm_tfidf_sec)
```

```
## [1] 10317
```

```
tim1 <- Sys.time()
n <- 59
results_sec <- rep(NA,n)
for(i in 1:n){
```

```
  split <- sample_cluster_wise(adjusted_format_sec, seed=i*2)
  results_sec[i] <- set_up_n_fit(split) %>% .$model %>%
    evaluate_model() %>% .$mispecified
}
tim2 <- Sys.time()
# evaluate fitting time
(u1_sec <- tim2-tim1)
```

```
## Time difference of 24.758 mins
```

```
tim1 <- Sys.time()
n <- 59
results_all_sec <- rep(NA,n)
for(i in 1:n){
  split <- sample_cluster_wise(adjusted_format_all_sec, seed=i*2)
  results_all_sec[i] <- set_up_n_fit(split) %>% .$model %>%
    evaluate_model() %>% .$mispecified
}
tim2 <- Sys.time()
# evaluate fitting time
(u2_sec <- tim2-tim1)
```

```
## Time difference of 31.37081 mins
```

```
tim1 <- Sys.time()
n <- 59
results_tfidf_sec <- rep(NA,n)
for(i in 1:n){
  split <- sample_cluster_wise(adjusted_format_tfidf_sec, seed=i*2)
  results_tfidf_sec[i] <- set_up_n_fit(split) %>% .$model %>%
    evaluate_model() %>% .$mispecified
}
tim2 <- Sys.time()
# evaluate fitting time
(u3_sec <- tim2-tim1)
```

```
## Time difference of 37.56571 mins
```

A overview of the results will be found in Table 4.

```
performance_matrix <- data.frame(freq2.embedding=c(results_sec%>%mean, u1_sec),
           all.embedding=c(results_all_sec%>%mean, u2_sec),
           tfidf=c(results_tfidf_sec%>%mean, u3_sec))
rownames(performance_matrix) <- c("misc. rate", "time")
performance_matrix %>%  stargazer(summary=FALSE, header=F,
                                  title="Performance of embeddings for Example 2",
                                 label="tab:sec_results")
```

Table 4: Performance of embeddings for Example 2

	freq2.embedding	all.embedding	tfidf
misc. rate	0.121	0.103	0.100
time	24.758	31.371	37.566