Homework Assignment 2: Applied Probabilistic Models

Analysis of the Book Structure

5273

1 Introduction

For this work, data is collected on the free eBooks library Project Gutenberg [1]. The chosen book for the analysis is: "The Autobiography of Benjamin Franklin". Data obtained from the Project Gutenberg are in `txt` format.

For the analysis, it is used the R software in its version 4.0.2 [3], and the code used is available on the GitHub repository [2]. This work is run on a MacBook Air with an Intel Core i5 CPU @ 1.8 GHz and 8 GB RAM.

2 Data

The book is downloaded directly from the web and in order to develop the analysis, the following code is used.

```
require(gutenbergr) #Download books from online library
require(tidytext) #Clean text
require(dplyr)  #Data Manipulation

library(textshape)

#Load the book: "The Autobiography of Benjamin Franklin"
book<-gutenberg_download(c(148))
```

a2.R

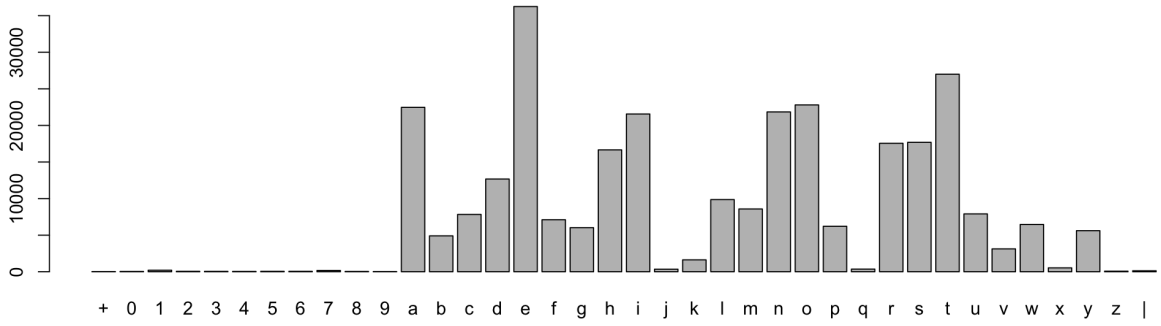
The book has a total of 294 003 characters (letters) and 66 520 words. This data is used to a further analysis of what the most important letters and words are according to its frequency. The next part of the code is set to this objective.

```r
#Variables used:
letters <- book %>% unnest_tokens(chars, text, "characters") #contains letters
words <- book %>% unnest_tokens(word, text, "words") #contains words

#Work with Letters:
png('barplot_letters.png',width = 2100, height = 768,res = 180)
barplot(table(letters$chars)) #Barplot of letters
dev.off()

freq_l <- as.data.frame(table(letters$chars))
names(freq_l) <- c('Letter', 'FrequencyL')

rl <- freq_l[freq_l$FrequencyL>500,] #Filter relevant letters
png('barplot_relevant_letters.png',width = 1366, height = 768,res = 150)
barplot(rl$FrequencyL, names.arg = rl$Letter)
dev.off()

rlo <- rl[order(rl$FrequencyL, decreasing=TRUE),]
png('barplot_relevant_letters_ordered.png',width = 1366, height = 768,res = 150)
barplot(rlo$FrequencyL, names.arg = rlo$Letter)
dev.off()

#Work with Words:
png('barplot_words.png',width = 1366, height = 768,res = 150)
barplot(sort(table(words$word), decreasing = TRUE)) #Barplot of words
dev.off()

freq_w <- as.data.frame(table(words$word))
names(freq_w) <- c('Word', 'FrequencyW')

muw <- freq_w[freq_w$FrequencyW > 100,]
png('barplot_most_used_words.png',width = 2100, height = 768,res = 180)
barplot(muw$FrequencyW, names.arg = muw$Word)
dev.off()

rw <- muw[muw$FrequencyW < 200,]
png('barplot_relevant_words.png',width = 2048, height = 768,res = 150)
barplot(rw$FrequencyW, names.arg = rw$Word)
dev.off()

rw_o <- rw[order(rw$FrequencyW, decreasing=TRUE),]
png('barplot_relevant_words_ordered.png',width = 2166, height = 768,res = 180)
barplot(rw_o$FrequencyW, names.arg = rw_o$Word)
barplot(rw_o$FrequencyW, names.arg = rw_o$Word, log = 'y')
dev.off()

places<-as.data.frame(table(grep("york|london|boston|newport|philadelphia|paris",
    words$word,value=TRUE)))
names(places) <- c('Place', 'FrequencyP')
rp <- places[places$FrequencyP>2,] #Filter relevant places
png('barplot_places.png',width = 1366, height = 768,res = 150)
barplot(rp$FrequencyP, names.arg = rp$Place)
dev.off()
```

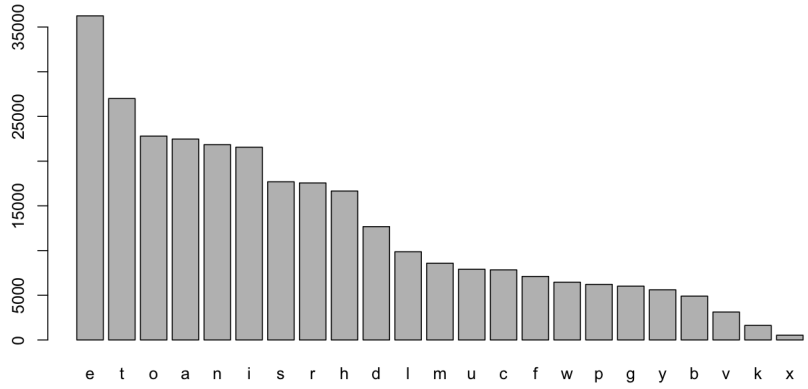
a2.R

2.1 Letters

For the analysis of the letters, barplots are generated. Here the horizontal axis represents all the used letters and the vertical one the corresponding frequencies. Figure 1 shows in (a) all the letters present in the document, sorted in decreasing order and by the same token in (b) it can be seen the most used



(a) Barplot of all letters present in the document



(b) Barplot of most used letters

Figure 1: Barplots of letters present in the document

letters. In this latter case frequency greater than 500 is the criteria to fall in this category. Frequencies of the six most used letters are shown in Table 1.

Table 1: Frequency of most used letters

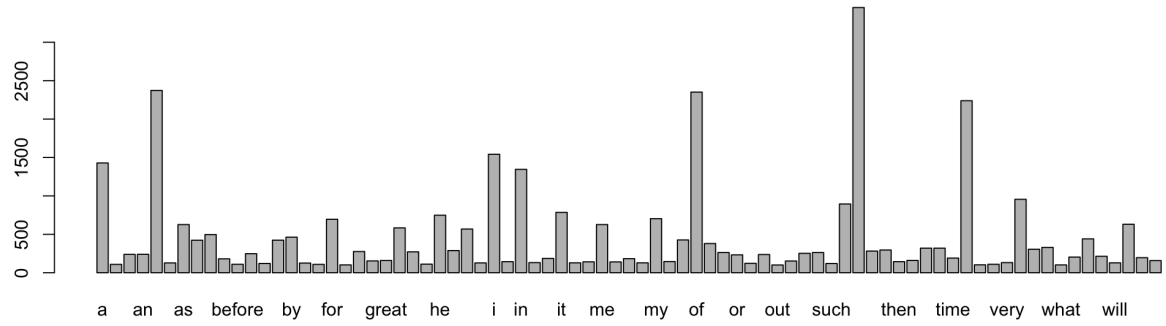
Letter	Frequency
e	36 252
t	27 001
o	22 803
a	22 472
n	21 845
i	21 561

2.2 Words

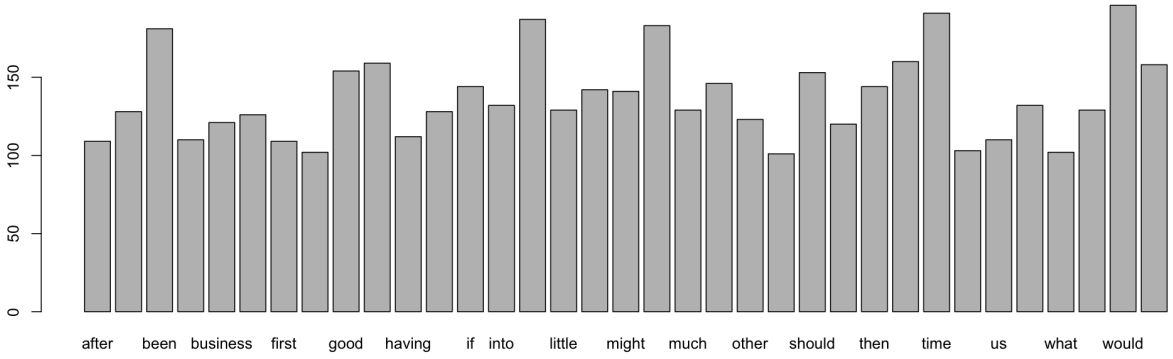
In a similar fashion, barplots are generated for the analysis of words. Here the horizontal axis represents all the used words in the document and the vertical one corresponds to the respective frequencies.

In the first attempt, it is hard to see a difference among words, that is why the following barplots are generated. Figure 2 shows the words filtered. In (a) it can be seen the frequency of the most used words. For this category, it is set as a requirement a frequency greater than 100. Furthermore, in (b), for this category, it is discarded the words with a frequency greater than 250 because articles and prepositions are included and these words are not very descriptive. Finally, those relevant words are sorted in decreasing order in Figure 3.

Last, in Figure 4 it is generated a barplot, which shows an analysis of several places mentioned in the autobiography, and it can be seen as the most relevant ones in the life of this character. From this barplot most of the events related to Franklin took place in the city of Philadelphia.



(a) Barplot of most used words present in the document



(b) Barplot of relevant words

Figure 2: Barplots of words present in the document

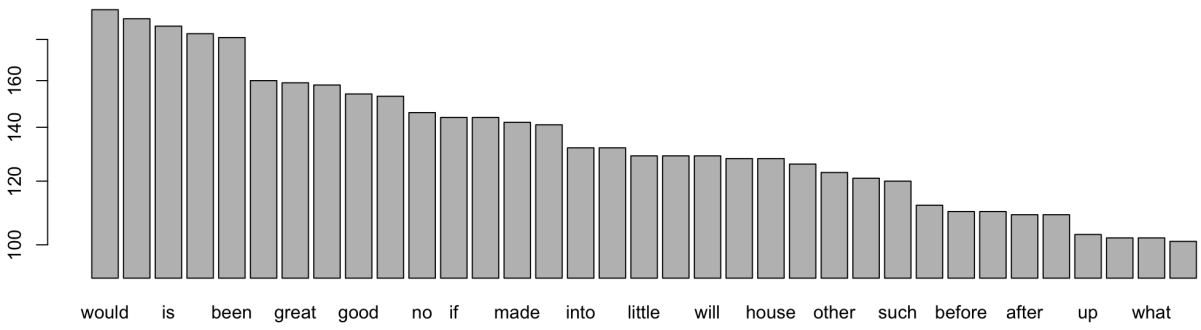


Figure 3: Barplot of relevant words present in the document (sorted)

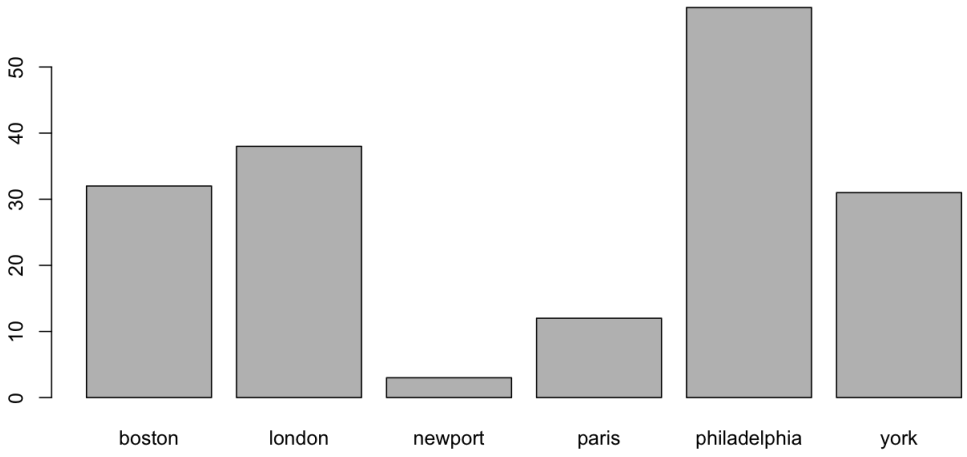


Figure 4: Barplot of relevant places mentioned in the document

References

[1] Michael Hart. Project Gutenberg, 1971. http://www.gutenberg.org/ebooks/, Last accessed on 2020-09-09.

[2] Hernandez, Oscar. Probability in R. https://github.com/oscaralejandro1907/probability-in-R/blob/master/assignment1/t1.R, 2020.

[3] The R Foundation. The R Project for Statistical Computing. https://www.r-project.org/, 2020.