

Using R to create a Wordcloud from text in an ebook

Judy Minichelli

November 2, 2017

Abstract

The article gives instructions on how to download text from an ebook and create a wordcloud in R. A wordclouds is a data visualization that shows the most commonly used words in a large text dataset where size is proportional to frequency; words with the highest count appear larger and in bold. The ebook used in this example is "The Wonderful Wizard of Oz"

The Wonderful Wizard of Oz was written by Frank Baum and published in 1900. It was the basis for the 1902 Broadway musical and the classic 1939 movie starring Judy Garland¹.

1 R Packages

The following packages were installed and brought in with library: dplyr, gutenbergr, stringr, tidytext, tm, wordcloud.

The Gutenberg Package is used to download the book text. The Wonderful Wizard of Oz is ebook number 55 as identified in the book's bibrec tab on the Project Gutenberg website (?). but you can also search by title to determine a book number using: `gutenberg_works(title == 'entertitlehere')`. Or, if you aren't sure of the title you can keywords: `gutenberg_works(title == str_detect(title, 'Frankenstein'))`

The function below stores the result of the book text download in the dataframe "wizard".

```
wizard<-gutenberg_download(55)

## Error in gutenbergr::gutenberg_download(55): could not find function "gutenberg_download"
```

This dataframe has two columns, one for each line in the book. For this exercise, it is not necessary to exclude chapter headings and the first few pages of text; however, the procedures below will accomplish the task.

¹https://en.wikipedia.org/wiki/The_Wonderful_Wizard_of_Oz.

2 How to Clean Data

```
library(stringr)

## Warning: package 'stringr' was built under R version 3.4.2

Wizard<-Wizard%>%
  filter(!str_detect(wizard$text, '^CHAPTER'))

## Error in eval(lhs, parent, parent): object 'Wizard' not found
```

The actual text begins on line 10 and ends on 4721 so the wizard data frame can be redefined to exclude the first 9 lines.

```
wizard<-wizard[10:4721,]

## Error in eval(expr, envir, enclos): object 'wizard' not found
```

3 The Wordcloud

To make the wordcloud, we first have to break up the lines into words. We can use a function from the tidytext package for this:

```
library(tidytext)

## Warning: package 'tidytext' was built under R version 3.4.2

words_df<-wizard%>%
  unnest_tokens(word, text)

## Error in eval(lhs, parent, parent): object 'wizard' not found

words_df

## Error in eval(expr, envir, enclos): object 'words_df' not found
```

We can remove common, unimportant words with the stop_words data frame and some dplyr:

```
words_df<-words_df%>%
  filter(!(word %in% stop_words$word))

## Error in eval(lhs, parent, parent): object 'words_df' not found

words_df

## Error in eval(expr, envir, enclos): object 'words_df' not found
```

Now, we need to calculate the frequencies of the words in the novel. Again, we can use standard dplyr techniques for this:

```
word_freq<-words_df%>%
  group_by(word)%>%
  summarize(count=n())

## Error in eval(lhs, parent, parent): object 'words_df' not found

word_freq

## Error in eval(expr, envir, enclos): object 'word_freq' not found
```

Finally, it's time to generate the wordcloud:

```
library(wordcloud)
library(tm)
wordcloud(word_freq$word,word_freq$count,min.freq=20)

## Error in wordcloud(word_freq$word, word_freq$count, min.freq = 20):
## object 'word_freq' not found
```