Treasure Island Wordcloud

Matthew Haines

December 17, 2017

Abstract

This PDF will contain a wordcloud and title of the book 'Treasure Island' by Robert Louis Stevenson.

Treasure Island

1 packages

This section will contain the packages which will then be used to load 'Treasure Island', manipulate string and form wordclouds.

```
> package<-c('dplyr')
> library(tidytext)
> library(tm)
> library(wordcloud)
> library(stringr)
> library(dplyr)
> library(knitr)
> library(gutenbergr)
```

The first step is to determine the id of Treasure Island:

```
> gutenberg_works()%>%
+   select(gutenberg_id,title,author)%>%
+   filter(title=='Treasure Island')

# A tibble: 1 x 3
  gutenberg_id           title                   author
         <int>           <chr>                    <chr>
1          120 Treasure Island Stevenson, Robert Louis
```

In the resulting tibble from the code above, one can pick out the id of the book; 120.

2 Chapter 1

Here I want to isolate the 'chapter 1' block of text

```
> library(stringr)
> df <- gutenberg_download(120)
> head(df[str_detect(df$text, '^CHAPTER'),],n=1)$text

character(0)
```

3 The Wordcloud

Next we will create a database of all the words in Treasure Island.

```
> words_df<-df%>%
+   unnest_tokens(word,text)
> words_df

# A tibble: 69,111 x 2
   gutenberg_id      word
          <int>     <chr>
 1          120  treasure
 2          120    island
 3          120        by
 4          120    robert
 5          120     louis
 6          120 stevenson
 7          120  treasure
 8          120    island
 9          120        to
10          120     s.l.o
# ... with 69,101 more rows
```

Using dplyr, we can remove stop and insigificant words. Then using the word cloud package we can compile the words into a wordcloud.

```
> words_df<-words_df%>%
+   filter(!(word %in% stop_words$word))
> words_free <- words_df%>%
+   group_by(word)%>%
+   summarise(count = n())%>%
+   arrange(-count)
> wordcloud(words_free$word, words_free$count, min.freq = 25)
```