Moby Dick Word Cloud

Jason Finkle

2025-07-18

library(rvest)  
library(tidyverse)

## Warning: package 'ggplot2' was built under R version 4.3.3

## Warning: package 'purrr' was built under R version 4.3.3

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.4 ✔ readr 2.1.5  
## ✔ forcats 1.0.0 ✔ stringr 1.5.1  
## ✔ ggplot2 3.5.2 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.3 ✔ tidyr 1.3.1  
## ✔ purrr 1.0.4   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ readr::guess\_encoding() masks rvest::guess\_encoding()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(dplyr)  
library(ggplot2)  
library("tm")

## Warning: package 'tm' was built under R version 4.3.3

## Loading required package: NLP

## Warning: package 'NLP' was built under R version 4.3.3

##   
## Attaching package: 'NLP'  
##   
## The following object is masked from 'package:ggplot2':  
##   
## annotate

library("SnowballC")

## Warning: package 'SnowballC' was built under R version 4.3.3

library("wordcloud")

## Warning: package 'wordcloud' was built under R version 4.3.3

## Loading required package: RColorBrewer

## Warning: package 'RColorBrewer' was built under R version 4.3.3

library("RColorBrewer")  
  
library(shinyWidgets)

## Warning: package 'shinyWidgets' was built under R version 4.3.3

library(readr)  
moby\_dick <- read\_csv("~/Desktop/mobydick.txt")

## Warning: One or more parsing issues, call `problems()` on your data frame for details,  
## e.g.:  
## dat <- vroom(...)  
## problems(dat)

## Rows: 19220 Columns: 2  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (2): The Project Gutenberg EBook of Moby Dick; or The Whale, by Herman  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

# Load the data as a corpus  
docs <- Corpus(VectorSource(moby\_dick))  
  
  
toSpace <- content\_transformer(function (x , pattern ) gsub(pattern, " ", x))  
docs <- tm\_map(docs, toSpace, "/")

## Warning in tm\_map.SimpleCorpus(docs, toSpace, "/"): transformation drops  
## documents

docs <- tm\_map(docs, toSpace, "@")

## Warning in tm\_map.SimpleCorpus(docs, toSpace, "@"): transformation drops  
## documents

docs <- tm\_map(docs, toSpace, "\\|")

## Warning in tm\_map.SimpleCorpus(docs, toSpace, "\\|"): transformation drops  
## documents

#clean the text  
  
# Convert the text to lower case  
docs <- tm\_map(docs, content\_transformer(tolower))

## Warning in tm\_map.SimpleCorpus(docs, content\_transformer(tolower)):  
## transformation drops documents

# Remove numbers  
docs <- tm\_map(docs, removeNumbers)

## Warning in tm\_map.SimpleCorpus(docs, removeNumbers): transformation drops  
## documents

# Remove english common stopwords  
docs <- tm\_map(docs, removeWords, stopwords("english"))

## Warning in tm\_map.SimpleCorpus(docs, removeWords, stopwords("english")):  
## transformation drops documents

# Remove your own stop word  
# specify your stopwords as a character vector  
docs <- tm\_map(docs, removeWords, c("blabla1", "blabla2"))

## Warning in tm\_map.SimpleCorpus(docs, removeWords, c("blabla1", "blabla2")):  
## transformation drops documents

# Remove punctuations  
docs <- tm\_map(docs, removePunctuation)

## Warning in tm\_map.SimpleCorpus(docs, removePunctuation): transformation drops  
## documents

# Eliminate extra white spaces  
docs <- tm\_map(docs, stripWhitespace)

## Warning in tm\_map.SimpleCorpus(docs, stripWhitespace): transformation drops  
## documents

# Text stemming  
docs <- tm\_map(docs, stemDocument)

## Warning in tm\_map.SimpleCorpus(docs, stemDocument): transformation drops  
## documents

#Term Document Matrix  
  
dtm <- TermDocumentMatrix(docs)  
m <- as.matrix(dtm)  
v <- sort(rowSums(m),decreasing=TRUE)  
d <- data.frame(word = names(v),freq=v)  
head(d, 10)

## word freq  
## whale whale 1316  
## one one 906  
## now now 738  
## “ “ 646  
## like like 585  
## upon upon 564  
## seem seem 469  
## ship ship 465  
## sea sea 455  
## old old 441

#Word Cloud  
  
set.seed(1234)  
wordcloud(words = d$word, freq = d$freq, min.freq = 1,  
 max.words=200, random.order=FALSE, rot.per=0.35,   
 colors=brewer.pal(8, "Dark2"))

## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.words =  
## 200, : instant could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.words =  
## 200, : open could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.words =  
## 200, : sudden could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.words =  
## 200, : reason could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.words =  
## 200, : remain could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.words =  
## 200, : howev could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.words =  
## 200, : stood could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.words =  
## 200, : wind could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.words =  
## 200, : dark could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.words =  
## 200, : whether could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.words =  
## 200, : inde could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.words =  
## 200, : perhap could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.words =  
## 200, : touch could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.words =  
## 200, : carri could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.words =  
## 200, : sun could not be fit on page. It will not be plotted.

