Module2Homework

# import libraries  
library(tidyverse)

## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.3 v purrr 0.3.4  
## v tibble 3.1.0 v dplyr 1.0.5  
## v tidyr 1.1.3 v stringr 1.4.0  
## v readr 1.4.0 v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(tidytext)   
library(readr)

# Read in the data  
amazonbook\_review <- read\_csv("~/MIS 506/Module 2/Homework/amazonbook\_review.csv")

##   
## -- Column specification --------------------------------------------------------  
## cols(  
## ReviewerID = col\_character(),  
## ReviewerName = col\_character(),  
## ReviewerText = col\_character()  
## )

names(amazonbook\_review)

## [1] "ReviewerID" "ReviewerName" "ReviewerText"

head(amazonbook\_review)

## # A tibble: 6 x 3  
## ReviewerID ReviewerName ReviewerText   
## <chr> <chr> <chr>   
## 1 "\"A10000012B7C~ "\"Adam\"" "\"Spiritually and mentally insp~  
## 2 "\"A2S166WSCFIF~ "\"adead\_poet@hotmail.com ~ "\"This is one my must have book~  
## 3 "\"A1BM81XB4QHO~ "\"Ahoro Blethends \\\"Ser~ "\"This book provides a reflecti~  
## 4 "\"A1MOSTXNIO5M~ "\"Alan Krug\"" "\"I first read THE PROPHET in c~  
## 5 "\"A2XQ5LZHTD4A~ "\"Alaturka\"" "\"A timeless classic. It is a ~  
## 6 "\"A3V1MKC2BVWY~ "\"Alex Dawson\"" "\"Reading this made my mind fee~

# Problem ONE

## Answer the following questions about the dataset:

### a)Describe this dataset.

Dataset 1 is Amazon Book Reviews consisting of ReviewerID, ReviewerName, and ReviewerText

### b)How many variables and observations are there?

There are 3 variables. (1) ReviewerID - character string representing the reviewer’s unique ID (2) ReviewerName - character string representing the reviewer’s screen name (3) ReviewerText - character string of the text review that was given by the reviewer.

There are 1000 observations/rows

### c)Which column contains the text data that you are going to analyze?

We’ll be analyzing the ReviewerText column.

# Break it up into tokens and display count  
tidy\_amazonbook <- amazonbook\_review %>%  
 unnest\_tokens("word", ReviewerText)   
tidy\_amazonbook %>%  
 count(word) %>%  
 arrange(desc(n))

## # A tibble: 3,101 x 2  
## word n  
## <chr> <int>  
## 1 the 1368  
## 2 i 1070  
## 3 a 821  
## 4 book 721  
## 5 and 681  
## 6 this 674  
## 7 to 609  
## 8 it 590  
## 9 of 559  
## 10 is 421  
## # ... with 3,091 more rows

# Problem 2

## Tokenize the text andtransform it to a tidy data structureand count the most popular words in the text.

The most popular words are “the” (1368), “i” (1070), “a” (821), and “book” (721).

# Prepare for analysis  
# First, get rid of stop words and common words  
data("stop\_words")  
undesirable\_words <- c("book", "read", "story", "reading", "books", "written")  
tidy\_amazonbook <- tidy\_amazonbook %>%  
 anti\_join(stop\_words) %>%  
 filter (!word %in% undesirable\_words)

## Joining, by = "word"

# Remove Numbers, whitespace, and special characters  
tidy\_amazonbook <- tidy\_amazonbook[-grep("\\b\\d+\\b", tidy\_amazonbook$word),]  
tidy\_amazonbook $word <- gsub("\\s+","", tidy\_amazonbook$word)  
tidy\_amazonbook $word <- gsub("[^a-zA-Z]","", tidy\_amazonbook$word)

# Problem 3

## Prepare the text for analysis by removing the stop words, undesirable words, numbers, whitespaces, special characters, and any other necessary steps.

After performing the removal, we went from 3101 words to 2540.

tidy\_amazonbook %>%  
 count(word) %>%  
 arrange(desc(n))

## # A tibble: 2,540 x 2  
## word n  
## <chr> <int>  
## 1 circus 146  
## 2 loved 96  
## 3 life 91  
## 4 movie 86  
## 5 love 79  
## 6 enjoyed 77  
## 7 time 74  
## 8 elephants 73  
## 9 jacob 72  
## 10 characters 71  
## # ... with 2,530 more rows

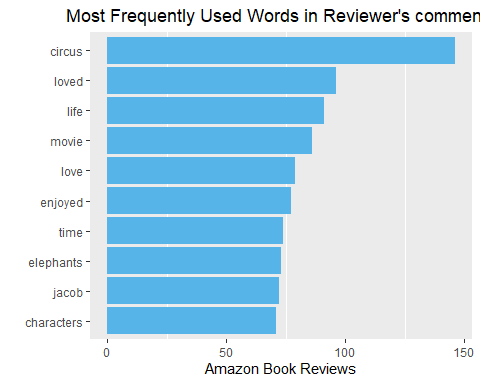
# Problem 4

## Again, count the most popular words in the text.

The most popular words are “circus” (146), “loved” (96), “life” (91) and “movie” (86)

my\_colors <- c("#E69F00", "#56B4E9", "#009E73", "#CC79A7", "#D55E00")  
  
tidy\_amazonbook %>%  
 count(word, sort = TRUE) %>%  
 top\_n(10) %>%  
 ungroup() %>%  
 mutate(word = reorder(word, n)) %>%  
 ggplot() +  
 geom\_col(aes(word, n), fill = my\_colors[2]) +  
 theme(legend.position = "none",   
 plot.title = element\_text(hjust = 0.5),  
 panel.grid.major = element\_blank()) +  
 xlab("") +   
 ylab("Amazon Book Reviews") +  
 ggtitle("Most Frequently Used Words in Reviewer's comments") +  
 coord\_flip()

## Selecting by n



# Problem 5

## Create a visualization of the most common words in the textand explain your results.

This looks like just a sample of reviews of Amazon books. For reasons unexplained, “circus” is the most frequently used word in this sample. Perhaps these reviews were pulled from the same book or author that featured a circus, movies, elephants, and a mysterious character named Jacob.