**Assessment Part 2: Dates, Times, and Text Mining**

In this part of the assessment, you will walk through a basic text mining and sentiment analysis task.

Project Gutenberg is a digital archive of public domain books. The R package **gutenbergr** facilitates the importation of these texts into R. We will combine this with the **tidyverse** and **tidytext** libraries to practice text mining.

Use these libraries and options:

library(tidyverse)  
library(gutenbergr)  
library(tidytext)  
options(digits = 3)

You can see the books and documents available in **gutenbergr** like this:

gutenberg\_metadata

**Question 6**

1/1 point (graded)

Use str\_detect to find the ID of the novel *Pride and Prejudice*.

How many different ID numbers are returned? correct

6 Loading

You have used 1 of 10 attempts Some problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

**Question 7**

1/1 point (graded)

Notice that there are several versions of the book. The gutenberg\_works function filters this table to remove replicates and include only English language works. Use this function to find the ID for *Pride and Prejudice*.

What is the correct ID number?

Read the gutenberg\_works documentation to learn how to use the function.

correct

1342 Loading

You have used 1 of 10 attempts Some problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

**Question 8**

1/1 point (graded)

Use the gutenberg\_download function to download the text for Pride and Prejudice. Use the **tidytext** package to create a tidy table with all the words in the text. Save this object as words.

How many words are present in the book? correct

122204 Loading

You have used 4 of 10 attempts Some problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

**Question 9**

1/1 point (graded)

Remove stop words from the words object. Recall that stop words are defined in the stop\_words data frame from the **tidytext** package.

How many words remain? correct

37246 Loading

You have used 1 of 10 attempts Some problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

**Question 10**

1/1 point (graded)

After removing stop words, detect and then filter out any token that contains a digit from words.

How many words remain? correct

37180 Loading

You have used 1 of 10 attempts Some problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

**Question 11**

3/3 points (graded)

Analyze the most frequent words in the novel after removing stop words and tokens with digits.

How many words appear more than 100 times in the book? correct

23 Loading

What is the most common word in the book? correct

How many times does that most common word appear? correct

597 Loading

You have used 2 of 10 attempts Some problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

**Question 12**

3/3 points (graded)

Define the afinn lexicon:

afinn <- get\_sentiments("afinn")

Note that this command will trigger a question in the R Console asking if you want to download the AFINN lexicon. Press 1 to select "Yes" (if using RStudio, enter this in the Console tab).

Use this afinn lexicon to assign sentiment values to words. Keep only words that are present in both words and the afinn lexicon. Save this data frame as afinn\_sentiments.

How many elements of words have sentiments in the afinn lexicon? correct

6065 Loading

What proportion of words in afinn\_sentiments have a positive value? correct

0.563 Loading

How many elements of afinn\_sentiments have a value of 4? correct

51 Loading

You have used 6 of 10 attempts