Due: 02/21/2026

## Homework Assignment: Normalizing Word Frequencies by Document Length

## Background

In Week 02, we compared word frequencies across two texts using **raw counts**. We found that "trade" appears 233 times in *The Circle of Commerce* (Text A) and 185 times in *Free Trade* (Text B). But does this mean Text A talks about trade more? Not necessarily—Text A might simply be longer!

Raw counts can be misleading when comparing documents of different lengths. To make fair comparisons, we need to **normalize** our word frequencies.

## How to Submit:

Post your code on your GitHub. On **Canvas**: post your response to the interpretive questions in the Discussion section (Week 2: Basics). Make sure to include your visualization **and a link to the code**. Finally, if Canvas, GitHub, or something similar is not functioning for this first assignment: don't panic! We will figure it out.

---

## Your Task

You will modify the Week 02 word frequency analysis to account for document length by calculating **relative frequencies** (proportions).

**We will break this down into steps:**

### I. Create a diagnostics table (before stopword removal)

Add a new section **immediately after** you create texts, before word_counts <- ....

Create a tibble called corpus_diagnostics with one row per document that includes:

- doc_title
- n_chars = number of characters in each document
- n_word_tokens = number of word tokens *before* stopword removal
- n_word_types = number of unique word types *after* lowercasing, *before* stopword removal

Your diagnostics must be computed using the tidy workflow used on the website (i.e., unnest_tokens() for word tokens; then str_to_lower(); then a summary that yields counts). unnest_tokens() is the key function for converting a "one row per document" tibble into "one row per token."

## II. Interpret the diagnostics (short prose to be shared on Canvas)

In **4–6 sentences**, answer:

- Are Text A and Text B comparable in length? (Use your diagnostics numbers.)
- If they differ substantially, what does that imply for interpreting **raw frequency** comparisons?

**Summary for you:**

The corpus integrity check exists to ensure that:

- you know what your corpus contains,
- your comparisons are methodologically sound,
- and your interpretations are constrained by evidence rather than assumptions.

It is not a technical hurdle—it is an **epistemic safeguard**.

## III. Compare normalized "trade" across the texts

Using the word_counts tibble from Week 02, calculate the total number of words (after stopword removal) in each document.

**Steps:**

1. Group by doc_title
2. Sum the word counts to get total words per document
3. Store this in a tibble called doc_lengths

**Hint:** Use group_by() and summarise() with sum().

Add a column to word_counts that shows each word's frequency as a **proportion** of the total words in its document.

**Steps:**

1. Join word_counts with doc_lengths
2. Calculate: relative_freq = n / total_words
3. Store the result in word_counts_normalized

**Hint:** Use left_join() and mutate()

**Then,** Using your normalized frequencies:

1. Filter for the word "trade" in both texts

2. Compare the **raw counts** vs. **relative frequencies**

Recreate the Week 02 word frequency visualization but using **relative frequencies** instead of raw counts.

**Requirements:**

- Use the same top 20 words as Week 02
- Show relative frequencies (not raw counts) on the x-axis
- Keep the side-by-side faceted layout
- Update axis labels and title to reflect normalization

**Hint:** Adapt the Week 02 plotting code, replacing n with relative_freq and updating labels.

**Answer these questions (short prose to be shared on Canvas):**

- Does Text A or Text B use "trade" more *proportionally*? And how does this compare to what the raw counts suggested?
- We normalized by dividing each word count by the total words in that document (after stopword removal). How would your results change if you normalized by the *original* document length (before stopword removal)? Would this be better or worse, and why? [This is a harder question than it would seem at first! Review the lecture notes].