Github: https://github.com/Shelf-life-without-water/IDS570_HW1_CODE

1. **Create a diagnostics table (before stopword removal)**

doc_title	n_chars	n_word_tokens	n_word_types
<chr>	<int>	<int>	<int>
1 text a	191605	33889	4773
2 text b	114211	19922	3332

2. **Interpret the diagnostics**

Are Text A and Text B comparable in length?

According to the diagnostics table, Text A and Text B are not comparable in length. Text A has 33,889 word tokens, while Text B has only 19,922, meaning Text A is about 70% longer.

If they differ substantially, what does that imply for interpreting raw frequency comparisons?

This substantial difference implies that raw frequency comparisons can be misleading: a higher raw count in Text A (like "trade" appearing 232 times vs. 185 times in Text B) may simply reflect its greater length rather than a greater focus on that concept. Therefore, to make fair comparisons, we must normalize word frequencies to account for document length.

3. **Compare normalized "trade" across the texts**

Does Text A or Text B use "trade" more *proportionally*? And how does this compare to what the raw counts suggested?

doc_title	word_n	word	n	relative_freq	raw_freq
<chr>	<dbl>	<chr>	<dbl>	<dbl>	<dbl>
1 text a	14694	trade	232	0.0158	0.0340
2 text b	8004	trade	185	0.0231	0.0271

Proportional analysis shows that Text B uses "trade" more densely: its relative frequency is 0.0231, compared to 0.0158 in Text A. This contradicts the raw counts, which suggested Text A used "trade" more (232 vs. 185 occurrences). Normalization reveals that Text B actually discusses trade more intensively per unit of text.

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?

If we normalized by the original document length (before stopword removal), Text A's relative frequency for "trade" would be about 0.00685 (232/33889) and Text B's about 0.00929 (185/19922). Text B would still have a higher proportion, but both values would be lower. Normalizing by the total words after stopword removal is better because stopwords (like "the", "and") carry little substantive meaning. Removing them gives a clearer picture of how prominent a word is within the meaningful content. Different texts may have different proportions of stopwords; normalizing by the original length would dilute the frequency of keywords in texts with more stopwords, potentially skewing comparisons. Thus, normalizing after stopword removal provides a more accurate measure of a word's importance in the actual content.

Plot: