……..

**Using tibbles** (Optional):

I mentioned in class that some packages now use something called ‘tibbles’ – a new type of data structure like S3 and S4. Tibbles are used in the tidy functions.

You will need to convert the corpus you have to a tibble structure.

Install the packages “dplyr”, “tidytext”, and “ggplot2”

Make them libraries.

To convert to a tidy corpus, you can do the following:

tidyBoiling <- tidy(BoilingStopNew);

You can tokenize the words and produce a tibble using:

BoilingWords <- tidyBoiling %>%

unnest\_tokens(word,text) %>%

select(id,word) %>%

mutate(word\_length = nchar(word)) %>%

arrange(desc(word\_length));

Note: I have written this out to show the individual methods, but you could type it all on one line.

BoilingWords %>% filter(id == BoilingStopNew);

Then, you can see the words arranged in order of length:

BoilingWords

You can use the same approach for finding the longest sentences, except that the unnest\_tokens argument would be replaced by:

unnest\_tokens(sentence, text, token = “regex”, pattern = “(?<!\\b\\p[L]r)\\.”)