Instructions

Purpose is to use a text file provided to create a tibble below.

Section 1) split the large string via pattern

section 2) extract dates

section 3) create tibble

Assignment should take 15~30 minutes

## # A tibble: 131 x 2

## date text  
## <date> <chr>   
## 1 2015-02-01 "MSNBC Febru~  
## 2 2015-02-01 "MSNBC Febru~  
## 3 2015-02-02 "MSNBC Febru~  
  
## # ... with 128 more rows

First downloadmsnbc\_text.TXT and load it with readr package. Save it to single string variable called text﻿

1) Split the string based on pattern

Rightnow, text variable should be a giant single string that has multiple documents.

If you open the msnbc\_text notice how each document start with with something like 1 of 131 DOCUMENTS, 2 of 131 DOCUMENTS and so on. This is a pattern that separates each document in the file.

Instead of one big string, split the string (which should be in a variable called text at this point) on the pattern that separates each document and save it as a character vector.

You can do this by writing regular expression that captures this pattern and then use str\_split(text, pattern) %>% unlist() to split the single string you read in with readr::read\_file() into separate documents

Check the length of your new character vector (make sure you have a character vector and not a list). You should have 132 items in your vector, but this is strange bc we have 131 documents. If you did this correctly, R will have created a string with only whitespace (“ and”" are whitespace characters) as the first element, check to make sure this is the case. If not, you did something wrong. If so, then subset the vector so we only include items 2 on from the text vector and save it back into the varaible text.

Lastly, trim whitespace from both sides of each document in the vector

Extract the dates

You should notice another pattern in the text for each document, the date appears at the top with a specific pattern. Use this pattern to extract the date from each document and save this in a variable called dates

Create a tibble with the data

create a tibble with all these variables in order (date, text).

and call it df. Each document’s data should be a row in the tibble