Problem Set 2

Programming and Data Management

Overview

In this problem set we will take a text data file which is messy and cannot be analyzed and transform it into a clean tibble that we can analyze. The text file looks like this and we obviously cannot do anything meaningful with the data in this condition.

```
##
##                                 1 of 131 DOCUMENTS
##
##
##                                       MSNBC
##
##                              February 1, 2015 Sunday
##
##                      SHOW: UP with STEVE KORNACKI 8:00 AM EST
##
## UP WITH STEVE KORNACKI for February 1, 2015
##
## BYLINE: Steve Kornacki, Ed Schultz, Ed Rendell, Joy Reid, Craig Melvin, Kevin
## Tibbles
##
## GUESTS: Wesley Lowery, Liz Mair, Mike Pesca, Jacob Jacobs, Mike Freeman, Drew
## Magary
##
## SECTION: NEWS; Domestic
##
## LENGTH: 14936  words
##
##
##
## HIGHLIGHT: From the scandal to the showdown in the NFL. A look at the upcoming
## Super Bowl, the controversy around the Patriots, and betting on Super Bowl.
## Judging Bowe Bergdhal. Cold weather in Chicago.
##
##
## STEVE KORNACKI, MSNBC HOST: From the scandal to the showdown. Good morning to
## everyone out there. Thanks for getting up with us on this chilly first day in
## February, Super Sunday. Winter`s looking more like winter with every passing
## minute. The latest details on this week`s big storm in a moment.
##
## Another big storm
```

Throughout this problem set, we will take this messy data and transform it into a managable tibble with the information we need which looks likes this. In this form, we can work with the data more eaisily and actually do anlysis.

```
## # A tibble: 131 x 4
##    date       show      word_lengths text
##    <date>     <fct>            <dbl> <chr>
##  1 2015-02-01 Up               14936 "MSNBC February 1, 2015 Sunday SHOW: ~
##  2 2015-02-01 Other            14055 "MSNBC February 1, 2015 Sunday SHOW: ~
##  3 2015-02-02 The Ed S~         7374 "MSNBC February 2, 2015 Monday SHOW: ~
##  4 2015-02-02 Hardball          8786 "MSNBC February 2, 2015 Monday SHOW: ~
##  5 2015-02-02 Other             6876 "MSNBC February 2, 2015 Monday SHOW: ~
##  6 2015-02-02 Other             8177 "MSNBC February 2, 2015 Monday SHOW: ~
##  7 2015-02-02 Other             7682 "MSNBC February 2, 2015 Monday SHOW: ~
##  8 2015-02-02 Other             7311 "MSNBC February 2, 2015 Monday SHOW: ~
##  9 2015-02-03 Hardball          8376 "MSNBC February 3, 2015 Tuesday SHOW:~
## 10 2015-02-03 The Ed S~         6944 "MSNBC February 3, 2015 Tuesday SHOW:~
## # ... with 121 more rows
```

Instructions

Download the text file `msnbc_text.TXT` from blackboard. You will be using this text file for this problem set

Load in the data (5pts)

Use the `read_file()` function from the `readr` package to read in the text file. It will be read in as a single string. Save it in a variable called `text`.

Split the string on the common pattern (10pts)

Open the text file, this can be done in notepad (windows) or textedit (mac) or any other text editor application.

Notice that each document in the file starts 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.

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 (15pts)

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`

Extract the shows (15pts)

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

Extract the length (15pts)

You should notice another pattern in the text for each document, the length appears in the transcript with a specific pattern starting with `LENGTH`. Use this pattern to extract the length from each document and save this in a variable called `lengths`. Convert this lengths variable to a numeric varaible rather than a string.

Replace whitespace (10pts)

Write code to replace any more than one whitespace character with a single whitespace character. For example `Hello,     my name is Robert` should be `Hello, my name is Robert`

Create a tibble with the data (5pts)

create a tibble with all these variables in order (date, show, word_length, text) and call it df. Each document’s data should be a row in the tibble.

Convert strings to factors (25pts)

The shows are strings but they should be factors, convert the show variable into a factor such that there are only 4 categories (Up, The Ed Show, Hardball, and an ‘Other’ category)