DSO 545: Statistical Computing and Data Visualization

Manipulation and Visualizion of Text Data using stringr Package

Fall 2017- LAB

In this lecture we will learn how to manipulate text data in R, and in particular we'll learn the structure of emails, and some string processing basics. Strings are not glamorous, high-profile components of R, but they do play a big role in many data cleaning and preparations tasks. This lecture is based on Hadely Wickham's notes and paper **"stringr: modern, consistent string processing"**.

We will be using a dataset that include email data ("email.rds"). For more information about RDS files click here (http://stackoverflow.com/questions/21370132/r-data-formats-rdata-rda-rds-etc).

```r
library(stringr)
emails <- readRDS("email.rds")
```

1. Create a character string that contains just one quotation mark: `"`
2. Create the following in R:

```
:-\
```

```
(^_^")
```

```
@_'-'
```

```
\m/
```

3. Load `stringr` R package, and search for `str_locate()` and `str_sub()`.
4. Locate the character "a" in "great", "fantastic", and "super".
5. Extract the substrings "tes", "ting", and "test" from "testing".
6. What do you think this code is doing?

```r
input <- c("abc", "defg")
str_sub(input, c(2, 3))
```

7. Use `str_locate()` to identify the location of the blank line that seperates the header from the body in the first email. (Hint: a blank line is a newline immediately followed by another newline)
8. Extract the header and body for the first email.
9. Split the header of the first email message into its corresponding metadata? (Hint, each part of the metadata is on a seperate line. Look for `str_split()`)
10. Extract the header and body for all email messages. Store the header and body in `header` and `body` variables respectively.