Text Analysis Lab

Liam Niehus-Staab

10/29/2018

# Intro Strings Questions

veggies = c("carrot", "bean", "peas", "cabbage", "scallion", "asparagus")  
  
# + Find those strings that contain the pattern "ea".  
str\_detect(veggies, "ea")

## [1] FALSE TRUE TRUE FALSE FALSE FALSE

# + Find those strings that end in "s".  
str\_detect(veggies, "s$")

## [1] FALSE FALSE TRUE FALSE FALSE TRUE

# + Find those strings that contain at least two "a"'s.  
str\_detect(veggies, ".\*a.\*a")

## [1] FALSE FALSE FALSE TRUE FALSE TRUE

# + Find those strings that begin with any letter except "c".  
str\_detect(veggies, "^[^c]")

## [1] FALSE TRUE TRUE FALSE TRUE TRUE

# + Find the starting and ending position of the pattern "ca" in each string.  
str\_locate\_all(veggies, "ca")

## [[1]]  
## start end  
## [1,] 1 2  
##   
## [[2]]  
## start end  
##   
## [[3]]  
## start end  
##   
## [[4]]  
## start end  
## [1,] 1 2  
##   
## [[5]]  
## start end  
## [1,] 2 3  
##   
## [[6]]  
## start end

1. The regular expression "^[Ss](.\*)(t+)(.+)(t+)" matches "scuttlebutt", "Stetson", and "Scattter", but not "Scatter." Why? The regex matches either S or s as first char, then any number of any char, then 1 or more t, then 1 or more of any character, then 1 or more t.

It doesn't match "Scatter" because after it matches the first 't' with t+, then the next 't' with .+, and it cant match the last t+ because there are only 2 't's in Scatter.

## Text Mining

#read in blogs, twitter and news  
Twitter1 <- readLines("~/Shared/F18MAT295/en\_US.twitter.txt",encoding = "UTF-8", skipNul = TRUE, warn = FALSE)  
  
# 1. Are strings involving the words "love" longer or shorter than strings with the word "hate"?  
#Strings with love appear to be a little be shorter on average.

# \* How many rows contain the word hate (include upper and lower case letters)?  
hate <- sum(grepl("[Hh][Aa][Tt][Ee]", Twitter1))

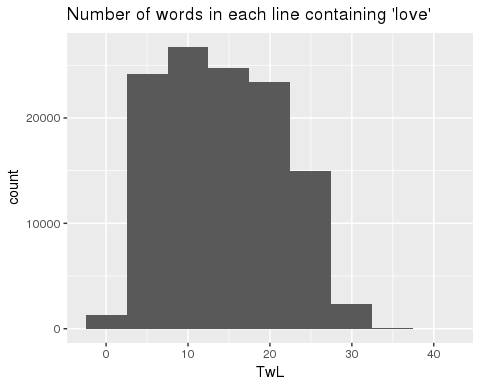
# \* Give a 5 number summary for the number of characters in texts that include the word hate.  
hate2 = grep("[Hh][Aa][Tt][Ee]", Twitter1)  
hateTwit = Twitter1[hate2]  
summary(nchar(hateTwit))

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 6.00 44.00 75.00 76.19 109.00 140.00

# \* Give a 5 number summary for the number of characters in texts that include the word love.  
love2 = grep("[Ll][Oo][Vv][Ee]", Twitter1)  
loveTwit = Twitter1[love2]  
summary(nchar(loveTwit))

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 6.00 43.00 73.00 74.28 106.00 140.00

# \* Create bar charts for both text files.  
TwL = stri\_count\_words(loveTwit)   
qplot(TwL, binwidth = 5, main = "Number of words in each line containing 'love'")



TwH = stri\_count\_words(hateTwit)   
qplot(TwH, binwidth = 5, main = "Number of words in each line containing 'hate'")

